RESEARCH REPORT

Practical Considerations for Pay for Success Evaluations

Sarah Gillespie   *Akiva Liberman*   *Janine Zweig*   *Devlin Hanson*

Mary Cunningham   *Michael Pergamit*

December 2016

:::URBAN
INSTITUTE · ELEVATE · THE · DEBATE

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.

Contents

Acknowledgments

Executive Summary

Evidence is at the core of the pay for success (PFS) movement, which pushes government to rigorously evaluate programs and pay only for those that achieve positive outcomes. Evidence, however, is only as good as the evaluation that produces it. As such, evaluators are an important part of any PFS project from beginning to end. In these projects, the evaluator implements activities that assess program outcomes to determine success payments from government to investors.

This report introduces practical considerations for evaluators when integrating established, rigorous evaluation methods into the PFS project structure. Considerations include protecting evaluator objectivity and independence; assessing projects for feasibility; balancing competing interests; getting buy-in on rigorous evaluation design; building evaluation flexibility into the PFS contract; working with an independent validator; navigating institutional review board considerations; and expanding the evaluation. By understanding and addressing these factors, evaluation partners can ensure that evaluation design strengthens the PFS project. Our experiences suggest the following recommendations for those involved in PFS evaluations:

- **Bring the evaluation partner on board as early as possible** to help ensure that important decisions in the early development phases are grounded in evidence.

- **Provide time and support for an evaluation planning period** so that the most rigorous evaluation design possible can be created with full buy-in from all key PFS partners.

- **Provide time and support for an implementation pilot period** to minimize implementation failure and to test key implementation assumptions, such as referral processes, program flow, and take-up rates.

- **Refer to other PFS contracts** to adapt language that builds in evaluation flexibility and to learn from other evaluation structures.

- **Consult with the independent validator as early as possible** to address any potential analytical or institutional review board concerns before implementation begins.

- **Create a process for constant communication between the evaluator and other PFS partners** to coordinate on implementation or evaluation challenges as soon as they arise, especially during early implementation.

Practical Considerations for Pay for Success Evaluations

Evidence is at the core of the pay for success (PFS) movement (box 1), which pushes government to rigorously evaluate programs and pay only for those that achieve positive outcomes. Evidence, however, is only as good as the evaluation that produces it. As such, evaluators are an important part of any PFS project from beginning to end. In these projects, the evaluator assesses program outcomes to determine success payments from government to investors. Evaluators should be engaged early in PFS project development. Their early involvement can inform overall project design in meaningful ways, such as by increasing the ability to rigorously assess outcomes and the confidence in payment schemes. For readers newer to evaluation and the PFS process, the variety of roles evaluators assume during PFS project development and implementation are described in detail in a companion paper (box 2).

BOX 1

What Is PFS?

Pay for success, or PFS, is an innovative financing mechanism that shifts risk for a new or scaling 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/.

This report introduces practical considerations for evaluators when integrating established evaluation methods into the structure of a PFS project. Although PFS evaluations should embrace the most rigorous design possible, the following considerations may differ or be inflated because of the unique PFS context:

- Protecting evaluator objectivity and independence

- Assessing for feasibility

- Balancing competing interests

- Getting buy-in on rigorous evaluation design

- Building evaluation flexibility into the PFS contract

- Working with an independent validator

- Navigating institutional review board (IRB) considerations

- Expanding the evaluation

As evaluators who have been part of PFS projects, we hope to share the lessons we have learned about these practical considerations. This paper is written for an audience of evaluators and others steeped in the details of program evaluation (box 3). We focus on the unique ways the PFS model influences evaluation approaches and on considerations that evaluators must weigh in ways they normally might not. By sharing these lessons, we hope to smooth the path forward for others engaging in PFS projects.

BOX 3

Key Terms

For consistency, we provide our definitions of several key terms used in this report.
Evaluation design: The written document that outlines how evaluation activities will be conducted, including an analysis plan for how program impacts will be estimated.
Program impacts: The change in outcomes that can be attributed to the program studied—often called *program effects*.
Program flow: The pace at which individuals enter the program, usually dictated by a referral process.
Take-up rate: The number of individuals offered the program who actually enroll in the program.
Random assignment: An experimental research method by which individuals referred to the program are randomly assigned to either the treatment group, which receives program services, or the control group, which does not.

Protecting Evaluator Objectivity and Independence

Perhaps the most important characteristic of an evaluation partner is objectivity. Unlike other PFS partners, evaluators have no financial stake in the outcome—their agenda is thorough data collection and evaluation. Because outcomes as reported by the evaluators will trigger or preclude payments from one party to another, questions about the outcomes' validity may set PFS partners' financial interests at odds with each other. In a PFS project, these financial interests heighten the demand for a clear understanding of an evaluator's objectivity and independence.

Because of its objectivity, the evaluator's perspective on the feasibility and risks of the project should play an important role well before implementation begins. Once the project is set and contracts are signed with established metrics in place, the evaluator may be in a poor position to raise critical questions, such as "Why do you believe the program can achieve an effect of that size?" Evaluators should have the opportunity to share concerns about feasibility before the project goes forward. Assessing several types of feasibility is discussed in more detail in the following section.

In a traditional evaluation, those involved in the program being evaluated are likely aware of the program's past performance and what type of success they might expect. But in PFS projects, partners at the table will have less firsthand knowledge of the program and will need a clear understanding of the program's probability of success based on analysis of available evidence. The evaluator is a critical advisor in this process.

To protect their objectivity and independence throughout the project, evaluators should advocate for publicly releasing the PFS contract, data, and reports on outcomes and findings in the evaluation. This increases transparency and protects the evaluators from pressure in the face of the PFS partners' financial interests.

Assessing for Feasibility

Evaluators involved in a PFS project should consider three types of feasibility: feasibility of the project producing the desired public benefits, feasibility of implementation as intended, and feasibility of demonstrating this success through a rigorous evaluation.

Feasibility of Program Success

Many PFS projects are designed to produce budgetary savings to government that will offset the cost of the implemented program. Sometimes this is because "business as usual" is very expensive; in criminal justice, this may mean incarceration, detention, or other out-of-home placements; for mental health, this may mean hospital stays. Under assumptions about business as usual and the expected impact of the proposed PFS program, the money that government saves might offset the cost of the program funded by the PFS project. By helping articulate these assumptions, evaluators can clarify the program's logic model for the target population, explore the minimum program impacts needed to generate outcome payments, and review how those impacts compare with existing evidence.

Feasibility of Implementation Success

Even though PFS practice urges evidence-based program design, the evidence is often incomplete and does not account for the risk of implementation failure. Experienced evaluators know that many programs fail in implementation, which then undermines impact analyses. How close is the PFS implementation context to the context from which evidence of previous impacts were taken? Does the PFS service provider have the experience and capacity to implement this program model with fidelity in this context? Answers to such questions are often lacking.

Yet the structure of PFS pushes for rapid initiation of an impact evaluation. The financial arrangements are often not well suited to a simulation or pilot phase to assess the feasibility of implementation as planned, because only program impacts will generate success payments. A simulation is less intensive than a pilot and uses existing data to simulate intended program or evaluation processes, such as identifying individuals who meet certain eligibility criteria for referral to the program. A pilot phase requires more resources and actual implementation of the program and evaluation processes, such as outreach to and enrollment of eligible individuals, but such implementation happens before partners are "on the clock" for program performance. Our experience shows that moving to implementation without such a phase may save money in the short term but cost more in the long term. A simulation or pilot phase is often advisable for two reasons.

First, almost all programs have a start-up period during which implementation issues surface; in successful cases, these hiccups are addressed, arrangements are altered, and a stable program emerges. Second, a simulation or pilot phase can considerably refine and correct program assumptions (or provide data-driven estimates where none existed) about the feasibility of intended referral practices.

Both a simulation and pilot phase can help estimate the pace at which individuals may enter the program (program flow), though only a pilot phase will help estimate the rate at which individuals choose to enroll in the offered program (take-up rate). To be useful in these ways, a simulation or pilot phase must be long enough to allow implementation assumptions to be observed. In our view, such a phase should concern implementation issues but should not be used to estimate program impacts, given small samples, weak counterfactuals, and room for potential error. If the pilot is the strongest basis for making assumptions about the size of program impact, then perhaps the program is not ready for a PFS project until more rigorous evidence is available. Simulations and pilot phases, as well as the importance of accurate program flow and take-up estimates, are discussed in detail in a later section.

Feasibility of Strong Evaluation

Feasibility requirements for demonstrating impact through rigorous evaluation are familiar to evaluators. These requirements include assumptions about the size of anticipated effects and their basis; the program's theory of change; considerations around counterfactuals; and issues of measurement, power, and statistics. Especially if involved early in the project, the evaluator should critically assess other partners' assumptions about program effect size. Although PFS was originally conceived for programs backed by established evaluation literature, which allows for highly predictable effect sizes for the target population, PFS projects in practice often lack a strong basis for hypothesizing the effect size. Absent an evaluator's objective assessment of program effects, nonevaluation partners may be led to base payment benchmarks on aspirations or simple round numbers rather than the evidence base, which sets the program up for failure. The evaluator can serve an important role if brought in early enough for evaluation planning and design.

Finally, there is the feasibility of creating a strong counterfactual, preferably through random assignment. In this context, feasibility often depends on assumptions about the rate at which individuals offered a program agree to participate (the take-up rate). In real-life circumstances with eligible program populations of fixed size, when take-up rates turn out to be lower than anticipated, filling program slots may require dipping into the intended control population or changing the randomization ratio in ways that jeopardize the feasibility of a random assignment study. While designing a PFS project, it is typical to consider and allow for the flexibility to implement backup evaluation plans in the case that random assignment needs to be abandoned because of slow program flow, low take-up rates, or other unanticipated impediments. This is more difficult in PFS projects than in most other evaluation contexts, in our experience. Is such a backup plan feasible? Are there other strong counterfactuals?

PFS projects are sometimes implemented in response to changing environments or are part of a larger effort at system reform, which may bedevil attempts to identify strong nonexperimental counterfactuals. These questions are considered further in the section on building evaluation flexibility.

Balancing Competing Interests

Competing interests can easily arise in a PFS project, given the diversity of stakeholders involved. Although all stakeholders want the project to be successful, each brings different motivations. The government aims to improve outcomes for a given target population, preferably with accessible cost savings that pay for the program. Private investors may seek a return on their funds that at least equals what they could attain elsewhere. And philanthropic investors seek to fund a more effective means of providing services to vulnerable populations.

Service providers face some of the biggest demands that can put them somewhat at odds with other stakeholders, especially during the evaluation design phase. The target populations in PFS projects are generally individuals who are hard to serve. For example, supportive housing programs target individuals who are chronically homeless and have varied and complex needs. These complex needs produce a reliance on high-cost public systems, such as jails and emergency health care. Successfully implementing an intervention could generate dramatic cost savings, reducing or covering the cost of the intervention, which motivates the PFS project. But engaging these individuals can be challenging, requiring an assertive process to reach out to eligible individuals rather than, as is usual, waiting for individuals to seek out services.

In a PFS contract, service providers have a stake in formulating eligibility criteria because they must commit to target enrollment or take-up rates specified in PFS service contracts that are linked to potential success payments. To maximize their ability to meet these enrollment targets, service providers may seek to alter either the eligibility criteria or the enrollment targets. However, such changes can jeopardize the ability to measure program impacts, including those linked to success payments. Altering the eligibility criteria often produces a treatment group composed of less-vulnerable individuals. Consequently, there may be less room to observe improved outcomes produced by the program. Similarly, altering the enrollment targets could reduce the number of treatment cases for which improved outcomes can be observed. Failure to find impacts will mean that investors do not get a return, the government likely doesn't achieve its goals, and the service provider fails to demonstrate efficacy. Although it is the investor who bears the financial risk in PFS, the service provider

assumes considerable reputational risk in a PFS project. For these reasons, evaluators must balance competing interests and design the most rigorous evaluation feasible with full buy-in from service providers and other key stakeholders. Strategies for achieving buy-in on evaluation design decisions are discussed in more detail in the next section.

Another place competing interests may arise is in the analysis plan, which dictates how success will be measured for payment purposes. We have experienced this tension within a random assignment evaluation design. In this context, many program evaluations focus primarily on intent-to-treat (ITT) analysis, where all randomized individuals are included in the impact estimates regardless of whether they actually enrolled in the program. However, low enrollment or take-up rates make finding impacts less likely in an ITT analysis because individuals are counted in the treatment group even when they do not receive program services. The alternative is to base success payments on analysis of treatment-on-the-treated (TOT), an estimate of impact based only on the individuals actually receiving the treatment.[1]

Different stakeholders may have different views on whether to use ITT or TOT estimates. ITT estimates are generally of more interest to policymakers because they describe what would be expected to occur if the program were replicated elsewhere for a similar target population. TOT estimates are generally of interest to programs because they reflect the impacts for those the program actually served, but this may overstate the true effect of the program. The evaluator can help stakeholders understand the differences in these estimation methods and the implications of using one over the other.

Other statistical issues relevant to the PFS partners' discussion could raise additional competing interests. A traditional evaluation would assert an impact only if statistical significance is achieved. Even traditional evaluations, however, face differences in opinion on the level of significance required to assert that an impact is "real." Small sample sizes and low take-up can make achieving statistical significance problematic even when large impacts are present. If the evaluation cannot detect an impact even when one is present, the government will not make repayments to investors. The evaluator can help other PFS partners understand the role of statistical significance in determining the credibility of a measured impact, including how the level of confidence in the finding relates to statistical significance. Additionally, PFS partners will likely discuss how statistical significance relates to the associated governmental savings tied to success payments.

The conflict in competing interests may be more nuanced than it appears during discussions of high-level issues, such as the use of ITT versus TOT and the use of statistical significance. For example, the

number of decimal places the evaluator reports and how those are rounded off could make a substantial difference if impact estimates are close to the payoff thresholds. Many PFS projects use a payment scheme with different payment amounts linked to different outcome levels. Say, for example, the PFS contract states that no payment is made to investors if negative outcomes are not reduced at least 20 percent; at 20 percent reduction, investors receive $3 million. If the impact is slightly below 20 percent, such as 19.956 percent, how the number gets rounded makes a substantial difference. Rounding to two decimal places yields no payment while rounding to one decimal place yields a substantial payment. Evaluators can help raise these issues during conversations about evaluation design, especially if the evaluator is brought on before implementation to support evaluation planning.

Getting Buy-In on Rigorous Evaluation Design

Evaluators often work with funders and implementation partners to design a program evaluation, whether or not it is linked to a PFS project. In our experience, however, PFS stakeholders may be more involved in the evaluation design process given its direct link to success payments. Additionally, many PFS evaluation contracts begin with implementation. This structure makes it challenging for PFS projects to build in an evaluation planning and design period, creating a barrier to an inclusive process that secures buy-in from important stakeholders. Many PFS investors and government partners seek the most rigorous method for determining success, which is usually a randomized controlled trial. Evaluators must work closely with these stakeholders and others directly involved in PFS implementation, such as the service providers, to determine the most rigorous design possible.

As many evaluators know, challenges to the evaluation design can arise when partners feel that they were not part of design decisions but are asked to support the evaluation, such as through referrals or data. In PFS projects, such challenges can be especially problematic because not receiving the necessary data or encountering a glitch in the referral process could lead to a breach of contract. Our experiences have generated several useful practices for obtaining buy-in at an early stage of PFS project structuring. Evaluators should advocate for time and support for evaluation planning well before implementation. During this planning phase, the evaluator will often conduct the type of feasibility analyses described earlier, as well as analysis of the target population and eligibility criteria. Service providers considering a random assignment design are often concerned about the ethical implications of denying services to those in need in the control group. By clearly defining the target population and the scope of need, evaluators can show how random assignment creates a lottery process, which is often the fairest way to distribute such scarce resources as housing assistance, slots in

supported work programs, or high-quality child care services. The evaluator can outline evaluation design options and actively involve stakeholders, especially service providers, in the consideration and feasibility of each option, including whether random assignment will be blind and how consent procedures will be navigated. Many of these decisions will also factor into IRB review and approval, discussed in a later section. We have sometimes obtained grant supports outside of the project to allow for an evaluation planning phase, ensure the project is grounded in evidence, and ensure the design of the most rigorous evaluation possible with buy-in from all stakeholders. Evaluators can raise the issue of evaluation planning with other PFS partners to discuss the advantages of and opportunities to funding such a process.

Once evaluation design decisions are finalized, another important consideration is securing buy-in for a simulation or pilot phase. This phase is in addition to the evaluation planning period discussed previously. As mentioned, this phase allows evaluators and other stakeholders to test the assumptions made in the evaluation design, such as the way the referral and enrollment process will work and the expected flow of individuals into the program. These phases can be challenging in PFS projects because stakeholders must decide whether funding for these phases will be part of project financing. If program resources are scarce, a pilot phase could be problematic if it limits resources available during implementation. A simulation is still feasible in this case because it can be done relatively quickly without draining program resources. Our experiences underscore the importance of preimplementation work and the evaluator's role in helping PFS partners consider the trade-offs, time, and resources necessary for a simulation or pilot phase.

Building Evaluation Flexibility into the PFS Contract

The evaluation design dictates in detail when and how investors will be repaid and is a critical part of the legally binding PFS contract. Significant changes to the evaluation design will often require the approval of all stakeholders. It is considered a best practice for evaluators to propose an evaluation design and analysis plan before beginning implementation. These plans can typically be revised as evaluators receive more information. However, the PFS contract may put evaluators in the unusual position of having to make detailed analysis decisions on controls and methodologies in advance, which may have substantial consequences for the end result and thus cannot be easily changed. This lack of flexibility may lead to an unideal ex post analysis plan that may not reliably detect meaningful program impacts, which are the primary interest for other PFS stakeholders. This is especially true when considering the data that will be used. Often, knowing the quality and completeness of a proposed

dataset is impossible. Consequently, evaluators may want to advocate for ways to build flexibility into the evaluation design to allow for updates as new information becomes available.

Evaluators can advocate for flexibility in the PFS contract through several ways. They can simply include such language as "These are the variables we intend to control for in our regression analysis. However, we will decide on the final set of controls to be used in the analysis upon receipt of the final data. This decision will be based on the quality, completeness, and collinearity of the controls." Further, evaluators may include certain conditions that must be met in specific methodologies.[2]

Flexibility could be built into the contract through several other ways:

- **A simulation or pilot phase**. This would occur before finalizing the evaluation design. As discussed, a simulation or pilot phase helps work out any kinks in the referral and random assignment process. It also helps the evaluators understand the target population characteristics, end sample size, the timeline for reaching the end sample size, and program flow.

- **Data collection to support possible changes to eligibility criteria.** If program referrals or enrollment are slower or smaller than anticipated, collecting additional information at intake allows program partners to later expand eligibility criteria (if possible) to generate the minimum referrals necessary for the evaluation.

- **Future timeline for sharing regression specifications.** Specifying in the evaluation design that finalized regression specifications will be made at a later date allows for flexibility that is not subject to complicated contract changes.

- **Midway "checkpoints" to examine data.** Early data pulls that help ensure data are complete and high quality before the final data receipt will help evaluators plan for any changes to the evaluation design and ensure reporting deadlines are met.

- **Analysis methods in the event of early termination of the PFS contract.** If the PFS contract is terminated early, an early analysis plan should be laid out in the evaluation design. This will allow for early analysis suitable for the smaller sample size and period of observation.

- **Alternative analysis methods if the primary method is not feasible.** As discussed, the primary method of analysis might not be feasible. An evaluation design may specify a random assignment evaluation with an ITT estimate as the final estimate, but this may not be the ideal design if crossovers[3] are too great. In such a case, having an alternative analysis plan (along with

the conditions under which the alternative plan would be triggered) outlined in advance is important.

Working with an Independent Validator

Another unique feature that appears in some PFS projects is the role of an independent validator. Any party to the PFS contract, such as the government or an investor, may choose to engage an independent validator as part of the PFS project. Engaging such an entity allows an external party to confirm that the evaluation conducted to assess the outcomes of the project is implemented accurately and rigorously, and that outcomes payments (or nonpayments) are legitimate.

Independent validators may play different roles in the PFS project depending on who is conducting the evaluation. In one PFS project, evaluation activities are conducted by a branch of the government partner. In this scenario, an independent validator—such as a nongovernmental research firm or consulting group—will validate the activities of the government-based agency. In other PFS projects, evaluation activities are conducted by independent researchers (based in research and consulting firms or at universities). In such a scenario, another nongovernmental research firm or consulting group is additionally employed to validate the activities the independent researchers are implementing.

Independent validators may be asked to validate several parts of the evaluation process. For example, they may be asked to monitor the recruitment and referral process for study participants, working to determine if inclusion and exclusion criteria are accurately applied to the recruitment and referral process, to confirm that study participants are properly randomly assigned to treatment or control groups, or to review whether outcomes are assessed accurately and in alignment with the contract. Independent validators may monitor this work by reviewing analytic strategies in evaluation plans, checking statistical programming code for errors, replicating the creation of target sampling frames, rerunning analyses, and duplicating statistical analyses for comparison with results achieved by project evaluators.

Along with the other unique dynamics of PFS evaluations, the inclusion of the independent validator considerably changes how the evaluator functions during a study. We have experienced both limitations and strengths of this PFS structure. The primary limitation is that evaluators' expert judgment throughout the project is more constrained than in typical evaluations. As discussed, for example, PFS evaluation plans are determined before data collection, and lessons learned when analyzing data typically cannot change the course of the analytic strategy. In a non-PFS evaluation,

evaluators can make decisions along the way about how to clean and code data, how to deal with missing data, how to weight data, what covariates to include in an analysis, and so forth. Because such decisions are largely prespecified in PFS projects, independent validators are looking to confirm that the evaluators followed the plan exactly as specified. If more flexibility was built into the contract as discussed, however, an evaluator would likely need to gain consensus with the independent validator in the case of a disagreement about how to finalize a particular analytic activity.

The major strength of including an independent validator is that it is similar to an in-depth peer review. The extra assessment of evaluation rigor and accuracy can manage any risk the PFS evaluator faces. All partners would agree that identifying an error after the fact, during the assessment of outcomes and success payments, would be a failure of the process. An independent assessment from a validator guards against such mistakes. Independent verification of findings also can protect an evaluator's objectivity by requiring transparency in data collection and analysis. In our experience, the best way to minimize the limitations and take advantage of the strengths of an independent validator is for the evaluator and independent validator to discuss the evaluation design as early as possible, coming to agreement on any discrepancies before implementation begins.

Navigating IRB Considerations

An IRB should review all research projects that include human subjects, including evaluations that are a part of PFS projects. Similar to research conducted outside of PFS projects, the evaluation should create protocols that protect human subjects, including for obtaining consent when appropriate, developing confidentiality guidelines, instituting data security procedures, and reporting results. The evaluation partner may need to build special considerations into the PFS contract that guide data use, especially for the investors and government partner, who may require data transparency and sharing that mitigate the risk of their investment.

The PFS contract should clearly define the use of the data collected for the evaluation. Do the investors own the data, or do the government partners? Who has access to the data during the life of the contract? What happens to the data in the event of an early contract termination or the selection of a new evaluator? These questions should be clearly answered in the PFS contract.

The PFS contract will likely require that the independent validator follow the IRB protocols established for the evaluation and should sign a pledge to that effect. IRB questions to consider specific to the role of the independent validator include "Who will be the independent validator?" "Where will

they validate results?" "Will their procedures adhere to those outlined in the IRB protocol?" As suggested, if the evaluator and independent validator discuss these IRB considerations early on in the PFS project, any potential concerns can be appropriately addressed.

Expanding the Evaluation

The final unique consideration for PFS evaluations concerns the scope of the evaluation. PFS contracts will create a narrow focus on outcomes tied to success payments, specifying in detail how they will be measured. Often, success payments will be tied to the outcomes with the strongest evidence base. Evaluators can help bring attention to opportunities to expand the evaluation to include other key outcomes and further build the evidence base for the program and perhaps future PFS projects. For example, existing supportive-housing PFS projects have focused payment measures on housing stability and, in one case, reductions in jail bed–days. The literature, however, shows that supportive housing likely has a large impact on health outcomes, though the evidence is not yet robust enough to use such outcomes as a performance benchmark. Including health outcomes in PFS evaluations, even outside of the payment structure, could help build the evidence such that future projects could use health outcomes as additional payment outcomes.

Additionally, evaluators should make PFS partners aware of the importance of a process evaluation to document the implementation of a program, exploring whether the program was implemented as intended and identifying any barriers or improvements to program implementation. Although the evaluator's primary role is to determine the impact of the program, the evaluator will almost certainly be asked why the program did or did not achieve its expected outcomes. Absent a process evaluation, little information will be available to help stakeholders understand the PFS project results. For example, a PFS project that does not achieve the agreed-upon outcomes may seem to suggest the program did not work, but a process evaluation could reveal that the program was not implemented with fidelity to its model, meaning the program may still work if implemented as intended. When PFS outcomes are achieved by the program, process evaluation findings can help provide the information necessary for those who may want to replicate or expand the program and its success.

Sometimes, additional outcomes and a process evaluation can be part of the evaluation design funded by the PFS evaluation contract. Raising additional funds outside of the PFS contract to support expanded data collection and analysis may also be possible. Funders are often eager to leverage the PFS investments and ensure the project captures important data on outcomes of interest to their specific

work. To mitigate any challenges that may arise from a somewhat bifurcated evaluation design, coordination is important among all evaluation funders to clarify expectations on issues such as data ownership, periods of observation, sample sizes, and reporting.

Conclusion

This paper discusses several unique practical considerations for PFS evaluations. Navigating these considerations is part of the role of the evaluation partner throughout the project. By understanding and addressing these practical considerations, evaluation partners can help ensure the evaluation design strengthens the PFS project. PFS evaluations, and the authors' own experiences, are still few enough that the lessons discussed in this paper are not yet necessarily best practices or models; there are certainly more lessons to be learned. However, documenting and learning from the lessons of each project is important so that evaluators continue to strengthen and smooth the path forward for future PFS evaluations.

Our experiences with the practical considerations discussed in this paper suggest the following recommendations for those involved in PFS evaluations:

- **Bring the evaluation partner on board as early as possible** to help ensure that important decisions in the early development phases are grounded in evidence.

- **Provide time and support for an evaluation planning period** so that the most rigorous evaluation design possible can be created with full buy-in from all key PFS partners.

- **Provide time and support for an implementation pilot period** to minimize implementation failure and to test key implementation assumptions, such as referral processes, program flow, and take-up rates.

- **Refer to other PFS contracts** to adapt language that builds in evaluation flexibility and learn from other evaluation structures.

- **Consult with the independent validator as early as possible** to address any potential analytical or IRB concerns before implementation begins.

- **Create a process for constant communication between the evaluator and other PFS partners** to coordinate on implementation or evaluation challenges as soon as they arise, especially during early implementation.

Notes

1. In TOT, we include the local average treatment effect, which is an estimation procedure that controls for the selection bias present in standard TOT analysis.

2. For example, if the analysis plan proposes a linear probability model for a binary outcome, assuming that the outcome of interest will not have a mean value in the tails, it would be prudent to specify that a linear probability model will be used if the outcome measure has a mean between .10 and .90, but a logit model will be used if the outcome measure has a mean outside of that range.

3. Individuals in the control group who receive the treatment.

About the Authors

Sarah Gillespie is a research associate in the Metropolitan Housing and Communities Policy Center at the Urban Institute. Her research focuses on housing and homelessness, place-based initiatives, and performance measurement. Her work currently includes an evaluation of a supportive housing social impact bond for frequent users of the criminal justice system in Denver and a five-site national evaluation of supportive housing for families involved in the child welfare system.

Akiva Liberman is a senior fellow in the Justice Policy Center at the Urban Institute. His primary interests concern interventions to reduce juvenile delinquency, evaluation methods, and the conditions of effective program implementation. He currently leads multisite evaluations of Juvenile Second Chance Act Reentry Demonstration Projects and OJJDP's Juvenile Justice Reform and Reinvestment Initiative, as well as an evaluation of Restorative Justice in Rhode Island Schools. He has been involved in two PFS projects with juvenile justice youth, as coinvestigator for Urban's evaluation of a PFS-funded intervention for reentry youth in Massachusetts, and as principal investigator for evaluation planning for New York State's community-based alternative to out-of-home placement for adjudicated youth.

Janine Zweig is a senior fellow in the Justice Police Center at the Urban Institute. Her work focuses on vulnerable populations and addresses issues related to violent victimization (particularly intimate partner and sexual violence), substance use, populations returning to communities from incarceration settings, and adolescent and young adult development. She has evaluated provisions of the Juvenile Second Chance Act, the Violence Against Women Act, and the Prison Rape Elimination Act, as well as many other policy issues. She is a co-investigator for Urban's evaluation of a PFS-funded intervention for reentry youth in Massachusetts.

Devlin Hanson is a research associate in the Center on Labor, Human Services, and Population at the Urban Institute. Hanson has substantial experience analyzing complex data sets and conducting impact evaluations. She has worked on both randomized controlled trial and quasi-experimental evaluations of supportive housing for child welfare–involved families. She also currently serves as the outcomes study

lead for the randomized controlled trial evaluation of the Denver Supportive Housing Social Impact Bond Initiative.

Mary Cunningham is a senior fellow in the Metropolitan Housing and Communities Policy Center at the Urban Institute. Her research focuses on homelessness, housing instability, and concentrated poverty. Cunningham leads studies examining the impact of housing vouchers on child welfare involvement, the impact of supportive housing on high-need families in the child welfare system, and outcomes from a homeless prevention program for at-risk veterans. She is currently coprincipal investigator of the evaluation of a PFS project providing supportive housing to homeless frequent users of services in Denver.

Michael Pergamit is a senior fellow in the Center on Labor, Human Services, and Population at the Urban Institute. He is a labor economist whose research is focused on vulnerable youth and families, particularly youth aging out of foster care, runaway and homeless youth, and disconnected (opportunity) youth, as well as access and integration of services. His research uses experimental, quasi-experimental, and nonexperimental methods. In particular, he has extensive experience leading randomized controlled trial program evaluations, including as coprincipal investigator of the evaluation of a PFS project providing supportive housing to homeless frequent users of services in Denver.