ISSUE BRIEF 6

Measuring Success

SOCIAL FINANCE

Given that payment in Pay for Success (PFS) projects is linked to performance metrics, including outcomes, choosing the right method to measure those metrics is crucial to project design. In this brief, we weigh the merits of different types of evaluations for PFS projects.

About This Issue Brief Series

This issue brief is the first in a 10-part series written for government officials interested in learning how to use Pay for Success tools and principles.

The series summarizes best practices and lessons learned at Social Finance from a decade of designing, launching, and managing Pay for Success projects. It includes guidance on each step of the process, from deciding whether Pay for Success is a good fit to actively managing a project post-launch.

Access the complete issue brief series here.

Choosing an Evaluation Method that Matches Project Context and Goals

The type of evaluation chosen for a PFS project will dictate how performance metrics are measured; ultimately, these results will inform how much a government—and therefore, its taxpayers—will pay for the project. But the right choices when it comes to measurement are not often obvious: Should we track changes among program participants over time? Should we measure the project's results against a historical baseline, or compare to another group of similar individuals? Are there ethical considerations that make certain methods more or less feasible? Are there metrics to measure for learning purposes beyond those linked to payment?

An effective selection process should pragmatically weigh the learning priorities and operational tradeoffs of the project partners in determining the most appropriate evaluation design.

PFS Project Evaluation Methods

PFS project developers have a wide range of evaluation methodologies to choose from, ranging from low-cost, nonexperimental designs to more rigorous and costly experimental ones. Regardless of the methodology selected, project developers should ensure that they put in place appropriate safeguards and think through backup methodologies in the case of unforeseen measurement obstacles or changes.

NONEXPERIMENTAL DESIGNS

Nonexperimental designs seek to calculate participants' performance metrics without measuring them against a comparison group. Nonexperimental designs are often simpler, less time-consuming, and less expensive to implement; they are easy to explain and easy to conduct. On the other hand, nonexperimental evaluation designs cannot isolate the causal impact of the PFS project, i.e., the design cannot account for what would have happened in absence of the project. While nonexperimental designs are a valuable method to track and understand project performance, they also have limitations relative to more rigorous methods.

An example of a nonexperimental design is pre-post analysis, in which evaluators calculate the change in outcomes for program participants from before the intervention to after the intervention is completed. This method can suggest a change to a participant's life trajectory, but does not demonstrate causality of program impact, control for factors external to the intervention that may be affecting participant outcomes (such as changes in economic conditions), or account for any changes that would have occurred naturally over time in absence of the intervention.

Another nonexperimental design method is validating implementation and outcomes' results relative to program targets. The involves defining metrics along a program's theory of change—from inputs to outputs to outcomes—and the associated expected targets for each metric. The evaluator would then calculate the level of achievement for each of those metrics. This method helps identify whether a program is meeting expectations along each stage of its theory of change but does not provide causal evidence as to a program's efficacy.

QUASI EXPERIMENTAL DESIGNS

Quasi-experimental designs (QEDs) leverage a non-randomized comparison group that is, ideally, nearly identical to the treatment group on observable characteristics such as age, race, gender, education level, income, etc. These types of studies generally experience fewer operational challenges compared to randomized designs. However, they cannot account for differences in unobservable characteristics between groups (such as perseverance or motivation). For example, most programs are voluntary; willingness to participate in a program could itself bias participation toward favorable outcomes.

An example of a quasi-experimental design is a matched-comparison group design, which consists of pairing program participants with non-treated comparison group members using demographic variables and other baseline data. Matching can be conducted either contemporaneously (at the same time as the program) or retrospectively (after a program is completed) and does not require real-time operational changes to program enrollment. However, it does require a large dataset of potential control group members to identify comparable matches for each treatment group member, and the ability to follow those control group members over time to collect outcomes data.

Sidecar Evaluations

Some PFS projects include sidecar evaluations, where an evaluator runs an RCT to further the project's learning agenda, but no payment is attached to RCT outcomes.

EXPERIMENTAL DESIGNS

Randomized controlled trials (RCTs) are considered to be the most reliable way to determine a program's causal impact. Here, eligible individuals are randomly assigned to either a treatment group where they receive the intervention or a control group where they do not. Through an RCT, evaluators estimate the average causal effect of an intervention by calculating the difference in outcomes between the two groups. Unlike other evaluations, RCTs can control for both observable and unobservable characteristics. RCTs work best in oversubscribed interventions where a fair lottery is the most ethical way to decide who can access services. RCTs also lend themselves well to natural experiments, or situations where randomization occurs naturally.

Despite their benefits, RCTs can introduce additional costs and operational complexities. They may even introduce ethical questions if one group of people is denied access to the standard of care in the presence of adequate funding or resources. Additionally, RCTs can require a longer time horizon to calculate outcomes, which can be unappealing for project stakeholders. To circumvent this problem, interim payments can be made on shorter-term metrics such as enrollment or program completion.

Unintended RCT Impacts

Many early SIBs used RCTs under the assumption of minimal operational changes. This approach sometimes led to unintended operational consequences, such as requiring program staff to turn away eligible clients after explaining a new service, which may in turn disincentivize other eligible participants from engaging.

Dynamic policy and service delivery environments are also a hinderance to RCTs. The Covid-19 pandemic made it impossible to use the completed RCT data for the Ventura County Project to Support Reentry, a Pay For Success project launched in 2017 designed to help formerly incarcerated people reenter their communities, given the extent that both policy and service delivery changed over the life of the project. Learn more.

Evaluation Selection Factors

When considering the evaluation design for a PFS project, project stakeholders should consider policy and project priorities and adhere to first principles, as discussed in Issue Brief 4—Getting Started. In addition, project teams should give weight to several other factors.

FACTORS TO CONSIDER IN PFS EVALUATION DESIGN

Learning Objectives	The choice of an evaluation methodology should be driven by the learning priorities of project partners. Project partners should align on the key questions they want answered that are feasible to address given project parameters.
Operations	Project partners must consider whether an evaluation may cause disruptions to the normal course of service delivery, which could impact project participants and the ability of the service provider to achieve desired outcomes. The evaluation must also be appropriate for the intervention and the geography of the study. For example, some evaluations require more intensive data collection and larger target populations, which might not be realistic in certain project contexts (e.g., a sparsely populated rural setting).
Data	Stakeholders should understand which data are required for different evaluation types. If additional data would need to be collected for the evaluation beyond what is currently reported (e.g., participant survey data), it will generally require more time, effort, and funding.
Ethics	In some evaluation designs, such as RCTs, eligible individuals are randomly assigned to either a treatment group where they receive the intervention or a control group where they do not. This model is therefore most appropriate when the intervention is already oversubscribed relative to the available budget, and/or there is a lack of evidence on intervention efficacy.
Project Size	Experimental designs such as RCTs generally require enough interest in the service in order to fairly enroll enough participants for both a treatment and control group, while other nonexperimental designs can be used with a smaller number of project participants.
Timeline	RCTs and QEDs require upfront time and resources to plan for and set up the evaluation; these may not be available due to budgetary or logistical constraints. In some cases, payment may need to be made on interim evaluation results to meet the needs of project stakeholders such as policymakers and outcomes funders.
Cost	Evaluations that involve a comparison group, such as RCTs, and/or require primary data collection, are generally more costly to conduct than nonexperimental methods.

Selecting an Evaluator

An evaluator is a PFS partner without a financial stake in the project who is engaged solely for the purpose of designing an evaluation, collecting data, and evaluating project results. Typically, as a first step, project partners release a solicitation, such as a Request for Qualifications (RFQ) or a Request for Proposals (RFP) to select an evaluator. Key criteria to look for when selecting an evaluator include:

- Analytical capability: ability to implement the chosen methodology
- Subject matter experience: a track record of evaluating similar interventions in the same sector using high-quality evaluations
- Operational expertise: capacity to support performance management
- Cost: budget that aligns with available resources

It is also beneficial if an evaluator has prior experience or knowledge of PFS projects.

Evaluation Design Example: Massachusetts Pathways to Economic Advancement

In 2017, Social Finance, the Commonwealth of Massachusetts, and Jewish Vocational Services (JVS), a Boston-based nonprofit, launched the Massachusetts Pathways to Economic Advancement Project, a PFS initiative with the goal of helping limited-English speakers and recent immigrants in Greater Boston obtain higher-wage jobs and access and persist in higher education. Project partners agreed to use nonexperimental and experimental evaluation methodologies to trigger outcome payments.

The project focuses on providing four different types of services, or tracks. Project partners decided that three of these tracks should measure results using nonexperimental designs based on two overarching concerns with randomization: ethics and sample size. The first track, Rapid Employment (RE), is designed to serve newly arrived immigrants, refugees, and political asylum seekers trying to secure their first jobs in the U.S. RE clients are referred to JVS through refugee resettlement agencies and are expected to accept any reasonable job offer after their training. Given the level of urgency of the RE program, project partners agreed that refusing services to refugees to randomize them into a control group could create ethical challenges. Tracks two and three are small, cohort-based programs serving 100 to 250 participants over three years as part of the project, and their small cohort sizes made establishing control groups infeasible.

The fourth track in the Pathways project is English for Advancement (EFA), an open entry and exit program for immigrants with low to intermediate English skills. It combines in-class vocational instruction with personalized job coaching. Project partners were interested in using a randomized design to assess how scaling EFA would impact participant earnings, which had not previously been extensively evaluated. JVS also believed that it would be able to recruit a large number of participants into this track through coordinated outreach efforts and partnerships with local organizations in these new cities. Based on these goals and the feasibility of using a randomized model, project partners opted to conduct an RCT with approximately 2,000 total participants in the combined treatment and control groups.

For Massachusetts and the Pathways partners, choosing the right evaluation meant tailoring methods to each component of the overall project—monitoring progress for three tracks, while building new, strong evidence through randomization in a fourth. Learn more.

Acknowledgements

This issue brief series was made possible with funding from the Robert Wood Johnson Foundation (RWJF) as part of their work to promote cross-sector alignment to better address the goals and needs of people and communities. The views expressed here do not necessarily reflect the views of the Foundation. To learn more about RWJF's work in cross-sector alignment, visit alignforhealth.org.

About Social Finance

Authors: Former Director Emily Carpenter, former Director Rachel Levy, and Managing Director Jake Segal.

Social Finance is a national nonprofit and registered investment advisor (SF Advisors, LLC). We work with the public, private, and social sectors to create partnerships and investments that measurably improve lives. Our Impact Investment team designs, launches, and manages impact-first investments. Our Advisory team partners with government and philanthropy leaders to implement data-driven programs for social impact. And through the Social Finance Institute, we aim to build the field and change systems through actionable research, communities of practice, and educational outreach. Since our founding in 2011, we have mobilized more than $350 million in new investments designed to help people and communities realize improved outcomes in workforce and economic mobility, health, and housing.

Learn more at socialfinance.org >>

Updated June 2024.