ISSUE BRIEF 6

MEASURING SUCCESS

Given that payment in Pay for Success (PFS) projects is linked to outcomes, choosing the right method to measure those outcomes 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 one of 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.

See here to access the complete issue brief series on our website.

CHOOSING AN EVALUATION THAT MATCHES PROJECT CONTEXT AND GOALS

THE TYPE OF EVALUATION chosen for a PFS project will dictate how outcomes 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 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 outcomes 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.

EVALUATION METHODS USED IN PFS PROJECTS

PFS project developers have a wide range of evaluation methodologies to choose from, ranging from low-cost, ***nonexperimental designs*** to more precise and costly ***experimental*** ones. All else equal, when it comes to measurement, more precise methodologies are generally better. 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' outcomes 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, their results can't exclude various threats to credibility, such as the results of policy or economic shifts, and can't account for what outcomes participants would have achieved in absence of the

program. Therefore, they produce the lowest level of confidence in program strength.

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.

QUASI-EXPERIMENTAL DESIGNS (QEDs)

QEDs leverage a non-randomized comparison group that is, ideally, nearly identical to the treatment group on *observable* characteristics such as age, race, gender, 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 ***matching,*** 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.

EXPERIMENTAL DESIGNS

Randomized controlled trials (RCTs) are generally considered to be the most reliable way to determine a program's causal impact. In an RCT, 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 are most appropriate when the intervention at hand is already oversubscribed, which makes the use of a fair lottery the most ethical way to decide who can access services. RCTs also lend themselves well to "natural experiments," or situations where randomization occurs naturally (e.g., patients involved in a waitlist).

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

Despite their benefits, RCTs can introduce additional costs and operational complexities. They may even introduce ethical questions if one group of people is

SELECTING AN EVALUATOR

The 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

denied access to the standard of care in the presence of adequate funding or resources. Additionally, RCTs (like most QEDs) can require a longer time horizon, which can be unappealing for project stakeholders (and particularly for funders). To circumvent this problem, interim payments can be made on shorter-term metrics such as enrollment or program completion earlier on in the process.

CHOOSING AN EVALUATION THAT MATCHES PROJECT CONTEXT AND GOALS

In 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 PAY FOR SUCCESS EVALUATION DESIGN

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 randomized controlled trials (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.
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 quasi-experimental designs (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.
Credibility	Evaluation methodologies should seek to refute threats to credibility that could be caused by motivational, confirmation, or other biases.

PFS IN PRACTICE: EVALUATION DESIGN FOR MASS. 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 (Pathways), a PFS initiative with the goal of helping limited-English speakers and recent immigrants in the Greater Boston area make successful transitions to employment, obtain higher-wage jobs, and access and persist in higher education. During the project development phase, project partners agreed to use **both nonexperimental and experimental evaluation methodologies** to trigger outcome payments.

The Pathways 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 asylees with low English proficiency seeking to secure their first jobs in the country. 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 a refugee in order to randomize them into a control group could create ethical challenges. Occupational Skills Training (Skills) and Bridges to College (Bridges) are both small, cohort-based programs that targeted serving 100 to 250 participants over three years as part of the PFS project; 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.

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

This brief was written by Narni Summerall, Senior Associate, with support from Emily McKelvey, Thomas Coen, Rachel Levy, and Jake Segal. Social Finance is a national nonprofit organization dedicated to mobilizing capital to drive social progress. Social Finance has pioneered Pay for Success, a set of innovative financing strategies that directly and measurably improve the lives of those in need. Read more about our work in Pay for Success at socialfinance.org.

BOSTON
10 Milk Street, Suite 1010
Boston, MA 02108

SAN FRANCISCO
650 California Street, Floor 7
San Francisco, CA 94108

AUSTIN
600 Congress Avenue
Austin, TX 78701

Support provided by
Robert Wood Johnson Foundation