Meg Massey
Kelly Walsh
Justin Milner
Teresa Derrick-Mills

Evaluation Design

Pay for Success Early Childhood Education Toolkit Report #6

What Is Pay for Success?

Pay for success (PFS) offers an alternative approach to investing in the future, including early childhood education. This innovative financing mechanism shifts financial risk from a traditional funder—usually government—to a new investor, who provides up-front capital to scale an evidence-based social program to improve outcomes for a vulnerable population. If an independent evaluation shows that the program achieved agreed-upon outcomes, then the investment is repaid by the traditional funder. If not, the investor takes the loss.

For more information on pay for success, please visit **pfs.urban.org**.

About the Early Childhood Education Toolkit

This toolkit is designed to guide jurisdictions and their partners through the core elements of a PFS project in early childhood education: the existing evidence for early childhood interventions, the role of data, the measurement and pricing of outcomes, program funding and financing, implementation, and evaluation design. The toolkit includes checklists, charts, and questions for consideration, to help direct and clarify thinking around the feasibility of pay for success to scale what works in early childhood education. Together, these briefs can help jurisdictions decide if pay for success is the right approach for them—and if so, how to get started.

Acknowledgments

The Pay for Success Early Childhood Education Toolkit was conceived by a working group led by the Urban Institute and its partners at the fifth annual Clinton Global Initiative America meeting in 2015 in Denver, Colorado.

The Urban Institute is incredibly grateful to our working group partners, who lent their time and talents to the development, research, and writing of this toolkit: Accenture, Bank of America Merrill Lynch, Enterprise Community Partners, Goldman Sachs, the Institute for Child Success, Nonprofit Finance Fund, The Reinvestment Fund, Salt Lake County, Social Finance US, and Third Sector Capital Partners.

Support for the Pay for Success Initiative at the Urban Institute is provided by the Laura and John Arnold Foundation. We thank our funders, who make it possible for Urban to advance its mission.

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

Evaluation Design

For this evaluation design report, we are particularly grateful to Erica Greenberg (Urban Institute). This report draws from two previous papers published by the Urban Institute's Pay for Success Initiative: "Measuring Success in Pay for Success: Randomized Controlled Trials as the Starting Point" (Milner and Walsh 2016) and "An Introduction to Evaluation Designs in Pay for Success Projects" (Walsh, Holston, TeKolste, and Roman).

Evaluation is perhaps the most critical element in a pay for success (PFS) project's design. Evaluation measures the impact of a program on the people it serves. This is important, not just to determine whether the project met the outcome targets that form the basis for repayment, but also to help estimate whether the program itself caused those outcomes, building the underlying evidence base of the intervention. Regardless of whether the intervention achieves its outcome targets, the evaluation gives us the knowledge we need to further understand what works and what does not.

This report describes evaluation design in PFS projects, with a focus on early childhood education (ECE). It is part of a larger toolkit for states, localities, and investors considering early childhood PFS projects and is based in part on stakeholders' experiences with ongoing PFS projects.[1]

In the first part of this report, we outline the role of evaluation in a PFS project and the importance of considering evaluation design early. The second part of the paper provides a high-level overview of three categories of evaluation design, including randomized controlled trials (RCTs), which are generally considered the most rigorous evaluation design. In the final section, we review considerations for planning an evaluation in the context of ECE programs.

Why Is Evaluation Important?

Two evaluation types are important for PFS: process evaluations and impact evaluations. **Process evaluations** examine the program's outputs (e.g., the number of children who regularly attended the program) and the processes, procedures, and actions through which the work was accomplished, including whether the original plans were followed or not and why. **Impact evaluations** look at outcomes, or how the program affected participants' lives in defined ways. PFS projects generally have an impact evaluation, sometimes paired with a process evaluation.

Process evaluations yield important information for replication and expansion because they describe how the work was done. They can also inform the interpretation of impact evaluation findings. However, in this report, **evaluation** refers exclusively to impact evaluations.

Evaluation is the backbone of the PFS model: with repayment contingent on program results, how best to measure results is a critical challenge for all partners in a PFS project. If a program succeeds, the lives of its participants improve, governments repay funders, funders may earn a return on investment, and service providers benefit from an objective validation of their model. Measuring impact has definitive consequences for all PFS participants.

Evaluation should be among the first things PFS project partners discuss—a step that should engage those with formal evaluation expertise—to integrate evaluation with all other aspects of project planning. An early approach gives project partners time to ensure they have the resources to carry out the evaluation and to consider and prepare for how the chosen evaluation design might shape planning and operations.

Rigorous ECE evaluations must account for two key ECE-specific issues: accounting for the normally increasing developmental functions of children and controlling for the nonrandom choice of families to enroll in available programs. Most children increase their skills and abilities as they get older, regardless of exposure to educational interventions. Experimental and quasi-experimental designs are most effective in controlling for this normal development because they involve a matched comparison group of children the same age as treatment group children attending ECE. The second issue, selection bias, relates to observed differences between the families who enroll in ECE and those who do not. When these differences are related to later child outcomes (e.g., through parent motivation or resources), they can undermine the validity of impact evaluation results. Although all research efforts face these issues, the central place of evaluation in PFS makes attending to them particularly important.

Many ECE programs have undergone experimental, quasi-experimental, and nonexperimental evaluations. A brief history of those evaluations—including the early demonstration projects that form the basis of our understanding of the long-term impact of ECE—is included in the first report of this ECE toolkit (box 1). Ongoing investments by the US Department of Education, state departments of education, school districts, foundations, and individuals in high-quality evaluations ensure that our understanding of what works in ECE will only continue to grow.

BOX 1

EVIDENCE AND EVALUATION

*The **evidence base** for a program or model is the set of **evaluations** that measure the program's process and impact in real-world situations. The evidence for preschool and other ECE interventions comes from decades of program evaluations and meta-analyses of those evaluations, which illustrate what impact the program has on the people who participate in it (see toolkit report #1,* The State of the Science on Early Childhood Education*).*

What Are the Evaluation Design Options?

Impact evaluations are built into all PFS projects and are not specific to the outcome area. Seven of the first 11 PFS projects include an **experimental** evaluation design. The Chicago ECE PFS project is one of two PFS projects to include a **quasi-experimental** design, while the Utah ECE PFS project relies on a **nonexperimental** evaluation design to measure success and trigger repayment. The sections on quasi-experimental and nonexperimental evaluation designs draw heavily from Walsh et al. (2016).

Rigorous program evaluations seek to create a comparison group—which does not receive the treatment—to compare with the treatment group. How—and how well—evaluation designs create this comparison group is the primary distinguishing factor between the evaluation designs described in this report.

Experimental Designs

A randomized controlled trial (RCT), or an **experiment**, randomly assigns eligible children to a treatment group that receives services through the program being tested and a control group that does not. If randomization is successful, differences in outcomes measured at the conclusion of the program reflect only the direct impact of the program allowing "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 [children]" (Milner and Walsh 2016). All other competing explanations are accounted for by an RCT design. This method produces causal results with the highest confidence and clarity. Numerous RCTs have been conducted on ECE programs, although not yet as part of a PFS project.

Given the scrutiny projects often receive, there is a "strong incentive for PFS project partners to use the best possible methodology for measuring success, reducing points of possible critique" (Milner and Walsh 2016). RCTs provide the most rigorous evaluation design option.

As described by Milner and Walsh (2016):

> RCTs enable us to easily present results, provide a fair way to distribute access to a service when resources are limited, balance a focus on precision during project formation with a simple analysis on the back end, and generally offer the best way to build strong evidence on a program's effectiveness.
>
> Further, although other methods can over- or underestimate program impact, RCTs generally provide the most accurate reflection of the program's effect by comparing performance to a randomized control group. The method is not without limitations… but the strengths of RCTs are particularly well suited to the requirements of a PFS project….
>
> [Additionally,] the amount of work required by researchers to analyze and communicate results may be less than that needed under other quasi-experimental designs that require complex methodologies.

Within the context of ECE projects, properly randomizing children is complex and requires significant planning. Generalizing evaluation results may be challenging if the group of children randomized into treatment and control groups is small or atypical. Local context may also limit the generalizability of RCTs, which is why replication studies of successful programs remain important.

That said, the strengths of the RCT model are well-suited to the requirements of a PFS project. RCTs

- generally provide the most accurate reflection of the program's effect;
- reduce bias in estimating program impact;
- offer a high degree of confidence in the results;
- provide the clearest answers about program impact;
- provide all eligible children with an equal chance of receiving services;
- can cost less than quasi-experimental designs, especially if evaluators have strong expertise and there are high-quality administrative data systems; and
- build the evidence base for a particular intervention.

Given these benefits, partners in a PFS project should begin their conversations on evaluation design by first considering an RCT. Only if context and other circumstances make RCTs impractical should partners consider less rigorous designs.

Quasi-experimental Designs

Quasi-experimental evaluation designs seek to create a comparison group as identical to the treatment group as possible, in the absence of randomization. These designs attempt to minimize bias from competing explanations and use various statistical methods to approximate an RCT's rigor. Compared with nonexperimental methods, quasi-experimental designs produce results with fewer biases and increase confidence that the treatment was related to better outcomes. The Chicago ECE PFS project uses a quasi-experimental design.

Quasi-experimental evaluations can take several approaches to create a nonrandomized comparison group. These designs can control for observable factors (e.g., race, gender, or family income level) that could bias the results. They include the following designs:

- **Regression discontinuity design**, a variant of regression analysis, compares two groups whose differences are based on an eligibility threshold for the program. In ECE, this threshold is often child birth date, creating a treatment group of eligible children and a comparison group of children too young or old for the program. Because the two groups are so similar before the program, differences measured after the program can generally be attributed to the program.

- **Propensity score design** uses weighting to artificially make the comparison group more similar to the treatment group.

- **Difference-in-difference design** compares change in outcomes over time for the treatment group and compares this with change in outcomes over time for the control group.

- **Instrumental variable design** looks at a third variable's impact on program participation to understand the program's impact on the participants.

BOX 2

CHICAGO'S QUASI-EXPERIMENTAL EVALUATION DESIGN

The study design for the Chicago ECE program uses a propensity matching technique where students in the Child-Parent Center prekindergarten programs are matched with one or two students from the same age group who did not attend prekindergarten, and one or two students from the same age group who attended other prekindergarten programs through Chicago Public Schools. Data on all Chicago Public School kindergarten students are cleared of identifying information and run through a nearest-neighbor matching algorithm that finds each Child-Parent Center prekindergarten student's closest student for comparison based on 18 criteria, including gender, ethnicity, neighborhood crime rates, and the percentage of single mothers in the student's neighborhood (Emanuel 2014). The Chicago ECE PFS project uses the propensity score design to measure program impacts on short-term outcomes (e.g., avoidance of special education by children who do not need it) and long-term outcomes (e.g., high school graduation and employment outcomes).

A quasi-experimental design can account for competing factors that could influence results. However, without randomizing volunteers, it is not possible to account for all factors, including unobservable ones, that might affect outcomes and introduce bias. Preschool, because it is not mandatory, is a good example of motivation creating a potential bias: if a child has involved or motivated parents who seek to enroll him or her in a voluntary preschool program, that child may be more likely to do well later on, even if the program itself has no or little impact on developmental outcomes. In an RCT design, randomizing volunteers would account for this potential bias.

TABLE 1

Evaluation Designs of the Two Existing ECE PFS Projects

CITY: PROJECT NAME	EVALUATION DESIGN	OUTCOME FOR PAYMENT	LENGTH OF EVALUATION
Salt Lake City: Utah High Quality Preschool Program	Longitudinal study	Preschool students avoid special education placement	Four-year service delivery term; 12-year repayment term and evaluation period
Chicago: Child-Parent Center Pay for Success Initiative	Quasi-experimental with propensity score matching	Decrease in special education, kindergarten readiness, achievement of reading at grade level in third grade	Four-year service delivery term; 17-year repayment term and evaluation period

Nonexperimental Designs

Pre-post analysis and benchmark designs are **nonexperimental** evaluations because they do not include an untreated comparison or control group but instead focus on the number of participants who experience targeted outcomes. The Utah ECE PFS project uses a longitudinal nonexperimental design for its evaluation. Because these designs do not compare the outcomes of the people treated by the program with people who were not, and because they do not account for competing explanations such as broader economic trends or changing neighborhood demographics, they may lead to conclusions that could be contradicted by a more sophisticated design.

Some PFS projects have used a benchmark to determine success payments. Some PFS projects have used a benchmark, rather than a design with a comparison group, to determine success payments. Originally designed for social impact bonds in the United Kingdom, these benchmark designs—also called **rate card** designs—are lists of solely treatment group outcomes that will trigger success payments. In these designs, if the treatment-group meets a target threshold, the government repays the funder based on an agreed-upon rate.

Although this approach can be straightforward, there is no comparison group to ensure that the outcomes were caused by the program or would have occurred even in the program's absence. For example, the metric used by the Utah project for repayment was the number of children who avoided special education who did not need it. The evaluation considered this number against historical data predicting outcomes for the target population (in the absence of the program) and established the rate for payment when those outcomes were met or exceeded. This evaluation's structure accounted for the statistical likelihood of children avoiding special education that did not need it, but it made it difficult to conclude whether those children would have avoided special education no matter what. Nonexperimental designs produce results with the lowest confidence and cannot account for most biases. This is important for governments that want to know whether the program was effective and should be expanded. PFS project partners measuring success with a non-experimental design should consider the following:

- Is a more rigorous evaluation design feasible?
- Does the design have anything to compare the outcomes measured against?
- Is it important to understand program impact, or is it sufficient to quickly and clearly measure outcomes achieved?
- To what extent do the evaluation design and outcome measures reflect children makingdevelopmental gains simply because they are getting older?

BOX 3

LIMITATIONS OF RATE CARDS

Rate cards have significant limitations. Chief among them is that they sidestep rigorous evaluation. Strong evaluations, ranging from randomized control trials[a] to a variety of quasi-experimental models,[b] are important not only to establish with confidence what results have been achieved but also to see whether these results can be attributed to the program or would have happened even without the program.

This question of attribution of impact is important because it ensures governments only pay for real results achieved by the funded program. For example, a weak evaluation design might find that students in a five-year program met test score targets, but because the evaluation did not control for other factors, it could obscure the fact that the results were caused by things other than the program itself (such as improved economic conditions in the community). In this case, government is actually paying for results that would have been achieved without the program. Strong evaluations can also help inform governments on whether they should invest in a program again in the future.

In addition, the more common approach of pricing outcomes based on establishing an outcome target for the entire population (i.e., 60 percent of participants achieve outcome A) has its own benefits. By setting systemwide targets rather than individual ones, this approach links projects to the government's bigger strategic plans, better insulates governments from making payments for results that cumulatively fall short of desired objectives, and gets government and other stakeholders in the habit of assessing existing evidence of program effectiveness to set realistic results targets.

Sources: Matthew Eldridge, "How the UK pays for success," *PFS Perspectives* (blog), Urban Institute, Pay for Success Initiative, May 23, 2016, http://pfs.urban.org/pay-success/pfs-perspectives/how-uk-pays-success.

[a] "Experiments," Urban Institute, accessed October 5, 2016, http://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/experiments.

[b] "Quasi-experimental Methods," Urban Institute, accessed October 5, 2016, http://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/quasi-experimental-methods.

Core Considerations for Planning an Evaluation

Evaluation planning, including the selection of an evaluation design, is integrated into PFS project planning from the beginning. The evaluation touches on everything from service provider selection to data collection to funding. It is critical to bring someone with evaluation expertise into the fold as early as possible.

Four core considerations emerge in early planning. These considerations are interrelated, and partners may need to address them iteratively throughout their planning process.

Reviewing Previous Evaluations

One criterion for selecting an intervention is a program model's evaluation history; partners in a PFS project should select programs with a strong evidence base from a history of rigorous evaluations. (As we discuss in more detail in toolkit report #5, *Project and Performance Management*, the service provider should also have experience implementing the chosen model.) With a baseline from previous evaluations of the same intervention, governments, service providers, and funders can have realistic expectations for how effectively the project might achieve the desired outcome. Past evaluations can also be helpful for selecting pragmatic and fair metrics, including the outcomes that can be measured and the proper instruments to do so (see toolkit report #3, *Outcomes Measurement and Pricing*). Funders are likely to consider an intervention's evaluation history to determine whether they will invest in a project (see toolkit report #4, *Program Funding and Financing*). Considerations for selecting a service provider to implement the chosen model are discussed at length in toolkit report #5, *Project and Performance Management*.

Differences between the population targeted in previous evaluations and the population targeted in the current intervention are important to consider because differences may diminish how the previous findings apply to the new intervention. Studies that include large, representative samples of children are said to have high **external validity**, meaning their results can generalize to samples in different demographic and policy contexts. Differences between the children the PFS-funded program intends to serve and the children in the previous studies could affect the external validity of the previous studies' findings. These differences include the proportions of children who are dual-language learners, are in families who are much poorer or much more financially well off, have had previous preschool experiences and similar preschool alternatives, are developmentally delayed or have physical disabilities, have parents with different levels of education or types of employment, have different cultural backgrounds, have experienced traumas or high levels of stress, and have less-stable housing situations.

If a project implements a program with limited evidence—for example, a program with few evaluations or a less rigorous evaluation design—there is a greater incentive to choose a strong evaluation design to assess outcomes. That evaluation can be the next step in building an evidence base.

Identifying the Target Population

Developing parameters for the children targeted by the evaluation—in most cases, a subset of the children eligible for the program—is informed by the size of the ECE program's target population, the service provider's resources and expertise, and other considerations specific to

the jurisdiction, including how ECE programs are administrated.

The target population for the evaluation is also informed by the broader pool of eligible children targeted by the program. ECE programs generally target 3- to 4-year-olds, but beyond that, many other factors come into play: Will the program focus on young children in one neighborhood, or children throughout a city or county who meet other criteria? Will the program focus on children from a low-income or a non-English-speaking household, or children with special needs?

Insufficient target populations are more likely in rural or sparsely populated areas and among children with a special need (e.g., an uncommon disability). It may also be difficult to recruit volunteers if parents are concerned about the confidentiality of data or testing results involving their children. Because informed consent is required for experimental evaluations, some eligible participants might decline to participate, a risk that should be factored into the projections for the number of participants.

If an evaluator determines that there is not a large enough population, the project's geographic boundaries, time period for measurement, or eligibility criteria for receiving services can be expanded. But changing eligibility criteria could affect expected outcomes. If the original target population consists of children who have never participated in a formal child care program, children in the treatment group would be expected to benefit from participating in a formal program, and fewer children would be needed to detect the change. If the target population is expanded to include children who did have some formal experience, the amount of change expected would be smaller, and more participating children would be needed to detect program impacts.

When identifying a target population, partners should consider the following factors that could bias the evaluation results:

Sample size: In program evaluation, the target sample size is the number of participants that need to enroll in the study to detect an effect of a given magnitude. The sample size has two parts: the number of children who receive the service (treatment group) and the number of similar children who do not receive the service (control or comparison group). If there are too few participants in the final study sample, a large margin of error will lead to inconclusive results or the false finding that a program that had a positive effect did not work. The number of participants needed depends on what the project is trying to measure and the instruments that will be used to measure them (see toolkit teport #3, *Outcomes Measurement and Pricing*).

Selection bias: If certain children are more likely to receive the intervention than others, the results could become biased. Selection bias can come from the parents of program participants for reasons associated with their children's later performance and from the programs themselves, which may give preference to applicants most likely to achieve the outcomes tied to repayment, a process known as **creaming**. Using kindergarten readiness as an outcome on which to base repayments is a strong example of both. Without an experimental design, the families who enroll in the program may be more motivated and have more resources than families who do not enroll; these motivations and resources might cause enrolled children to do better in school, regardless of their experiences in the intervention. Similarly, centers might implement screening processes that eliminate from consideration children with behavioral or other challenges—or centers might be located in relatively more advantaged neighborhoods—leading to the selection of children more likely to be kindergarten ready from the outset of the intervention. These issues can lead to **endogeneity**, wherein researchers confuse better outcomes with a difference between who gets into a program and who does not.

Properly randomizing: Random assignment in randomized controlled trials creates two groups that have the same observable and unobservable characteristics. Because random assignment must occur after the children have been deemed eligible for the program, the motivations of parents are similar (i.e., parents of participating and nonparticipating children wanted their children to have this opportunity) and participating and nonparticipating children met all the eligibility criteria. One key challenge to properly randomizing is getting the buy-in from service providers. In a random assignment process, the people screening recruits and applicants become part of the evaluation because they can bias the selection of children. These screeners must receive training about how they are part of the evaluation team and must understand how and why the randomization process is ethical and fair:

- Randomly determining who receives services is fairer than awarding slots only to those at the top of a waiting list. Whether the community randomizes or not, there will not be enough funds to provide the service to every family who wants it, so randomization does not decrease the number of children served. Anyone applying during the application period may or may not get into the program; in a traditional waiting list approach, the most organized or most eager parents will apply first and get their children into the program, leaving the least organized and sometimes the most needy on the waiting list.

- More children (not fewer) are being served because most PFS projects aim to increase the number of children receiving services. Children not receiving the services are not missing out; the children receiving the services are receiving something they would

not otherwise have received without the evaluation. Further, randomization creates a robust research base from which to argue for increased funding in an effective program. Evaluators should consider the typical enrollment process in the communities and organizations participating in the evaluation to determine the best way to integrate randomization into enrollment. If enrollment occurs in multiple schools, organizations, or neighborhoods, it is important to consider balancing randomization across those locations and relying on randomization tools.

Unobserved effects or omitted variable bias: Genetics, experience, or other factors not considered when selecting people to receive the intervention might predict successful outcomes regardless of a given intervention. Statistical models can only control for characteristics that can be quantified; much of what determines success is difficult to measure, including children's prior and contemporaneous early learning experiences, home resources, and parent motivation. Caution interpreting results is wise.

Defining Metrics to Be Tracked and Analyzed

Selecting outcomes is discussed at length in toolkit report #3, *Outcomes Measurement and Pricing*. Once outcomes are selected, it is important to identify the resources to properly measure them. Partners may decide to pair the impact evaluation with a process evaluation. This structure allows intermediate success payments to be made based on interim indicators or outputs (e.g., the number of children to complete a program) before the completion of the impact evaluation. This strategy can engage partners on short-term "wins" while emphasizing long-term outcomes.

Measurement issues to consider when assessing early childhood education outcomes include the following:

1. How much time is needed to demonstrate results?
2. How much of the intervention do the children need to receive to make a difference?
3. Who is doing the measuring before and after the intervention?
4. What kinds of training and inter-rater reliability checks will be employed?
5. Will measures be executed using pencil and paper, tablets, or another mode?
6. In quasi-experimental or non-experimental evaluations, how will a representative control group be identified and assessed?
7. What kinds of data review and cleaning procedures need to be established?
8. How will the evaluation team ensure the privacy and confidentiality of children assessed?

Process and impact evaluations rely on accurate and complete data collection and analysis. Clearly understanding local data systems and the reliability of data in those systems is critical for a successful evaluation. Toolkit report #2, *Using Data to Inform Decisionmaking* details strategies for identifying and using appropriate data systems to ensure that the evaluation is as accurate as possible.

Covering Evaluation Costs

Evaluations should cost anywhere from 0.5 to 15 percent of the program cost. These costs include the labor and expertise to plan the evaluation, collaborate with stakeholders, collect and analyze data, and report results.

Costs are inherent to all evaluation designs, not just the most rigorous ones. Guarding against **evaluation risk** (i.e., the odds that the project will fail to meet its planned outcomes) is a worthy investment. Selecting a rigorous evaluation design, ensuring data readiness, and contracting a reputable evaluator can mitigate evaluation risk and attract potential investors. The more rigorous the design, the more likely the cost will be outweighed by the benefits that come from increased clarity and confidence in the results. Particularly given the relative newness and attendant scrutiny of PFS projects, ensuring that the evaluation is as rigorous as possible allows the results to stand up to critique.

Evaluation costs can be covered by the project's funders, but often, additional nonprincipal investors—usually philanthropies—cover costs. For example, a grant from the Laura and John Arnold Foundation covered the costs of the randomized controlled trial for Santa Clara County's Project Welcome Home.[1] In the Chicago ECE project, the Finnegan Family Foundation provided funds for project evaluation.

Evaluation costs can be reduced by selecting an independent evaluator with expertise in that program and reducing the need for new primary data collection by relying on existing data systems (see toolkit report #2, *Using Data to Inform Decisionmaking*).

[1] Fact Sheet: Santa Clara County's Project Welcome Home, http://www.thirdsectorcap.org/wp-content/uploads/2015/08/150811_SCC-CH-PFS_Fact-Sheet.pdf

Conclusion

Because of its core role in PFS, a rigorous evaluation design is an imperative not be an afterthought. Evaluation design shapes many of the operational considerations that come with implementing a PFS project and thus must be discussed up front. Engaging an evaluator at the onset of a project design sets the stage for ensuring that the project's impact is captured as thoroughly and accurately as possible, and with minimal disruption to the other partners involved in the project's implementation.

Reference

Emanuel, Rahm. 2014. *Loan Agreement and Contract with IFF Pay for Success I LLC to Serve At-Risk Children to Increase School Readiness and Reduce Later Public School Spending*. Chicago: Office of the City Clerk. http://www.payforsuccess.org/sites default/files/o2014-8677.pdf.

Milner, Justin, and Kelly Walsh. 2016. "Measuring Success in Pay for Success: Randomized Controlled Trials as the Starting Point." Washington, DC: Urban Institute. http://urbn.is/2dg3oVl.

Walsh, Kelly, Rebecca TeKolste, Ben Holston, and John Roman. 2016. "An Introduction to Evaluation Designs in Pay for Success Projects." Washington, DC: Urban Institute. http://urbn.is/2deENh8.