From Evidence to Outcomes

Using Evidence to Inform Pay for Success Project Design

Justin Milner and Matthew Eldridge

May 2016

The public and private sectors are showing tremendous interest in a new financial mechanism called pay for success, or PFS.[1] At the heart of all PFS projects is testing 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 cost savings, families and society benefit from better outcomes, and social service providers strengthen the case for funding their model.

With the results of these projects determined by the performance of the chosen program, a fundamental question comes to the fore: how do we know what will work?

While there is no simple answer, a growing body of social science research points to one critical element: strong evidence. For public and private leaders interested in PFS, consulting the evidence base for compelling social programs is essential for developing a strong project.[2]

In this brief, we offer policymakers and PFS stakeholders principles to help them understand and interpret evidence and lay the groundwork for incorporating evidence into broader public decisionmaking. We explore the following questions:

- What is evidence?
- Why does evidence matter for PFS projects?
- How do you assess the quality of existing evidence and program evaluations?
- What do you do when the evidence base is limited?

While this brief looks retrospectively at the evidence base for social programs, two accompanying briefs look prospectively at selecting strong evaluation designs for a PFS project to determine payments and to continue building the evidence base for other programs (box 1).

BOX 1

Urban Institute Briefs Exploring Evidence and Pay for Success

- "From Evidence to Outcomes: Using Evidence to Inform Pay for Success Project Design" explains the role of evidence and its importance for program selection in PFS projects.
- "Understanding Evaluation Design in Pay for Success Projects" discusses the importance of rigorous evaluation designs to PFS projects and outlines key considerations stakeholders should keep in mind when selecting one.
- "Measuring *Success* in 'Pay for Success': Randomized Controlled Trials as the Starting Point" explores in greater depth the value of randomized controlled trials in the context of PFS projects and considers potential opportunities, challenges, and solutions related to their use.

What Is Evidence?

Evidence is "the available body of facts or information indicating whether a belief or proposition is true or valid."[3] In social science, that available body of facts (the **evidence base**) is built through testing and evaluating hypotheses linking a specific intervention with target outcomes.

Researchers look to establish evidence of impact through rigorous evaluations. As Tatian (2016, 3) writes, impact is "the net effect of a program relative to what would have happened had the program not existed. In other words, the impact should be changes in outcomes *produced by the program alone* and not caused by other factors" (italics in the original).

Impact evaluations provide evidence of a program's effect on the people, families, or communities it serves and contribute to a program's evidence base. As with all areas of scientific inquiry, an evidence base is not static; it is subject to interpretation and revision as new testing and evaluation update prior knowledge and understanding. Evaluations of current programs extend the reach of what we know.

The evidence bases for specific programs, or program categories, vary in strength. Most government-funded programs have not been rigorously evaluated to demonstrate whether their activities (e.g., after-school tutoring) lead directly to the intended outcomes (e.g., improved academic achievement).[4] This does not mean that such programs are ineffective, but it does mean that we are in the dark about their true impact.

In contrast, programs that have undergone rigorous evaluation develop an evidence base that helps us better estimate their likely impact (box 2). Evidence-based program clearinghouses distinguish those programs using such monikers as model programs, top tier, and effective.[5] Though the list of programs with strong evidence bases is still short, growing investments in evaluation from government and philanthropy will help increase our knowledge of what works.

BOX 2

Building the Evidence Base for the Nurse-Family Partnership

Nurse-Family Partnership (NFP) is an innovative nonprofit community health program that provides home visits by registered nurses to low-income first-time mothers. NFP aims to improve pregnancy outcomes, prevent child abuse and neglect, improve school readiness, and increase maternal well-being.

NFP has been extensively evaluated and is considered highly effective by evidence-based program clearinghouses. NFP has undergone multiple rigorous domestic randomized controlled trials over several decades. Because the studies tracked many outcomes over many years, the program has built a strong evidence base for the approach.

For more information on NFP's evidence base, see http://toptierevidence.org/programs-reviewed/interventions-for-children-age-0-6/nurse-family-partnership.

Why Does Evidence Matter for Pay for Success Projects?

The ability to harness evidence and knowledge about what works in order to make social welfare investments improves government decisionmaking. PFS projects are compelling because they push policymakers and stakeholders to consider evidence up front by outlining the target outcomes, project terms, and program design before funding.

Evidence should help drive PFS project development in three important ways:

- **Guide project scoping.** Stakeholders should understand their target population and its needs. In the early stages of PFS project development, stakeholders should explore opportunities to align the specific needs of vulnerable populations in their area with programs that can best address those needs, based on available evidence. Stakeholders should review local data to examine the issues (e.g., juvenile delinquency, truancy, or substance abuse) that a community, city, or state is working to address. A better understanding of the issues—including the prevalence and the demographics of the affected populations—helps shape discussions and inform program selection and project design. Programs for certain issues and populations that have strong evidence bases can start stakeholders thinking about which programs a PFS project might fund.

- **Inform program selection.** After identifying an issue, stakeholders should explore programs to improve target outcomes. Even the best social programs are unlikely to work for all recipients. Nonetheless, programs with stronger evidence increase the probability of overall effectiveness and decrease the risk of ineffectiveness. Stakeholders should focus on programs with strong track records and clear measurable outcomes. One helpful resource can be evidence-based program clearinghouses that collect resources on what does and does not work.[6] These clearinghouses distinguish between programs that have undergone rigorous evaluation (e.g., experimental or quasi-experimental designs) and those that have been examined through less rigorous approaches (e.g., pre- and posttests, difference-in-difference comparisons).

- **Influence terms of the deal.** Programs with existing evidence also allow for greater confidence when stakeholders estimate likely impact, thus helping set and price outcome targets for the PFS project. For example, if a program's evidence demonstrates it has reduced teen pregnancy rates by 40 percent compared with a control group,[7] an outcome target of 40 percent reduction is a good metric to start the conversation among stakeholders. When the same program has been evaluated multiple times, meta-analysis can be used to develop better impact estimates.

How Do You Assess the Quality of Existing Evidence?

Not all evidence is created equal. An informed consumer of evidence should be aware of issues that could undermine the conclusions of certain evaluations. Here, we'll focus on causality (did the program *cause* the outcomes?) and external validity (are the results from the evaluation specific to the context?).

In many cases, governments, funders, and others are presented with an existing social program and need to apply these concepts to determine whether the program has a sound evidence base (box 3).

Causality

Evaluations don't just measure whether outcomes are achieved; they try to establish causality, estimating with a degree of confidence that the outcomes produced are attributed to a specific programmatic intervention. How do evaluations achieve that confidence? Ideally, programs would assign the same people to two groups: the treatment group, which receives services from the program, and the comparison group, which does not. In reality, people cannot be in two places at once, and researchers have to create as strong a counterfactual as possible through different research designs.

The most rigorous way to test the impact of a program is to develop a strong comparison group. Under this type of design, the difference in outcomes between the treatment and comparison groups is the impact of the program. A companion brief will discuss the merits and weaknesses of different evaluation designs' approaches to creating this comparison group. In sum, the more the comparison group resembles the treatment group, the more certain we can be that we are measuring the true causal effect. Randomized controlled trials, for example, randomly select individuals to go into the treatment or control group. In such trials, researchers are able to better isolate a program's impact.

Findings from evaluations that use other methods to approximate the comparison group should be viewed with more caution because of bias. Two common identification problems are especially important.[8]

BOX 3

Questions to Assess the Strength of a Program's Evidence Base

Suppose you represent a local government and have been presented with a potential pay for success project. The program is a family-based therapy model for juvenile offenders to prevent reoffending, and you want to know if it will work in your jurisdiction. This program has had a nonexperimental evaluation that appears to show significant reductions in the number and nature of reoffending among participants relative to the comparison group.

Several key questions can help interpret the strength of this evidence to your context:

- Did the study control for self-selection bias (e.g., the offenders most serious about reforming opted in to the program, biasing the results)?
  - If the study did not control for self-selection bias, the sample may have only included the participants with the highest likelihood of success.
- Did the study develop an appropriate comparison group to measure the outcomes of the treatment?
  - If a strong comparison group is not selected, there is no good baseline to compare the outcomes of the treatment group against, and it's difficult to say whether the program *caused* the outcomes.

Additionally, you should consider the context where the previous study was conducted and ask questions to compare with your context:

- Is the target population the same as or similar to the population previously studied?
  - Different populations present different challenges and qualities that can affect the success or failure of a program. What works with one group may not work with another.
- Do the targeted problems have similar characteristics?
  - If the problem targeted by the evaluated program differs significantly from the problem in your locality, the required solution may also differ.
- Are the components of the program the same?
  - Some solutions may be superficially similar but differ in important ways (e.g., offering group counseling versus one-on-one counseling services).
- Does the implementing service provider have similar capacity?
  - Provider capacity is a critical determinant of program delivery success, particularly when the solution is a complex and resource-intensive program.

The more of these questions you can answer with "yes," the more confident you can be about the ability to replicate the results in your context.

First, many characteristics of programs and participants may be difficult to observe. These uncontrolled or unobserved variables can distort evaluation findings and statistical conclusions. For example, highly motivated parents may sign their children up for an after-school tutoring class. If those children perform better at school, the impact might be a result of the classes, or it might simply be the result of those children having highly motivated parents. This **self-selection bias** can complicate our ability to say that a program works or not.

Second, without strong testing, it can be hard to determine the direction of causality between a program and outcomes. **Reverse causality** occurs when the relationship between cause and effect may be contrary to expectation. Imagine a health insurance program with voluntary enrollment. Those who join the program may be those in greater need of health care (e.g., already ill or high-risk individuals). If an evaluation of health outcomes compares participants with nonparticipants, it might show worse outcomes for participants and you could incorrectly conclude that the program *caused* the worse health outcomes when, in reality, this is a case of reverse causality.

External Validity

Stakeholders should also consider whether the results of an evaluation were context specific. Can the results of a specific study be generalized to a broader population that could include more people, different people, or other geographic areas? Researchers consider results that can be generalized **externally valid**.

This question is critical for PFS because most projects borrow a program that may have worked in a different place or with a different population. For instance, a program with positive effects on adults may not work with adolescents. Similarly, a program initially delivered in a secure setting (e.g., jail, detention) may yield different results when delivered in a community setting. Stakeholders on the first PFS project in the United States at Rikers Island in New York City dealt with significant issues adapting their chosen program to the prison context (Berlin 2016).

What Do You Do When the Evidence Base Is Limited?

In an ideal world, stakeholders could choose from many effective programs and select the one that best matches their needs and context. However, in many cases, potential programs have limited evidence of effectiveness, and stakeholders are uncertain about the likelihood of the program's success. In these cases, you can take other steps to assess and manage risk.

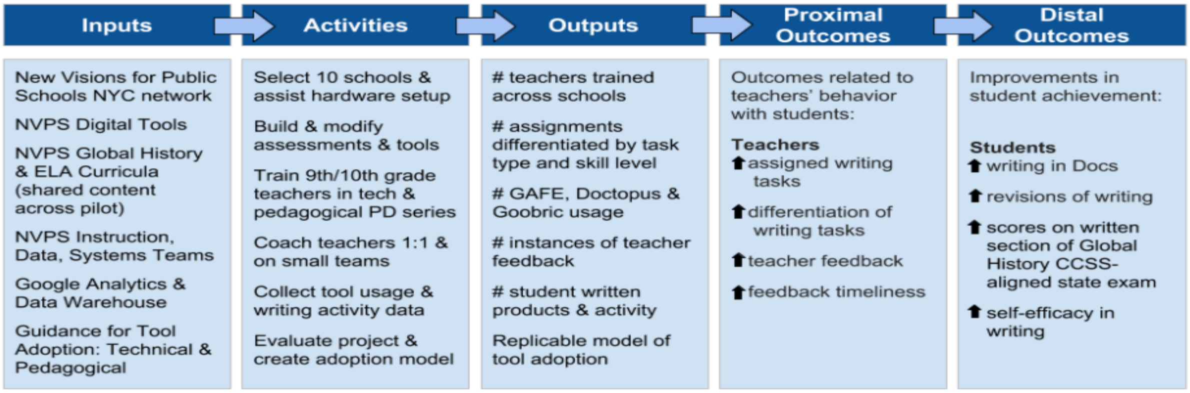
Assess Other Characteristics of Program Strength

Even when programs don't have a track record of performance as measured by strong evaluations, assessing a few key characteristics will improve your understanding of the program's strength and the potential for successful implementation. First, consider whether the program has a clear and compelling theory of change (or, more formally, a logic model). A theory of change articulates how the program intends to yield the desired outcomes through activities and outputs (box 4). Strong theories of change clearly answer questions about the target population, the target results, the period, the programmatic activities, the program context, and the core hypotheses and assumptions behind the work.[9]

BOX 4

Theory of Change for an Innovative Education Program

In partnership with the New York City Department of Education, the nonprofit New Visions for Public Schools developed a program to “leverage technology to support writing instruction aligned with the Common Core State Standards.” This program applied for and won a 2015 Department of Education Investing in Innovation Fund development grant to test and evaluate the model. The program, despite lacking an evidence base, won the funding partly by demonstrating that it was grounded in a logical model that clearly explains how it anticipated the program’s inputs and activities would lead to desired outcomes in the target population.



Inputs	Activities	Outputs	Proximal Outcomes	Distal Outcomes
New Visions for Public Schools NYC network NVPS Digital Tools NVPS Global History & ELA Curricula (shared content across pilot) NVPS Instruction, Data, Systems Teams Google Analytics & Data Warehouse Guidance for Tool Adoption: Technical & Pedagogical	Select 10 schools & assist hardware setup Build & modify assessments & tools Train 9th/10th grade teachers in tech & pedagogical PD series Coach teachers 1:1 & on small teams Collect tool usage & writing activity data Evaluate project & create adoption model	# teachers trained across schools # assignments differentiated by task type and skill level # GAFE, Doctopus & Goobric usage # instances of teacher feedback # student written products & activity Replicable model of tool adoption	Outcomes related to teachers’ behavior with students: **Teachers** ⬆assigned writing tasks ⬆differentiation of writing tasks ⬆teacher feedback ⬆feedback timeliness	Improvements in student achievement: **Students** ⬆ writing in Docs ⬆ revisions of writing ⬆ scores on written section of Global History CCSS-aligned state exam ⬆ self-efficacy in writing

Source: New Visions for Public Schools (n. d.).

Another key factor in the program’s potential for impact is the strength and capacity of the proposed service provider. As the entity delivering the program, the provider has a major influence on the program’s success. Strong providers possess characteristics such as a history of implementing similar programs with the same target population, demonstrated organizational capacity, and senior leadership commitment. These issues are important even for programs with strong evidence bases and a history of rigorous evaluation. In those cases, providers with the capacity to implement the program with fidelity and quality will be important to achieving the target outcomes.

Finally, programs that have not undergone rigorous evaluations with control groups can still have data on pre- and post-intervention conditions for the populations with whom they are working. While this pre- and post-tracking will not be able to discern the impact of the program, trend lines can signal that the program is working as expected.

Reevaluate the Risk Level of the Project

PFS projects are composed of various components working together to achieve an outcome for a target population. Those components (e.g., political will, government capacity, provider quality) each carry some risk, which makes up the overall risk that the project will not reach its target outcome. A fundamental source of risk in a PFS project is programmatic—that is, will a program work as intended for whom it is intended? One benefit of evidence is that it helps decrease program risk; if you have strong evidence that the program works, the risk that it won't perform is diminished (though not eliminated). The level of risk will have implications for attracting funders and for program design.[10]

Consider Effect Size

Where evidence is limited, stakeholders may want to place their PFS bet on a program or type of program that offers the largest potential impact (**effect size**). Rather than focusing on proven programs (though that is an important place to start), PFS projects should also be considered mechanisms for testing promising innovations that may deliver greater impact than existing programs. Programs with limited to no evidence *and* small potential effect sizes may not be impactful enough to merit the risk from the perspectives of both governments and funders.

Embed Strong Evaluation on the Back End

Although a program may lack a strong evidence base, PFS project leaders can build a strong evaluation of the program. This ensures the program's outcomes are rigorously evaluated to fulfill a commitment to all parties and to help build the program's evidence base.

PFS projects can help develop high-quality evidence through rigorous evaluation. Building strong evaluation designs (randomized controlled trials, in particular) helps expand our collective knowledge base about what works best, for whom, and under what circumstances.

Conclusion

Using evidence to make public welfare decisions improves government effectiveness and drives better outcomes for society. Pay for success can be on the front line of that change, helping improve the use and availability of evidence by making stakeholders consumers and generators of evidence.

For consumers, prior evidence on program impact should help guide project decisionmaking. Understanding what evidence exists and how to assess program strength is important for PFS projects and evidence-based policymaking more broadly. Stakeholders can also generate evidence by ensuring rigorous, objective, and transparent research designs are used to evaluate the impact of their programs. Information about effectiveness generated by such evaluation helps create or build a program's research base and continue a virtuous cycle of evidence building that can serve communities in the future. PFS projects have great potential to elevate and institutionalize the use of rigorous evidence.

Notes

1. PFS shifts risk from a traditional funder (usually a government) to a new funder (usually a private investor or philanthropy). That funding pays for services up front to improve outcomes for a vulnerable population. If an independent evaluation shows that the program achieved the target outcomes, the traditional funder repays the new funder's investment with interest.

2. In this brief, "PFS project" refers to a deal and its accompanying activities, "program" refers to a specific intervention addressing the social problem (e.g., permanent supportive housing), and "outcomes" refers to the specific social welfare improvements that determine a project's success.

3. *Oxford Dictionaries*, s.v. "evidence," accessed April 13, 2016, http://www.oxforddictionaries.com/us/definition/american_english/evidence.

4. See "Moneyball for Government: The Book," Results for America, accessed April 12, 2016, http://moneyballforgov.com/moneyball-for-government-the-book/.

5. For examples of the various terms, see Blueprints for Healthy Youth Development, www.blueprintsprograms.com; Top Tier Evidence, www.toptierevidence.org; and the Substance Abuse and Mental Health Services Administration's National Registry of Evidence-Based Programs and Practices, http://www.nrepp.samhsa.gov/01_landing.aspx.

6. These databases include a benefit-cost policy analysis tool produced by the Washington State Institute for Public Policy and an interactive clearinghouse developed through a partnership between Pew Charitable Trusts and the MacArthur Foundation. In addition, there are several issue-specific evidence-based policy clearinghouses, including Blueprints for Healthy Youth Development, California Evidence-Based Clearinghouse for Child Welfare, Coalition for Evidence-Based Policy, CrimeSolutions.gov, National Registry of Evidence-Based Programs and Practice, Promising Practices Network, What Works Clearinghouse, and What Works in Reentry Clearinghouse.

7. "Treatment Foster Care Oregon (formerly MTFC)–Top Tier," Coalition for Evidence-Based Policy, accessed April 12, 2016, http://evidencebasedprograms.org/1366-2/multidimensional-treatment-foster-care.

8. Other evaluations include regression discontinuity design, the difference-in-difference method (which compares treatment and control groups that are not randomly assigned), and historical baseline analyses (which compare the treatment group's outcomes to outcomes for a similar population in the past).

9. Matthew Forti, "Six Theory of Change Pitfalls to Avoid," *Stanford Social Innovation Review*, May 23, 2012, http://ssir.org/articles/entry/six_theory_of_change_pitfalls_to_avoid.

10. In financial transactions, funders expect higher returns to compensate for additional risk. Therefore, reducing programmatic risk may allow governments to offer lower rates of return. In PFS, this issue is complicated by other motivating factors on the part of funders and risk-distorting effects of philanthropic funders (i.e., through catalytic first-loss capital and credit enhancements). As the PFS landscape broadens and the number of projects increases, the terms of these projects should become standardized and the ability to price risk should improve.

References

Berlin, Gordon L. 2016. *Learning from Experience: A Guide to Social Impact Bond Investing.* New York: MDRC. http://www.mdrc.org/sites/default/files/Learning_from_Experience_SIB.pdf.

New Visions for Public Schools. n. d. "Personalization at Scale: Technology Integration to Drive Common Core Writing." Washington, DC: US Department of Education. http://www2.ed.gov/programs/innovation/2015dev/newvisionsnarr.pdf.

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

About the Authors

Justin Milner is a senior research associate in Urban Institute's Policy Advisory Group. He is also a managing director of the Pay for Success Initiative. His work focuses on the intersection of research, policy, and practice; supporting efforts to engage effectively with policymakers and practitioners in the application of research findings; and the development of new evidence. His past experience includes roles at the Annie E. Casey Foundation and the US Department of Health and Human Services. He received a BA in political science from Yale University and an MPA from the Woodrow Wilson School at Princeton University.

Matthew Eldridge is research products manager of the Urban Institute’s Pay for Success Initiative. He is interested in how finance and economics, specifically public-sector finance, interact with public policy. Before joining Urban, he worked at the World Bank and as a consultant on financial services policy and regulatory issues. He earned his BA from Virginia Tech and his MS from the London School of Economics – both in international development.

Acknowledgments

This brief was funded by the Laura and John Arnold Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

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

The authors thank Erika Poethig, John Roman, and Pamela Blumenthal for commenting on earlier versions of this brief.

URBAN INSTITUTE

2100 M Street NW
Washington, DC 20037

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.