Very Preliminary. Please do not quote or disseminate further without permission.

The Economics and Econometrics of Social Impact Bonds

Avi Feller and Jeffrey B. Liebman

March 2014

1. Introduction

Governments spend billions of dollars each year purchasing social services from service providers. In most cases, rigorous evaluations of these programs have produced far from stellar results. For example, Baron and Sawhill (2010) review the 10 broad-based federal social programs that have been evaluated through a randomized experiment since 1990. They report that 9 evaluations --of programs ranging from job training to early childhood education -- found "weak or no positive effects," while one, Early Head Start, showed "meaningful, though modest, positive effects." More systematic meta-analyses, in issue areas such as job training, recidivism, and welfare-to-work, have also found disappointing results.[1] While particular interventions have demonstrated their effectiveness on small populations (Perry Preschool served 58 children, the Nurse Family Partnership evaluation in Elmira served 216 people), there are very few examples in which governments have procured social services for broad populations and rigorously demonstrated the ability to move the dial on a significant outcome.[2] While it is likely that there are some highly effective programs that simply have not been evaluated, the enduring challenges of recidivism, disconnected youth, kindergarten readiness, early

[1] For example, see Greenberg, Michalopoulos and Robins (2006) and Ashworth et al (2004).

[2] In contrast, government appears to be quite effective at solving problems where writing checks is the solution. For example, Social Security has nearly eliminated poverty among the elderly (Meyer and Sullivan, 2010; Engelhardt and Gruber, 2006), and Medicare and Medicaid have reduced the risk associated with out-of-pocket health-care expenditures (Finkelstein and McKnight, 2008).

childhood education, diabetes, among many others, suggest the continued need for additional innovation in how we tackle social problems.

Since 2010, governments in the UK, US, and Australia have been experimenting with a new way of procuring social services. This new approach has two core features. The first is the use of high-powered "pay-for-success" performance contracts between the government and the private sector to obtain social services. Under these contracts, the government pays entirely or almost entirely based upon the outcomes achieved by the social services rather than paying for the services themselves. If impact on the outcomes is not achieved, the government does not pay. The second feature is that private investors, both philanthropic and commercial, provide the operating capital for the social service providers and absorb most of the financial risk associated with the uncertain performance payments. A "social impact bond" or SIB is the term used for the arrangement through which the private investors finance the service delivery in exchange for the right to receive the government performance-based payments if the intervention is successful.

In order to better understand how governments can foster social innovation and improve the results they obtain with their social spending, we, through the Harvard Kennedy School Social Impact Bond Technical Assistance Lab (SIB Lab), have been providing pro bono assistance to eight states and two cities that are developing pay-for-success/social impact bond initiatives. These include New York State which launched a $21 million initiative in December 2013 that is delivering job training services to men who have recently been released from state prison, and Massachusetts which launched a $38 million initiative in January 2014 that is delivering services to juveniles who are

involved in the criminal justice system. The New York and Massachusetts projects are the largest SIBs in the world to date and the first two whose impacts are being evaluated via randomized experiments.

While several authors, including ourselves, have described the SIB concept in policy papers,[3] there are fundamental aspects of these contracts that are ripe for more rigorous economic analysis. The aim of this paper is to supply that more rigorous analysis around four topics:

- Why is there underinvestment in social innovation and what government and market failures do SIBs potentially overcome?
- What tradeoffs are involved in designing pay-for-success payout schedules?
- How much risk is involved in these contracts and how should it be shared among contract parties?
- How can evaluation methodologies be designed that are robust enough to allow millions of dollars to flow based on the outcomes of the evaluations?

The paper begins by explaining the SIB model and summarizing the projects that are currently underway. Subsequent sections address each of the four topics.

2. Pay for Success Contracts Using Social Impact Bonds

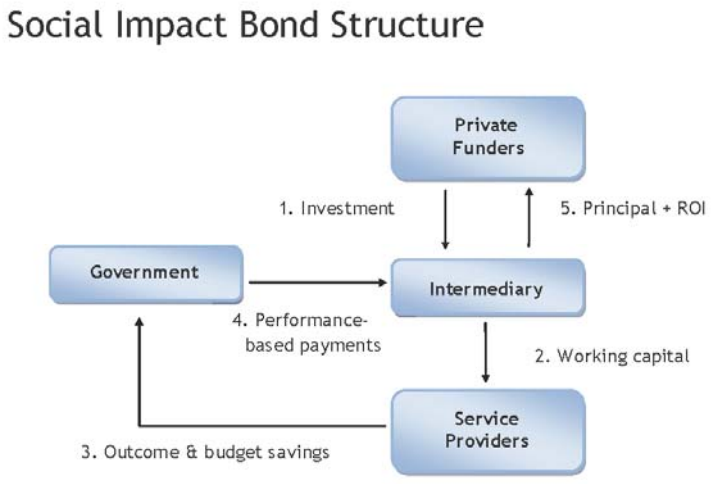
Under the most common social impact bond model, the government contracts with a private sector intermediary to obtain social services. The government pays the intermediary entirely or almost entirely based upon achievement of performance targets.

[3] See www.hks-siblab,org/publications/

Performance is rigorously measured by comparing the outcomes of individuals referred to the service provider relative to the outcomes of a comparison or control group that is not offered the services.

If the intermediary fails to achieve the minimum target, the government does not pay. Payments typically rise for performance that exceeds the minimum target, up to an agreed-upon maximum payment level. The intermediary obtains operating funds by raising capital from private commercial or philanthropic investors who provide upfront capital in exchange for a share of the government payments that become available if the performance targets are met. The intermediary uses these operating funds to contract with one or more service providers to deliver the interventions necessary to meet the performance targets. Figure 1 illustrates these relationships.

Figure 1



Governments are attracted to this approach because it shifts the risk of innovation from taxpayers to private investors, provides resources for preventive investments that may reduce future budgetary expenditures, and offers a way to make more rapid progress in achieving social policy goals. Service providers are attracted to this approach because

it provides stable multi-year funding and begins a relationship with the government that can enable operations to scale rapidly if the provider is able to demonstrate program effectiveness. Philanthropic investors are attracted to this model because it provides rigorous performance assessments of the initiatives they are funding and offers a way to massively scale these initiatives through government funding if they are proven successful. Commercial investors are attracted to the model because they see the opportunity to get involved in a promising new market, providing growth capital to social service providers, while also demonstrating that they are investing in the communities in which they are doing business.

The U.K. Ministry of Justice established the first SIB in the world, contracting with a nonprofit intermediary, Social Finance U.K., to provide services to prevent reoffending by 3,000 short-sentence male prisoners over six years at a prison in Peterborough, England. Social Finance raised $8 million from social investors to finance service delivery by another nonprofit, the St. Giles Trust. The government will make payments to Social Finance only if the reoffending rate falls by at least 7.5 percent compared to the recidivism rate in a comparison group of similar prisoners at other prisons that are not receiving the intervention. If payments are earned, they will be made in the fourth, sixth, and eighth years, based on outcomes achieved in working with prisoners during three consecutive two-year periods. Several additional SIB projects are now in operation in the U.K., including efforts to tackle a variety of family problems, reduce homelessness, and provide increased support for at-risk youth.

The first U.S. SIB was establish by the Bloomberg Administration in New York City in 2012. This initiative is providing services to sixteen- to eighteen-year-olds who are jailed at Rikers Island with the aim of reducing recidivism and related budgetary and

social costs. Services are being delivered to approximately 3,000 adolescent males per year for three years. MDRC is serving as the intermediary, overseeing day-to-day implementation of the project and managing the two non-profit service providers that are delivering the intervention. Goldman Sachs is funding the project's delivery and operations through a $9.6 million loan to MDRC. The city will make payments that range from $4.8 million if recidivism is reduced by 8.5 percent to $11.7 million if recidivism is reduced by 20 percent. Bloomberg Philanthropies is guaranteeing the first $7.2 million of loan repayment.

In recent months, New York State and Massachusetts have both launched pay for succes projects using SIBs. The New York project is providing job training and employment servics for 2,000 men who have been recently released from state prison with the goal of increasing their employment and reducing recidivism. The Massachusetts SIB is serving approximately 1,000 young males who are involved in the criminal justice system. Several other state and local governments are developing SIBs in policy areas ranging from diabetes prevention to early childhood education, and from addiction treatment for families involved in the child welfare system, to chronic homelessness.

3. The Challenge of Social Innovation

Consider a social planner trying to decide whether or not to test a new social intervention, such as a job training program for disconnected youth or home-visiting services for low-income pregnant mothers and their children. The social planner cares

both about the expected net benefits of the current project and about the payoff to scaling up the project to serve additional individuals if the current project is successful.

We assume the cost of the intervention, C_I, is known, but that the benefits are uncertain. To keep the model as simple as possible we further assume that the program either has benefits of B_I[4] such that $B_I > C_I$, or benefits of zero, where p is the probability that the current project has benefits of B_I vs. zero. The cost of evaluating the project is C_E. If the evaluation demonstrates that the intervention is successful, the social planner will expand the program to provide additional services in the future, where S is the scale of the broader population, and δ is a discount factor since the potential expansion would occur after the original project.

In this case, the expected net present value of the project from the standpoint of the social planner is:

$$ENPV_{SP} = pB_I - C_I - C_E + p\delta S(B_I - C_I) \qquad (1)$$

The social planner will undertake the project if $ENPV_{SP} > 0$. Note that even if the expected value of the current project ($pB_I - C_I$) is quite negative, the intervention can be worth testing so long as a successful intervention can be expanded to a sufficiently large scale (i.e., as long as δS is large). Intuitively, what this means is that even if the most promising early childhood intervention had only a one in ten chance of proving successful, it would be worth testing if a successful intervention could be taken nationwide.

The Government's Problem

[4] Think of this as the present discounted value of the stream of benefits yielded by the project.

We posit that governments are typically solving a problem that is different from the social planner problem.

First, a local government will typically place little weight on the value of other jurisdictions scaling up a successful intervention developed by the local government. Thus the local scaling factor, $S_{Local,}$ is less than the global scaling factor, S, and will reflect only the further expansion within the local jurisdiction. Without a subsidy from the national government or philanthropy there will be insufficient testing of new social interventions.

Second, governments rarely conduct rigorous evaluations of the effectiveness of their social spending. The result of this is that once programs get into the budget and accumulate constituencies, the programs become immortal, whether they are effective or not.

Third, governments discount future benefits too heavily. Because government officials serve finite terms, because voters are myopic, and because of the perceived urgency of solving immediate fiscal crises, policymakers tend to down-weight future benefits; this means that the government problem has an extra discount factor γ that does not appear in the social planner's problem. The result of this myopia is that governments underinvest in preventative social programs and sometimes fail to enact programs even when they are effectively self-financing, that is, even when up-front investment in preventable social problem leads to budgetary savings down the road.

Siloed decision making represents an additional channel through which future benefits get discounted by government decision makers. A preventative social program financed by one agency may generate benefits and budgetary savings for a different

agency (e.g. an investment by the education department in early childhood education may reduce future prison costs). The agency responsible for the potential preventative spending program may not fully value the benefits that accrue to a different agency and policy domain.

Taken together, these imply that, in deciding whether or not to fund a new intervention, a state budget office solves:

$$ENPV_G = p\gamma B_I - C_I + \delta S_{Local}(pB_I - C_I) \qquad (2)$$

and will only take on projects where the expected value of the current intervention is positive. The insufficient use of evidence in decision making and the difficulty of eliminating programs once they are created causes governments to lose the option value of experimenting with low probability, high value projects. In addition, because the government never discovers which programs are effective, it fails to scale up the effective ones — programs continue at their initial scale indefinitely and society loses the potential benefits from expanding them.

In addition to these three factors, which all fall under the category of "poor government decision making," we posit a fourth reason why current government practices lead to insufficient progress in tacking social problems -- the production process through which governments collaborate with private sector actors to produce social services is far from the frontier. Assume that the benefits of the intervention depend on government inputs, private sector inputs, and the technology by which government procures and manages its relationship with the private sector service providers: B_I*=F(private providers, government workers).* Often governments respond to budget pressures by protecting programmatic spending, while under investing in their own

human capital, especially in procurement offices and in the analytic and IT staff necessary to monitor contractor performance, so insufficient quality or quantity of government workers may hinder successful production of social services. In addition, the management techniques used may be suboptimal. For example, no one typically measures the impact of programs and no one within or outside of government is held accountable for program performance. If anything is measured, it is the number of slots filled or the number of people served, not the program's impact. Additionally, because of annual budgeting and frequent turnover of political leadership, it is next to impossible for governments to sustain a multi-year focus on collaborating with private sector actors to achieve an improved outcome. Finally, there may be insufficient use of performance incentives in contracts.

A. The Social Impact Bond As a Potential Solution: Part I

The Social Impact Bond offers a potential solution to each of the four explanations for why current government practices lead to insufficient progress in addressing social problems.

First, it provides a mechanism through which philanthropy and/or the national government can collaborate with a local government to subsidize the learning value of a project. In exchange for the local government agreeing to rigorously evaluate an intervention, the philanthropy or national government can cover some of the cost of the initial project, causing local governments to undertake projects that have positive

expected net present value for society due to their potential scalability, but negative NPV when viewed only from the local perspective.[5]

Second, the Social Impact Bond provides the government with a means to test a new intervention in a way that prevents ineffective programs from becoming immortal. Because impacts are rigorously assessed and because a project that fails to achieve the preset performance targets will be very publicly seen as a failure, the SIB greatly reduces the immortality risk. Similarly, a successful SIB will greatly increase the chance that a proven intervention is expanded to serve more people.

Third, SIBs appear to overcome the political obstacles to investing in prevention that arise from government actors discounting future benefits. The "money back guarantee for taxpayers" aspect of the model appears to be enough to overcome the reluctance to incur a cost today that might yield benefits in the future.

Fourth, the SIB model offers the potential to overcome some of the public sector human capital problems that hinder effective collaboration with private sector service providers. The SIB model brings additional expertise to the government, both in the form of government-side advisers and private sector intermediaries. More importantly, it establishes a contractual multi-year commitment between a government agency and service providers to work together on a sustained basis to achieve an outcome – something that is next to impossible under the ordinary operations of government. And the performance-based payments may focus attention and incentives around achieving improved outcomes in a way that traditional slot-based funding does not.

[5] The Social Impact Bond is not a unique tool for accomplishing this. A foundation could achieve the same end by directly financing a portion of an intervention that was procured through more-conventional means.

SIBs are not necessarily the unique solution to these problems. In theory, government could routinely evaluate programs and make funding decisions accordingly; decide to make more investments in preventative social programs; and reorient themselves to focus on establishing sustained multi-year efforts to tackle social problems. However, existing political incentives and public sector management practices do not appear to be accomplishing any of these things on a regular basis. The rationale for SIBs, therefore, is as a leadership strategy that a government can use to overcome the existing barriers.

B. The Provider's Problem

Consider a social service provider who has developed an intervention that yields positive social benefits and who wants to expand operations to serve additional people. If the social planner were the relevant decision maker, the provider could simply borrow money to make the investments necessary to expand capacity and know that the social planner would purchase the additional services once the capacity was created. But with government as the payer, the social service provider has no guarantee that the government will purchase the additional capacity if developed. This makes expansion excessively risky and private financing unlikely to be available. Indeed, a government's decision to purchase additional services from a service provider may be determined more by the effectiveness of the lobbying firm that the service provider hires than the quality of the evidence establishing its effectiveness.

C. The Social Impact Bond as a Solution: Part II

A social impact bond combines in a single transaction the private sector financing of the expansion of the service provider's operations and the government commitment to purchase for multiple years the additional quantity of services made possible by the expansion. Moreover, because investors get repaid only if promised impacts are achieved, the only service providers whose expansions will get funded are those with sufficiently promising or proven interventions to convince investors to back them. Thus the SIB introduces private sector discipline into government decision making about which social services to expand. This second rationale for SIBs is closely related to the poor government decision making rationale. It is because governments cannot be counted upon to always purchase social services whenever the social benefits of doing so exceed the social costs that capital markets are not able to finance the expansion of proven social service providers in the absence of SIBs.

4. Tradeoffs in the Design of Pay-For-Success Payment Schedules

This section begins by describing the characteristics of existing pay for success contracts. Then it turns to a more theoretical discussion of the bargaining positions that different parties to the contracts would be predicted to take.

A. Payment Schedules in the Initial Pay for Success Contracts

The payment schedules in the initial pay for success contracts have several common features. First, the *minimum payment threshold* is set such that the government pays only when results are sufficiently large that it can plausibly claim that, with high likelihood,

the impacts are not simply the result of chance. For example, the UK Peterborough SIB makes payments only when recidivism is reduced by at least 7.5 percent and the Massachusetts Juvenile Justice SIB makes payments only when recidivism is reduced by at least 5 percent. Statistical considerations can inform the choice of thresholds—for example, the government might want to know the observed impact at which, for the anticipated project size, it can claim a positive effect with 75 percent probability. In practice, however, the threshold is determined via negotiations between the government and private sector partners: The government balances its competing priorities of setting a high enough threshold to ensure sufficiently strong evidence before making payments and a low enough threshold to ensure that the private sector partners have positive performance incentives over as wide a range of outcomes as possible. In turn, the private sector partners seek higher rates of return for higher minimum payment thresholds.

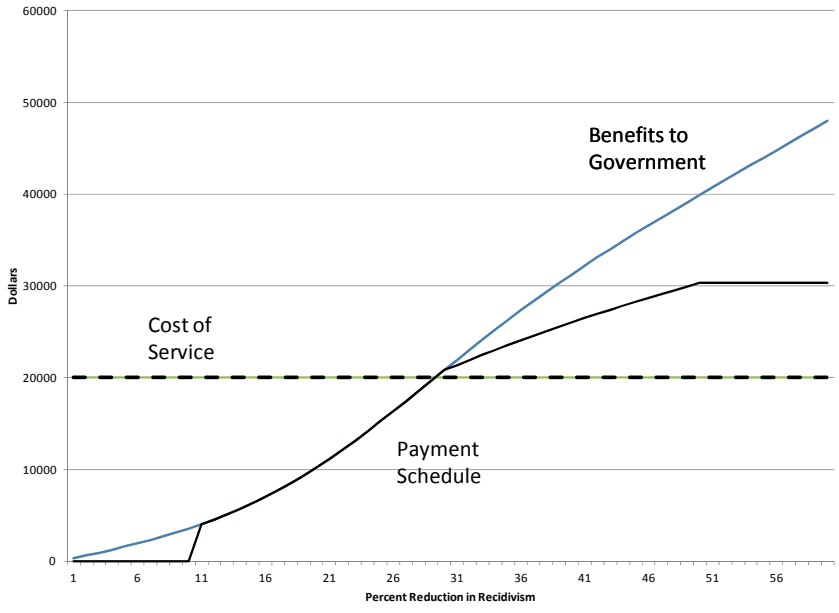
Second, once the minimum repayment threshold is reached, governments typically pay an amount equal to the full public benefits produced by their projects until the breakeven point -- the point at which investors have been repaid their principal. In some projects, governments have been willing to make payments only based upon the actual budget savings that flow from the project. In others, governments have also been willing to pay for non-monetizable social value.

Third, once impacts exceed the breakeven point, incremental benefits are shared between the government and the investors, often on a 50-50 basis. In other words, above the breakeven point, the government pays 50 cents for every dollar of incremental benefit. The flatter slope reduces the financial incentives for incremental performance but allows taxpayers to share in some of the net benefits of the project. In addition, there

is a maximum total payment that is determined by the amount the legislature appropriates for the project. The maximum total payment tends to be set only modestly above the expected outcome, since it is costly to obtain budgetary authority from the legislature. Therefore, a 50-50 split in payment between the breakeven point and the maximum payment allows for a positive financial incentive for better performance over a wider outcome range than would be possible if the government continued to pay 100 percent of the benefits up to the maximum.

Figure 2 shows the payment schedule for an illustrative recidivism project. Benefits to the government from reduced days of incarceration initially accrue relatively slowly, because with small reductions only pure marginal cost savings are achieved. At higher outcomes more fixed costs are avoided (a wing of a prison can be closed for example) and benefits accrue more rapidly. In this example, payments begin only when the reduction in recidivism reaches 8 percent. The breakeven point occurs at approximately 30 percent, and maximum payment occurs with a 50 percent reduction in recidivism.

Figure 2



Because payment schedules tend to be approximately linear over a wide range of outcomes, mean preserving spreads in the probability distribution of outcomes have relatively little impact on expected payout -- only the portions of the distribution that are pushed beyond the ends of linear payment range affect the expected outcome. However, the cap at the top and the zero payment range at the bottom imply that mean preserving spreads in the overall outcome distribution decrease expected payments. Because the private sector partners are not penalized for performance below zero, the risk that the project might have negative outcomes and thereby raise public sector costs is fully borne by the government.

B. Bargaining Positions

Investors will participate in a SIB only when

$$\int_{-\infty}^{\infty} P(y)f(y)dy + B_{NP} > I + RP \qquad (3)$$

where *y* denotes the outcome that determines payments, *P(y)* is the payment schedule as a function of the outcome, *f(y)* is the probability distribution of the outcome, B_{NP} are the non-pecuniary benefits to the investor from participating in a SIB (e.g. positive press coverage for a commercial bank or warm glow for a social-minded individual investor), *I* is the amount of operating capital provided to fund service delivery, and *RP* is the risk-premium required to compensate a risk-adverse investor for taking on risk.

A government will participate in a SIB only when

$$\int_{-\infty}^{\infty} B(y)f(y)dy + B_S > \int_{-\infty}^{\infty} P(y)f(y)dy \qquad (4)$$

and

$$\int_{-\infty}^{\infty} B(y)f(y)dy + B_S - \int_{-\infty}^{\infty} P(y)f(y)dy > \int_{-\infty}^{\infty} B(y_c)g(y_c)dy_c - I \qquad (5)$$

In equation 4, *B(y)* is the public sector benefits at each outcome level *y*, and B_S is the value the government gets from obtaining information via the SIB project that informs its future decision making. The equation shows that a government will invest in a SIB only if the expected public benefits from the current project plus the learning value of the project exceed the payments the government makes to investors. In the second equation y_c represents the outcomes that are achieved through conventional approaches to spending. This equation shows that the benefits to the government of the SIB net of payments to the investors need to be greater than the benefits that could be obtained if the government simply purchased social services from the provider through traditional funding mechanisms.

These participation constraints have several implications for SIB projects. First, substituting the investor constraint into the second government constraint, we see that a SIB is viable only when the expected increase in public sector benefits from the SIB exceeds the risk premium net of any non-pecuniary investor benefits:

$$\int_{-\infty}^{\infty} B(y)f(y)dy + B_S - \int_{-\infty}^{\infty} B(y_c)g(y_c)dy_c > RP - B_{NP} \qquad (6)$$

In other words, it is only worthwhile for a government to undertake a SIB if it can achieve better public sector outcomes -- otherwise it is not worth paying the risk premium. One often hears SIB advocates claim that SIBs are beneficial because they allow governments to shift performance risk to the private sector. This is nonsense. From a social planner perspective governments should be approximately risk neutral, since they can spread risk over the entire tax base. Therefore, governments should not pay a premium to transfer risk to risk-adverse private sector actors. But if collaborating with the private sector allows the government to produce better outcomes—by addressing

any of the government failures presented in section 2—then SIBs can be in the government's interest.

Second, this last equation makes it clear why there can be a rationale for a philanthropic or federal government role in these transactions. In cases in which the learning value to a particular government is smaller than the global learning value, such that the inequality is not satisfied. a third party can make a transaction viable by absorbing some of the risk (lowering the risk premium that the government needs to pay) or by making a portion of the government payments. This has occurred in all of the U.S. SIB transactions to date. For example, in the New York City Rikers Island project, the Bloomberg Foundation backstopped about 80 percent of Goldman Sachs's investment. In the Massachusetts juvenile justice project, philanthropic investors financed about half of the transaction using a mezzanine structure in which the commercial investors who financed the other half will be repaid first. In addition, the U.S. federal government made grants to Massachusetts and New York state covering about half of the performance payments in those projects.

Third, governments should rely on private investment in these projects only to the extent necessary to achieve improved outcomes. For example. if by financing 25 percent of an intervention with private dollars it becomes possible to overcome the political barriers to investing in prevention, ensure that outcomes are rigorously measured, and set up a six-year collaboration between government agencies and private sector providers focused on improving a social outcome. then there is no reason for the government to go to 100 percent private financing and pay the cost of the extra risk premium. Indeed, it seems likely that projects will often be able to overcome many of the government failures

discussed in section 2 with less than 100 percent private financing. But there may also be cases where only 100 percent private financing can overcome the political myopia that leads to under investment in prevention. And in cass in which the main benefit of a SIB is coming via investor attention to outcomes, more private financing may lead to improved oversight and management.

Fourth, government officials sometimes confound the portion of the investor payment necessary to compensate a risk neutral investor for potential losses and the portion necessary to persuade a risk adverse investor to take on a project. This confusion can make SIBs look expensive relative to simply paying for the service directly. Consider a SIB that raises $10 million in investor capital to finance service delivery and which pays back the investor four years later if performance targets are met. In keeping with our simple model in Section 2, assume that the project can only lead to two outcomes, success and failure, with a 70 percent chance of success. A risk neutral investor will need to be paid $14.3 million ($10 million/.7) in initial period discounted dollars in the successful state for this project to be viable. This means that payments in the fourth year will need to be $14.3 million x $(1+r)^4$, where r is the nominal interest needed to compensate the investor for the time value of money (but not for risk, which has already been incorporated). If r is .05 then the final payment will need to be $17.4 million, yielding a 14.8 percent annualized return if the project is successful.

This payment structure may give the impression of the government overpaying by 74 percent relative to simply purchasing the services up front for $10 million. But of course the expected payments by the government are in fact only $10 million in period 1 dollars, assuming that the government discounts at the same interest rate as the investor.

If the project fails, which occurs with 30 percent probability, the government actually comes out ahead by $10 million—a fact often ignored by those who view this approach as overpaying by 74 percent.

As equation (6) shows, from the public sector point of view, there is only a cost to using the SIB if the government ends up paying a risk premium -- in this example an annualized return above 14.8 percent. In some cases, the investor discount rate may be higher than the government discount rate. In the example above, if the government nominal discount rate was 4 percent then approximately $652,000 of the payments (or 1 percentage point of the 14.8 percent annualized return) would be the result of the government using more costly private sector financing and would need to be offset by better expected outcomes from using this funding mechanism if a SIB is to be justified.[6]

5. Uncertainty in the Distribution of Outcomes

Government decisions about how much to offer to pay investors and investor decisions about whether to accept the offer depend heavily on both the anticipated and the actual outcome distribution upon which payments are based.

There are two main sources of uncertainty in a Social Impact Bond project: uncertainty about the true impact of the project and uncertainty in the statistical measurement of the project's observed impact. For exposition only, assume that the observed benefit, *B*, comes from the following hierarchical Normal model:

$$B \mid \tau \sim N(\tau, \sigma_B^2)$$

$$\tau \sim N(\mu_\tau, \sigma_\tau^2)$$

[6] This is not to say that it is impossible for a government to grossly overpay in a SIB project. For example, if the government wrongly assessed the probability of success at 70 percent and paid accordingly, while the true probability of success was 95 percent, then the government would be overpaying.

where τ is the true impact of the intervention for this particular Social Impact Bond. Then σ_B is the statistical uncertainty associated with measuring τ; and σ_τ is the uncertainty in τ. Both σ_B and σ_τ have multiple components.

A. Uncertainty About the True Impact of the Program

Consider first the uncertainty about the true impact of the program, σ_τ. There are two key components of this uncertainty:

Historical Uncertainty. The first key component is the uncertainty about the historical evidence of program effectiveness, as viewed through a meta-analytic framework. This uncertainty largely depends on two main factors:

- *The number and size of previous evaluations:* In general, σ_τ will decrease as the number and sample size of previous evaluations increases.
- *Methodological uncertainty of previous evaluations*: Similarly, σ_τ will be larger if all prior evaluations were non-randomized studies than if they were well-designed RCTs. If the historical evidence was not from a RCT, the measured effect might be due to differences between the populations served and not-served rather than due to the program itself. Similarly, if the people served were in different locations or different time periods than the comparison people who were not served, the measured effect might be due to these contextual differences rather than from the true impact of the program. Stated differently, in a meta-analytic framework, we effectively inflate the reported standard errors of non-randomized

designs, since the reported results do not generally incorporate uncertainty associated with possible violations of the identifying assumptions. Doing so in a systematic way is difficult (Sekhon, 2010).

Implementation Uncertainty. The second key component is the uncertainty in extrapolating the results from historical studies to the particular implementation for the Social Impact Bond. For example, the program may serve a somewhat different population than the population for the historical estimates. Additionally, program operations may differ from those in the past (e.g. the process by which individuals are referred to the provider, who is managing the operations, what services are delivered, etc.)

In most cases in which a social impact bond is being seriously considered, there are a reasonable number of previous studies available. Nonetheless, these are unlikely to be well-conducted RCTs on an identical population. Therefore, it is difficult to pin down these sources of uncertainty. Fortunately, absent outside information, there is no particular reason to think that the methodological uncertainty is asymmetric or centered away from zero. The implementation uncertainty might, in addition to spreading out the probability distribution, shift one's assessment of likely impacts toward zero (if the context in which implementation is occurring is particularly challenging) or shift one's assessment upward (if the attention to outcomes and measurement in the SIB project is seen as increasing the effectiveness of the program model).

In addition to these two sources of uncertainty, there is a third component -- Project Selection Bias – that, if properly accounted for, will almost always shift the expected outcome distribution toward zero. The central concern is a classic example of

regression toward the mean: even if we perform an exact replication of a high-impact historical study, the measured impact for the replication will likely be closer to zero than the initial result. Social Impact Bond projects are selected from a very wide range of potential social interventions and are generally chosen based on existing evidence of strong impacts. As a result, the anticipated effects for SIB projects will be closer to zero than the average results from previous studies. This concern is even greater if one views positive results as more likely to be published or if a project is designed by focusing on a subgroup of the population for which the intervention appears to have particularly strong impacts. In this case, it is even more likely that the previous observed impact is a statistical anomaly.

The bottom line is that, in most cases, a proper probability distribution of true project impacts will be much wider than implied by the standard error on the historical evaluation study, it will be shifted toward zero, and a significant portion of the distribution will be to the left of zero. None of this should be surprising if one considers the results in past evaluations of social programs or the well-documented challenges that have arisen of replicated successful programs in new sites.[7]

B. Uncertainty in the statistical measurement of program outcomes

Whatever methodology is used to estimate the impact of the program will result in a measured impact that is a random draw around the true impact. If a non-RCT approach

[7] Evaluation expert and sociologist Peter Rossi was mostly being serious when he issued his "iron law" of evaluation ("the expected value of any net impact assessment of any large scale social program is zero") and his "stainless steel law" of evaluation ("the better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero"). Peter Rossi, "The Iron Law of Evaluation and other Metallic Rules," Research in Social Problems and Public Policy, 4:3-20 (1987).

is used, additional uncertainty will be introduced due to potential bias from contextual factors, self-selection into the treatment population, etc.

One the most challenging issues in designing a social impact bond is acquiring sample sizes large enough to produce reasonably precise estimates about program impacts. First, in some cases, very few people are a good fit for the intervention, especially if eligibility is restricted to the subset for whom it is most cost-effective to offer the intervention (often those predicted to have the worst outcomes without the intervention). In these cases, the pooling of samples of multiple years of service delivery is often necessary for evaluation purposes. Second, it is more challenging to raise sufficient capital to deliver services to 500 people a year than it is to raise sufficient capital to deliver services to 250 people a year.

C. Comparing Sources of Uncertainty

Even with moderate sample sizes, it is generally the case that the uncertainty about the true impact of the project, σ_τ, is much greater than the additional uncertainty introduced via the statistical measurement of program outcomes, σ_B. The expected payouts by the government to investors therefore tend to vary only modestly with the precision of the impact estimates. This fact creates tension in negotiations between the government for whom a substantial portion of the value of doing the SIB comes from learning about τ, the true impact of the project, and the commercial investors whose financial payoff is based only on B, the impact in the current project. Governments push for larger cohort sizes and longer project periods while commercial investors prefer smaller, shorter projects. When a significant portion of commercial investor motivation

comes from non-pecuniary interests that do not increase with project size, this negotiating tension is even more intense. Philanthropic investors often have interests in social learning that are aligned with the government interests.

6. Robust Evaluation Methodologies

As discussed above, RCTs are the centerpiece of rigorous Pay-for-Success evaluations in Massachusetts and New York State, maximizing the governments' confidence that they are paying for actual results. At the same time, the Pay-for-Success evaluation methodologies need to be robust enough for millions of dollars to change hands based upon the final estimates.

Like other RCTs, we therefore fully pre-specify the evaluation methodology so that there is no debate about the precise analysis steps at the conclusion of the project. Unlike other RCTs, however, we also specify a secondary evaluation methodology in the event that the results from the RCT are insufficiently precise to form a reasonable basis for payments. We refer to this secondary methodology as a *backstop methodology.* The design we have helped Massachusetts and New York State implement is a randomized experiment backstopped by a non-experimental difference-in-differences (in Massachusetts) or before-after (in New York) study. Taken together, the goal is to have a methodology that still allows results to be measured and payment to occur even if something unanticipated occurs along the way that impacts the quality of the evaluation.

A. Design and Analysis of the RCT

In social impact bond projects with rigorous evaluation methodologies, the government randomly refers specific individuals to the service provider. For example, in the New York State adult recidivism project, the government refers a random subset of men released from prison who are predicted to have a high risk of recidivating. This referral process avoids the "cream-skimming" or "self-selection" processes that might otherwise bias impact estimates of differences between treatment populations and comparison or control populations.

But under this intent-to-treat design, it is uncertain what fraction of those referred will actually end up receiving services from the provider and therefore how many individuals need to be referred to the provider to fill the available slots. In addition, because providers are already providing services in the community, there is uncertainty about what fraction of comparison or control group individuals will end up getting served by the provider -- diminishing the experimental-control contrast in the treatment. To deal with the uncertainty about program take-up rates, payments are based on an IV estimate calculated by dividing the intent to treatment estimate by the difference in provider take-up rates between the treatment and control group:[8]

$$\mathrm{IV}_{\mathrm{RCT}} = \frac{\text{ITT estimate}}{\hat{p}_T - \hat{p}_C}$$

This generates an impact estimate that is in "per person served" units. The payment schedules then multiply these estimated impacts by the payment rate and the number of treatment group members who receive services to determine payments.

[8] In some SIB projects, the IV estimate is calculated in a regression framework that allows for the increased precision that comes from covariate adjustment.

Under standard assumptions, the IV estimate measures the local average treatment effect (LATE) -- the impact of the intervention on those individuals who receive services when they are in the treatment group but not when they are in the control group. This LATE estimate is used to determine the payment rate for all treatment group members served by the service provider -- meaning both "compliers" (the individuals who determine the LATE estimates) and "always takers" (individuals who receive services regardless of which treatment status they are assigned to). Given that there is generally no *a priori* reason to think that impacts are strongly different for individuals referred to the providers through other mechanisms, the fact that payment rates are the same for always takers as for compliers does not seem like a major problem.

Paying based on the IV estimate produces an additional benefit. The providers are paid to serve a fixed number of individuals. By referring a larger number to them, but making the provider responsible for recruiting individuals to their program and deciding which individuals to serve, the providers have the proper incentive to allocate their fixed resources to those members of the referral group for whom the provider's intervention will have the largest impact.

B. Backstop Methodology

The biggest risk to the IV estimate is that a sizable fraction of control group members will end up receiving services, causing the treatment-control difference in exposure to the intervention to be small and the IV estimate to be imprecise. The back stop methodologies are designed to provide a way to make payments even when the IV estimate turns out to be uninformative.

This approach builds off recent work by Hartman et al. (2014), who use observational studies to adjust experimental estimates of the effect of pulmonary artery catheterization (PAC). It also follows from the meta-analysis literature, in particular, Rubin (1992), who argues that, above all else, the goal of meta-analysis is to estimate the effect for some idealized experiment on a desired population of interest.

Set-up and Notation

Since the exact design will depend on the circumstances for each SIB project, we present a stylized example here. To simplify the discussion, assume that we have two areas in the our jurisdiction, $j \in \{0,1\}$, that we observe for two time periods, $t \in \{0,1\}$. The proposed SIB evaluation is to conduct an RCT of the intervention in area $j = 1$ at time $t = 1$. See Figure 3.

Let W_i be an indicator for whether individual i is in an area and a time that receives the intervention (i.e., $j = 1$ and $t = 1$). Let Z_i be an indicator for whether individual i is randomly referred to the social service provider. Finally, let D_i be an indicator for whether individual i enrolls in the program (shown as grey in Figure 3). Since only individuals in area $j = 1$ and time $t = 1$ can access the treatment, $D_i = 1$ or $Z_i = 1$ can only hold if $W_i = 1$. At the same time, since the service provider will not refuse service to individuals, it is possible that $D_i = 1$ if $W_i = 1$ but $Z_i = 0$. Finally, let Y_{ijt} be the observed outcome for individual i in area j at time t, with corresponding potential outcome, $Y_{ijt}(W_i, Z_i, D_i)$.

Estimating LATE via RCT

First, we focus on the subset of individuals for whom $W_i = 1$, which is a textbook setup for RCT with non-compliance. Let $D_{i11}(W_i = 1, Z_i) = D_i(1, Z_i)$ be the potential outcome for treatment take-up, D_i. Under the assumption of monotonicity (i.e., no defiers) and SUTVA (Imbens and Rubin, 2014), we then have the three usual compliance types:

- *RCT Compliers:* $D_i(1,0) = 0$ and $D_i(1,1) = 1$
- *RCT Always Takers:* $D_i(1, z) = 1$ regardless of z
- *RCT Always Takers:* $D_i(1, z) = 0$ regardless of z

Next, we assume that Z_i is randomly assigned for individuals with $W_i = 1$ (in practice, Z_i will be assigned via stratified randomization), where

$$Z_i \perp \left(Y_i(1,0,0),\ Y_i(1,0,1),\ \ Y_i(1,1,0),\ Y_i(1,1,1),\ D_i(1,0),\ D_i(1,1)\right) \mid W_i = 1.$$

We then make the usual exclusion restrictions for the RCT:

- Exclusion Restriction For RCT Never Takers: $Y_i(1,1,0) = Y_i(1,0,0)$; and
- Exclusion Restriction For RCT Always Taker: $Y_i(1,1,1) = Y_i(1,0,1)$.

Finally, we assume that the instrument is relevant, i.e., random referral indeed induces some individuals to take up the treatment, $Pr(D_i(1, z) = z) \equiv p_{RCT,c} > 0$, where $p_{RCT,c}$ is the proportion of compliers in the RCT.

Under these assumptions, we use the typical IV estimator,

$$\widehat{IV}_{RCT} = \frac{\widehat{ITT}_{RCT}}{\hat{p}_{RCT,c}}$$

or the regression-adjusted Two-Stage Least Squares version, which estimates the treatment effect for RCT Compliers (Angrist, Imbens, and Rubin, 1996):

$$LATE_{RCT} = E[Y_i|W_i = 1, D_i = 1, \text{RCT compliers}] - E[Y_i|W_i = 1, D_i = 0, \text{RCT compliers}].$$

Estimating LATE via Quasi-Experimental Design

We now leverage the fact that we also observe comparable individuals in other areas and at different times. First, we assume that we can conduct an additional non-randomized study (NRS) to estimate the effect of "rolling out" the SIB project in a given jurisdiction at a given time, $W_i = 1$ vs. $W_i = 0$. Possible designs include difference-in-differences or matching estimators. For example, the difference-in-differences design for New York State compares target jurisdictions in New York City and Rochester before and after the SIB launch to comparable jurisdictions elsewhere in the State before and after the SIB launch.

We are agnostic as to the exact empirical strategy, but assume that the given approach yields estimates of the following ITT estimate:

$$ITT_{NRS} = E\big[Y_{ijt}\big|W_i = 1\big] - E[Y_{ijt}|W_i = 0].$$

Such a setup also assumes (generally implicitly) W_i is as-if randomly assigned, either conditional on covariates (i.e., matching) or for a certain parametric form (e.g., difference-in-differences).

To estimate a policy-relevant quantity, we need to make some additional assumptions. First, we make another exclusion restriction-type argument that potential outcomes only depend on treatment take-up:

- $Y_i(0,0,0) = Y_i(1,0,0) = Y_i(1,1,0) = Y_i(D_i = 0)$;
- $Y_i(1,1,1) = Y_i(1,0,1) = Y_i(D_i = 1)$.

In other words, the outcomes only depend on whether an individual actually enrolls in a program, not how that individual initially starts the program. This assumption is analogous to the consistency assumption in Hartman et al. (2014) and the parallel design assumption in Imai et al. (2013).

This implies that we have two additional compliance types:

- *SIB Compliers:* $D_i(W_i = 0, 0) = 0$ and $D_i(W_i = 1, z) = 1$, regardless of z; and
- *SIB Never-Takers:* $D_i(w, z) = 0$, regardless of w and z.

Note that the set of SIB Compliers is the union of RCT Compliers and RCT Always Takers. In other words, we are interested in the individuals who would enroll in the SIB program, regardless of how they initially enroll—i.e., because they are randomly assigned or because they would show up regardless.

Let $p_{SIB,c}$ be the proportion of SIB Compliers. Since W_i is as-if randomly assigned by assumption, this proportion is the same in expectation for both $W_i = 0$ and $W_i = 1$. We can therefore estimate this quantity via the observed proportion of individuals in $W_i = 1$ who take up the treatment: $p_{SIB,c} \equiv E[D_i = 1 | W_i = 1]$.

Then, with some abuse of notation, we can construct the Non-Randomized Study IV estimator,

$$\widehat{IV}_{NRS} = \frac{\widehat{ITT}_{NRS}}{\hat{p}_{SIB,c}},$$

which is an estimate for:

$$LATE_{SIB} = E[Y_i(1)|\text{SIB Compliers}] - E[Y_i(0)|\text{SIB Compliers}]$$
$$= E[Y_i(1)|\text{RCT Compliers or RCT Always Takers}]$$
$$- E[Y_i(0)|\text{RCT Compliers or RCT Always Takers}].$$

As before, we could use the regression-adjusted version if desired. If there are no RCT Always Takers (i.e., no crossovers in the RCT), then the two LATE estimands, $LATE_{RCT}$ and $LATE_{SIB}$, are equivalent. If this is not the case, we need one additional assumption:

$$E[Y_i(1) - Y_i(0)|\text{RCT Compliers}] = E[Y_i(1) - Y_i(0)|\text{RCT Always Takers}].$$

In other words, the effect of the intervention is the same for individuals induced to take up the treatment by the randomization and for those who would take up the treatment whenever it is available, regardless of randomization. Note that this second quantity, $E[Y_i(1) - Y_i(0)|\text{RCT Always Takers}]$, looks unusual at first. However, the *availability* of the program can still have an effect even if the randomly assigned offer has no effect for this subgroup (i.e., the usual exclusion restriction holds).

Combining Estimators

Given the above assumptions, $LATE_{RCT} = LATE_{SIB}$, so both IV_{RCT} and IV_{NRS} estimate the same quantity of interest. The goal is to optimally combine the two estimators.

First, the optimal combination depends on the correlation between IV_{RCT} and IV_{NRS}. Since this will depend on the precise empirical strategy for the non-randomized study, we focus on the case when these two estimators are approximately independent.

Based on our experience in Massachusetts and New York, this appears to be a sensible approximation for a surprising number of non-randomized designs, including difference-in-differences estimation. Nonetheless, it is straightforward to extend these results to the dependent case. Finally, since both estimators are approximately Normal in large samples, un-correlation implies independence.

The textbook solution to combining two such estimators is the precision-weighted average (i.e., Fisher weighting). Therefore, the key question is the variance estimates to use for each estimator. If we fully believe the identifying assumptions for the non-randomized design, then the relevant standard errors are those for the usual IV estimators:

$$se\left(\widehat{IV}\right) \approx \frac{se(\widehat{ITT})}{\hat{p}_c}$$

with appropriate extensions for robustness, clustering, etc.

In practice, however, we worry that, even if IV_{NRS} is approximately unbiased, the nominal standard errors do not accurately reflect the uncertainty in the estimate, due to uncertainty in the identifying assumptions (e.g., Manski, 2013). To account for this, we artificially inflate the standard error on IV_{NRS}, with the precise amount dependent on expert judgment and consultation with other SIB parties. Formally, this corresponds to a simple variance components model with an extra additive, mean-zero error component.

Figure 4 shows example weights for a stylized example. Importantly, as $p_{RCT,c}$ increases, the weight on IV_{RCT} increases. In this example, when the proportion of compliers is greater than 30 percent, nearly all of the weight is on the RCT.

Figure 3: Stylized SIB Evaluation Example

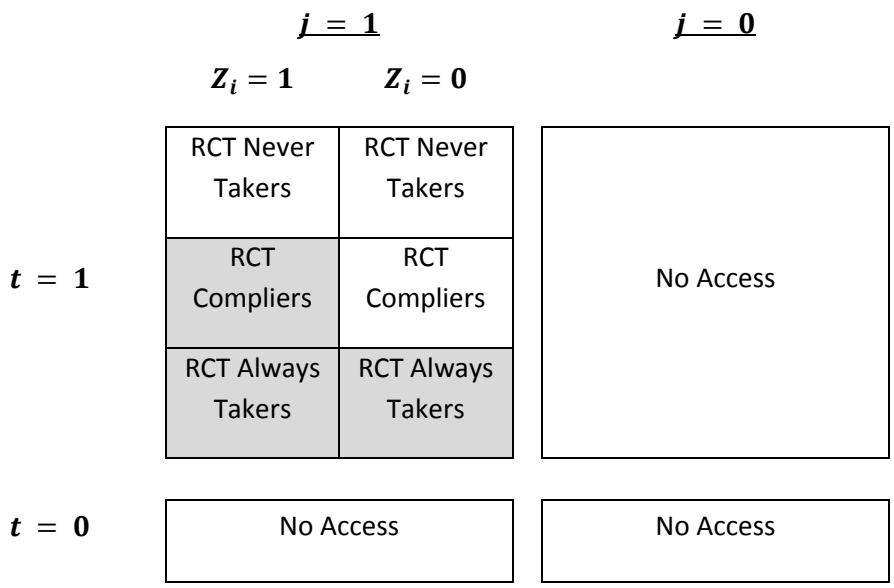
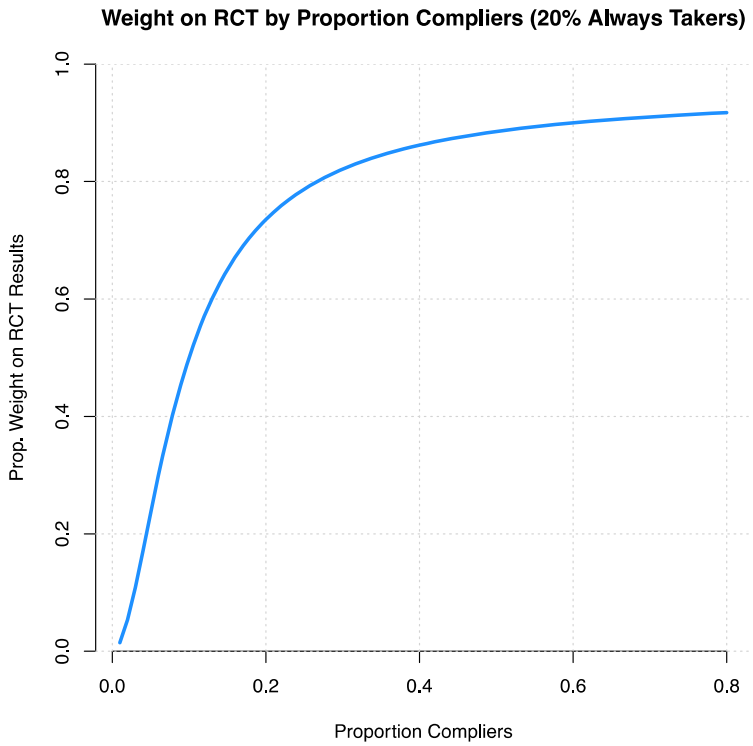


Figure 4: Example Weights



7. Conclusion

Pay for success contracts using social impact bonds are an experimental method for government to procure social services. They have the potential to produce additional investments in things that already work, create incentives for better performance, generate more learning about which social interventions work, foster better government decision making, and establish more effective multi-year outcome focused collaborative efforts. But there are also plenty of ways in which these transactions could fail. It may prove impossible to raise sufficient investment dollars because projects are too risk for the returns the government is willing to pay. The transaction costs may be too high, and the projects may fail to overcome enough of the barriers to effective government management of social service procurement. Moreover, it may simply be

that sufficiently successful social interventions don't yet exist. With more than a dozen U.S. governments developing pay for success projects, we will soon know a lot more about this model's potential.