Title: ***Why is Impact Measurement Abandoned in Practice? Evidence use in evaluation and contracting for five European Social Impact Bonds***

Running Head: Why is Impact Measurement Abandoned in Practice? Five Social Impact Bonds

Article Category: Research Article

Authors:

Debra Hevenstone
Social Work, Bern University of Applied Sciences, Bern, Switzerland

Alec Fraser
Government & Business, King's College London, UK

Lukas Hobi
Social Work, Bern University of Applied Sciences, Bern, Switzerland

Gemma Geuke
Social and Behavioral Sciences, Utrecht, Netherlands

Abstract

Despite broad consensus on the importance of measuring "impact," the term is not always understood as estimating *counterfactual* and *causal* estimates. We examine a type of public sector financing, "Social Impact Bonds" (SIBs), a scheme where investors front money for public services, with repayment conditional on impact. We examine five cases in four European countries of SIBs financing active labor market programs (ALMPs), testing the claim that SIBs would move counterfactual causal impact evaluation to the heart of policy. We examine first how evidence was integrated in contracts and second the overall evidence generated. Third, finding that neither contracts nor evaluations used counterfactual definitions of impact, we explore stakeholders' perspectives to better understand the reasons why. We find that although most stakeholders wanted the SIBs to generate impact estimates, beliefs about public service reform, incentives, and the logic of experimentation led to the acceptance of non-causal definitions.

Keywords: Pay by Results, Impact, Active Labor Market Program, Social Impact Bonds

1. Introduction

There has been a trend towards more (quasi) experimental, causal, counterfactual approaches to measuring impact in economics and policy research over the past decades (Banerjee et al., 2016), with many policy makers re-assessing their approaches (Gaffey, 2013). Simultaneously, there has been push-back from academics (Deaton and Cartwright, 2018) and policy makers (Stern et al., 2012) arguing that the appropriate evaluation strategy depends on many different factors. Moreover, it can be difficult to identify when counterfactual causal estimates are used, given different understandings of terms like "impact" and "attribution" (Heinrich, 2007; White, 2010).

One key factor in whether (quasi-) experimental counterfactual impact evaluation (i.e., randomized control experiments (RCTs) or causal inference techniques attempting to approximate RCTs— from this point on simply "impact measurement") is used, is whether the program has a contract form implying its necessity. It has been argued for almost half a century that when using Pay by Results (PbR, contracts including incentive payments) impact must be measured (Wedel, 1976), though in practice PbR has seldom done so (Martin, 2005). It has been argued that a relatively new contract form, the Social Impact Bond (SIB) or Pay for Success financing could entail a stronger requirement for impact estimates (Fox and Morris, 2021; Fraser et al., 2020) as a SIB uses investors to front money for public services. As such investor contracts paying more than the government bond rate without evidence of financing impact is a waste of taxpayer money. Thus, one would expect the estimation of programming and financing impact in SIBs.

In examining the choice of evaluation approach the type of intervention is also relevant (Stern et al., 2012). For some policy areas, process measures and monitoring might be the focus (e.g., homeless shelters), but for services like active labor market programs (ALMPs), impact measurement is central to understanding services delivered. Further, ALMPs have clearly measurable outcomes, recognized causal pathways, and an accessible toolbox of counterfactual approaches well-suited to impact estimation – evidenced by meta-reviews covering almost 400 such evaluations (Card et al., 2010, 2018). Therefore, ALMPs are a common programming area for both PbR (Heinrich, 2007; Koning and Heinrich, 2013) and SIBs (GOLab, 2022).

We examine five European SIB-financed ALMPs, four of which were early experiments in their national contexts. Given the financing and intervention, one would expect that these programs generated impact estimates, either for use in assessing contract fulfillment, for evaluation purposes, or both. We ask three questions. First, whether and how were impact estimates integrated in contracts and second, beyond contract terms, did these projects produce impact estimates? Third, what were the reasons?

We begin by examining the literature on performance contracts and evidence, presenting arguments why SIBs might encourage impact evaluation. We then consider potential confounding factors, focusing on ideological motivations and interests, the logic of early policy experiments, and the role of fiscal federalism. In the empirical section we describe the cases, their evaluation strategies, and the integration of evidence in contracts. We then examine the perspectives and choices of key stakeholders. In the discussion we summarize the factors influencing evidence use and, in the conclusion, reflect upon how policy makers might create the conditions to improve evidence generation and the research agenda to test whether these conditions improve evidence.

2. Theory

2.1 Evidence-informed policy and contracts

Many see evaluation as part of a collaborative learning process (Lowe et al., 2021; Sullivan, 2018) – a promising strategy to motivate impact evaluations given that social service managers and policy makers actively look towards research to learn how to improve programming (Jennings and Hall, 2012; McBeath et al., 2015). However, in practice, there are obstacles to generating impact evaluations for "learning purposes." Practitioners might lack the resources or expertise to commission high-quality studies (Sullivan, 2018) and government funding streams might not cover evaluation—though this has improved (Haskins and Margolis, 2015). Further, actors have differing perspectives and goals (Boaz et al., 2019) and commissioners may knowingly adopt inappropriate or simplistic methods because of a lack of time and flexibility (Cox and

Barbrook-Johnson, 2021) or might be unaware when they create pressure on evaluators to deviate from best practice (Pleger and Hadorn, 2018).

Some argue that one solution is integrating impact estimates into contract terms (Haskins, 2018) or a combined approach where evaluation primarily has a learning purpose, but some low-level incentives are integrated into contracts (Overholser, 2018). This belief was one factor behind decades of experimentation with PbR contracts, which pay social service providers partly based on impact.

PbR experiments have partially disappointed this hope. While they have increased the use of evaluation and auditing, they have not motivated impact evaluations (Martin, 2005; Pattyn, 2014). There are many reasons for this. First, social service providers can seldom wait for payment or absorb high economic risk. Second, while RCTs offer solid and easily communicated impact estimates (Duflo, 2020), they are costly, sometimes ethically dubious, denying potentially helpful interventions to eligible populations (Huitema et al., 2018; Teele, 2014), and can be limited when interested in understanding underlying mechanisms or heterogeneity, or limited in scale-up (Bonell et al., 2012). Third, although quasi-experimental approaches can approximate RCTs, particularly for ALMPs (Card et al., 2018), they can be difficult for policy makers to interpret (Bitler et al., 2019). Further, given multiple and complex goals (Heinrich and Marschke, 2010) and the difficulty of including all relevant outcomes in contracts (Kok et al., 2017), properly integrating impacts into contracts is challenging.

The use of non-impact measurement in contracts has proven to be problematic. Non-impact measurements are often uncorrelated with impacts (Barnow, 2000; Heckman et al., 2002), generate undesired strategic behavior (Heinrich, 2007; Heinrich and Choi, 2007; Heinrich and Marschke, 2010; Koning and Heinrich, 2013), pressure front line workers to act against their professional judgement (Pihl-Thingvad, 2016), generate a loss of regulatory control and ethical standards (Adams and Balfour, 2010), and inflate government costs (Webster and Harding, 2001). This has generated wariness about PbR (Ngan and Robinson, 2015) and suggestions for "soft" contract management tools like public recognition (Heinrich and Marschke, 2010), monitoring, praise/reprimands, mentoring (Brown and Potoski, 2003; Girth, 2017), and active contract management (Heinrich and Choi, 2007).

Some have considered SIBs to be a potential better mechanism to encourage impact evaluations.

Social Impact Bonds

SIBs, or the idea of investors fronting money for social programs, emerged in the UK in 2010 during the Great Recession and government-imposed austerity. Original proponents (Cohen, 2011) emphasized that SIBs would encourage innovative partnerships between public, private and philanthropic sector stakeholders, deliver cashable savings for government, and link investors returns to impacts with *attribution proven through rigorous evaluation* (Fraser, Tan, Lagarde, et al., 2018).

Several arguments underlie the hypothesis that SIBs might better motivate impact estimates. First, investors have more reserves and can wait for payments, passing improved incentives onto providers (Pauly and Swanson, 2013). Second, investors' quantitative know-how could introduce an additional layer of "rigor and scrutiny," bringing quantitative empiricism to service delivery (Burand, 2012). Third, SIBs could improve evidence because evidenced programs are more likely to attract investors. Finally, contract conditions require evidence for payout (Fraser et al., 2020) because impact estimates ensure that outcome payments are earned in a valid, attributable way (Fox and Albertson, 2011). This brings us to our first hypotheses:

H1a. SIB contracting will motivate impact evaluations.
H1b. SIB contracts will directly incorporate impact evaluations into contract terms.

2.2 Why not impact?

Alternatively, there are reasons why SIBs might not improve upon PbR evaluations. SIBs face many of the same obstacles as PbR like ethical considerations around RCTs or evaluation costs. New obstacles may be introduced like unclear or complex data ownership or less data sharing due to competition (Fraser, Tan, Lagarde, et al., 2018b). Beyond these, we identify four potential reasons that SIBs might not motivate impact estimates or their integration into contract terms.

Strong *ideologies and beliefs* can motivate involved stakeholders to avoid the risk of estimating impacts, which might show SIB-financed programs to fail—while stakeholders believe that irrespective of a single project's result, the idea has merit. This may apply to either the impact of the SIB-financing mechanism (i.e., stakeholders may desire a SIB-financed project to be seen as "successful" so future SIB-financed projects can be developed) or the impact of the intervention (i.e., stakeholders may want their chosen intervention to be seen as "successful" so it may be replicated elsewhere).

The SIB literature exhibits a strong *public sector reform narrative* (Fraser, Tan, Lagarde, et al., 2018) building on New Public Management principles to argue that public and non-profit organizations have shortcomings in terms of service design, delivery, and accountability, and have so far failed to find solutions to entrenched social problems. This is grounded on two hypothesized mechanisms. First it is argued the *market* can improve public services through increased contact with the private sector, exposure to entrepreneurial ideologies, market mechanisms to manage performance, and improved accountability (Cohen, 2011). The inclusion of financial sector actors and the disciplinary potential of financial investment is attractive to those with these beliefs (Cohen, 2011). Second, the public sector reform narrative is linked to beliefs about *inter-sectoral collaboration*. In the tradition of corporatist institutional norms (Rhodes, 2001), many believe that public-private-philanthropic partnerships offer the best solutions to intractable issues. Those with strong beliefs might feel that in the long run the idea has merit irrespective of the impact of one single project.

Similarly, political goals may undermine impact evaluation. Evaluation is a tool of legitimation suggesting policy makers take evidence seriously and make pragmatic and objectively justified decisions (Ahonen, 2015). Simultaneously, using an academic impact evaluation, with a real risk of programs showing no impact, might be politically risky. From this perspective, running evaluations that are non-impact estimates, biased towards showing success, are the most attractive option. Biased approaches might be possible if those with a stake in success are the ones designing contracts and evaluations (Cooper et al., 2016; Warner, 2013).

Further, SIBs are often seen as experimental tools – apposite to piloting new interventions (Fraser et al., 2020). The logic of piloting might undermine evidence generation (Ettelt et al., 2015). In experimental contexts, learning or the appearance of learning can be important—so alternative forms of evaluation that offer learning opportunities without the risk of "failure" are attractive. If a program is a pilot, the unwillingness to conduct an impact evaluation might also be considered a short-term condition with planned improvements in evaluation for future iterations.

Finally, the structure of fiscal federalism might undermine impact evaluation. SIBs can be vehicles to transfer funds between levels of governments—with poorly coordinated transfers motivating irrational spending (Keen and Kotsogiannis, 2002; Oates, 2005) and inflating budgets (Rodden, 2003). When SIBs offer unconditional transfers to local government, they are attractive to local government irrespective of impact. Local stakeholders might wish to avoid the risk of showing failure in an impact evaluation, which would lead to the curtailment of future transfers.

This literature suggests hypotheses in direct opposition to H1:

H2a. SIB contracts will not motivate impact evaluations
H2b. SIB contracts will not incorporate impact evaluations into contract terms.

Additionally, the literature implies specific reasons that could underline H2a and H2b.

H3. Ideology and beliefs, political goals, the logic of piloting, and fiscal federalism underlie SIBs' failure to motivate impact estimates and their integration into contracts.

Early evidence from the UK suggests SIBs have not motivated impact measurement with evaluations rarely using (quasi-)experimental designs (Carter et al., 2018; Fox and Albertson, 2011; Fraser, Tan, Lagarde, et al., 2018) and often conflating outcomes with impacts (Fox and Morris, 2021). Further, evaluations focus entirely on programs, not financing. Although SIBs have motivated additional data collection, expertise still comes from social service providers (Fraser et al., 2020) with investors reacting to contract incentives by intervening more in operations than in data and evaluation (Cooper et al., 2016). Regarding the integration of impact measurement in contracts, evidence from the UK suggests that SIBs have largely used non-impact evaluations

in contracts to offer safe investments with high returns (Berndt and Wirth, 2018). Recently there has been a shift towards using rate cards in the UK, where payments are based on prices from general correlations of outcomes with government costs, making program impact (and costs) irrelevant to contract terms. It is currently unknown: 1. Whether SIBs have also failed to generate impact estimates and their integration into contracts in non-UK contexts and 2. The reasons for this—both in the UK and elsewhere.

3. Methods

We examine five cases of SIBs funding ALMPs. To select the cases, we contacted all 39 SIBs in Europe targeting employment outcomes at the time we initiated the study. Twenty-nine rejected our proposal; 10 verbally agreed to participate; 6 signed letters of support. We worked with five of these cases, excluding one where employment was a marginal outcome. Our cases might not be representative of all SIBs. Each program had its own reason for participating (e.g., one provider had a strong interest in the quantitative analysis that the study promised). This does not invalidate the inferences from our cases; it means results should be interpreted considering the case selection process, which is potentially biased towards cases with an interest in evidence.

The analysis uses comparative case study methods to explore the perceptions and broader narratives offered by informants reflecting their experiences of designing and delivering services through a SIB. Interviews were conducted until "data saturation" (Glaser and Strauss, 1967). The interviews were transcribed, and transcripts were coded using NVivo 12. The data were interrogated repeatedly to understand key emergent issues using the principles of "constant comparison" (Glaser, 1965). The analytical approach used both inductive and deductive reasoning – exploring emergent issues alongside insights from wider policy, management, and economic theory (Langley, 1999).

INSERT TABLE 1 HERE

In addition to the interview data, we used documentary data from the sites. For instance, most SIB financed programs underwent a local evaluation and wherever possible we used these documents. We were not granted access to contracts between parties. We therefore explored the nature of the contractual requirements through evaluations, audit reports, and interviews. We created case study reports for each site and graphically mapped the governance and financial relationships between partners. These graphics were shared with informants to confirm accuracy.

Our analysis proceeds in two steps. The first section uses documentary evidence and interviews to understand the evaluations conducted and the integration of evaluation into contracts. The second section uses interviews to understand the dynamics around evidence generation.

4. Analysis

4.1 Cases, Evaluation, and Contracts

We examine five SIB-financed ALMPs beginning 2013 to 2017 and ending 2016 to 2021, in the UK, Netherlands, Switzerland, and Germany—the first wave of SIBs in Continental Europe, but the second in the UK.

INSERT TABLE 2 HERE

Germany

The German ALMP was offered by a for-profit educational institution with significant experience. The program served 69 youths who were disconnected from society (discontinued training, unemployed, no contact with public services) of an estimated 300-450 eligible individuals. Government informants reported the eligibility criterion was defined such that the program would be a complement, not substitute for, or compete with, existing programs. The program offered counselling, job search, and recreation.

Furthering evidence-informed policy was a key motivation mentioned by all interviewed parties but there was some ambiguity in what that meant. The intermediary emphasized that there are many ways to measure impact, and the most important thing was to collect data so they could "show what they had achieved" while an investor argued a comparison group was unnecessary, as the counterfactual employment for such a disadvantaged group could be assumed to be null. Both do not make sense in a counterfactual causal framework. A government official suggested that the program was too small to generate statistically significant results-- which is possibly true depending on the size of the effect.[1]

Against this backdrop, the project chose to use an audit checking the number of participants achieving certain outcomes and negotiated flat targets (e.g., 20 youths placed in work, apprenticeship, or education). Flat targets theoretically make it easier to "game the contract" by recruiting more participants or by moving clients from other programs at the same provider, though the provider reported having trouble finding eligible participants. Recruitment was stopped directly after achieving the targeted placements—a choice that would have biased impact estimates had they been run.

Contract incentives were absolute and unbalanced: in the case of failure (e.g., placing 19 instead of 20 youths in work) the investors would lose their entire investment and in the case of success they would earn a 1% annual interest. Investors felt that they had no bargaining position given the low (essentially 0%) interest rate on government bonds at this time. Government representatives felt that high returns could be politically problematic.

Switzerland

The Swiss ALMP was run by an experienced non-profit provider. They were funded by the same government agency to simultaneously run two ALMPs for refugees, one as a SIB and the other with per-participant financing. The SIB served 241 refugees, with the more work-ready enrolled in the SIB-funded program. While both programs offered language classes, work training, and job search, the SIB offered more post-employment support and more resources, resulting in more intensive support like lower student-teacher ratios in language classes.

Measuring impact was an important motivator for the SIB with the investor arguing that it was important to make a public statement about evidence in policy. The government agreed, originally planning an RCT, but they ultimately cancelled it, enrolling the control group to serve more clients under the generous SIB funding which had 50% more funding per participant, even *after* enrolling the control group.

Stakeholders used diverse definitions of "impact." Interviewed government officials referred to the raw number of employed participants as evidence that the program had increased employment. In contrast, investors implied that impact requires a comparison with non-participants, critiquing the government's target as somewhat high *given the general employment rate of refugees in the canton*. The evaluator felt both, a negotiated benchmark or one based on the average employment rate, were insufficient and proposed a quasi-experimental approach using administrative data after the RCT was cancelled, which was rejected. Ultimately, the published local evaluation focused on raw outcomes, though it included some unadjusted comparisons to the non-SIB group at the same provider, using data provided by this research project.

Contracts used negotiated flat targets including payment by participant (e.g., serve at least 120 individuals) as well as payments on outcomes like employment.[2] The enrollment of the RCT control group assured achievement on the participation target and made flat employment targets more achievable. Further, targets around work subsidies were dropped after it became evident that employers did not want them. In the online supplementary materials, we show that these changes pushed the program over the threshold to "success."

Investor returns were capped at a loss/win of 1% annual interest. As in Germany, the investor felt they had limited negotiation power with negative government bond rates. The government described the contract as

[1] With the observed employment rate and a comparison group of 300 individuals, an employment rate some 15 percentage points lower than the observed employment rate would yield a statistically significant effect.

[2] Permanent positions with at least 21 hours/week for 20 high-skill clients and 24 middle-skill clients (weight: 18% for each), 2. 24 high-skill and 32 middle-skill individuals in a standard job irrespective of part-time or contract type (weight: 15% for each), and 3. 30 high-skill and 52 middle-skill individuals that do not break off an entrance into education or employment (weight: 12% for each).

low-return / low-risk. With SIBs being so complex, however one government employee did not seem to understand repayment terms falsely stating that the SIB allowed the government to "lead a program without paying."

UK

In the UK the non-profit provider had decades of experience offering programs for disadvantaged youths and managed different funding streams with overlapping clientele and programming. The program served 1,300 youths ages 14-20 at risk of becoming NEET (not in education, employment, or training), offering in-school supports like mentors and group work. The provider reported SIB funding increased pressure but also increased flexibility, due to more resources. "Creaming" and "parking" were both themes in interviews, with front-line workers reporting pressure to serve those clients who were likely to achieve payable outcomes (creaming) and to stop serving those who had already achieved payable outcomes (parking). The SIB was largely financed through the central government with a small local match. At the time of the SIB, the local government faced acute austerity measures.

As in Germany and Switzerland, evaluations took an audit approach, counting the number of participants achieving goals. The provider was tasked with collecting documentation, with front-line workers complaining about the administrative burden and the additional "transactional" layer in client interactions. Further, some participants found it embarrassing for their social worker to contact their employer. The central government initially planned an impact evaluation but ultimately cancelled it. The evaluator reported that the government felt it was too difficult to get the data, complex and expensive. At the same time, the government agency commissioning the program had signed individual consent from participants to link their data for research purposes and owned the necessary administrative data sets on educational and employment outcomes. The evaluator and provider expressed regret that impact was not measured.

The contract used a "rate card" to assign payments to outcomes. Rate card prices are likely based on the correlation of outcomes (e.g., an educational certification) with government costs (e.g., future benefits). Outcomes included 7 "hard" targets related to employment and education and 4 "soft" targets related to school attendance and attitude. One would expect uncertainty around potential payouts given that rate card payments are unrelated to the cost or original investment, but (perhaps because of the soft targets) one investor reported that they understood from the start that they would receive a high interest rate.

While investor returns are not public, they were many times the government bond rates. Some believe these high payments were related to missteps in the rate card design. All interviewed parties felt the contract was not designed to deliver savings, with some local government officials and provider informants feeling that necessary social service resources were wasted on administrative costs and investor payouts.

Netherlands 1

In the first Dutch case the small for-profit provider had less experience in ALMPs before running the SIB. They served 160 youths ages 17-27 on social benefits. The program originally focused on mentoring as an intervention and self-employment as an outcome, but with SIB funding they broadened services to include job search and training and outcomes to include employment and education.

All parties reported that impact evaluation motivated the SIB but were generally dissatisfied with the evaluation. A local company estimated a Cox Hazard model predicting social assistance exit based on 40 years of benefit data. The auditor used the estimated model to calculate expected benefit days for participants. Because benefits had been a long-term support until activation policies were introduced in the 1990s and scaled up in the 2000s (Cremers, 2018), the data basis was not a match and most controls could not be considered in the prediction estimates due to missing data for SIB participants.[3] The company that estimated the model outlined these limitations and expressed concerns about striking a balance between solid modelling and comprehensibility, as well as how quickly the model was adopted.

[3] Having a family member or partner on benefits, benefits receipt history, age, number of children, education, region of origin, and whether the case was post 2009 (due to the recession).

Government informants understood the limitations that had been highlighted to them, but the company was right to be concerned about comprehensibility. In interviews a government employee talked about the recession dummy capturing an increase in assistance days, while in fact the coefficients were in the opposite direction, and audit reports have a key mathematical error in the calculations for baseline assistance durations, potentially accidentally guaranteeing maximum payments.[4]

Dutch investor returns were continuous. There was a flat payment of 40 euros assuming a baseline reduction of 154 days in median assistance followed by an additional payment of 15 Euros for every additional day to a maximum of 210 days, yielding a maximum investor annual rate of return of 12% for a total possible of 32%. Government commissioners reported they were willing to shoulder high costs to establish the model.

Netherlands 2

The second Dutch case was delivered by a for-profit provider with significant experience and diverse funding prior to and following the SIB. The program served Dutch vocational workers on benefits, helping them find employment in Germany. The official target group was those with completed vocational education, though less well-trained participants were referred to the program. Services included 2 weeks of German classes, case management, certifications, 4 weeks of German during an apprenticeship, and a 6 month to one-year trial contract supported by weekly and then monthly check-ins with the employer and employee. When the project was set up, the unemployment rate in the Netherlands was higher than in Germany, though after 15 months, the situation reversed, with 4 times as many workers commuting from Germany to the Netherlands as vice versa.[5] With only 33 participants enrolled halfway (compared to a target of 69 at that point in time), the program was discontinued.

As in the other cases, impact measurement was an important motivation for all parties. With many of the same actors as in the first Dutch SIB, the project used a simpler evaluation strategy comparing participants' outcomes to a control group of individuals on social welfare in the same municipality. The control group was generally more employable, with fewer production sector workers, more women, more prime-age workers, a higher average education, and less long-term benefit dependency. As such, without statistical adjustment, the evaluation and related contract terms were biased against success. Despite this bias, the evaluation found significant and strong reductions in benefit days and long-term benefit receipt.

Returns were similar to the first Dutch SIB with a maximum return of 10% per year. Provider contracts again guaranteed a certain minimum number of clients though in practice they referred fewer. Both the provider and investors reported that the locality ran its own lower-cost ALMP to which they referred more easily placeable clients. Given an economic shift towards more employment opportunities on the Dutch side of the border and the low number of referrals, the government decided to cancel the program early. While final payments were not concluded at the time of this study, respondents guessed that only 30% of payments would be made to the provider and investors, and the provider organization reported that they likely lost money.

Comparison

In all five cases improving evidence was a key stated motivation for the SIB, and though evaluators pressed for impact estimates, across the board other approaches were used. In most cases there was a basic audit, reporting the number of participants achieving certain goals. These audits created significant burdens on social service providers, even when the government owned the relevant administrative data. In the Dutch cases there was a greater attempt to estimate impact, though in both cases the approach was flawed, including undetected mathematical errors and misunderstandings among government officials. In the Swiss case, due to data collected for the same study funding this paper, the final report offered a comparison with clients in a non-SIB funded program, but without statistical adjustment.

[4] Auditing reports predict expected days on social assistance using $\frac{m}{e^{\beta_1 X_1} * e^{\beta_1 X_1}}$ (where m is median and the X's refer to the control variables) rather than $m * e^{\beta_1 X_1 + \beta_2 X_2}$. An example prediction in the auditing report for a woman age 25-34, with an unknown education level from Suriname (an individual whose hazard ratios are all less than one) is reported to have an expected median duration of 1390 days on benefit (compared to the median of 334 days) while the correct calculation would have been 80 days.

[5] https://opendata.grensdata.eu/#/InterReg/de/

Contracts used rate cards (UK) negotiated flat targets (Switzerland and Germany), and imperfect "impact" estimates (the Netherlands). Most contracts were acknowledged to be low-risk and biased towards success—except for the second Dutch SIB, which was biased towards failure. The level of returns varied dramatically across sites. Key actors in Germany and Switzerland oriented towards low government bond rates, while they were irrelevant to Dutch and British stakeholders.

We reject H1a and H1b that SIB funding increases impact evaluation and its integration in contracts, and find evidence for H2a and H2b, that the SIB *failed to* motivate impact evaluation and its integration into contracts. The next section explores why.

4.2 Why not impact evaluation?

Why was high-quality evidence ignored in practice, despite claims among all actors that impact measurement was a key reason to use a SIB and solid advice from evaluators on how to achieve that goal?

First, we would note that while all actors mentioned the importance of measurement, there was feedback that having *high-quality* evidence was not of utmost importance. For example, in Germany, it was stressed that impact measurement was marginal to other goals.

> *Nobody really cared if they are measuring their impact or not, they get their funding maybe for some other reasons or because people believe that this is a good program, but like this possibility to really show what you can achieve didn't really – or let's say was not in the focus of the public funding partners.*
> Intermediary Germany

What then were the primary goals?

Ideology

Perhaps, more than anything else, across all five sites, informants drew on discourses that closely corresponded with the public sector reform narrative and anti-bureaucratic sentiments. They use the word "impact," without necessarily meaning impact measurement. Foremost amongst these were investors and intermediaries but also government commissioners. One quote from the Dutch site conveys this perspective.

> *We also want to change the culture in local government, a culture which is now still a lot based on input and throughput but not on output and impact… So, we want to use Impact Bonds to strengthen the culture far more on …* ***impact****.*
> Investor, Netherlands 2

In all the cases we saw optimism around public sector reform narratives relevant to cross-sectoral collaborative partnerships, explicitly prioritizing this over solid empirical evidence.

> *One of the benefits was actually to have all the players in the field to solve a problem at the same table, so private - so investors, private entrepreneurs, the government and the social institution.*
> Investor, Switzerland

> *When we found out that the [evaluation] model was not very useful, there was not a problem because everybody thought we needed to experiment, we needed to show ourselves and the Netherlands that you can work together with a social entrepreneur with a social investor and data collector.*
> Government, Netherlands 1

Actors supporting public sector reform narratives saw their commitment to the idea of the SIB innovation as greater than these individual ALMPs, wanting to show that SIBs could improve poor public service delivery. They had a long-term commitment to create an alternative model and measuring impacts was not only less important than establishing the model, but negative impact, which could be random for any individual program, could threaten that goal.

Provider perspectives were more ideologically diverse. The head of the UK provider might be said to have a public sector reform narrative, reporting that finance brought a new level of scrutiny to the project, shaking up

the status quo, "*a lot of the questions we get from [the investor] are actually related to the outcomes rather than the finance.... And I find their input really helpful, they support and challenge."* But other provider CEOs, as in Germany saw it as irrelevant, *"it was another source of money and we didn't rebuild our structure or process or something else… [The SIB led to] a different kind of controlling but we did what we did before."*

Among practitioners there was mixed evidence in relation to the public sector reform narrative. Some assessed the SIB positively, but others found it had no impact while still others stressed negative effects, illustrated by the following two quotes:

> *The numbers [and financial incentives] are not that important for us… The bonus system will change maybe something if it goes in my pocket!*
> Front-line worker Switzerland

> *The investors, they fluctuated between being really caring about the young people, but just going, thinking yeah but where's the money coming from, coming in you know…they'd set a target which was totally unrealistic, but they wanted a hundred thousand every month …[we] were running around like headless chickens really just trying to get outcomes.*
> Manager, UK

Support for the public sector reform narrative among the government, investors, and intermediaries was more relevant to contracts and evaluations than amongst providers.

Interests

Politicians, government commissioners, investors and intermediaries all reported a strong desire to avoid perceptions of policy failure.

> *A politician, he's [sic] there to sell success, so he loves [to] say 'I set up a new mechanism and we helped, like, 200 citizens and it was a big success' – that's a good story. Not… 'OK, we tried to help 200 people, it didn't work out but fortunately investors are [paying for this].*
> Investor Netherlands 2

> *The SIB experiment is done, and some credit may be taken for being innovative, whilst maintaining ultimate control and limiting the risk of political exposure / perception of scandal.*
> Government Switzerland

Part of the desire to avoid failure had to do with relationships. In every site we looked at, stakeholders knew they would continue to work together in the future. Although some respondents were negative about the SIB approach, in the short run, no one wanted to embarrass their project partners. Rather, actors wanted the individual program evaluations to show "success," while impact was more relevant to private planning of future contracts.

For example, in Switzerland, in the short-term, the government made choices in contract terms that guaranteed this specific SIB's success, but, in private, employment rates that were lower than anticipated were noted, with a Swiss government commissioner reporting he had "*no political desire or institutional need to build on this learning going forward*."

In NL2, which had the evaluation closest to an impact estimate, commissioners took choices based on other criteria. With a shift in economic climate, the high cost of the SIB-funded intervention, and the low skills of program referrals, the local commissioner made the politically difficult decision to abort the program. The decision, however, had less to do with impact and more to do with the need for the program.

We realized OK is this going to change in the coming years, the economy is probably going to grow, companies are really interested in technical skilled people so… normal employers wouldn't take them but they said they might, were too heavy cases, they didn't show up … so those two combined made us realize that it's not going to be profitable for people involved and for the government ... OK we're going to use it, the possibility to terminate it.

Government, Netherlands 2

In most of our cases, there was a will to show a successful project in the short-term, which meant avoiding impact estimates). In the long run, then lacking program-specific impact estimates, the government looked to other indicators (such as a comparison to overall refugee employment rates or unemployment rates) to make their decisions.

In the UK at the local level, these calculations were different, as national funds heavily subsidized local programming. The SIB was thus seen as a tool to tap into additional funding. Whether SIB financing had impact was thus irrelevant.

> *Local authority's a tough environment to work with, has been for the last ten years, does feel like you're kind of the axe man really. And I guess [that's where] our interest in Social Impact Bonds and the whole Social Investment world probably five or six years ago started to emerge…to respond to what was coming…which was that huge budget challenge for local authorities, so we were looking around for alternative ways of doing public services and drawing in money.*
> Local Government Commissioner, UK

> *We knew SIBs were the new kid on the block, so I think our thinking was – and I think this was borne out actually – was if we get in there now, we get learning how to do these things then when the next round of money comes out from government… then we're in a good position [to get further funds].*
> Service Provider, UK

In the UK, central "outcomes funds," offering subsidies for local services in an era of austerity, reduced the importance of evidence. At the federal level it is still unclear why impact the evaluator's proposals for impact estimates were rejected, as national government actors declined to be interviewed. We can say that the UK central government has promoted SIBs for more than a decade with broad political support, moving away from early (quasi-)experimental approaches towards rate cards.

Experimentation

Parties at every site except NL2 focused on experimentation and learning, with a desire to successfully conclude an initial experiment.

A Dutch commissioner reported he was recruited because their primary goal was to generate innovation *"And one of the reasons [the municipality] wanted me there was that I had some experience with Social Impact Bonds and they were looking for new solutions in a rather traditional field, social affairs."*

In Germany informants reported that they were conscious of the different criteria used in an experiment compared to a full-scale program.

> *We did this project as a model [pilot] project. We looked for a topic that is currently not being worked on in-house, we had a special topic, we took money from an area where there especially is budget for experiments and innovations and so it worked quite well. If you do this again now, or do it regularly as a project, a whole series of further questions will naturally arise. Here in the experimental area, you simply have more leeway than if you introduce it regularly.*
> Government Commissioner, Germany

We see this also in the first Dutch case, where despite an awareness of flaws in evaluation and target design, actors took a longer-term view towards learning and developing the SIB model, with impact evaluation being a long-term goal.

> *We didn't have [the best] data so… the vice mayor, our director, he said 'well we will just work with something because otherwise you never get the perfect model', we now have proof that the Social Impact Bond comes [from] all parties working together, getting their strength and improving, you could say, programs, social programs, it's for us more important to be [accepting] numbers if we know how much money we can save, but if you do Social Impact Bond again, then there have to be huge improvement in the measuring model… And that's why we, in our second Social Impact Bond … we used a control group.*

Commissioner, Netherlands

Across our cases there was an experimental *raison d'être* which impacts upon how "success" may be interpreted, and impact was not necessary to claim success. This is perhaps particularly so for Switzerland, Germany, and our first Dutch cases– where we are looking at the first SIBs in each respective country. It is harder to argue that the UK case was an experiment, as it was by far not the first national SIB, but the SIB was still the first SIB in the local area, and the idea was framed as innovative and experimental.

5. Discussion

In this paper we examined whether a relatively new type of contract, the SIB, motivated impact estimates in evaluations, whether impact estimates were used in contract terms, and why.

There were several findings that were uniform across cases. Although we found universal enthusiasm for using SIBs to motivate impact estimates in principle, in practice while SIBs motivated evaluation, they did not motivate impact evaluation. One exception was the second Dutch case, in which actors used their earlier experiences to improve both evidence collection and contract design, though the final approach was still flawed. Perhaps most importantly, even though the Netherlands saw significant improvements in measuring program impact, they still did not consider measuring investor impact—the relevant metric for investor contracts.

In most cases we found evidence confirming concerns voiced by Heinrich and Choi (2007) and White (2010), that there is great variability in the understanding of what solid evidence is. Stakeholders had different understandings of impact and counterfactuals, with some considering any job placement to be an indication of impact. Further, lacking statistical knowledge led to errors and misinterpretations.

We also found evidence supporting Ahonen (2015 and Cox and Barbrook-Johnson (2021) suggesting that government commissioners have other concerns, and that evaluation can be more about legitimation than finding out what works. Further, stakeholders (excluding some providers) believed in the need for *public sector reform* (Fraser, Tan, Lagrade, et al., 2018). Actors were not interested in the impact of a single case per se, but rather in in showing the viability of a model promising improvements through complex combinations of market, entrepreneurial, and collaborative mechanisms (Fraser et al., 2022). Further, we found significant evidence that viewed as a pilot (Ettelt et al., 2015), actors were enthusiastic about showing success, to further future learning opportunities—considering rigorous impact estimates to be relevant for later expansions.

Unique to the UK case was the impact of inter-governmental transfers. In both Germany and the UK, a higher level of government funded the bulk of a more local intervention. In Germany this was not reported to play a role—potentially due to sufficient local financing regardless. However, in the UK, the local government and provider stressed that the SIB was attractive irrespective of impact, because it was a way to tap into federal funding, reflecting concerns with financial accountability typical to the literature on fiscal federalism (Oates, 2005; Rodden, 2003).

Overall, the dynamics on the ground were complex with interests, ideology and beliefs, relationships, and inter-governmental transfers all playing a role. In a world without these factors, most stakeholders were interested in impact estimates, but with them, impact was simply not a priority.

6. Conclusion

Evaluators and policy makers should be wary of arguments that contracts might encourage evidence-informed policy. There are many perspectives, interests, and understandings at play in these negotiations— real world policy making is messy and contested. SIB financing increased the amount of data being collected but did not generate impact estimates and data collection created unnecessary administrative burdens on providers. Stakeholders tend not to be well-trained to understand evidence generation and if a bad outcome means a loss of face, contradicts strongly held beliefs, inhibits the further testing of a model they believe in, or threatens funding for a program area they find important, many appear ready to sacrifice the goal of evidence

generation. Further, incentives encouraged providers to engage in practices that biased estimates and to engage in gaming practices long observed in the PbR literature.

What does this mean for the potential to encourage impact evaluation and evidence-informed policy? Although we found that high stakes (both financial and in profile) might undermine evidence generation, key stakeholders would have embraced low-profile impact estimates to inform policy and programming decisions. These results can be interpreted as implying that a collaborative and non-threatening approach where evaluators work hand in hand with policy makers and practitioners to give them the answers they need—without the potential for negative headlines or funding cuts—might be a more promising road to finding what works. This may align with a Human Learning Systems approach (Lowe et al., 2021).

An important avenue for future investigation is whether a collaborative learning model of impact evaluation can be effective, looking at whether the proper level of resources would be committed when actors have no stake in results. Further, it is unclear how the results of such evaluations would feed back or be disseminated to policy and programming. A low-stakes evaluation designed to be used only by direct stakeholders could fail to inform the practices of other governments or providers. A key research agenda is to examine whether collaborative low-stake evaluations motivate more impact estimates, and how such evaluations inform a community greater than those directly involved.

7. Bibliography

Adams GB and Balfour DL (2010) Market-Based Government and the Decline of Organizational Ethics. *Administration & Society* 42(6): 615–637.

Ahonen P (2015) Aspects of the institutionalization of evaluation in Finland: Basic, agency, process and change. *Evaluation* 21(3): 308–324.

Banerjee AV, Duflo E and Kremer M (2016) The influence of randomized controlled trials on development economics research and on development policy. In: *The State of Economics, the State of the World*, pp. 482–488.

Barnow BS (2000) Exploring the relationship between performance management and program impact: A case study of the job training partnership act. *Journal of Policy Analysis and Management* 19(1): 118–141.

Berndt C and Wirth M (2018) Market, metrics, morals: The Social Impact Bond as an emerging social policy instrument. *Geoforum* 90: 27–35.

Bitler M, Corcoran S, Domina T, et al. (2019) *Teacher Effects on Student Achievement and Height: A Cautionary Tale*. 26480, Working Paper, November. National Bureau of Economic Research.

Boaz A, Davies H, Fraser A, et al. (2019) *What Works Now? Evidence Informed Policy & Practice*. Bristol: Policy Press.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health intervention. *Social Science & Medicine* 75: 2299–2306.

Brown TL and Potoski M (2003) Managing contract performance: A transaction costs approach. *Journal of Policy Analysis and Management* 22(2): 275–297.

Burand D (2012) Globalizing Social Finance: How Social Impact Bonds and Social Impact Performance Guarantees Can Scale Development Articles and Comments Arising from the 2012 Fall Conference: The Law & Finance of Social Enterprise. *New York University Journal of Law and Business* 9: 447–502.

Card D, Kluve J and Weber A (2010) Active Labour Market Policy Evaluations: A Meta-Analysis*. *The Economic Journal* 120(548): F452–F477.

Card D, Kluve J and Weber A (2018) What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association* 16(3): 894–931.

Carter E, Fitzgerald C, Dixon R, et al. (2018) *Building the tools for public services to secure better outcomes: Collaboration, Prevention, Innovation*. Available at: /knowledge/resources/evidence-report/ (accessed 14 January 2019).

Cohen R (2011) Harnessing social entrepreneurship and investment to bridge the social divide. In: *EU Conference on the Social Economy*. Brussels: European Economic and Social Committee:

Cooper C, Graham C and Himick D (2016) Social impact bonds: The securitization of the homeless. *Accounting, Organizations and Society* 55: 63–82.

Cox J and Barbrook-Johnson P (2021) How does the commissioning process hinder the uptake of complexity-appropriate evaluation? *Evaluation* 27(1): 32–56.

Cremers J (2018) Social assistance in the Netherlands. Chronique Internationale.

Deaton A and Cartwright N (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210: 2–21.

Duflo E (2020) Field Experiments and the Practice of Policy. *American Economic Review* 110(7): 1952–1973.

Ettelt S, Mays N and Allen P (2015) Policy experiments: Investigating effectiveness or confirming direction? *Evaluation* 21(3): 292–307.

Fox C and Albertson K (2011) Payment by results and social impact bonds in the criminal justice sector: New challenges for the concept of evidence-based policy? *Criminology & Criminal Justice* 11(5): 395–413.

Fox C and Morris S (2021) Evaluating outcome-based payment programmes: challenges for evidence-based policy. *Journal of Economic Policy Reform* 24(1): 61–77.

Fraser A, Tan S, Lagarde M, et al. (2018) Narratives of Promise, Narratives of Caution: A Review of the Literature on Social Impact Bonds. *Social Policy & Administration* 52(1): 4–28.

Fraser A, Tan S, Boaz A, et al. (2020) Backing what works? Social Impact Bonds and evidence-informed policy and practice. *Public Money & Management*, 40(3): 195–204.

Fraser A, Knoll L and Hevenstone D (2022) Contested Social Impact Bonds: Welfare Conventions, Conflicts and Compromises in Five European Active-Labor Market Programs. *International Public Management Journal*.

Gaffey V (2013) A fresh look at the intervention logic of Structural Funds. *Evaluation* 19(2): 195–203.

Girth AM (2017) Incentives in Third-Party Governance: Management Practices and Accountability Implications. *Public Administration Review* 77(3): 433–444.

Glaser BG (1965) The Constant Comparative Method of Qualitative Analysis. *Social Problems* 12(4): 436–445.

Glaser BG and Strauss AL (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter.

GOLab (2022) Social Impact Bonds Database. Oxford University.

Haskins R (2018) Evidence-Based Policy: The Movement, the Goals, the Issues, the Promise. *The Annals of the American Academy of Political and Social Science* 678(1): 8–37.

Haskins R and Margolis G (2015) *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Brookings Institution Press.

Heckman JJ, Heinrich C and Smith J (2002) The Performance of Performance Standards. *The Journal of Human Resources* 37(4): 778–811.

Heinrich CJ (2007) False or fitting recognition? The use of high-performance bonuses in motivating organizational achievements. *Journal of Policy Analysis and Management* 26(2): 281–304.

Heinrich CJ and Choi Y (2007) Performance-Based Contracting in Social Welfare Programs. *The American Review of Public Administration* 37(4): 409–435.

Heinrich CJ and Marschke G (2010) Incentives and their dynamics in public sector performance management systems. *Journal of Policy Analysis and Management* 29(1): 183–208.

Huitema D, Jordan A, Munaretto S, et al. (2018) Policy experimentation: core concepts, political dynamics, governance and impacts. *Policy Sciences* 51(2): 143–159.

Jennings ET and Hall JL (2012) Evidence-Based Practice and the Use of Information in State Agency Decision Making. *Journal of Public Administration Research and Theory* 22(2): 245–266.

Keen MJ and Kotsogiannis C (2002) Does Federalism Lead to Excessively High Taxes? *American Economic Review* 92(1): 363–370.

Kok L, Tempelman C, Koning P, et al. (2017) Do Incentives for Municipalities Reduce the Welfare Caseload? Evaluation of a Welfare Reform in the Netherlands. *De Economist* 165(1): 23–42.

Koning P and Heinrich CJ (2013) Cream-Skimming, Parking and Other Intended and Unintended Effects of High-Powered, Performance-Based Contracts. *Journal of Policy Analysis and Management* 32(3): 461–483.

Langley A (1999) Strategies for Theorizing from Process Data. *Academy of Management Review* 24(4): 691–710.

Lowe T, French M, Hawkins M, et al. (2021) New development: Responding to complexity in public services—the human learning systems approach. *Public Money & Management* 41(7): 573–576.

Martin LL (2005) Performance-Based Contracting for Human Services. *Administration in Social Work* 29(1): 63–77.

McBeath B, Jolles MP, Carnochan S, et al. (2015) Organizational and Individual Determinants of Evidence Use by Managers in Public Human Service Organizations. *Human Service Organizations: Management, Leadership & Governance* 39(4): 267–289.

Ngan P and Robinson H (2015) *Outcome-based payment schemes: government's use of payment by results - National Audit Office (NAO) Report*. HC-86. National Audit Office. Available at: https://www.nao.org.uk/report/outcome-based-payment-schemes-governments-use-of-payment-by-results/ (accessed 14 January 2019).

Oates WE (2005) Towards a Second Generation Theory of Fiscal Federalism. *International Tax and Public Finance* 12: 349–373.

Overholser GM (2018) Pay for Success Is Quietly Undergoing a Radical Simplification. *The Annals of the American Academy of Political and Social Science* 678(1): 103–110.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348–367.

Pauly M and Swanson A (2013) *Social Impact Bonds in Nonprofit Health Care: New Product or New Package?* 18991. NBER Working Paper. Available at: https://www.nber.org/papers/w18991 (accessed 14 January 2019).

Pihl-Thingvad S (2016) The Inner Workings of Performance Management in Danish Job Centers: Rational Decisions or Cowboy Solutions? *Public Performance & Management Review* 40(1): 48–70.

Pleger LE and Hadorn S (2018) The big bad wolf's view: The evaluation clients' perspectives on independence of evaluations. *Evaluation* 24(4): 456–474.

Rhodes M (2001) The Political Economy of Social Pacts. In: *The New Politics of the Welfare State*. Oxford University Press.

Rodden J (2003) Reviving Leviathan: Fiscal Federalism and the Growth of Government. *International Organization* 57(04): 695–729.

Stern E, Stame N, Mayne J, et al. (2012) Broadening the range of designs and methods for impact evaluations. Available at: http://www.ids.ac.uk/publication/broadening-the-range-of-designs-and-methods-for-impact-evaluations (accessed 22 June 2022).

Sullivan JX (2018) The Role of Nonprofits in Designing and Implementing Evidence-Based Programs. *The Annals of the American Academy of Political and Social Science* 678(1): 155–163.

Teele DL (2014) *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. Yale University Press.

Warner ME (2013) Private finance for public goods: social impact bonds. *Journal of Economic Policy Reform* 16(4): 303–319.

Webster E and Harding G (2001) Outsourcing Public Employment Services: The Australian Experience. *Australian Economic Review* 34(2): 231–242.

Wedel KR (1976) Government contracting for purchase of service. *Social Work* 21(2): 101–105.

White H (2010) A Contribution to Current Debates in Impact Evaluation. *Evaluation* 16(2): 153–164.

Captions

Figure A1: How target achievement changed with key administrative decisions

Tables

Table 1: Qualitative interview informant details

	Providers	*Government Commissioners*	*Investor/ intermediaries*	*Evaluators*	**TOTALS**
UK	10	1	1	1	**13**
Netherlands (2)	4	1	2	2	**9**
Switzerland	6	2	1	1	**10**
Germany	2	2	3	2	**9**
TOTALS	**22**	**6**	**7**	**6**	**41**

Table 2: Case Study Overview

	DE	*CH*	*UK*	*NL1*	*NL2*
Intervention	ALMP w/ counseling & youth work	ALMP w/ place then train	In-school support, some ALMP	ALMP w/ mentoring	ALMP w/ DE language & recertifications
Target group	NEET w/o contact to social services, work, education	Refugees	At-risk of NEET youth	Young adults on benefits	Unemployed vocational workers
Size	69	241	1300	160	33 (discontinued)