



Study on the benefits of using social outcome contracting in the provision of social services and interventions – a cross-country comparative assessment of evolving good practice in cross- sectoral partnerships for public value creation

Final Study Report

Written by Klimavičiūtė L. (PPMI), Chiodo V. (Polimi), De Pieri B. (Polimi), Gineikytė V. (PPMI)
January – 2021

PPMi



**POLITECNICO
DI MILANO**

EUROPEAN COMMISSION

Directorate-General for Employment, Social Affairs & Inclusion

Directorate Social Affairs Directorate

Unit C.1 — Social Investment Unit

Contact: Monika Chaba

E-mail: Monika.Chaba@ec.europa.eu

*European Commission
B-1049 Brussels*

**Study on the benefits of using
social outcome contracting in the
provision of social services and
interventions – a cross-country
comparative assessment of
evolving good practice in cross-
sectoral partnerships for public
value creation**

Final Study Report

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Authors: Luka Klimavičiūtė (PPMI), Veronica Chiodo (Polimi), Benedetta De Pieri (Polimi), Vaida Gineikytė (PPMI)

Scientific supervisors: Egidijus Barcevičius (PPMI), Mario Calderini (Polimi)

Acknowledgements: The authors would like to thank PPMI colleagues Laurynas Stankevičius, Gabija Skardžiūtė, Ugnė Pilkionytė, Loreta Matulevič and Polimi colleagues Alice Borrello, Federica Falomi, Francesco Gerli for their research, contributions in drafting the Annexes of this report, and peer review.

Manuscript completed in February 2021

This document has been prepared for the European Commission however it reflects the views only of the authors, and the European Commission is not liable for any consequence stemming from the reuse of this publication. More information on the European Union is available on the Internet (<http://www.europa.eu>).

Luxembourg: Publications Office of the European Union, 2021

© European Union, 2021



The reuse policy of European Commission documents is implemented based on Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Except otherwise noted, the reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of elements that are not owned by the European Union, permission may need to be sought directly from the respective rightholders.

PDF ISBN 978-92-76-30286-5

doi: 10.2767/492722

KE-01-21-065-EN-N

Table of Contents

Executive summary	8
Rationale	8
Methodology	8
Findings	8
Recommendations	12
Résumé analytique	16
Fondement	16
Méthodologie	16
Conclusions	17
Recommandations	21
Zusammenfassung	25
Begründung	25
Methodik	25
Ergebnisse	26
Empfehlungen	30
Introduction	34
1. Concepts and definitions	36
2. Study process and methods	41
2.1. Methodology	41
2.2. The programmes analysed	43
2.2.1. SOC schemes	43
2.2.2. Traditionally financed equivalents	48
3. Outcomes measurement	53
3.1. Definition of outcomes and data management	53
3.1.1. Definition of outcomes and metrics	53
3.1.2. Data collection and management	62
3.2. Evaluation methods used in SOC schemes	63
3.2.1. Non-experimental approaches without comparison	64
3.2.2. Non-experimental approaches with comparison	65

3.2.3. Quasi-experimental approaches	67
3.2.4. Discussion of overall evaluation design.....	68
3.3. Summary of findings on outcomes measurement.....	70
4. Effectiveness.....	71
4.1. Achievement of outcomes	71
4.2. Added value with regard to the benefits and drawbacks of SOC compared with TF models	80
4.2.1. Benefits	80
4.2.2. Drawbacks.....	82
4.2.3. Debated points	84
4.3. Scalability and replicability	86
4.3.1. Replicated and/or scaled SOC cases.....	87
4.3.2. Lessons learned about scaling and replication	88
4.4. Summary of findings regarding SOC effectiveness	89
5. Efficiency.....	90
5.1. Intervention costs	91
5.2. Operational costs.....	98
5.2.1. Comparison of set-up costs.....	98
5.2.2. Comparison of management costs.....	102
5.3. Summary of findings regarding SOC efficiency	104
6. Designing SOC schemes	106
6.1. Design elements.....	106
6.2. Risk sharing between commissioners and investors/service providers	111
6.2.1. Risk transfer in SIBs	111
6.2.2. Risk transfer in PbR schemes	113
6.3. Impact on third-sector organisations and social economy enterprises.	114
6.4. Impact on the public sector	116
6.5. Summary of findings regarding SOC design.....	117
7. Conclusions and recommendations	119
7.1. Conclusions	119
7.1.1. Effectiveness and added value	119
7.1.2. Efficiency	121
7.1.3. Design of social outcomes contracts	121
7.1.4. Outcome measurement.....	122
7.2. Recommendations	123

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

List of abbreviations

AA Pilots	Activity Agreement Pilots
ABLE	Adolescent Behavioural Learning Experience
CBO	Commissioning Better Outcomes Fund
CPA	Contract package area
CRC	Community Rehabilitation Companies
DIB	Development impact bonds
DWP	Department for Work and Pensions (UK)
EU	European Union
FTE	Full-time equivalent
HM	Her Majesty
IPS	Individual placement and support
LASARS	Local Area Single Assessment and Referral Services
MHEP	Mental Health and Employment Partnership
MoJ	Ministry of Justice (UK)
NAO	National Audit Office (UK)
ND25PL	New Deal for those aged 25+
NDYP	New Deal for Young People
NEET	Not in education, employment or training
NFA	No fixed abode
NGO	Non-governmental organisation
NPO	Non-profit organisation
NPS	National Probation Service (UK)
PbR	Payment-by-Results
PL	Provider-led
RCT	Randomised controlled trial
SBB	Social benefit bond
SIB	Social impact bond
SOC	Social outcomes contracting
SPV	Special purpose vehicle
TF	Traditionally financed
TR	Transforming Rehabilitation
TUPE	Transfer of Undertakings (Protection of Employment)
YC	Youth Contract

Executive summary

Rationale

Social outcome contracts (SOCs) is a relatively new type of mechanism in public service procurement. It focuses on harnessing the resources of the public, private, philanthropy and civil society sectors, with the goal of jointly implementing effective interventions in the public domain. The tool is said to provide potential to modernise public services and European welfare regimes: by encouraging a culture of performance measurement, supporting cross-sector partnerships and allowing new intervention models to be piloted and scaled — thereby fostering learning and social innovation. Nevertheless, little empirical research exists that compares SOC models with traditional financing (TF) mechanisms such as subsidies, grants, fee-for-service contracts, block contracts and in-house delivery. The study therefore aims to assess the effectiveness and efficiency of outcomes-based contracts in comparison to interventions delivered under traditional financing, as well as evaluating the outcome measurement methods applied to determine their impact. The study focuses exclusively on social services rather than public services more generally.

The study investigates two types of SOCs: Social impact bonds (SIBs) and other Payment-by-Results (PbR) schemes. We define SIBs as funding mechanisms in which investors provide up-front funding for a programme. The investors bear the risk of losing their investment if the programme fails to achieve its target outcomes, but are repaid in full, with an additional interest payment from the commissioner in the event that the SIB is successful. Some SIBs also include an intermediary, which helps to find the service providers, investors, and/or evaluators. In contrast, PbR programmes do not involve an external investor or an intermediary. Rather, service providers only receive payments from commissioners if the providers achieve the target outcomes. PbR models typically vary in terms of how much of the total payment to service providers is based on outcomes. In pure PbR models, 100% of the payment is based on outcomes. This is not the case in mixed PbR models, in which part of the payment is financed traditionally.

Methodology

Fifteen SOC cases were selected for in-depth analysis during the study, capturing diverse SOC models, geographical areas, and social services. To be included in the selection, the programmes had to be completed (or nearly completed), and evaluations had to be available to enable data collection. For this reason, a number of the SIBs selected were the first to be implemented in their respective countries. For each SOC model, comparable TF programmes were identified. The comparisons between SOC and TF aimed to assess the extent to which both programmes achieved the targets set for them, and at what cost. The comparisons also sought to determine the added value of SOC programmes in terms of benefits and drawbacks compared with TF, as well as identifying design elements that enabled replicability, scalability and supported third-sector organisations and social economy enterprises. The data were gathered via desk research, relying heavily on programme evaluation reports and audit inquiries, as well as 54 interviews with commissioners, service providers, evaluators, intermediaries, experts and investors.

Findings

Effectiveness and added value

To compare the effectiveness of SOC and TF programmes, we assessed the extent to which each programme was able to achieve the outcomes set at its outset. These outcomes

included the number of participants employed, reductions in re-offending rates, the number of children who avoided out-of-home care, and others, depending on programme objectives and service areas. **There is insufficient evidence to claim that outcomes-based contracts are more effective than traditional financing or vice versa.** This is partly due to the fact that in the majority of the comparable TF programmes, no outcome targets were set, and no rigorous evaluation was conducted.

However, our analysis also revealed that the achievement of targets did not necessarily define the overall effectiveness and success of a scheme. In fact, **some schemes that did not achieve their targets were nevertheless considered successful, since they were able to address the needs and expectations of the various stakeholders involved.** Thus, they were extended or replicated. This was particularly true in the case of SIBs, where SIBs were used to fill a gap in funding for a particular group of participants, or when they were used as instruments to test whether an innovative intervention was effective and could be scaled/replicated. Interviewees noted some **additional benefits** of SOC programmes, including the development of a measurement infrastructure, providing evidence for policy making, sharing knowledge among stakeholders, unlocking financial resources, and enhancing flexibility for service providers.

Although SOC appears to focus more frequently on securing employment for specific categories of disadvantaged persons, such as those affected by mental health problems, migrants, and former prisoners, no specific target group proved to be more positively impacted than others by either SIB or PbR models. **The effectiveness of SOC related more to the context and the design of an intervention than to specific types of beneficiaries or areas of social services.**

In line with the existing literature, **we found that PbR contracts encouraged episodes of so-called ‘creaming’¹, and as well as the ‘parking’² of harder-to-help participants.** The evidence collected shows that this problem stemmed from two causes. First, the PbR payment mechanism pushed providers to prioritise easier-to-help participants, particularly when programme funding was limited and targets were difficult to attain, because otherwise the programme would not have been financially viable for providers. Second, where programme participation was mandatory, some participants were unwilling to participate or did not think that the programme’s goals (e.g. employment) were within their reach, due to the multiple barriers they faced in achieving the target outcomes (e.g. childcare responsibilities, disability, lack of transportation, etc.). In such cases, service providers found it appropriate to ‘park’ such participants. The study identified fewer cases of parking in SIBs, although cases of creaming occurred at the point at which participants were recruited in SIBs as well.

In four of the cases analysed, contracts were formally re-negotiated or contract managers informally applied different performance criteria when providers failed to meet target expectations due to changing macroeconomic conditions. The added pressure for providers to reach unattainable targets also encouraged perverse incentives and added to the complexity of contract management.

Although some of the analysed PbR schemes were not very effective with regard to the groups that were hardest to help, they were quite effective for others. For example, the Work Programme performed worse for people claiming unemployment benefits due to

¹ In this study, ‘creaming’ refers to the process of selecting those participants for the programme that are most likely to achieve its outcomes, in order for service providers or investors to receive payments.

² ‘Parking’ refers to a process by which providers try to keep costs down by doing little to serve those with the poorest anticipated outcomes, while instead focusing resources on more able clients with (for instance) better employment prospects.

disability, but it exceeded its targets for most other jobseekers. A similar trend was observed by the commissioner for JobPath. This suggests that PbR programmes are perhaps best utilised to address target groups that are relatively easy to help, but not so easy that the outcomes could be achieved without the intervention (e.g. the long-term unemployed). **Nevertheless, when we compared the performance of PbR programmes for these groups with the performance of similar TF programmes, we found mixed results.** Furthermore, in a number of cases, direct comparisons of the programmes' effectiveness could not be made because outcomes had not been consistently tracked in the TF programmes.

Efficiency

We compared the cost per outcome in each matched pair of SOC and TF programmes to assess the relative efficiency of SOC and TF programmes. Cost per outcome represents, for example, the cost per participant who found a job; who remained in employment for a certain duration of time; who completed the entire rehabilitation programme, etc., depending on the goals of the programme concerned. Sufficient information was available to make comparisons between intervention costs in seven out of the 15 SOC-TF pairs. Importantly, our comparisons were limited in that we could not control for intervening factors such as participant characteristics, so **we cannot claim that SOC programmes were more or less efficient specifically due to the funding mechanism alone.**

Among those cases in which comparisons could be made, **SIBs tended to cost more per outcome achieved than similar traditionally funded programmes. The commissioners nevertheless saw these additional costs as being justified:** rather than being used to deliver the programme more efficiently, SIBs were used to test interventions in different contexts or on a different scale from that at which the same programme had previously been implemented. Furthermore, cases were also encountered in which commissioners issued no payments to the investors when the SIBs failed to reach their target outcomes, meaning that the programme cost nothing to the taxpayer (discounting the resources invested in setting the programme up and overseeing it). **SIBs were thus useful in that when a programme failed to achieve its outcome targets, it could be easily shut down.** This contrasts with some TF programmes, which continue to run despite a lack of evidence that they are effective.

With regard to PbR schemes, a commonly cited argument for their use is that they are more efficient than traditionally funded programmes because their focus on outcomes drives service providers to innovate and eliminate inefficient practices. Indeed, the two PbR programmes in our study whose costs per outcome could be compared against those of TF programmes showed either a similar or lower cost than equivalent TF programmes. However, it is impossible to draw comparisons with regard to the other selected PbR programmes, due **insufficient information about both the costs of PbR programmes and the outcomes of the TF programmes.** Furthermore, given concerns about the perverse incentives that may arise in PbR contracts, it is important to investigate whether the quality of services is maintained at a lower cost.

Lastly, **both SIBs and PbR programmes generally entailed greater operational (set-up and management) costs than TF programmes,** for example, **sometimes taking between a year and two years to set up.** Available information is insufficient to assess whether these additional costs offset the savings that PbR programmes potentially achieve by promoting the efficiency of interventions. **There is also little transparency as to the costs involved in the design and oversight of SIBs.**

Social Outcome Contract Design

The cases analysed include various sub-types of SIB and PbR models, with each model having different implications in terms of the level of risk transferred from commissioner to investors or service providers. By 'risk', we mean the financial consequences in the event that the intervention is not successful in achieving its target outcomes. In both types of SOC, we found that **no model design could reduce the risks for all parties involved: if the risk was reduced for the commissioner, it increased for investors or service providers**. This contrasts with the existing rhetoric that outcomes-based contracts – SIBs in particular – are win-win situations in which all parties stand to benefit.

The amount of risk transferred from commissioners to investors in SIBs varied, depending on a number of factors. Ambitious **programme targets** increased the risk for investors to lose their investment, yet lowered the risk for commissioners that they would have to fund ineffective interventions. Similarly, a greater **number of outcomes** assessed was associated with a lower risk for investors that none of the outcomes would be achieved. The risk for investors was further lowered if philanthropies or commissioners **guaranteed part of the investment**. The more **frequent the payments** were, the lower the risk was for investors to lose all of their investment. On the other hand, more frequent payments implied the risk for commissioners that they would partly re-fund the intervention whose results could not be sustained. Finally, commissioners carried substantial risk when they were responsible for **recruiting** a contractually-specified number of participants, irrespective of other payment arrangements with investors.

Among the PbR programmes studied, the greater the proportion of a contract's value that was based on outcomes, the more difficult it was for small third-sector organisations and social economy enterprises to bid for the contract as prime providers (as opposed to subcontractors). Third-sector organisations lacked sufficient funds to shoulder the upfront costs of the programme, and were unable to bear the financial risks if outcomes were not achieved. It was generally easier for small third sector organisations and social economy enterprises to participate in SIBs, because such schemes posed less financial risk to service providers if the interventions failed. Nevertheless, both SIBs and PbR programmes were generally positively perceived by commissioners due to the benefits they generated, even taking into account the additional costs.

Outcome measurement

The evaluation of outcomes is one of the key characteristics of SOC schemes, which ultimately determines payments to investors or service providers. Among the cases analysed, **the way in which outcomes were defined had a significant influence in terms of encouraging or discouraging perverse incentives such as creaming and parking**. Outcomes were often defined using a **binary or a frequency approach**. When binary outcomes were chosen, providers had to achieve an absolute target, and no payment was granted for achieving lesser results. An example of a binary outcome would be whether or not a person found a job. With frequency-based schemes, rewards were staggered according to the agreed frequency of results, with payments increasing as results increase. For example, providers or investors would receive additional payments for every three months that programme participants remained employed. Of these two models, **binary outcomes were more often associated with creaming and parking** because the target outcome (e.g. employment) was often out of reach for some programme participants, motivating providers to focus on other clients.

Furthermore, **investors argued that frequency-based measures better reflected the true success or failure of a programme.** Some investors questioned whether a programme can truly be considered to have failed if the target is achieved in most instances – for example, for 19 out of 20 target participants (a situation that would be deemed a failure in the case of a binary model). The downsides of frequency measures were that outcomes were more difficult to track, were less intuitive to communicate to the public, and that they limited the risk that could be transferred from commissioners to investors or service providers. To accommodate the differing interests of the stakeholders involved, a third of the cases analysed chose to evaluate outcomes in terms of both frequency and binary measures.

Importantly, **we encountered instances in which involving the service provider in the choice of outcomes helped to prevent mission drift and ensured that medium-term impacts were considered in addition to short-term impacts** (see Annex 1 regarding the co-design phase in the Drug and Alcohol Recovery Pilots as well as DUO for a JOB SIB). Involving service providers also allowed their expertise to be drawn upon in determining what outcomes were best able to ensure the well-being of the beneficiaries.

With regard to evaluation methods, in a number of the programmes studied, service providers were responsible for gathering the evidence to prove that targeted outcomes were achieved. This process was often challenging because providers lacked the expertise to gather the necessary information. The information that was supposed to be used as evidence (e.g. employment contracts; education records; proof of participant eligibility) was not available, meaning that providers could not claim payments for some of the outcomes they achieved. Furthermore, the process was more time-consuming than providers anticipated, so they were able to dedicate less time than expected to working with beneficiaries. Finally, conflicts of interest were apparent with regard to the data reported by service providers in PbR programmes, because these data determined the payments to the providers.

Although randomised controlled trials (RCTs) are considered the gold standard in positivist evaluations, **none of the 15 analysed cases employed an RCT.** One programme—the ABLE SIB—attempted an RCT of a behavioural therapy programme aimed at jailed youth offenders, but eventually decided against this approach because clear boundaries could not be maintained between treatment and control groups when participants had to change housing units. Service providers from other programmes also felt that it would be unethical to deny services to some potential participants for the sake of forming a control group, particularly in cases where no other providers or state institutions could provide the same services outside the SOC programme. Lastly, RCTs are also expensive to run.

As a result, the most rigorous methods encountered among the cases analysed were quasi-experimental, yet these were only used in seven cases. Out of these cases, only two schemes used the quasi-experimental design to evaluate the outcomes related to the payment mechanisms. This means that rigorous evaluation methods, which causally link the intervention to its effects, triggered the payment only in two schemes. **In most cases, the absence of a control group meant that the outcomes achieved in SOC programmes could not be definitively attributed to the intervention. Thus, even though the SOC schemes had been implemented for some time now, many still lacked robust evaluations that would allow for causally determining their impact, as well as effectiveness, efficiency and added value.**

Recommendations

Recommendation 1: SIBs should be commissioned in areas in which there is a gap in funding, and for the purpose of testing whether an innovative intervention is effective and/or scalable.

On the basis of the findings outlined above, we recommend that SIBs should be commissioned in areas in which there is a gap in funding, and for the purpose of testing whether an innovative intervention is effective and/or scalable. Such SIBs were perceived as most successful by the stakeholders interviewed, even when the SIBs failed to achieve their target outcomes. However, it is difficult to justify the use of SIBs when they fall outside these criteria, due to their additional management costs as well as interest payments to investors. Feasibility studies should be conducted prior to launching SIBs to ensure that that the funding instrument brings added value in return for the additional costs involved.

Recommendation 2: The commissioning of PbRs should be avoided when services are being contracted for groups that face multiple barriers to achieving the target outcomes, particularly where participation in the programme is mandatory.

Given the perverse incentives identified in the PbR programmes studied, we suggest that PbR contracts should not be utilised when contracting services for groups that face multiple barriers to achieving the target outcomes, particularly when participation in the programme is mandatory. Rather, we recommend that services for groups with complex needs should be funded in a traditional way. It is important to note that SIBs, as opposed to PbR programmes, could be an effective way of testing interventions targeting beneficiaries with multiple, complex needs. Nevertheless, SIBs entail additional costs, so programmes should not continue to be funded through SIBs once there is sufficient evidence that they are effective in a particular context.

Recommendation 3: Assess the effectiveness and efficiency of PbR programmes for easier-to-help groups by tracking outcomes in TF programmes and using control groups.

Given the mixed results observed on the effectiveness of PbR and TF models, coupled with a lack of information about their respective efficiency, we argue that more data is needed on the outcomes of TF programmes. We suggest two approaches to tracking outcomes in TF programmes. First, outcomes of TF programmes could be assessed using or linking data via existing administrative databases (e.g. datasets from tax authorities, social security agencies, child support agencies and others). Second, the traditionally funded model could be used as the control group for the PbR model. For example, in the latest Work and Health Programme in the UK, the Department for Work and Pensions contracted out the delivery of the programmes to private providers on a PbR basis, but also continued to deliver the programme in-house through local employment offices, called Jobcentre Plus. In addition to providing key data to assess the impact of SOC compared to TF, this set-up helps to assess value for money and set performance targets for the providers.

Recommendation 4: Track and report operational costs to ensure that public procurement is transparent.

Assessing the efficiency of both PbR and SIB models is particularly challenging due to the limited information available on the costs incurred when designing these contracts, setting up programmes, monitoring performance and evaluating results. These costs should be included in the total reported programme cost, even though this is not currently standard

practice. Such costs should be publicly available to ensure that public procurement is transparent.

Recommendation 5: To ensure that small third sector organisations can participate in outcomes-based contracts, use the PbR mechanism as a premium for results accomplished beyond the expectation, base a small portion (e.g. up to 20%) of the PbR contract on outcomes, or implement SIBs.

Among the PbR cases analysed, third-sector organisations lacked sufficient funds to shoulder the upfront costs of the programme, and could not bear the financial risks if outcomes were not achieved. This was particularly true in the case of programmes where a large portion of the payments was based on outcomes. Therefore, we recommend using the PbR mechanism as a premium for results achieved beyond the expectations of the targets set by programme commissioners. An alternative would be to base a small portion (e.g. up to 20%) of the PbR contract on outcomes. This would mean that most of the payments in the contract would be based on outputs. Another alternative would be to use SIBs instead. For both SIBs and PbR models, setting the length of the contract to at least three years (or ideally longer) would also give more stability to service providers in terms of hiring and project management.

Recommendation 6: Contracts should provide some flexibility with regard to performance targets by tying performance expectations to the performance of the control group, setting different targets reflecting different macroeconomic environments, setting payment caps and floors, and/or including early discontinuation clauses.

To avoid the risk of having to renegotiate agreements, SOC contracts should build in flexibility with regard to performance targets, so that these targets can alter depending on changing macroeconomic conditions, which in turn affect the outcomes achieved. Performance expectations could be tied to the performance of a control group, which would receive services at the same time as participants in the SOC programme. Alternatively, different performance targets could be specified prior to the start of the programme, according to whether or not the economy experiences a recession during the programme's delivery. Payment caps would safeguard the commissioner against having to pay out excessive profits in the event that the programme performs better than expected. On the flip side, payment floors could be implemented to ensure that SOC's are financially viable for providers even when participant volumes are low. Finally, early discontinuation clauses would limit the risk to both commissioners and investors in cases where the programme's early performance fails to meet minimum thresholds.

Recommendation 7: Combine both binary³ and frequency outcome measures, and consider the distance travelled by participants as well as customer satisfaction to prevent perverse incentives.

³ With binary outcomes, providers have to achieve an absolute target, and no payment is granted for achieving lesser results. An example of a binary outcome would be whether or not a person found a job. In frequency schemes, rewards are staggered according to the agreed frequency of results, with payments increasing as results increase. For example, providers or investors would receive additional payments for every three months that programme participants remain employed.

To discourage perverse incentives and accommodate the interests of different stakeholders, we recommend combining frequency outcome measures with binary when designing SOC schemes. However, even when frequency outcome measures are combined with binary ones, some programme participants may still be too far from achieving even the lowest outcome threshold, which often results in creaming and parking. For example, if a programme aims to place participants into employment and sustain that employment for a number of months, some participants will not be able to find a job. Therefore, we further suggest including 'distance travelled' outcomes into PbR contracts outcomes such as agreement on an action plan, CV preparation, interview attendance and others, which would make it more financially viable for providers to work with less job-ready clients. Another way to limit creaming and parking is to monitor performance among different groups of individuals (i.e. those closer to and further from achieving the target outcomes) and surveying participant satisfaction while the programme is ongoing.

Recommendation 8: Involve service providers in defining outcome measures and evaluation design.

Given that the involvement of service providers helped to ensure that the most relevant outcomes were chosen in a number of the SOC programmes analysed, we recommend that service providers should be included in the process of defining outcomes. Furthermore, given the burden that evaluations impose on service providers, they should be also be consulted regarding evaluation design. For example, during the initial phases of the programme, a test-run of data collection techniques could be implemented that would clarify what information will be used to gather evidence for payment-related outcomes. If necessary, service providers should be trained in how such data should be handled.

Résumé analytique

Fondement

Les contrats basés sur les résultats sociaux (ou SOC, de l'anglais *Social outcome contracts*) sont un type de mécanisme relativement nouveau dans les marchés publics de services. Ils se concentrent sur l'exploitation des ressources des secteurs public, privé, philanthropique et de la société civile, dans le but de mettre en œuvre conjointement des interventions efficaces dans le domaine public. Cet outil pourrait permettre de moderniser les services publics et les régimes sociaux européens en encourageant une culture de la mesure des performances, en soutenant les partenariats intersectoriels et en permettant de piloter et d'étendre de nouveaux modèles d'intervention, ce qui favoriserait l'apprentissage et l'innovation sociale. Néanmoins, il existe peu de recherches empiriques comparant les modèles SOC aux mécanismes de financement traditionnels (FT) tels que les subventions, les bourses, les contrats à l'acte, les contrats de bloc et la livraison en interne. L'étude vise donc à évaluer l'efficacité et l'efficience des contrats basés sur les résultats par rapport aux interventions réalisées dans le cadre du financement traditionnel, ainsi qu'à évaluer les méthodes de mesure des résultats appliquées pour déterminer leur impact. L'étude se concentre exclusivement sur les services sociaux plutôt que sur les services publics en général.

L'étude porte sur deux types de SOC : les contrats à impact social (ou CIS) et les autres systèmes de paiement au résultat (ou PbR, de l'anglais *Payment by Result*). Nous définissons les CIS comme des mécanismes de financement dans lesquels des investisseurs fournissent un financement initial pour un programme. Les investisseurs assument le risque de perdre leur investissement si le programme n'atteint pas les résultats escomptés, mais sont remboursés intégralement, avec un paiement d'intérêts supplémentaire de la part du commissaire, en cas de succès du CIS. Certains CIS incluent également un intermédiaire qui aide à trouver les prestataires de services, les investisseurs et/ou les évaluateurs. En revanche, les programmes PbR n'impliquent pas d'investisseur externe ou d'intermédiaire. Dans ce cas, les prestataires de services ne reçoivent des paiements des commissaires que si les résultats visés sont atteints. Les modèles PbR varient en fonction de la part de paiement aux prestataires de services qui est basée sur les résultats. Dans les modèles PbR purs, 100 % du paiement est basé sur les résultats. Ce n'est pas le cas avec les modèles PbR mixtes, pour lesquels une partie du paiement est financée de manière traditionnelle.

Méthodologie

Quinze cas de SOC ont été sélectionnés et analysés en détail dans le cadre de l'étude. Pour ce faire, divers modèles SOC, zones géographiques et services sociaux ont été choisis. Pour être inclus dans la sélection, les programmes devaient être achevés (ou presque) et des évaluations devaient être disponibles afin de permettre la collecte de données. De ce fait, un certain nombre des CIS sélectionnés sont les premiers à avoir été mis en œuvre dans leurs pays respectifs. Pour chaque modèle SOC, des programmes FT comparables ont été identifiés. Les comparaisons entre SOC et FT ont visé à évaluer dans quelle mesure les deux programmes ont atteint les objectifs fixés, et à quel coût. Les comparaisons ont également visé à déterminer la valeur ajoutée des programmes SOC en termes d'avantages et d'inconvénients par rapport au FT, ainsi qu'à identifier les éléments de conception permettant la reproductibilité, l'extensibilité et la participation des organisations du secteur tertiaire et des entreprises sociales. Les données ont été recueillies au moyen d'une recherche documentaire, reposant en grande partie sur les rapports d'évaluation et les enquêtes d'audit des programmes, ainsi que sur 54 entretiens

avec des commissaires, des prestataires de services, des évaluateurs, des intermédiaires, des experts et des investisseurs.

Conclusions

Efficacité et valeur ajoutée

Afin de comparer l'efficacité des programmes SOC et FT, nous avons évalué dans quelle mesure chaque programme a pu atteindre les résultats fixés au départ. Ces résultats incluent le nombre de participants ayant trouvé un emploi, la réduction des taux de récidive, le nombre d'enfants ayant évité la prise en charge extrafamiliale, etc., en fonction des objectifs du programme et des domaines de service. **Les preuves ne sont pas suffisantes pour affirmer que les contrats basés sur les résultats sont plus efficaces que le financement traditionnel ou inversement.** Cela s'explique en partie par le fait que, dans la majorité des programmes FT comparables, aucun objectif de résultat n'a été fixé et aucune évaluation rigoureuse n'a été menée.

Toutefois, notre analyse a également révélé que la réalisation des objectifs ne définissait pas nécessairement l'efficacité et le succès global d'un programme. De fait, **certains programmes n'ayant pas atteint leurs objectifs ont néanmoins été considérés comme des succès, car ils ont pu répondre aux besoins et aux attentes des différents acteurs concernés.** Ainsi, ils ont été étendus ou reproduits. Ceci est particulièrement vrai dans le cas des CIS, lorsqu'ils ont été utilisés pour combler un manque de financement pour un groupe particulier de participants, ou lorsqu'ils ont été utilisés comme instruments pour tester si une intervention innovante était efficace et pouvait être étendue/reproduite. Les personnes interrogées ont noté certains **avantages supplémentaires** des programmes SOC, notamment le développement d'une infrastructure de mesure, la fourniture de données probantes pour l'élaboration des politiques, le partage des connaissances entre les parties prenantes, le déblocage de ressources financières et l'amélioration de la flexibilité pour les fournisseurs de services.

Bien que les SOC semblent se concentrer plus fréquemment sur l'obtention d'un emploi pour des catégories spécifiques de personnes défavorisées, telles que les personnes souffrant de problèmes de santé mentale, les migrants et les anciens détenus, aucun groupe cible spécifique ne s'est avéré plus favorablement touché que les autres par les modèles CIS ou PbR. **L'efficacité des SOC est davantage liée au contexte et à la conception d'une intervention qu'à des types de bénéficiaires ou des domaines spécifiques de services sociaux.**

Conformément à la littérature existante, **nous avons constaté que les contrats PbR encouragent les pratiques d'« écrémage »⁴ et de « mise à l'écart »⁵ des participants les plus difficiles à aider.** Les preuves recueillies montrent que ce problème a deux causes. Premièrement, le mécanisme de paiement PbR pousse les prestataires à donner la priorité aux participants les plus faciles à aider, en particulier lorsque le financement du programme est limité et que les objectifs sont difficiles à atteindre, car sinon le programme ne serait pas financièrement viable pour les prestataires. Deuxièmement, lorsque la participation au programme est obligatoire, certains participants ne sont pas disposés à y

⁴ Dans cette étude, le terme « écrémage » fait référence au processus de sélection des participants qui ont le plus de chances d'atteindre les résultats visés par le programme, afin que les prestataires de services ou les investisseurs puissent recevoir leurs paiements.

⁵ Le terme « mise à l'écart » désigne un processus par lequel les prestataires tentent de réduire les coûts en ne faisant pas grand-chose pour servir les personnes dont les résultats escomptés sont les plus faibles, tout en concentrant les ressources sur les participants plus aptes et offrant (par exemple) de meilleures perspectives d'emploi.

participer ou ne pensent pas que les objectifs du programme (par exemple l'emploi) sont à leur portée, en raison des multiples obstacles auxquels ils sont confrontés pour atteindre les résultats visés (par exemple les responsabilités en matière de garde d'enfants, le handicap, le manque de transport, etc.). Dans de tels cas, les prestataires de services ont jugé approprié de « mettre à l'écart » ces participants. L'étude a identifié moins de cas de mise à l'écart dans les CIS, bien que des cas d'écroulement se soient produits lors du recrutement des participants dans les CIS également.

Dans quatre des cas analysés, les contrats ont été officiellement renégociés ou les gestionnaires de contrats ont appliqué de manière informelle des critères de performance différents lorsque les fournisseurs n'ont pas atteint les objectifs fixés en raison de l'évolution des conditions macroéconomiques. La pression supplémentaire exercée sur les fournisseurs pour qu'ils atteignent des objectifs irréalisables a également encouragé des incitations perverses et ajouté à la complexité de la gestion des contrats.

Bien que certains des programmes PbR analysés n'aient pas été très efficaces pour les groupes les plus difficiles à aider, ils se sont avérés assez efficaces pour d'autres. Par exemple, le *Work Programme* (programme Travail) n'a pas donné de bons résultats pour les personnes en demande d'allocations chômage pour cause d'invalidité, mais il a dépassé ses objectifs pour la plupart des autres demandeurs d'emploi. Une tendance similaire a été observée par le commissaire en charge du programme *JobPath*. Ceci suggère que les programmes PbR seraient mieux adaptés pour les groupes cibles relativement faciles à aider, mais nécessitant toutefois une intervention pour atteindre ces résultats (par exemple, les chômeurs de longue durée). **Néanmoins, lorsque nous avons comparé les performances des programmes PbR pour ces groupes avec les performances de programmes FT similaires, nous avons constaté des résultats mitigés.** En outre, dans un certain nombre de cas, il n'a pas été possible de comparer directement l'efficacité des programmes, car les résultats n'ont pas été suivis de manière cohérente dans les programmes FT.

Efficiences

Nous avons comparé le coût par résultat dans chaque paire de programmes SOC et FT afin d'évaluer l'efficacité relative des programmes SOC et FT. Le coût par résultat représente, par exemple, le coût par participant qui a trouvé un emploi, qui est resté en activité pendant une certaine durée, qui a achevé l'ensemble du programme de réhabilitation, etc., selon les objectifs du programme en question. Les informations disponibles ont été suffisantes pour permettre de comparer les coûts d'intervention dans sept des quinze paires SOC-FT. Il est important de noter que nos comparaisons ont été limitées dans la mesure où nous n'avons pas pu contrôler certains facteurs d'intervention tels que les caractéristiques des participants. **Nous ne pouvons donc pas affirmer que les programmes SOC ont été plus ou moins efficaces sur la base du seul mécanisme de financement.**

Parmi les cas où des comparaisons ont pu être faites, **les CIS ont eu tendance à coûter plus cher par résultat obtenu que les programmes similaires financés de manière traditionnelle. Les commissaires ont néanmoins estimé que ces coûts supplémentaires étaient justifiés** : plutôt que d'être utilisés pour exécuter un programme de manière plus efficace, les CIS ont servi à tester des interventions dans des contextes différents ou à une échelle différente. En outre, il existe également des cas où les commissaires n'ont effectué aucun paiement aux investisseurs lorsque les CIS n'ont pas atteint les résultats escomptés, ce qui signifie que le programme n'a rien coûté au contribuable (en excluant les ressources investies dans la mise en place et la supervision du programme). **Les CIS ont donc été utiles dans le sens où lorsqu'un programme n'a pas atteint ses objectifs de résultats, il a pu être facilement arrêté.** Cette situation contraste avec certains programmes FT, qui continuent à fonctionner malgré le manque de preuves de leur efficacité.

Concernant les programmes PbR, un argument courant en faveur de leur utilisation est qu'ils sont plus efficaces que les programmes FT parce que l'accent mis sur les résultats incite les prestataires de services à innover et à éliminer les pratiques non efficaces. En effet, les deux programmes PbR de notre étude dont les coûts par résultat ont pu être comparés à ceux des programmes FT ont montré un coût similaire ou inférieur à celui des programmes FT équivalents. Toutefois, il est impossible d'établir des comparaisons s'agissant des autres programmes PbR sélectionnés, en raison du **manque d'informations sur les coûts des programmes PbR et les résultats des programmes FT**. De plus, étant donné les préoccupations relatives aux incitations perverses pouvant apparaître dans le cadre des contrats PbR, il est important d'examiner si la qualité des services est maintenue à un coût moindre.

Enfin, **les programmes CIS et PbR ont généralement entraîné des coûts opérationnels (mise en place et gestion) plus importants que les programmes FT**, par exemple, **leur mise en place prenant parfois entre un et deux ans**. Les informations disponibles sont insuffisantes pour évaluer si ces coûts supplémentaires sont compensés par les économies que les programmes PbR peuvent permettre de réaliser en favorisant l'efficacité des interventions. **La transparence quant aux coûts liés à la conception et à la surveillance des CIS est également très faible**.

Conception des contrats basés sur les résultats sociaux

Les cas analysés incluaient divers sous-types de modèles CIS et PbR, chaque modèle présentant des implications différentes en termes de niveau de risque transféré du commissaire aux investisseurs ou aux prestataires de services. Par « risque », nous entendons les conséquences financières dans le cas où l'intervention ne parviendrait pas à atteindre les résultats escomptés. Pour les deux types de SOC, nous avons constaté qu'il **n'existait pas de conception de modèle permettant de réduire les risques pour toutes les parties concernées : si le risque était réduit pour le commissaire, il augmentait pour les investisseurs ou les prestataires de services**. Ceci contraste avec la rhétorique actuelle selon laquelle les contrats basés sur les résultats – les CIS en particulier – sont des situations gagnantes pour toutes les parties.

Le niveau de risque transféré des commissaires aux investisseurs dans le cas des CIS variait en fonction d'un certain nombre de facteurs. Des **objectifs de programme** ambitieux augmentaient le risque que les investisseurs perdent leur investissement, tout en réduisant le risque pour les commissaires de devoir financer des interventions inefficaces. De même, l'évaluation d'un plus grand **nombre de résultats** était associée à un risque moindre pour les investisseurs qu'aucun résultat ne soit atteint. Le risque pour les investisseurs était encore réduit si des philanthropes ou des commissaires **garantissaient une partie de l'investissement**. Plus la **fréquence des paiements** était élevée, moins les investisseurs risquaient de perdre la totalité de leur investissement. D'un autre côté, des paiements plus fréquents impliquaient le risque pour les commissaires d'avoir à refinancer en partie une intervention dont les résultats n'étaient pas durables. Enfin, les commissaires couraient un risque important lorsqu'ils étaient chargés de **recruter** un nombre de participants défini par contrat, indépendamment des autres modalités de paiement établies avec les investisseurs.

Parmi les programmes PbR étudiés, plus la part du contrat basée sur les résultats était importante, plus il a été difficile pour les petites organisations du secteur tertiaire et les entreprises sociales de prendre part au contrat en tant que fournisseurs principaux (par opposition aux sous-traitants). Les organisations du secteur tertiaire ne disposaient pas de fonds suffisants pour assumer les coûts initiaux du programme et

n'étaient pas en mesure de supporter les risques financiers si les résultats n'étaient pas atteints. Il a été généralement plus facile pour les petites organisations du secteur tertiaire et les entreprises sociales de participer aux CIS, car ces programmes présentent moins de risques financiers pour les prestataires de services si les interventions échouent. Néanmoins, tant les programmes CIS que PbR ont été généralement perçus positivement par les commissaires en raison des avantages qu'ils génèrent, même en tenant compte des coûts supplémentaires.

Mesure des résultats

L'évaluation des résultats est l'une des principales caractéristiques des programmes SOC, qui détermine en fin de compte les paiements aux investisseurs ou aux prestataires de services. Parmi les cas analysés, **la manière dont les résultats ont été définis a une influence significative en termes d'encouragement ou de découragement des incitations perverses telles que l'écramage et la mise à l'écart**. La plupart du temps, les résultats ont été définis par une **approche binaire ou fréquentielle**. Dans le cas d'une approche binaire, les fournisseurs doivent atteindre un objectif absolu, et aucun paiement n'est accordé pour des résultats inférieurs. Par exemple, un résultat binaire consiste à savoir si une personne a trouvé un emploi ou non. Dans le cas d'une approche fréquentielle, les récompenses sont échelonnées en fonction de la fréquence de résultats convenue, les paiements augmentent donc à mesure que les résultats augmentent. Par exemple, les prestataires ou les investisseurs reçoivent des paiements supplémentaires pour chaque trimestre où les participants au programme restent en activité. Parmi ces deux modèles, **les résultats binaires ont plus souvent été associés à des pratiques d'écramage et de mise à l'écart** parce que le résultat visé (par exemple, l'emploi) était souvent hors de portée pour certains participants au programme, ce qui a motivé les prestataires à se concentrer sur d'autres participants.

En outre, **les investisseurs ont fait valoir que les mesures fréquentielles reflètent de manière plus juste le succès ou l'échec d'un programme**. Certains investisseurs se sont demandé si un programme peut vraiment être considéré comme ayant échoué si l'objectif est atteint dans la plupart des cas – par exemple, pour 19 des 20 participants cibles (une situation considérée comme un échec dans le cas d'un modèle binaire). Les inconvénients des mesures fréquentielles sont que les résultats sont plus difficiles à suivre, qu'ils sont moins intuitifs à communiquer au public et qu'ils limitent le risque pouvant être transféré des commissaires aux investisseurs ou aux prestataires de services. Afin de tenir compte des intérêts divergents des parties prenantes, un tiers des cas analysés a choisi d'évaluer les résultats à la fois par des mesures binaires et fréquentielles.

Il est important de noter que **nous avons rencontré des cas où l'implication du prestataire de services dans le choix des résultats a permis d'éviter la dérive de la mission et de garantir la prise en compte des impacts à moyen terme en plus des impacts à court terme** (voir l'annexe 1 concernant la phase de co-conception des *Drug and Alcohol Recovery Pilots* – projets pilotes pour la guérison des addictions à l'alcool et aux drogues –, ainsi que le programme CIS *DUO for a JOB*). L'implication des prestataires de services a également permis de faire appel à leur expertise afin de déterminer les résultats les plus susceptibles d'assurer le bien-être des bénéficiaires.

Concernant les méthodes d'évaluation, dans un certain nombre de programmes étudiés, les prestataires de services ont été chargés de rassembler les éléments attestant de la réalisation des objectifs visés. Ce processus s'est souvent avéré difficile, car les prestataires ne disposaient pas de l'expertise nécessaire pour rassembler les informations requises. Les informations devant servir de preuve (par exemple, les contrats de travail, les dossiers scolaires, la preuve de l'éligibilité des participants) n'étaient pas disponibles, ce qui signifie que les prestataires n'ont pas pu demander de paiements pour certains résultats obtenus. En outre, le processus a pris plus de temps que ce qui était attendu, de sorte que les prestataires ont pu consacrer moins de temps que prévu au travail avec les

bénéficiaires. Enfin, des conflits d'intérêts sont apparus concernant les données communiquées par les prestataires de services dans le cadre des programmes PbR, puisque ce sont ces données qui déterminaient les paiements aux prestataires.

Bien que les essais contrôlés randomisés (ECR) soient considérés comme l'étalon-or des évaluations positivistes, **aucun des 15 cas analysés n'a fait appel à un ECR**. Un seul programme – le CIS ABLE – a tenté un ECR d'un projet de thérapie comportementale destiné aux jeunes délinquants emprisonnés, mais a finalement décidé de ne pas suivre cette approche parce qu'il était impossible de maintenir des frontières claires entre le groupe de traitement et le groupe témoin lorsque les participants devaient changer d'unité de logement. Les prestataires de services d'autres programmes ont également estimé qu'il serait contraire à l'éthique de refuser des services à certains participants potentiels dans le but de constituer un groupe de contrôle, en particulier dans les cas où aucun autre prestataire ou institution publique ne pourrait fournir les mêmes services en dehors du programme SOC. Enfin, les ECR sont également coûteux à gérer.

En conséquence, les méthodes les plus rigoureuses rencontrées parmi les cas analysés étaient quasi-expérimentales, et elles n'ont été utilisées que dans sept cas. Parmi ces cas, seuls deux programmes ont utilisé le modèle quasi-expérimental pour évaluer les résultats en lien avec les mécanismes de paiement. Cela signifie que des méthodes d'évaluation rigoureuses, qui établissent un lien de causalité entre l'intervention et ses effets, n'ont déclenché le paiement que dans deux programmes. **Dans la plupart des cas, l'absence de groupe de contrôle a signifié que les résultats obtenus dans le cadre des programmes SOC ne pouvaient être attribués de manière définitive à l'intervention. Ainsi, bien que les programmes SOC aient été mis en œuvre depuis un certain temps déjà, nombre d'entre eux ne faisaient toujours pas l'objet d'évaluations solides permettant de déterminer les causes de leur impact, ainsi que leur efficacité, leur efficience et leur valeur ajoutée.**

Recommandations

Recommandation n° 1 : Mettre en place les CIS dans les domaines où il existe un déficit de financement et dans le but de tester si une intervention innovante est efficace et/ou évolutive.

Sur la base des conclusions exposées ci-dessus, nous recommandons que des CIS soient mis en place dans les domaines où il existe un déficit de financement, et dans le but de tester si une intervention innovante est efficace et/ou évolutive. Les parties prenantes interrogées estiment que ces CIS sont les plus performants, même lorsqu'ils n'atteignent pas les résultats escomptés. Toutefois, il est difficile de justifier l'utilisation des CIS lorsqu'ils ne répondent pas à ces critères, en raison de leurs coûts de gestion supplémentaires ainsi que des intérêts versés aux investisseurs. Des études de faisabilité doivent être menées avant le lancement des CIS afin de s'assurer que l'instrument de financement apporte une valeur ajoutée en contrepartie des coûts supplémentaires qu'il implique.

Recommandation n° 2 : Éviter de mettre en place des PbR lorsque les services sont contractés pour des groupes qui se heurtent à de multiples obstacles pour atteindre les résultats visés, en particulier lorsque la participation au programme est obligatoire.

Compte tenu des incitations perverses identifiées dans les programmes PbR étudiés, nous suggérons de ne pas utiliser les contrats PbR lorsque des services sont contractés pour des groupes qui se heurtent à de multiples obstacles pour atteindre les résultats visés, en particulier lorsque la participation au programme est obligatoire. Nous recommandons plutôt un financement traditionnel des services destinés aux groupes présentant des besoins complexes. Il est important de noter que les CIS, par opposition aux programmes PbR, peuvent être un moyen efficace de tester les interventions qui ciblent des bénéficiaires présentant des besoins multiples et complexes. Néanmoins, les CIS entraînent des coûts supplémentaires, de sorte que les programmes ne devraient pas continuer à être financés par les CIS une fois que les preuves de leur efficacité dans un contexte particulier sont suffisantes.

Recommandation n° 3 : Évaluer l'efficacité et l'efficience des programmes PbR pour les groupes les plus faciles à aider en réalisant un suivi des résultats des programmes FT et en utilisant des groupes de contrôle.

Étant donné les résultats mitigés observés sur l'efficacité des modèles PbR et FT, associés à un manque d'informations sur leur efficience respective, nous soutenons que davantage de données sont nécessaires sur les résultats des programmes FT. Nous suggérons deux approches pour suivre les résultats des programmes FT : premièrement, évaluer les résultats des programmes FT en utilisant ou en reliant les données via les bases de données administratives existantes (par exemple, les données des autorités fiscales, des organismes de sécurité sociale, des organismes d'aide à l'enfance et autres) ; deuxièmement, utiliser le modèle FT comme groupe de contrôle pour le modèle PbR. Par exemple, dans le cadre du dernier *Work and Health Programme* (programme Travail et santé) au Royaume-Uni, le Département du Travail et des Retraites a sous-traité la mise en œuvre des programmes à des prestataires privés sur une base PbR, mais a également continué à mettre en œuvre le programme en interne par l'intermédiaire des agences locales pour l'emploi, les Jobcentre Plus. En plus de fournir des données clés pour évaluer l'impact des SOC par rapport aux FT, ce dispositif permet d'évaluer le rapport qualité-prix et de fixer des objectifs de performance pour les fournisseurs.

Recommandation n° 4 : Suivre et rendre compte des coûts opérationnels afin de garantir la transparence des marchés publics.

L'évaluation de l'efficience des modèles PbR et CIS est particulièrement difficile en raison du peu d'informations disponibles sur les coûts engagés lors de la conception de ces contrats, de la mise en place des programmes, du suivi des performances et de l'évaluation des résultats. Ces coûts devraient être inclus dans le coût total du programme déclaré, même si ce n'est pas la pratique courante actuellement. Ces coûts devraient être accessibles au public afin de garantir la transparence des marchés publics.

Recommandation n° 5 : Afin de permettre aux petites organisations du secteur tertiaire de participer à des contrats basés sur les résultats, utiliser le mécanisme PbR comme une prime pour les résultats obtenus au-delà des attentes, baser une petite partie (par exemple jusqu'à 20 %) du contrat PbR sur les résultats ou mettre en place des CIS.

Parmi les cas de PbR analysés, les organisations du secteur tertiaire ne disposaient pas de fonds suffisants pour assumer les coûts initiaux du programme et ne pouvaient pas supporter les risques financiers si les résultats n'étaient pas atteints. Ceci était

particulièrement vrai dans le cas des programmes pour lesquels une grande partie des paiements était basée sur les résultats. C'est pourquoi nous recommandons d'utiliser le mécanisme PbR comme une prime pour les résultats obtenus au-delà des objectifs fixés par les commissaires du programme. Une alternative serait de baser une petite partie (par exemple jusqu'à 20 %) du contrat PbR sur les résultats finaux. Cela signifie que la plupart des paiements prévus dans le contrat seraient basés sur les résultats intermédiaires. Une autre solution serait d'utiliser les CIS à la place. Pour les modèles CIS et PbR, le fait de fixer la durée du contrat à au moins trois ans (ou idéalement plus) donnerait également plus de stabilité aux prestataires de services en termes d'embauche et de gestion de projet.

Recommandation n° 6 : Les contrats devraient prévoir une certaine flexibilité concernant les objectifs de performance en liant les attentes de performance aux performances du groupe de contrôle, en fixant des objectifs différents reflétant des environnements macroéconomiques différents, en fixant des plafonds et des planchers de paiement, et/ou en incluant des clauses d'interruption anticipée.

Pour éviter le risque de devoir renégocier les accords, les contrats SOC devraient prévoir une certaine flexibilité concernant les objectifs de performance, de sorte que ces objectifs puissent être modifiés en fonction de l'évolution des conditions macroéconomiques, qui à son tour affecte les résultats obtenus. Les attentes de performance pourraient être liées aux performances d'un groupe de contrôle, qui recevrait les services en même temps que les participants au programme SOC. Une autre solution consisterait à définir différents objectifs de performance avant le début du programme, selon que l'économie connaît ou non une récession pendant la mise en œuvre du programme. Le plafonnement des paiements permettrait au commissaire de ne pas avoir à verser des bénéfices excessifs si le programme donne de meilleurs résultats que prévu. D'un autre côté, des planchers de paiement pourraient être mis en place afin de garantir que les SOC soient financièrement viables pour les fournisseurs, même lorsque le volume des participants est faible. Enfin, des clauses d'interruption anticipée limiteraient le risque pour les commissaires et les investisseurs dans les cas où les premiers résultats du programme n'atteindraient pas les seuils minimums.

Recommandation n° 7 : Combiner les mesures de résultats binaires⁶ et fréquentielles, et tenir compte de la « distance parcourue » par les participants ainsi que de la satisfaction des participants pour éviter les incitations perverses.

Afin de décourager les incitations perverses et tenir compte des intérêts des différentes parties prenantes, nous recommandons de combiner les mesures de résultats binaires et fréquentielles lors de la conception des programmes SOC. Cependant, même lorsque les mesures de résultats binaires et fréquentielles sont combinées, certains participants au programme peuvent encore être trop loin d'atteindre le seuil de résultats le plus bas, ce qui

⁶ Avec des résultats binaires, les fournisseurs doivent atteindre un objectif absolu, et aucun paiement n'est accordé en cas d'obtention de résultats inférieurs. Par exemple, un résultat binaire consiste à savoir si une personne a trouvé un emploi ou non. Dans le cas d'une approche fréquentielle, les récompenses sont échelonnées en fonction de la fréquence de résultats convenue, les paiements augmentent donc à mesure que les résultats augmentent. Par exemple, les prestataires ou les investisseurs reçoivent des paiements supplémentaires pour chaque trimestre où les participants au programme restent en activité.

se traduit souvent par un écrémage et une mise à l'écart. Par exemple, si un programme vise à placer les participants dans un emploi et à maintenir cet emploi pendant un certain nombre de mois, certains participants ne parviendront pas à trouver un travail. Par conséquent, nous suggérons également d'inclure dans les résultats des contrats PbR les résultats liés à la « distance parcourue », tels que l'accord sur un plan d'action, la préparation du CV, la participation à des entretiens et autres. De cette manière, il serait plus viable financièrement pour les prestataires de travailler avec des participants moins aptes à l'emploi. Une autre façon de limiter l'écrémage et la mise à l'écart consiste à surveiller les performances de groupes d'individus différents (c'est-à-dire plus ou moins proches de parvenir aux résultats visés) et à sonder la satisfaction des participants pendant la durée du programme.

Recommandation n° 8 : Impliquer les prestataires de services dans la définition des mesures de résultats et la conception de l'évaluation.

Étant donné que la participation des prestataires de services a contribué à garantir le choix des résultats les plus pertinents dans un certain nombre de programmes SOC analysés, nous recommandons que les prestataires de services soient inclus dans le processus de définition des résultats. En outre, étant donné la charge que les évaluations imposent aux prestataires de services, ils devraient également être consultés sur la conception de l'évaluation. Par exemple, au cours des phases initiales du programme, un essai de techniques de collecte de données pourrait être mis en œuvre afin de clarifier les informations à utiliser pour recueillir les preuves des résultats déclencheurs de paiements. Si nécessaire, les prestataires de services devraient être formés à la manière dont ces données doivent être traitées.

Zusammenfassung

Begründung

Social Outcome Contracts (SOCs) sind Verträge, bei denen das Ergebnis einer Sozialleistung maßgeblich ist, und stellen eine relativ neue Form von Mechanismus im öffentlichen Auftragswesen dar. Dabei liegt der Fokus darauf, Ressourcen von öffentlichen und privaten Trägern, Förderern und der Zivilgesellschaft gebündelt zu nutzen, um wirksame Interventionen im öffentlichen Bereich durchzuführen. Sie gelten als potenzielles Instrument, mit dem das öffentliche Auftragswesen und die europäischen Sozialsysteme modernisiert werden können: Ergebnismessung und sektorübergreifende Partnerschaften werden angeregt, neue Interventionsmodelle getestet und skaliert und somit eine Lernkultur und soziale Innovation gefördert. Allerdings gibt es wenig empirische Studien, in denen SOC-Modelle mit traditionellen Finanzierungsmechanismen (TF) wie Subventionen, Fördermitteln, Dienstleistungsverträgen, Rahmenverträgen oder interner Durchführung verglichen werden. In dieser Studie sollen daher die Wirksamkeit und Wirtschaftlichkeit von ergebnisbasierten Verträgen im Vergleich zu traditionell finanzierten Interventionen und ebenso die hierfür angewandten Methoden zur Messung von Ergebnissen bewertet werden. Im Fokus stehen dabei ausschließlich Sozialleistungen und nicht öffentliche Dienstleistungen im Allgemeinen.

In der Studie werden zwei Arten von SOCs betrachtet, bei denen Wirkung und Ergebnis maßgeblich sind: Social Impact Bonds (SIBs) und Payment-by-Results (PbR) Programme. SIBs werden dabei als Finanzierungsmechanismen definiert, bei denen Investoren die vorab benötigten Finanzmittel für ein Programm bereitstellen. Dabei tragen die Investoren das Risiko, die investierten Mittel zu verlieren, sollten die angestrebten Programmergebnisse nicht erzielt werden. Ist das SIB-Programm jedoch erfolgreich, werden sämtliche Investitionen rückerstattet und eine zusätzliche Zinszahlung vom Auftraggeber gewährt. Bei einigen SIB-Programmen gibt es Vermittler, die bei der Suche von Dienstleistern, Investoren und/oder Prüfern behilflich sind. Im Gegensatz dazu sind bei PbR-Programmen keine externen Investoren und Vermittler involviert. Dienstleister erhalten nur dann Zahlungen vom Auftraggeber, wenn die angestrebten Ergebnisse erzielt wurden. PbR-Modelle variieren darin, zu welchem Anteil die Gesamtzahlung an den Dienstleister ergebnisabhängig ist. Bei reinen PbR-Modellen sind 100% der Zahlung abhängig von den Ergebnissen. Dies ist nicht der Fall bei gemischten PbR-Modellen, bei denen die Zahlung zu einem gewissen Anteil traditionell finanziert wird.

Methodik

Fünfzehn SOC-Fälle wurden für die umfassende Analyse im Rahmen der Studie ausgewählt, mit unterschiedlichen SOC-Modellen, geographischen Gebieten und Sozialleistungen. Dabei wurden nur (nahezu) abgeschlossene Programme berücksichtigt, bei denen Bewertungen vorlagen und eine Datenerhebung möglich war. Aus diesem Grund zählen einige der ausgewählten SIBs zu den im jeweiligen Land ersten derartigen Programmen. Für jedes SOC-Modell wurden vergleichbare Programme mit traditioneller Finanzierung (TF) identifiziert. Der Vergleich zwischen SOC- und TF-Programmen sollte dazu dienen, zu bewerten, in welchem Maß bei beiden die angestrebten Ziele erreicht wurden und mit welchem finanziellen Aufwand. Dabei sollte auch der Mehrwert von SOC-Programmen in Bezug auf Vor- und Nachteile im Vergleich zu TF-Programmen ermittelt werden und Gestaltungselemente identifiziert werden, die eine Reproduzierbarkeit und Skalierbarkeit ermöglichten und für Dienstleistungs- und Sozialunternehmen eine Unterstützung darstellten. Die Datensammlung erfolgte im Rahmen einer Sekundärstudie,

bei der in erster Linie Programmbewertungsberichte und Prüfanfragen sowie 54 Befragungen mit Auftraggebern, Dienstleistern, Prüfern, Vermittlern, Experten und Investoren berücksichtigt wurden.

Ergebnisse

Wirksamkeit und Mehrwert

Um die Wirksamkeit von SOC- und TF-Programmen zu beurteilen, wurde bewertet in welchem Maße bei jedem Programm die vorab festgelegten Ergebnisse erzielt werden konnten. Dazu zählten die Anzahl von Teilnehmern die eine Beschäftigung fanden, eine geringere Rückfallquote, die Anzahl Kinder, die nicht in eine Pflegeeinrichtung eingewiesen werden mussten, und weitere Aspekte abhängig von Programmzielen und Leistungsbereichen. **Es gibt keine ausreichenden Beweise dafür, dass ergebnisbasierte Verträge wirksamer sind als solche mit traditioneller Finanzierung oder umgekehrt.** Dies liegt zum Teil daran, dass beim Großteil vergleichbarer TF-Programme weder zu erzielende Ergebnisse festgelegt wurden, noch eine stringente Bewertung stattfand.

Die Analyse ergab jedoch auch, dass das Erreichen von Zielen nicht unbedingt für die Wirksamkeit und den Erfolg eines Programms insgesamt maßgeblich ist. **Tatsächlich wurden einige Programme trotz nicht erreichter Ziele als erfolgreich bewertet, da sie den Bedürfnissen und Erwartungen der verschiedenen Beteiligten gerecht wurden.** Folglich wurden sie ausgeweitet und reproduziert. Dies war besonders bei SIBs der Fall, die dazu eingesetzt wurden, eine Finanzierungslücke für eine bestimmte Gruppe von Teilnehmern zu schließen oder zu testen, ob eine innovative Intervention wirksam ist und skaliert/reproduziert werden kann. Teilnehmer der Befragung nannten **weitere Vorteile** von SOC-Programmen, einschließlich Entwicklung von Messinfrastruktur, Bereitstellung von Beweisen für die Politikgestaltung, Wissensaustausch unter den Beteiligten, Erschließung finanzieller Ressourcen sowie mehr Flexibilität für Dienstleister.

Zwar scheinen SOC's häufiger dafür eingesetzt zu werden, bestimmte Kategorien benachteiligter Personen wie psychisch Erkrankte, Migranten und ehemalige Häftlinge in Beschäftigung zu bringen. Es wurde jedoch festgestellt, dass bei keiner spezifischen Zielgruppe mit Hilfe von SIB- oder PbR-Modellen eine positivere Wirkung erzielt wurde als bei einer anderen. **Die Wirksamkeit von SOC's hing stärker vom jeweiligen Kontext und der Gestaltung einer Intervention ab als von bestimmten Leistungsempfängertypen oder Bereichen von Sozialleistungen.**

In Übereinstimmung mit der vorhandenen Literatur **ergab die Studie, dass PbR-Verträge Vorfälle von Bevorzugung (Creaming-Effekt⁷) sowie Benachteiligung (Parking-Effekt⁸) von schwer zu helfenden Personen bewirkten.** Die erfassten Daten belegen, dass dieses Problem zwei Ursachen hat. Zum einen trieb der Zahlungsmechanismus Dienstleister dazu, leicht zu helfenden Personen zu bevorzugen, insbesondere bei eingeschränkter Programmfinanzierung und schwer zu erreichenden Zielen. Das Programm wäre für sie sonst nicht tragbar gewesen. Zum anderen waren einige Leistungsempfänger nicht bereit, an Programmen mit verpflichtender Teilnahme teilzunehmen, oder waren nicht davon überzeugt, die Programmziele (z. B. Beschäftigung) erreichen zu können aufgrund multipler Hindernisse (z. B. Kinderbetreuung, Behinderung, fehlende Transportmittel etc.). In solchen Fällen fanden es die Dienstleister angemessen, diese Teilnehmer zu „parken“. Die Studie ergab weniger solcher Fälle bei SIBs, jedoch

⁷ In der Studie ist mit Creaming-Effekt die Auswahl von Programmteilnehmern gemeint, die am wahrscheinlichsten die angestrebten Ergebnisse erreichen und somit Dienstleistern oder Investoren ermöglichen, Zahlungen zu erhalten.

⁸ Mit Parking-Effekt ist ein Verfahren gemeint, durch das Dienstleister versuchen, Kosten gering zu halten, indem sie wenig tun für Personen mit geringen Aussichten auf erfolgreiche Ergebnisse und sich stattdessen auf Personen mit besseren Erfolgsaussichten konzentrieren, z. B. mit besseren Beschäftigungsaussichten.

traten hier Fälle von Bevorzugung in einem Ausmaß auf, dass sogar Teilnehmer für SIBs rekrutiert wurden.

In vier der untersuchten Fälle wurden die Verträge formell neu verhandelt oder der Vertragsmanager wandte informell andere Erfolgskriterien an, wenn Dienstleister Zielvorgaben aufgrund von veränderten makroökonomischen Bedingungen nicht erreichen konnten. Die zusätzliche Belastung für Dienstleister, unerreichbare Ziele zu erreichen, begünstigte zudem falsche Anreize und verkomplizierte die Vertragsverwaltung.

Obwohl einige der untersuchten PbR-Programme nicht sehr wirksam für schwer zu helfenden Teilnehmergruppen waren, erwiesen sie sich als ziemlich wirksam für andere. Das Beschäftigungsprogramm erwies sich zum Beispiel als weniger erfolgreich für Antragsteller für Arbeitslosenhilfe mit Behinderung, übertraf jedoch die gesetzten Ziele für die meisten anderen Arten von Arbeitssuchenden. Ein ähnlicher Trend wurde vom Auftraggeber von JobPath beobachtet. Hieraus lässt sich vermuten, dass PbR-Programme wahrscheinlich am besten für Zielgruppen geeignet sind, denen leicht zu helfen ist, jedoch nicht so leicht, dass die angestrebten Ergebnisse auch ohne Intervention erzielt werden könnten (z. B. Langzeitarbeitslose). **Der Erfolg von PbR-Programmen im Vergleich mit dem Erfolg von TF-Programmen ergab allerdings eine gemischte Bilanz.** Darüber hinaus war in mehreren Fällen ein direkter Vergleich der Programmwirksamkeit nicht möglich, da bei TF-Programmen Ergebnisse nicht konsequent erfasst wurden.

Wirtschaftlichkeit

Bei jedem SOC-TF-Programmvergleich wurden die Kosten pro Ergebnis betrachtet, um die relative Wirtschaftlichkeit derartiger Programme zu bewerten. Mit Kosten pro Ergebnis sind je nach Zielen des betreffenden Programms zum Beispiel die Kosten pro Teilnehmer, der eine Beschäftigung gefunden hat, der für eine bestimmte Zeit beschäftigt geblieben ist, der das Rehabilitationsprogramm abgeschlossen hat, etc. gemeint. Bei sieben von insgesamt 15 SOC-TF-Programmvergleichen waren ausreichende Informationen über Interventionskosten verfügbar. Es ist wichtig darauf hinzuweisen, dass dabei beeinflussende Faktoren wie Besonderheiten von Teilnehmern nicht kontrolliert werden konnten. **Es kann daher nicht ausgesagt werden, dass SOC-Programme allein aufgrund des Finanzierungsmechanismus wirtschaftlicher oder unwirtschaftlicher waren.**

Für die Fälle, bei denen Vergleiche möglich waren, waren die Kosten pro Ergebnis bei SIBs meist höher als bei ähnlichen traditionell finanzierten Programmen. Die Auftraggeber sahen die zusätzlichen Kosten jedoch als gerechtfertigt an: SIBs wurden dabei nicht in erster Linie dafür eingesetzt, um ein Programm auf wirtschaftlichere Weise umzusetzen, sondern um bestehende Programme in unterschiedlichen Kontexten oder mit variierender Skalierung testen zu können. Außerdem traten Fälle auf, in denen Auftraggeber keine Zahlung an Investoren leisteten, wenn die SIBs nicht die angestrebten Ergebnisse erzielten, und so für den Steuerzahler keine Kosten entstanden (abgesehen von den investierten Mitteln für Aufbau und Verwaltung des Programms). **Bei einem erfolglosen Programm stellten sich SIBs folglich als hilfreich heraus, um dieses schnell einstellen zu können.** Anders ist dies bei einigen TF-Programmen, die trotz fehlender Beweise für ihre Wirksamkeit weiterlaufen.

Was PbR-Programme betrifft, ist ein oft genanntes Argument für ihren Einsatz, dass sie wirtschaftlicher als traditionell finanzierte Programme sind, da der Fokus auf Ergebnisse Dienstleister zu Innovation und Abschaffung unwirtschaftlicher Praktiken antreibt. Tatsächlich wiesen in der Studie die zwei PbR-Programme, deren Kosten pro Ergebnis mit denen ähnlicher TF-Programme verglichen wurden, entweder vergleichbare oder geringere

Kosten auf. Bei den anderen ausgewählten PbR-Programmen ist es indes nicht möglich einen Vergleich zu erstellen, da hier nur **unzureichende Informationen sowohl über Kosten von PbR-Programmen als auch Ergebnisse von TF-Programmen** vorliegen. Darüber hinaus ist es aufgrund der bestehenden Bedenken hinsichtlich falscher Anreize bei PbR-Verträgen wichtig zu prüfen, ob die Qualität von Leistungen bei geringerem Kostenaufwand beibehalten werden kann.

Sowohl SIB- als auch PbR-basierte Programme waren allgemein mit größeren operationellen Kosten (für Aufbau und Verwaltung) verbunden als TF-Programme. Zum Beispiel **dauerte deren Aufbau manchmal zwischen einem und zwei Jahren.** Die verfügbaren Informationen sind nicht ausreichend, um bewerten zu können, ob die zusätzlichen Kosten durch Einsparungen aufgewogen werden, die PbR-Programme potenziell erzielen können, indem sie die Wirtschaftlichkeit von Interventionen fördern. Zudem besteht wenig Transparenz hinsichtlich der Kosten für Programmgestaltung und -überwachung bei SIBs.

Vertragsgestaltung

Zu den untersuchten Fällen zählen auch mehrere Unterarten von SIB- und PbR-Modellen, mit jeweils unterschiedlich hoher Risikoübertragung vom Auftraggeber an die Investoren oder Dienstleister. Mit „Risiko“ sind die finanziellen Folgen gemeint, falls bei einer Intervention die angestrebten Ergebnisse nicht erzielt werden. Bei beiden Arten von SOC **konnte das Risiko für alle Beteiligten nicht durch die Vertragsgestaltung verringert werden: wenn das Risiko für den Auftraggeber verringert wurde, erhöhte es sich für die Investoren oder Dienstleister.** Dies widerlegt die bestehende Aussage, dass ergebnisbasierte Verträge, insbesondere SIBs, für alle Beteiligten von Vorteil sind.

Der Risikoanteil, der bei SIBs vom Auftraggeber an die Investoren übertragen wurde, variierte aufgrund mehrerer Faktoren. Ambitionierte **Programmziele** erhöhten das Risiko für Investoren, ihre Investition zu verlieren, und verringerten zugleich das Risiko der Auftraggeber, unwirksame Interventionen finanzieren zu müssen. Gleichmaßen war die Bewertung einer größeren **Anzahl von Ergebnissen** mit einem geringeren Risiko für Investoren verbunden, dass keines der Ergebnisse erzielt würde. Das Risiko von Investoren wurde weiter verringert, wenn Förderer oder Auftraggeber einen **Teil der Investition garantierten. Je regelmäßiger die Zahlungen** vorgesehen waren, desto geringer war das Risiko für Investoren, die gesamte Investition zu verlieren. Regelmäßigere Zahlungen bedeuteten allerdings für Auftraggeber das Risiko einzugehen, eine Intervention anteilig zu finanzieren, bei der die angestrebten Ergebnisse nicht erreicht werden konnten. Letztlich trugen Auftraggeber ein bedeutendes Risiko, wenn sie unabhängig von weiteren Zahlungsvereinbarungen mit Investoren dafür verantwortlich waren, eine vertraglich festgelegte Anzahl von Teilnehmern zu **rekrutieren.**

Bei den untersuchten PbR-Programmen verhielt es sich so, dass je größer der ergebnisbasierte Anteil eines Vertrags war, desto schwieriger war es für kleinere Dienstleistungs- und Sozialunternehmen ein Angebot als Hauptdienstleister (im Gegensatz zu Unterauftragnehmern) einzureichen. Dienstleistungsunternehmen fehlten die erforderlichen Finanzmittel, um die Kosten zur Einrichtung des Programms zu decken. Sie waren auch nicht in der Lage, das finanzielle Risiko im Falle nicht erzielter Ergebnisse zu tragen. Im Allgemeinen war es für kleinere Dienstleistungsunternehmen und Sozialunternehmen einfacher sich an SIBs zu beteiligen, da derartige Modelle bei fehlendem Erfolg mit einem geringeren finanziellen Risiko für den Dienstleister verbunden waren. Auftraggeber erachteten SIB- und PbR-Programme aufgrund der mit ihnen verbundenen Vorteile und trotz zusätzlicher Kosten insgesamt als positiv.

Messung von Ergebnissen

Die Bewertung von Ergebnissen ist ein Hauptbestandteil von SOC-Programmen, da hierdurch bestimmt wird, ob Investoren oder Dienstleister Anrecht auf eine Zahlung haben. Bei den untersuchten Fällen **wirkte sich die Art, wie Ergebnisse definiert wurden, maßgeblich darauf aus, ob falsche Anreize für Creaming- und Parking-Effekte gefördert oder verhindert wurden**. Ergebnisse wurden häufig mit einem **binären oder frequenzbasierten Ansatz** definiert. Bei binären Ergebnissen mussten Dienstleister ein absolutes Ergebnis erreichen und es gab keine Zahlung bei geringeren Ergebnissen. Ein Beispiel hierfür ist, ob eine Person eine Beschäftigung gefunden hat oder nicht. Bei frequenzbasierten Modellen wurden die Zahlungen gemäß der vereinbarten Frequenz der Ergebnisse stufenweise gewährt, wobei sich die Zahlungen mit zunehmenden erfolgreichen Ergebnissen erhöhen. Dienstleister oder Investoren erhielten zum Beispiel zusätzliche Zahlungen für je drei Monate, die ein Programmteilnehmer beschäftigt blieb. Bei Modellen **mit binären Ergebnissen traten häufiger Creaming- und Parking-Effekte auf**, da das angestrebte Ergebnis (z. B. Beschäftigung) für einige Teilnehmer nicht zu erreichen war und Dienstleister sich deshalb auf andere Personen fokussierten.

Außerdem **argumentierten Investoren, dass die Messung frequenzbasierter Ergebnisse besser den wahren Erfolg oder Misserfolg eines Programms zeigte**. Einzelne Investoren zweifelten an, dass ein Programm als gescheitert betrachtet werden kann, wenn das angestrebte Ziel in hohem Maße erreicht wurde – z. B. 19 von 20 Teilnehmern (was bei einem binären Modell als Misserfolg bewertet würde). Der Nachteil von frequenzbasierten Ergebnissen war, dass diese schwieriger zu erfassen waren, sie der Öffentlichkeit weniger intuitiv zu vermitteln waren und der Risikoanteil, der vom Auftraggeber an Investoren oder Dienstleister übertragen werden konnte, begrenzt war. Um den unterschiedlichen Interessen aller Beteiligten gerecht zu werden, wurden bei einem Drittel der untersuchten Fälle sowohl binäre als auch frequenzbasierte Ergebnisse gemessen.

Es gab Fälle, in denen die **Einbeziehung von Dienstleistern bei der Auswahl von Ergebnissen dabei half, ein Abweichen von der wahrzunehmenden Aufgabe zu verhindern, und sicherstellte, dass nicht nur kurzfristige, sondern auch mittelfristige Ergebnisse berücksichtigt wurden** (siehe Anhang 1 bezüglich der kooperativen Gestaltung bei Drogen- und Alkohol-Rehabilitationspilotprojekten sowie „DUO for a JOB SIB“). Die Einbindung von Dienstleistern ermöglichte ebenfalls, ihre Fachkompetenz zu nutzen, um Ergebnisse festzulegen, mit denen das Wohlbefinden der Leistungsempfänger am besten gesichert wurde.

Im Hinblick auf die Bewertungsmethoden lässt sich sagen, dass bei mehreren untersuchten Programmen die Dienstleister dafür verantwortlich waren, die Beweise dafür zusammenzutragen, dass die angestrebten Ergebnisse erzielt wurden. Dies war oft eine Herausforderung, da die Dienstleister nicht über die Kompetenz verfügten, die erforderlichen Beweise zu erfassen. Informationen, die als Beweise dienen sollten (z. B. Arbeitsverträge, Bildungsnachweise, Teilnahmeberechtigungsnaehweise), lagen nicht vor und die Dienstleister konnten deshalb für einige der erzielten Ergebnisse keine Zahlungen fordern. Hinzu kommt, dass die Ergebniserfassung zeitaufwändiger war, als Dienstleister vorgesehen hatten, und sie deshalb weniger Zeit als erwartet für Leistungsempfänger hatten. Bei PbR-Programmen traten Interessenkonflikte hinsichtlich der von Dienstleistern erfassten Daten auf, weil diese maßgeblich für die Zahlungen waren.

Obwohl randomisierte kontrollierte Studien (Randomised Controlled Trials - RCTs) als Goldstandard für positivistische Bewertungen gelten, wurde **bei keiner der 15 untersuchten Fälle eine RCT angewandt**. Bei einem Programm (ABLE SIB) wurde die RCT eines Verhaltenstherapieprogramms für jugendliche Häftlinge versucht, jedoch

letztlich davon abgesehen, da keine klaren Abgrenzungen zwischen Behandlungs- und Kontrollgruppen vorgenommen werden konnten, wenn Teilnehmer in eine andere Wohneinheit wechselten. Dienstleister von anderen Programmen erachteten es zudem als unmoralisch, potenziellen Teilnehmern eine Leistung zu verweigern, nur um eine Kontrollgruppe zu bilden – insbesondere, wenn keine anderen Dienstleister oder staatliche Institutionen dieselben Leistungen außerhalb eines SOC-Programms anbieten. Zudem sind RCTs sehr teure Verfahren.

Die konsequentesten Methoden, die bei den untersuchten Fällen beobachtet wurden, waren quasi-experimentell, fanden aber nur bei sieben Fällen Anwendung. Von diesen wurde nur bei zwei Programmen die quasi-experimentelle Gestaltung dafür genutzt, die Ergebnisse im Zusammenhang mit dem Zahlungsmechanismus zu bewerten. Das bedeutet, dass konsequente Bewertungsmethoden, bei denen die Intervention kausal mit ihren Auswirkungen in Verbindung gesetzt wurde, nur bei zwei Programmen für die Zahlung maßgeblich war. **In den meisten Fällen bedeutete das Fehlen einer Kontrollgruppe, dass die in SOC-Programmen erzielten Ergebnisse nicht eindeutig der Intervention zugeschrieben werden konnten. Obwohl die SOC-Programme bereits seit einiger Zeit durchgeführt wurden, fehlten bei vielen immer noch belastbare Bewertungen, um einen kausalen Zusammenhang bei Auswirkungen sowie Wirksamkeit, Wirtschaftlichkeit und Mehrwert feststellen zu können.**

Empfehlungen

Empfehlung 1: SIBs sollten in Bereichen eingesetzt werden, wo eine Finanzierungslücke besteht, und um zu testen, ob eine innovative Intervention wirksam und/oder skalierbar ist.

Auf Grundlage der beschriebenen Studienergebnisse sollten SIBs in Bereichen eingesetzt werden, wo eine Finanzierungslücke besteht, und um zu testen, ob eine innovative Intervention wirksam und/oder skalierbar ist. Derartige SIBs wurden von den befragten Beteiligten als am erfolgreichsten betrachtet, sogar wenn die SIBs die angestrebten Ergebnisse nicht erzielten. Der Einsatz von SIBs ist allerdings schwer zu rechtfertigen, wenn sie den oben genannten Kriterien nicht entsprechen, da höhere Verwaltungskosten und Zinszahlungen an Investoren anfallen. Vor der Einführung von SIBs sollten Durchführbarkeitsstudien erfolgen, um sicherzustellen, dass das Finanzierungsinstrument einen Mehrwert zum Ausgleich für die zusätzlichen Kosten schafft.

Empfehlung 2: Die Auftragsvergabe in Form von PbRs sollte vermieden werden, wenn es bei dem Vertrag um Leistungen für Gruppen geht, für die es schwer ist die angestrebten Ergebnisse zu erzielen, weil sie mit multiplen Hindernissen konfrontiert sind, insbesondere bei Programmen mit verpflichtender Teilnahme.

In Anbetracht der falschen Anreize, die in den PbR-Programmen beobachtet wurden, sollten PbRs nicht für Verträge über Leistungen für Gruppen eingesetzt werden, für die es schwer ist die angestrebten Ergebnisse zu erzielen, weil sie mit multiplen Hindernissen konfrontiert sind, insbesondere bei Programmen mit verpflichtender Teilnahme. Leistungen für Gruppen mit komplexen Bedürfnissen sollten auf traditionelle Weise finanziert werden. Es ist hierbei wichtig anzumerken, dass SIBs im Gegensatz zu PbR-Programmen ein wirksames Verfahren zum Testen von Interventionen für Leistungsempfänger mit multiplen und komplexen Bedürfnissen sein könnten. SIBs bringen allerdings zusätzliche Kosten mit sich. Deshalb sollten Programme nicht weiter durch SIBs finanziert werden, sobald ausreichend Beweise vorliegen, dass sie in einem bestimmten Kontext wirksam sind.

Empfehlung 3: Wirksamkeit und Wirtschaftlichkeit von PbR-Programmen für leicht zu helfenden Gruppen durch Erfassen von Ergebnissen in TF-Programmen und mit Hilfe von Kontrollgruppen bewerten

In Anbetracht der gemischten Bilanz hinsichtlich Wirksamkeit von PbR- und TF-Modellen sowie fehlenden Informationen über ihre Wirtschaftlichkeit sind mehr Daten über die Ergebnisse von TF-Programmen erforderlich. Bei TF-Programmen sind zwei Ansätze zur Ergebniserfassung zu empfehlen. Zum einen könnten die Ergebnisse von TF-Programmen bewertet werden, indem Daten aus bestehenden administrativen Datenbanken (z. B. Datensätze von Steuerbehörden, Sozialversicherungsbehörden, Jugendämtern etc.) verwendet oder verlinkt werden. Zum anderen könnten die traditionell finanzierten Modelle als Kontrollgruppen für PbR-Modelle genutzt werden. Im Rahmen des letzten Beschäftigungs- und Gesundheitsprogramms im Vereinigten Königreich zum Beispiel beauftragte das zuständige Ministerium private Dienstleister auf PbR-Basis mit der Durchführung einzelner Programme und führte diese weiterhin auch intern durch lokale Arbeitsämter (Jobcentre Plus) durch. Neben der Bereitstellung von Schlüsseldaten, die es ermöglichen die Wirkung von SOC's im Vergleich mit TF zu bewerten, hilft dieses Verfahren dabei, das Kosten-Nutzen-Verhältnis zu bewerten und Zielvorgaben für Dienstleister festzulegen.

Empfehlung 4: Operationelle Kosten erfassen und melden, um sicherzustellen, dass die öffentliche Auftragsvergabe transparent ist

Die Wirtschaftlichkeit von PbR- und SIB-Modellen zu bewerten, ist aufgrund der wenigen verfügbaren Informationen über Kosten, die bei der Gestaltung dieser Verträge, dem Aufbau der Programme, der Ergebniserfassung und -bewertung angefallen sind, besonders schwierig. Diese Kosten sollten in den gemeldeten Gesamtkosten des Programms enthalten sein, obwohl dies derzeit keine gängige Praxis ist. Derartige Kosten sollten der Öffentlichkeit zugänglich sein, um sicherzustellen, dass die öffentliche Auftragsvergabe transparent ist.

Empfehlung 5: Um sicherzustellen, dass sich kleinere Dienstleistungsunternehmen an ergebnisbasierten Verträgen beteiligen können, sollte der PbR-Mechanismus als Prämie für Erwartungen übertreffende Ergebnisse angewendet werden, alternativ könnte ein geringer Anteil (z. B. bis 20 %) des PbR-Vertrags ergebnisbasiert sein oder SIBs eingesetzt werden.

Bei den untersuchten PbR-Fällen fehlten Dienstleistungsunternehmen die erforderlichen Finanzmittel, um die Kosten zur Einrichtung des Programms zu decken. Sie waren auch nicht in der Lage das finanzielle Risiko im Falle nicht erzielter Ergebnisse zu tragen. Das war insbesondere bei Programmen der Fall, in denen ein großer Anteil der Zahlungen ergebnisbasiert waren. Daher sollte der PbR-Mechanismus als Prämie für Ergebnisse eingesetzt werden, welche die von den Programmauftraggebern gesetzten Ziele übertreffen. Alternativ könnte ein geringer Anteil (z. B. bis 20 %) des PbR-Vertrags ergebnisbasiert sein. Das würde bedeuten, dass die meisten vertragsgemäßen Zahlungen leistungsbasiert wären. Eine weitere Alternative könnte der Einsatz von SIBs sein. Eine Vertragsdauer von mindestens drei Jahren (oder idealerweise länger) bei SIB- und PbR-

Modellen könnte für Dienstleister mehr Planungssicherheit hinsichtlich Personalfragen und Projektmanagement schaffen.

Empfehlung 6: Verträge sollten im Hinblick auf Zielvorgaben anpassbar sein und Erfolgserwartungen in Verbindung mit dem Erfolg einer Kontrollgruppe gesehen werden. Dabei sollten unterschiedliche Ziele abhängig vom makroökonomischen Kontext gesetzt und Ober- sowie Untergrenzen für die Zahlung und/oder Klauseln für frühzeitige Vertragsbeendigung vereinbart werden.

Um das Risiko zu vermeiden, Vereinbarungen neu verhandeln zu müssen, sollten SOC hinsichtlich der Zielvorgaben anpassbar sein, so dass Ziele abhängig von den makroökonomischen Bedingungen – welche die erzielten Ergebnisse beeinflussen – unterschiedlich sein können. Erfolgserwartungen könnten in Verbindung mit dem Erfolg einer Kontrollgruppe gesehen werden, die zur gleichen Zeit eine Leistung empfangen, wie die Teilnehmer des SOC-Programms. Alternativ könnten vor Programmbeginn unterschiedliche Zielvorgaben festgelegt werden, für den Fall, dass die Wirtschaft während der Programmdurchführung eine Rezession durchläuft oder nicht. Zahlungsobergrenzen würden den Auftraggeber davor schützen, übermäßige Prämien zu zahlen, falls das Programm besser als erwartet abschneidet. Als Ausgleich könnten Zahlungsuntergrenzen vereinbart werden, um sicherzustellen, dass SOC für Dienstleister finanziell tragbar sind, auch wenn die Teilnehmerzahl gering ist. Schließlich könnten Klauseln für eine frühzeitige Vertragsbeendigung das Risiko sowohl für Auftraggeber als auch Investoren vermindern, wenn Mindestschwellenwerte zu Beginn des Programms nicht erreicht werden.

Empfehlung 7: Erfassung von binären⁹ und frequenzbasierten Ergebnissen kombinieren und die Fortschritte von Teilnehmern sowie ihre Zufriedenheit berücksichtigen, um falsche Anreize zu vermeiden

Um falsche Anreize zu vermeiden und den Interessen unterschiedlicher Beteiligter gerecht zu werden, sollte bei der Gestaltung von SOC-Programmen die Erfassung binärer und frequenzbasierter Ergebnisse kombiniert werden. Aber auch wenn frequenzbasierte zusammen mit binären Ergebnissen gemessen werden, kann es vorkommen, dass einige Programmteilnehmer immer noch nicht die Mindestschwellenwerte erzielen. Dies führt häufig zu Creaming- und Parking-Effekten. Beispielsweise bei einem Programm, dessen Ziel es ist, dass Teilnehmer eine Beschäftigung finden und diese für einige Monate behalten, sind einige Teilnehmer nicht in der Lage eine Beschäftigung zu finden. Daher ist es ratsam, in PbR-Verträgen Ergebnisse festzulegen, die sich auf die Fortschritte des Teilnehmers beziehen, wie Erstellung eines Aktionsplans oder Lebenslaufs, Teilnahme an einem Vorstellungsgespräch etc. Dies stellt sicher, dass es für Dienstleister wirtschaftlich tragbar ist, mit Teilnehmern zusammenzuarbeiten, die noch weiter vom Ziel der Beschäftigung entfernt sind. Eine andere Möglichkeit um Creaming- und Parking-Effekte zu verringern, besteht darin, den Erfolg innerhalb der unterschiedlichen Gruppen zu überprüfen (Teilnehmer, für die es leichter, und Teilnehmer, für die es schwerer ist, die angestrebten Ergebnisse zu erreichen) und die Zufriedenheit der Teilnehmer während der Programmlaufzeit zu untersuchen.

⁹ Bei binären Ergebnissen mussten Dienstleister ein absolutes Ergebnis erreichen und es gab keine Zahlung bei geringeren Ergebnissen. Ein Beispiel hierfür ist, ob eine Person eine Beschäftigung gefunden hat oder nicht. Bei frequenzbasierten Programmen werden Zahlungen gemäß der vereinbarten Frequenz der Ergebnisse stufenweise gewährt, wobei sich die Zahlungen mit zunehmenden erfolgreichen Ergebnissen erhöhen. Dienstleister oder Investoren erhalten zum Beispiel zusätzliche Zahlungen für je drei Monate, die ein Programmteilnehmer beschäftigt bleibt.

Empfehlung 8: Dienstleister bei der Festlegung zu erfassender Ergebnisse und Bewertungsverfahren einbeziehen

In Anbetracht der Tatsache, dass die Einbeziehung von Dienstleistern bei mehreren der untersuchten SOC-Programme dabei half, die relevantesten Ergebnisse auszuwählen, ist es ratsam, Dienstleister bei der Festlegung von Ergebnissen einzubeziehen. Darüber hinaus sollten sie auch hinsichtlich Bewertungsverfahren zu Rate gezogen werden, da Bewertungen eine große Verwaltungslast für Dienstleister darstellen. So könnte in den frühen Phasen des Programms zum Beispiel ein Testversuch für die Erfassung von Daten durchgeführt werden, um deutlich zu machen, welche Informationen benötigt werden, um Ergebnisse belegen zu können, die für eine Zahlung erzielt werden müssen. Dienstleister sollten falls erforderlich darin geschult werden, derartige Daten zu erfassen.

Introduction

Social outcome contracting (SOC) is a relatively new type of mechanism in public service procurement. It focuses on harnessing the resources of the public, private, philanthropy and civil society sectors, with the goal of jointly implementing effective interventions in the public domain. The tool is said to provide potential to contribute to the modernisation of public services and European welfare regimes more generally: by encouraging a culture of performance measurement, supporting cross-sector partnerships and allowing to new effective intervention models to be piloted and scaled — thereby fostering learning and social innovation. More specifically, SOC schemes are expected to address many issues relating to traditional financing (TF) models of social services provision (such as subsidies, grants, fee-for-service contracts, block contracts and in-house delivery). These include:

- a low level of incentives and few opportunities to innovate and look for new, better and more effective and efficient solutions;
- a lack of outcomes measurement, and therefore insufficient evidence with which to evaluate the effectiveness of existing services. as well as This can also weaken or prevent the learning process;
- the potentially inefficient use of public finances in terms of the outcomes achieved.

Policy makers at national and EU levels are increasingly exploring opportunities to apply SOC models within public services. The goals of providing quality public services, the modernisation of welfare states and social innovation, as well as a focus on impacts/results – key aspects that SOC promise to deliver – are highlighted in a number of EU policy documents, including the Social Investment Package 'Towards Social Investment for Growth and Cohesion'¹⁰, the European Pillar of Social Rights¹¹, the EU Budget Focused on Results initiative, the InvestEU Programme¹², and the revised Public Procurement Directive. Finding ways to effectively fund public services has become increasingly important in the context of the Covid-19 pandemic, which has put many Europeans out of work, accompanied by a rise in domestic violence and limited access to public services among those who depend on them most.

Nonetheless, even though SOC schemes continue to spread across Europe and the world, little robust evidence exists as to their actual potential for innovation, as well as their impact, added value, and any drawbacks they may have. While SOC schemes have recently been the subject of increasing research interest from scholars and international organisations. no consensus has yet been reached as to their utility. A considerable number of authors regard SOC schemes as win-win situations, fostering both public and private/financial sector reforms¹³, improving the performance management of service delivery, as well as harnessing private resources to solve social issues. Others, meanwhile, tell a more cautionary story, highlighting the potential risks and drawbacks relating to their complex set-up, performance monitoring and evaluation, as well as their potential for scaling up and the related risk of 'creaming' beneficiaries. More importantly, there is also a lack of evidence about SOC, as most published articles and grey literature are best described as

¹⁰ Available at <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2013:0083:FIN:en:PDF>

¹¹ Available at https://ec.europa.eu/commission/sites/beta-political/files/social-summit-european-pillar-social-rights-booklet_en.pdf

¹² Available at https://eur-lex.europa.eu/resource.html?uri=cellar:319a131d-6af6-11e8-9483-01aa75ed71a1.0002.03/DOC_1&format=PDF

¹³ A literature review on SIBs conducted by Tan et al. (2019) suggests that the current discourse on SIBs consists not only of a reform narrative concerning the public sector, but also the financial sector. The latter suggests "that private sector actors, in particular, management consultancies and specialist intermediary organisations, can effect socially worthwhile change through social entrepreneurship whilst simultaneously pursuing commercial interests". See Tan, S., Fraser, A., McHugh, N. & Warner, M. (2019). Widening perspectives on social impact bonds. Taylor & Francis. Retrieved from: <https://www.tandfonline.com/doi/full/10.1080/17487870.2019.1568249>

commentary¹⁴. This lack of data concerns key aspects of SOC, such as the appropriateness of evaluation, measurement of outcomes and the attribution of causality (i.e. to what extent a specific intervention actually contributed to the desired effects) methods, better or worse practices, and effects of the legal and institutional contexts.

The objective of the study was therefore to fill some of the knowledge gaps and examine existing SOC schemes and evolving practices. The key research questions included:

- What are the most appropriate methods to monitor, measure and evaluate the outcomes of SOC schemes, and to allow the causality of impact to be established?
- In which areas of interventions and among which target groups do SOC deliver better or worse results than traditionally financed (TF) interventions?
- What are the features of SOC schemes in which the cost of intervention is higher/lower than for the same intervention delivered under TF?
- Which SOC design models are cost-effective?
- Which SOC design models are easily replicable and scalable?
- How can SOC schemes be designed to encompass social economy enterprises, not-for-profit service providers, and philanthropy organisations?
- In which design models has blending with other financial instruments been used successfully?

The study is structured as follows:

- Chapter 1 clarifies the main concepts and definitions used in the study.
- Chapter 2 outlines the methodology, which is presented in greater detail in Annex 2. This chapter presents a brief overview of the 15 SOC schemes selected for in-depth analysis in the study. Please note that individual analysis of the SOC schemes, together with comparisons with similar TF programmes, is presented in Annex 1.
- Chapter 3 discusses the methods used to measure outcomes in the SOC schemes analysed.
- Chapter 4 focuses on the effectiveness and added value of the SOC schemes, in comparison with similar TF programmes. In this chapter, we also present examples of SOC schemes that have been replicated and/or scaled.
- Chapter 5 compares the efficiency of SOC and TF programmes, taking into consideration their intervention and operational costs.
- Chapter 6 outlines the lessons learned from the cases analysed about various SOC design models, including the impact these design decisions have on third-sector organisations and social economy enterprises.
- Chapter 7 provides conclusions and recommendations with respect to the main research questions.

Chapters 3-6 include summaries of the main findings at the end of each chapter.

¹⁴ Fraser et al (2018). 4-28.

1. Concepts and definitions

With the emergence of various SOC schemes, in particular the launch of social impact bonds (SIBs) in the UK in 2010, interest in and uptake of these service delivery models has grown around the world. Non-profit organisations, such as the Finnish Innovation Fund SITRA and Social Finance Global Network in the UK, the Netherlands, USA, Israel and India, have become drivers advancing innovative financing mechanisms worldwide¹⁵. The UK appears to be the country that has implemented the most SOC operations, both within Europe and globally, but a number of cases can also be identified in Australia and the US. Other countries, such as the Netherlands, Portugal, Finland, Canada and Japan, are also experimenting with the development of SOC schemes. Meanwhile, other EU Member States have been more hesitant due to legal uncertainties or the fact that these concepts simply unfamiliar within their administrative systems¹⁶. Some, however, (e.g. Slovenia¹⁷) are seriously considering the feasibility and opportunities offered by SOC.

With its innovative form and potential for as-yet-undiscovered models, the terminology used to describe SOC schemes remains inconsistent, with various terms used interchangeably, or the same terms being used to denote slightly different things. We have identified a number of terms for models linked to the social outcomes contracting, such as payment by Results (PbR), social impact bonds, pay for Success (P4S or PFS), development impact bonds (DIBs), results-based financing (RBF), performance-based financing (PBF), and pay for performance (P4P). In the literature, these terms are used inconsistently, leaving the concepts fluid and hard to distinguish. The implementation of this assignment, therefore requires the mapping and conceptualisation of terms relating to SOC and their interrelationships, in order to guide our study process (see the box below). We then go on to present the main elements and characteristics of each.

Note on terminology

SOC is frequently used as an umbrella term, with PbR and SIB treated as two types of SOC. However, in other sources the terms PbR and SOC are nominally equivalent. Often, PbR is also used as an umbrella term, with SIB regarded as a type of PbR.¹⁸ Terms such as pay for success¹⁹, pay for performance, results-based financing²⁰ or performance-based financing²¹, and outcome-based payment are also used in this way, as a broader term, with SIB (or, in the context of development aid, DIB) being a sub-type.

We found that slightly different terminology tends to be used in certain contexts. For example, the term 'pay for success' seems to be more widely used in the US²².

¹⁵ Social Finance (n.d.). Retrieved from: <https://www.socialfinance.org.uk/who-we-are/global-network> & EVPA (2017).

SITRA – Impact Investing. Retrieved from: https://evpa.eu.com/uploads/publications/Policy_Nexus_FI_SITRA_2017.pdf

¹⁶ SITRA (2019). Social outcomes contracting and public procurement in various EU Member States. Retrieved from: <https://media.sitra.fi/2019/06/07135808/sitrasocsummary-findings-social-outcomes-contracting.pdf>

¹⁷ Kump Nataša, Kavaš Damjan, Črnigoj Matjaž (n.d.). Implementing Social Impact Bonds in Slovenia. Institute for Economic Research. Retrieved from: http://alpsib-project.eu/media/1059/discussion_paper_final_ier.pdf

¹⁸ OECD Working Paper (2016). Understanding social impact bonds. Retrieved from: <http://www.oecd.org/cfe/leed/UnderstandingSIBsLux-WorkingPaper.pdf>

¹⁹ Nonprofit Finance Fund (n.d.). Basics. Retrieved from: <https://www.payforsuccess.org/learn/basics/>

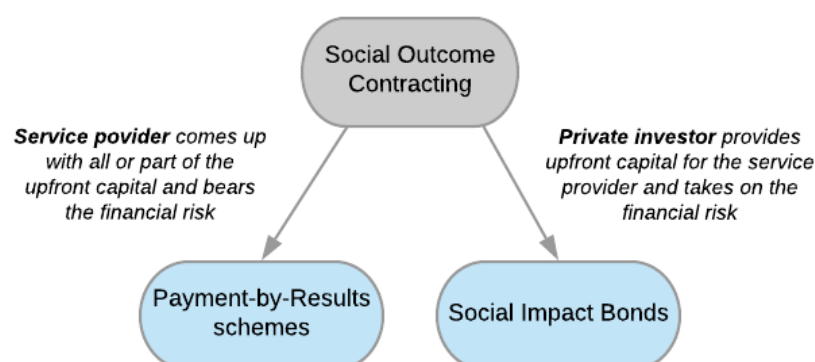
²⁰ Sida (2015). A methodological introduction. Results based financing approaches (RBFA) – what are they? Sida. Retrieved from: <https://www.sida.se/contentassets/1b13c3b7a75947a2a4487e2b0f61267c/18235.pdf>

²¹ USAID (n.d.). Performance-based financing. Retrieved from: <http://www.fpf financingroadmap.org/learning/specific-topics/performance-based-financing>

²² Albertson, K., Fox, C., O'Leary C., & Painter, G. (2018). Payment by results and social impact bonds. Retrieved from https://books.google.lt/books?id=1NFMDwAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Results-based financing²³ and performance-based financing appear to be the preferred terms in the context of international aid. In the UK, the term PbR is also linked to output-based commissioning initiatives within the healthcare sector²⁴, which fall outside the scope of this assignment. In Australia, SIBs are referred to as social benefit bonds²⁵. Such variety is, however, the result of the novelty of this phenomenon and the lack of a widely acknowledged vocabulary, rather than objective differences between the schemes.

Given that no single widely accepted and consistent terminology exists, and for the sake of conceptual clarity, we use SOC as an umbrella term, while PbR denotes non-SIB SOC schemes.



It is important to note that other forms of SOC may exist, although they are more rarely applied in the provision of public services. For example, a prize-based challenge is an open bid competition that awards a financial prize to the best innovative solution developed within a predefined timeframe. Rather than the best proposal being judged, competing solutions are assessed on the basis of the results they deliver, and competing actors must finance the innovations themselves unless they win²⁶. In preparation of this study, we did not identify any notable completed schemes of this kind in the area of public services.

Social outcomes contracting (SOC) is a broad term denoting the procurement of services based on outcomes rather than outputs. In such contracts, the commissioner (central or local government) and service provider agree on the pre-defined, desired end results (i.e. outcomes) of an intervention, on which the final payment is based²⁷. Once completed, an intervention based on an SOC scheme is considered successful if the service provider

²³ Sida (2015).

²⁴ Albertson, K. et al (2018).

²⁵ UNDP (n.d.). Social and development impact bonds (results-based financing). Retrieved from: <https://www.sdfinance.undp.org/content/sdfinance/en/home/solutions/social-development-impact-bonds.html>

²⁶ Instiglio (2017). A practitioner's Guide to Results-Based Financing. Retrieved from: https://www.payforsuccess.org/sites/default/files/resource-files/RBF_PractitionersGuidebook_Instiglio_18Oct2017.pdf

²⁷ National Audit Office (2015). Outcome-based payment schemes: government's use of payment by results. NAO. Retrieved from: <https://www.nao.org.uk/wp-content/uploads/2015/06/Outcome-based-payment-schemes-governments-use-of-payment-by-results.pdf>

achieves the pre-defined outcomes, and the commissioner therefore pays the agreed amount in return for these results. To decide whether the pre-defined outcomes have been achieved, a designated evaluator monitors the service provider's progress, conducts an impact assessment and reports on the intervention's performance. If the desired outcomes have not been achieved, the payment is not made. In this way, SOC's offer a way to share risk between the public and private sectors in the event of an unsuccessful intervention.

Evaluation or outcomes/impact measurement is a key feature of SOC interventions. Besides providing the basis for the payment for services, evaluation brings other side benefit that enable the improvement of social interventions more generally. These include the spread of performance measurement culture, accountable reporting, fairer distribution of funding, opportunities for learning and improvement, increased opportunities for future fundraising, and new collaboration opportunities.²⁸ However, the implementation and quality of evaluations can also pose some significant challenges in the implementation of SOC's.

SOC financing instruments²⁹ and models of payment are various. For example, in binary models, the provider must achieve an absolute target, and no payment is granted for achieving lesser results. In frequency schemes, rewards are staggered according to an agreed frequency of results, with payments increasing as results increase. Meanwhile, hybrid grants cover the cost of delivering a service, but additional payments are rewarded as bonuses if 'additional' impacts can be demonstrated at the end of a programme³⁰. The key component in all of these models is that the commissioning authority is not the body supplying service providers with upfront investment to implement the intervention.

The main distinction by which SOC schemes can be further differentiated is the involvement of a third party – a **private investor**. Here, we can most clearly distinguish two models, namely **social impact bonds** (SIB), which involve a private investor; and the **payment by results** (PbR) schemes, which do not (i.e. the service provider itself comes up with the upfront capital, with or without contributions from the commissioning authority). Some researchers and practitioners treat PbR as a broader category covering what we term 'SOC' in the context of this study, with SIB as a sub-type³¹. However, to provide a clearer analytical distinction, we treat SIB and PbR as two mutually exclusive sub-types of SOC.

The main characteristics and forms of **PbR** are discussed above in general features of SOC's. From a design viewpoint, PbR is a mechanism under which all or part of the payment from the commissioning authority depends on the provider achieving outcomes specified by the commissioner. Service providers therefore **need to make an upfront investment**, and some form of upfront payment or 'fee for service'. Providers are, to a greater or lesser extent, free to choose the interventions required to secure the desired outcomes, and are motivated to ensure successful performance.

The deferral of payment is generally part of the attractiveness of PbR for commissioners; however, it creates risk for providers, who need to finance the upfront investment in the interim. There are reported instances of smaller welfare-to-work providers withdrawing from contracts due to this time lag between investment and payment. One approach to overcome this and make schemes more attractive to potential bidders is to make a portion of the

²⁸ inFocus (2016). 7 Steps to effective impact measurement. inFocus Enterprises Ltd. Retrieved from: [https://www.undp.org/content/dam/ukraine/docs/DG/socinnov/7%20Steps%20to%20Effective%20Impact%20Measurement_v3_13.12.16%20\(1\).pdf](https://www.undp.org/content/dam/ukraine/docs/DG/socinnov/7%20Steps%20to%20Effective%20Impact%20Measurement_v3_13.12.16%20(1).pdf)

²⁹ Instiglio (2017).

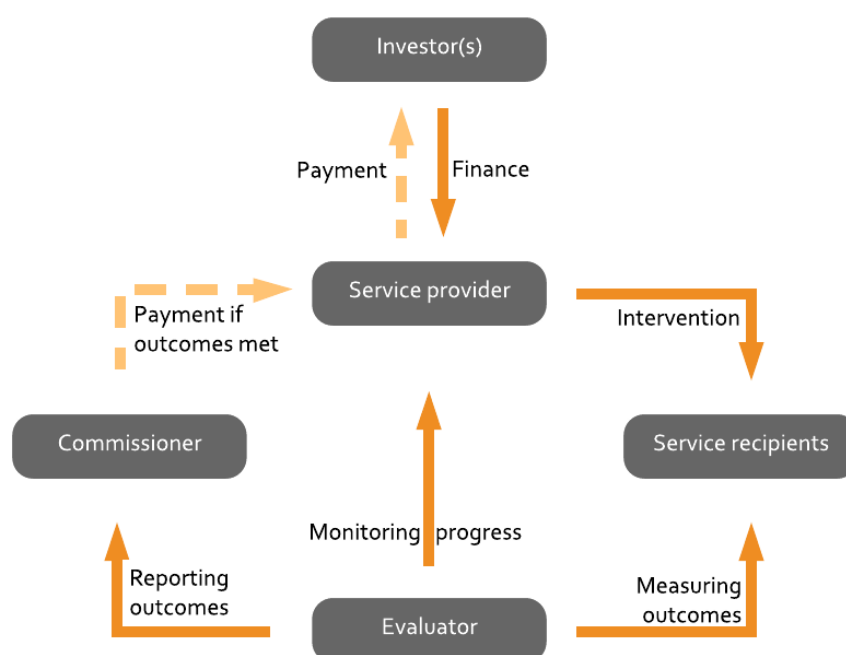
³⁰ NICVA (n.d.). Payment by Results. Retrieved from: https://www.nicva.org/sites/default/files/d7content/attachments-articles/public_sector_reform_-_payment_by_results.pdf; Publishing Service of the UK Government. (n.d.). Payment by Results. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/249928/Payment-by-Results.pdf

³¹ See, for instance, National Audit Office (2015).

payment upfront, rather than being delayed and contingent on the achievement of a specified outcome. Another approach is to use SIB schemes.

In the SIB form of SOC, a **separate private investor** pays for the activities of a social service provider, thereby taking the financial responsibility for tackling a social issue (see the figure below). For the commissioner, this provides a means to harness private resources through joint delivery, and to shift the risk onto an investor in the event that the intervention fails. For the social service provider, SIB provides upfront capital, shifting the financial risk onto the investor. An independent evaluator measures target group outcomes periodically, and reports on the SIB's progress and final results. Based on the final results of the SIB, which are predefined in a contract signed by all stakeholders, the commissioner (e.g. government or local authority) pays part of its cost-savings to the investor³², which should compensate for the upfront capital plus interest.

Figure 1. Basic model of social impact bonds



Private funding for SIBs can be issued in two main ways: through **SIB funds** and **individual SIBs**. The main difference is that SIB funds issue multiple contracts focusing on the same social issue, whereas individual SIBs release one payment contract at a time. Furthermore, individual SIBs take one of the following forms:

- In a direct SIB, a delivery contract is signed between the outcomes-payer, the investor and the service provider. The investor carries the risk if the required results are not achieved.
- Intermediated SIBs also involve an intermediary to liaise between investors, service providers and the commissioner. Social Finance Global Network, a multinational non-governmental organisation (NGO) mentioned above, is one of the most notable actors playing this role in a number of countries. In an intermediated SIB, the delivery contract is signed between the outcomes payer and an investor-owned

³² Van Es, B. C., Houben, O. J., & à Nijeholt, M. P. L. (2016). Social Impact Bonds: an Innovative Way to Combat Youth Unemployment. *Journal of Positive Management*, 7(1), 45. Retrieved from: <https://apcz.umk.pl/czasopisma/index.php/IPM/article/view/IPM.2016.003>

special purpose vehicle (SPV), which contracts the service provider, supports the performance management process and specifies the financial model. Risk is shared among investors within the SPV, and repayment depends on the SIB's structure.

- Finally, a managed SIB is also signed between the outcomes-payer and the intermediary, but the intermediary usually manages the entire process.³³

At least theoretically, SIBs offer opportunities to all the parties involved. The government shares the financial risk of setting up a social intervention with private investors (an issue we explore in depth in Section 6.2). 'No cure, no pay' ensures that governments only pay for proven results, which also clarifies the effectiveness of the intervention. Moreover, different departments, agencies and arms of government can use this instrument to pool funds in cases where interventions address issues that cut across different mandates, and to achieve outcomes that are of interest to more than one departmental sponsor. The instrument is not limited to annual budgets, so can be used for multi-year interventions, avoiding the need for repeated funding applications and decisions that also pose a risk to the continuity of the intervention³⁴. Meanwhile, private investors gain the chance to pursue corporate social responsibility, and may obtain returns on social investments. Finally, social service providers are offered a long-term investment that may be used to design, test and implement their innovative interventions on a larger scale – in other words, they are provided with better environment for scaling up.

In general, supporters of SIBs praise them for pulling together diverse actors and expertise from different domains, fostering cooperation and innovation, and promoting a culture of performance measurement. Consequently, this model is becoming increasingly popular around the world (see the figure below). For example, as of July 2020, the Brookings Institution has mapped a total of 194 impact bonds in 33 countries³⁵ — marking an increase of around 45% in 18 months (as of January 2019, Brookings found a total of 134 SIBs³⁶). Although most of these are in the UK, the US and other high-income countries, the model is also emerging in developing regions.

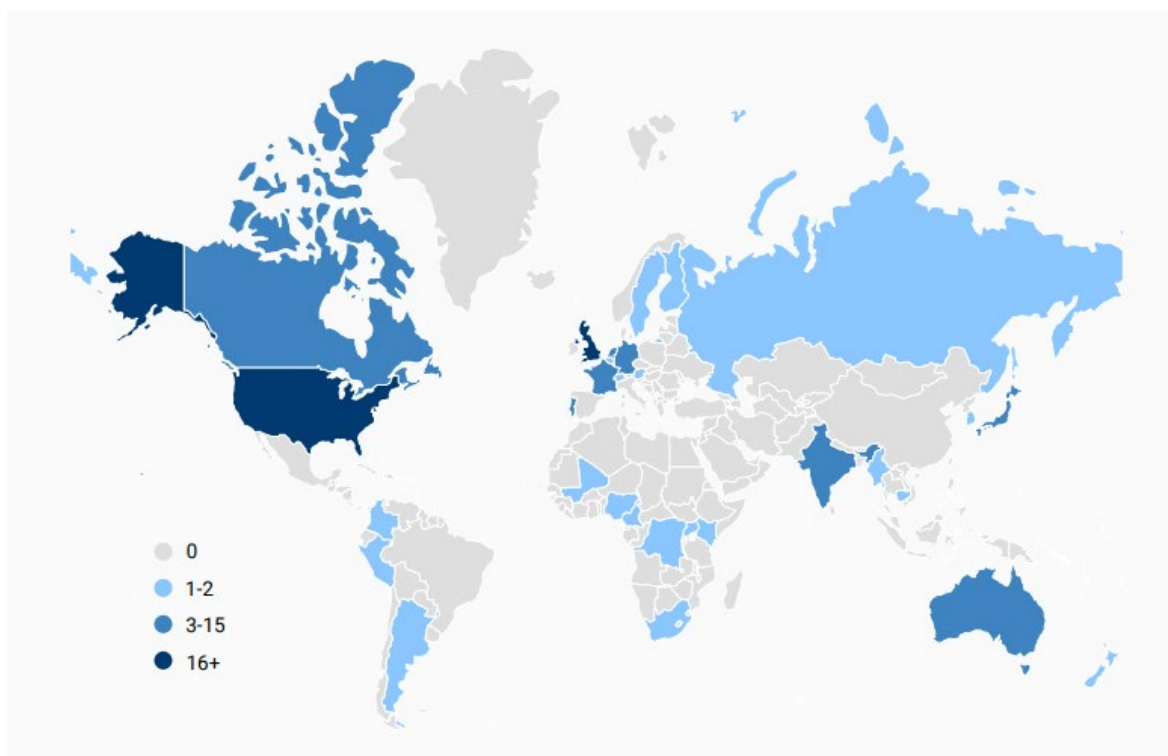
³³OECD. (2016). Understanding Social Impact Bonds. OECD. Retrieved from: <http://www.oecd.org/cfe/leed/UnderstandingSIBsLux-WorkingPaper.pdf>.

³⁴ Marks, M. B. and Weaver P. M. (2017), Are Social Impact Bonds a Viable Resource for Social Innovations? A Brief Discussion Paper. TRANSIT Working Paper #13, July 2017

³⁵ Brookings. (2020). Social and development impact bonds by the numbers. Retrieved from: <https://www.brookings.edu/research/social-and-development-impact-bonds-by-the-numbers/>

³⁶ Global Economy and Development at Brookings. (2019). Brookings Impact Bonds Snapshot – 1 January 2019. Retrieved from <https://www.brookings.edu/wp-content/uploads/2019/01/Impact-Bonds-Snapshot-January-2019.pdf>

Figure 2. SIBs around the world



Source: Brookings Institution Global Impact Bond Database, July 2020.

Nonetheless, SIBs are also often related in the literature to various risks such as technical issues, considerable administrative burden, and transactional costs (sometimes even outweighing the possible government savings), as well as perverse incentives (parking³⁷, creaming, cherry-picking³⁸, etc.). Other critics emphasise the normative nature of some pro-SOC arguments, arguing that SIBs reduce the central features of social interventions to a by-product of investment by turning citizens into commodities³⁹.

2. Study process and methods

The following chapter first summarises the methods used to compare SOC and TF programmes, following which we present the cases selected for analysis. We also outline the limitations regarding their comparability.

2.1. Methodology

To fully answer the research questions outlined in the Introduction, the study followed a mixed methods approach, relying on extensive desk research and literature review; interviews with relevant stakeholders; content analysis; and quantitative methods such as

³⁷ Parking refers to a process, by which providers try to keep costs down by doing little to serve those with the poorest anticipated outcomes, while instead focusing resources on more able clients with better employment prospects.

³⁸ Creaming and cherry picking both refer to the selection of the easiest-to-help participants, in order to ensure that providers can achieve the required outcomes, and providers and/or investors receive payments.

³⁹ Roy, Michael J., Neil McHugh & Stephen Sinclair (2018). A Critical Reflection on Social Impact Bonds. Stanford Social Innovation Review. Retrieved from: https://ssir.org/articles/entry/a_critical_reflection_on_social_impact_bonds

descriptive statistics and cost-effectiveness analysis. Details regarding each stage of the research process, together with the limitations of the study, are presented in Annex 2. Below, we briefly summarise the methodology.

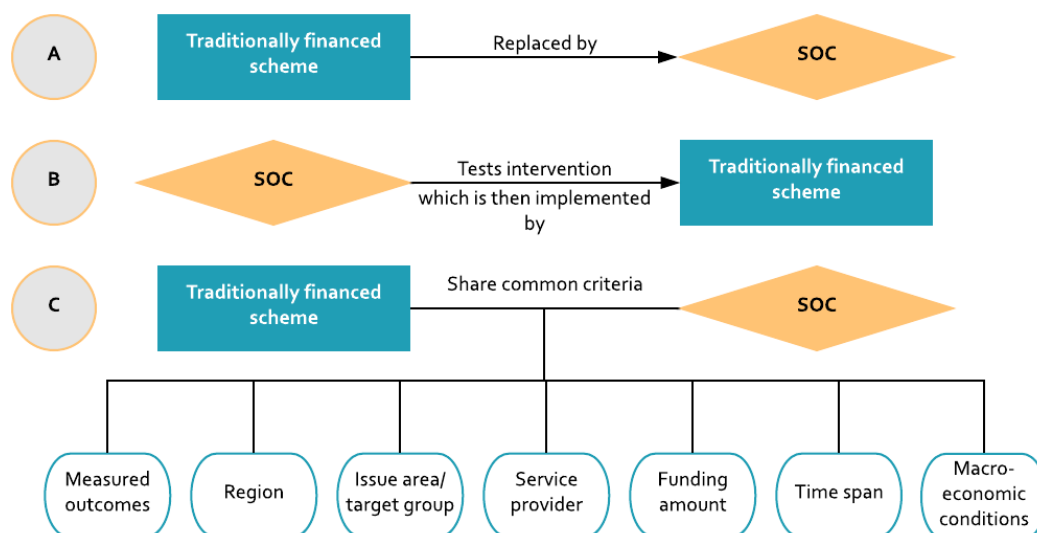
We began our research by compiling a long list of 64 diverse SOC schemes and models completed worldwide, presented in Annex 3. We then reviewed these SOC operations to identify the schemes for which we could find evaluations. From this narrowed list of SOC programmes, we selected 15 cases for in-depth study and comparison with traditionally financed programmes. Selection of the 15 cases was carried out on the basis of the availability of data, the presence of control groups in the evaluations reviewed in order to identify comparable TF programmes, geographic diversity, model diversity (various types of SIBs and PbR schemes), as well as reflecting a diversity in the types of social services considered.

Given that the programmes had to be completed and evaluated in order to be included in the study, a number of SIBs selected had been the first to be implemented in their respective countries. This is important, because any lessons learned about programme design may have been incorporated into subsequent SIBs commissioned by the same authority, thus lowering the set-up costs of such programmes. Nevertheless, little research exists to assess these potential savings.

For each shortlisted SOC scheme, we utilised one of three options to find a comparable TF to pair it with (see the figure below):

- A. equivalent TF social interventions (in terms of goals and activities) that were in place before the introduction of a particular SOC scheme;
- B. TF social interventions that substituted successful SOC schemes;
- C. a broader approach to match SOC and TF schemes implemented in similar contexts on the basis of several similarity criteria, as illustrated in the figure below.

Figure 3. Different options for matching SOC and TF schemes



Source: PPMI.

It is important to note that these comparisons are subject to a number of limitations. For example, in some cases the service delivered in the SOC and TF pairs was the same, but it was delivered by different service providers. In other cases, outcomes were tracked differently, or the TF scheme lacked an evaluation, making it difficult to compare its overall effectiveness with that of a similar SOC scheme. In some of our SOC-TF pairs, target groups were similar, but not exactly the same. Due to these limitations, we are cautious

about arguing that the differences observed in terms of programme performance or costs are solely *because of* their funding mechanism, as other factors might also have influenced their effectiveness and efficiency. The main limitations regarding the comparison of each SOC scheme with its paired TF are outlined in Table 3 and Table 9, and are discussed in greater depth under the analysis of each scheme in Annex 1.

In total, we conducted **54** interviews. As the table below illustrates, we conducted most of these interviews with commissioners, service providers and evaluators. This is because these three stakeholder groups are common to both SIBs and PbR schemes. Unfortunately, we were unable to conduct interviews with every stakeholder group from every scheme. Despite reaching out to interviewees multiple times, conducting fieldwork during the COVID-19 pandemic meant that some stakeholders – commissioners and service providers in particular – were not available. This was particularly true for stakeholders who had changed jobs since they had worked on the selected programmes. Where interviews could not be conducted with such stakeholders, we contacted other experts. These included representatives from the same commissioning authority who were not directly involved in the selected SOC programme but who were familiar with the case, as well as scholars who had studied the selected case.

With regard to desk research, we reviewed the evaluations and audit reports on the selected SOC and TF programmes that had been published before September 2020. We also examined the results of congressional hearings regarding specific SOC schemes and academic articles that focused either on the selected cases, or on specific concepts explored in this study. In total, 273 sources were reviewed.

Table 1. Number of interviews completed with different stakeholder types

Stakeholder Type	Number of Interviews
Commissioner	14
Service provider	14
Evaluator	13
Investor	6
Intermediary	4
Expert	3
Total	54

2.2. The programmes analysed

2.2.1. SOC schemes

The 15 selected schemes analysed in the study were geographically diverse, representing the UK, the Netherlands, Germany, Austria, Belgium, Ireland, the US, and Australia. They were also split almost evenly between SIB (8) and PbR models (7) (see the table below). For an explanation of the reasons why each case was selected for in-depth study, please see Table 69 in Annex 2. Table 2 overleaf provides a summary of the programmes and their objectives. Please note that detailed outcome targets are presented in Chapter 4.

With regard to the areas of intervention covered, six schemes addressed labour market-related issues. Five dealt with integrated services, meaning that areas of intervention overlapped. For example, they combined crisis and emergency (Perspektive:Arbeit) or disability (Provider-led Pathways) services with interventions relating to the labour market. Three schemes focused on social exclusion, and one on caring obligations.

The reasons why various SOC models were chosen to fund these programmes (as specified in the evaluation reports reviewed, as well as interviews with commissioners) were quite varied. SIB commissioners were often interested in testing out new solutions to existing problems. Furthermore, the prospect of acquiring additional funding to finance social interventions was noted in half of the selected SIBs. PbR commissioners were also interested in service innovation, as well as delivering services more efficiently. Other reasons included:

- the desire to test out outcomes-based contracting;
- to assess whether an intervention is cost-effective prior to funding it traditionally on a larger scale;
- involving private and third-sector organisations in the delivery of social services; and
- utilising the flexibility that is inherent in the private sector compared with the public sector, meaning that, for example, more employment advisers can be hired in times of high unemployment, and fewer when unemployment is low.

According to the definitions used in this study, PbR schemes do not involve an investor, while SIBs do. In half of the SIBs selected, these investors were philanthropic organisations; the other half were a mixture of philanthropic and private investors. Notably, not a single SIB selected for the study relied solely on for-profit investors, probably illustrating that SIBs carry a greater risk for investors than other types of investment, in comparison with the returns they offer. This risk depends in large part on how investors are repaid. Six of the 15 schemes used binary-only repayment models⁴⁰ (see Chapter 2 and the table below) and four rewarded investors on the basis of frequency-only outcomes. One of the selected schemes employed a hybrid remuneration model, in that bonus payments were granted for results achieved beyond expectation. The latter case belonged to a group of five programmes that combined multiple repayment models. This occurred when schemes had different repayment arrangements attached to different outcomes measured.

⁴⁰ In binary models, the provider has to achieve an absolute target, and no payment is granted for achieving lesser results. In frequency schemes, rewards are staggered according to an agreed frequency of results, with payments increasing as results increase (see Chapter 1).

Table 2. SOC schemes analysed in the study

SOC	Type	Country	Social Service Area	Payment model	Objective
Eleven Augsburg	SIB	Germany	Labour market-related	Binary	Jobs/apprenticeships for difficult-to-reach youth
Perspektive:Arbeit	SIB	Austria	Crisis and emergency Labour market-related	Binary	Sustained employment of survivors of domestic violence
DUO for a JOB	SIB	Belgium	Labour market-related	Binary	Employment of non-EU immigrant youth
Buzinezzclub	SIB	Netherlands	Labour market-related	Frequency	Reducing the duration of unemployment benefit claims among the youth
BOAS Werkt	SIB	Netherlands	Labour market-related	Binary	Employment in Germany of those unemployed in the Netherlands
ABLE Programme	SIB	USA	Social exclusion	Frequency	Reducing recidivism among 16 to 18-year-olds
Benevolent Society	SIB	Australia	Caring obligations	Frequency	Reducing entries into out-of-home care by children
Mental Health and Employment Partnership	SIB	UK	Disabilities Labour market-related	Binary & Frequency	Sustainable employment of people with mental health issues
JobPath	PbR	Ireland	Labour market-related	Frequency	Sustainable employment of the long-term unemployed
Provider-led (PL) Pathways to Work	PbR	UK	Disabilities Labour market-related	Binary	Sustainable employment of those claiming unemployment benefits due to health reasons
Work Programme	PbR	UK	Labour market-related Disabilities Social exclusion	Binary, Frequency & Hybrid	Sustainable employment of various groups of the unemployed
Youth Contract (YC) (16 to 17-year-olds)	PbR	UK	Labour market-related	Binary	Jobs/apprenticeships/education for difficult-to-reach youth
Drug and Alcohol Recovery Pilots	PbR	UK	Social exclusion	Binary & Frequency	Reduction in the misuse of drugs and alcohol and reintegration into community
Transforming Rehabilitation	PbR	UK	Social exclusion	Binary & Frequency	Reduction in re-offending for short sentence offenders

SOC	Type	Country	Social Service Area	Payment model	Objective
Troubled Families (phase II)	PbR	UK	Social exclusion Specific problems	Binary & Frequency	Reducing crime, anti-social behaviour, truancy, unemployment, mental health problems and domestic abuse among target families

The selected schemes also represent a diverse set of service providers, including non-profits, for-profit companies, public authorities and a mixture of the three. Delivery by non-profit organisations was more prevalent in the SIB schemes selected. This is probably because these were mostly on a smaller scale. The PbR schemes selected also involved the third sector, but more often in combination with delivery by private-sector organisations. Outcomes-based public sector delivery was implemented in cases when central government issued funding to local authorities, as was the case in the Drug and Alcohol Pilots, the Youth Contract and Troubled Families.

On the basis of the schemes analysed, we also distinguish between direct and intermediary beneficiaries. In two of the schemes (Youth Contract and Mental Health and Employment SIB), payments were issued to employers as an incentive to hire programme participants (direct beneficiaries). While employers were not the target groups for the intervention, they nevertheless benefitted from the programme.

For detailed descriptions of each SOC scheme, see Annex 1. The two boxes below provide a brief summary of the DUO for a JOB programme as an illustrative example of how a SIB was set up, as well as a summary of the Transforming Rehabilitation (TR) intervention to illustrate a PbR scheme. Please note that while the SIB in these examples was successful while the PbR scheme was not, this is not representative of all SIBs and PbR schemes: among the cases analysed, we encountered both success and failure to achieve targets in types of programme.

Example of a social impact bond scheme: DUO for a JOB SIB

DUO for a JOB (or DUO in short) is a non-profit organisation created in Belgium in November 2012, which provides an intergenerational mentorship programme. The programme pairs young immigrants with volunteer Belgian mentors aged 50 and older for a period of six months, with the aim of helping immigrant youth find jobs. Young immigrants receive insights into life and work in Belgium from people with a deep knowledge about the country.

Between 2014 and 2016, the mentorship programme was partly funded through a SIB. The SIB emerged in response to high gap in labour market participation between those born within the country and non-European immigrants, as well as a high youth unemployment rate in general⁴¹. Actiris, the agency for employment, while acknowledging the absence of a specific employment policy targeting non-EU youth in Brussels, also recognised budgetary pressures. It therefore sought alternative ways to fund targeted employment policies. Among other organisations, DUO was invited to apply for SIB funding – and subsequently acquired it. Other parties involved in the project included an undisclosed foundation which served as an investor; KOIS (an intermediary); and two evaluators: Observatoire Bruxellois de l'Emploi (a public agency, part of Actiris) and the Centre for Social Economy at the University of Liège.

The main objective of the SIB was to achieve a 10% higher relative rate of employment among the DUO for a JOB group than the control group. The control group consisted of migrants registered at Actiris who had similar demographic characteristics, but who did not enrol in the DUO for a JOB programme. The employment rate was measured one year after the end of the intervention.⁴²

In each of the years during which the programme was implemented as a SIB, the objective was achieved. In total, 322 mentee-mentor duos were established, which resulted in 133 job placements for DUO's mentees. For each year the objective was

⁴¹ KOIS (n.d.). DUO for a JOB: the first Social Impact Bond in continental Europe. Retrieved from: <https://www.koisinvest.com/post/duo-for-a-job-the-first-social-impact-bond>

⁴² Dermine, T., Le Grelle, M., & Simonart, F. (2016). Social Impact Bonding. *Stanford Social Innovation Review*, 2016 (Fall), p. 62.

fulfilled, investors earned a 4% interest on their investment. The SIB represents a binary repayment model, because the investors' reward depends on whether or not 10% more participants find jobs, compared with the control group.

Example of a payment by results scheme: Transforming Rehabilitation

In 2014, the UK government's Transforming Rehabilitation reform created 21 privately owned Community Rehabilitation Companies, or CRCs, as well as the public sector National Probation Service (NPS). The CRCs managed offenders who presented a low or medium risk of serious harm in the community. CRCs were requested to deliver innovative rehabilitative support and mentoring to offenders. They were required to provide services concerning education, substance misuse, housing and employment, but were given a lot of freedom over the means used to drive down reoffending rates.⁴³ Meanwhile, offenders posing a high risk of serious harm to the public were directly managed by the NPS.

Payments to the CRCs included both fees for the services provided and a PbR component. PbR represented around 10% of the total predicted payments to the CRCs. The fee-for-the service component primarily covered the CRCs' operating costs for the mandated activities (e.g. delivering the sentence of the court and licence conditions, resettlement services, and the mandatory activity days required by the court).

Two reoffending measures were used to assess the performance of the CRCs (and the NPS), and to determine the outcomes-based payment component:

- the binary rate: proportion of offenders who reoffend;
- the frequency rate: the average number of re-offenses per reoffender.

While reoffending has decreased overall, the CRCs have not achieved the targets set by the Ministry of Justice. The Ministry expected CRCs to reduce reoffending by 3.7 percentage points over the life of the contracts. However, by March 2017, mid-way through the reform, there was an overall 2.5 percentage point reduction in the proportion of proven reoffenders since 2011. Between 2011 and March 2017, there was a 22% overall increase in the average number of re-offenses per reoffender, and just six out of 21 CRCs consistently achieved their targets for reducing reoffending.⁴⁴ As a result, CRC activities were to be terminated in December 2020, switching back to the traditionally financed model in place prior to the reform.⁴⁵

2.2.2. Traditionally financed equivalents

The TF programmes chosen for comparison with SOC schemes are summarised in the table below, which also includes information on whether their key characteristics are the same, similar to or different from the comparable SOC's. The 'match method' column indicates the method used to identify TF programmes, as outlined in Figure 3.

⁴³ User Voice (2016). Transforming Rehabilitation: The operational model from the service user's perspective. Retrieved from: <https://www.nao.org.uk/wp-content/uploads/2016/04/User-Voice-report-Transforming-rehabilitation-The-operational-model-from-the-service-users-perspective.pdf>

⁴⁴ MoJ (2020). Final and interim proven reoffending statistics for Community Rehabilitation Companies and the National Probation Service. January to March 2018 and 2017/18 annual cohort. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/882276/crc-nps-final-results-bulletin-apr20.pdf; NAO (2019). Transforming Rehabilitation: Progress Review – Transcript. Retrieved from: <https://www.nao.org.uk/wp-content/uploads/2019/02/Transforming-Rehabilitation-Progress-review.pdf>

⁴⁵ NAO (2019).

Most of the SOC programmes targeted the same or a similar population of beneficiaries, and implemented similar activities in comparable locations to their TF equivalents. An example of similar target populations would be the Youth Contract and Activity Agreement Pilots. But while both programmes targeted 16-17 year-old youth not in education, employment or training (NEETs), Youth Contract focused on the most difficult cases.

More differences emerge with regard to the commissioners of the SOC or TF schemes: five out of 15 pairs were commissioned by different commissioning bodies. Providers were also different in half of the programmes; however, this is to be expected given that we are comparing SOC and TF models and, as discussed in Chapter 7, the two models are suited to different types of providers. Finally, information on the annualised number of participants and annualised costs was not available for some of the schemes selected. Where information was available, the two indicators appear different in most cases. We took this into account in Chapter 5 by calculating the cost per participant.

The TF programmes are described together with SOC programmes in Annex 1. Meanwhile, in the boxes beneath Table 3 below, we provide brief summaries of the DUO for a JOB programme after it was funded traditionally following the discontinuation of the SIB, as well as the Probation Trusts – the TF equivalent of the Transforming Rehabilitation programme, introduced in Section 2.2.1.

Table 3. SOC schemes analysed in the study

SOC	TF	Match Method	Target Population	Activities	Location	Commissioner	Provider	Annualised No. of Participants	Annualised Cost
Eleven Augsburg	AMA-Zukunft	C	Similar	Similar	Same	Different	Similar	Different	Different
Perspektive:Arbeit	Perspektive:Arbeit	B	Same	Same	Similar	Different	Similar	N/A	Same
DUO for a JOB	DUO for a JOB	B	Same	Same	Similar	Similar	Same	N/A	Different
Buzinezzclub	Buzinezzclub	B	Same	Same	Same	Same	Same	N/A	N/A
BOAS Werkt	Transfer Point	C	Same	Different	Similar	Different	Different	Different	Different
ABLE Programme	RESTART	C	Similar	Similar	Similar	Different	Different	Different	Different
Benevolent Society	Intensive Family Preservation	C	Similar	Same	Same	Same	Different	Same	Different
Mental Health and Employment Partnership	Camden Individual Placement Support	A	Same	Same	Similar	Different	Same	Different	Different
JobPath	Intreo Local Employment Services	C	Different Similar	Similar Similar	Same Same	Same Same	Different Different	N/A Similar	Similar Different
PL Pathways to Work	Jobcentre Plus Pathways	C	Same	Similar	Different	Same	Different	Similar	Similar
Work Programme	NDYP/ND25pl	C	Similar	Similar	Same	Same	Different	Different	N/A
Youth Contract (16-17 year olds)	Activity Agreement Pilots	A	Similar	Similar	Similar	Similar	Different	Different	Same
Drug and Alcohol Recovery Pilots	Lincolnshire Drug and Alcohol Programme	A	Same	Same	Same	Same	Same	Same	N/A
Transforming Rehabilitation	Probation Trusts	A	Similar	Similar	Same	Same	Different	Different	Different

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

SOC	TF	Match Method	Target Population	Activities	Location	Commissioner	Provider	Annualised No. of Participants	Annualised Cost
Troubled Families (phase II)	Earned Autonomy Model	B	Same	Same	Same	Same	Same	Same	N/A

Traditional Financing provided to DUO for a JOB after the SIB

Following the SIB, DUO was allocated a first ‘bridging’ subsidy for two years (2017-2019) by Actiris on the basis of the preliminary results of the SIB. This was done because DUO had delivered good results and the organisation had to wait for two years to apply to the quinquennial call for tenders (for traditional procurements). In 2019, DUO acquired funding from Actiris for an additional five years (until 2024) after a regular tendering procedure⁴⁶. Therefore, in this study we compare the time periods in which the DUO mentorship programme was implemented as a SIB (2014-2016) vs using TF (2017 onwards), reflecting the different funding streams it received.

It is important to note that after the bridging subsidy, the project was scaled up. During the SIB, DUO for a JOB was granted relatively little funding because it was the commissioner’s first experience with a SIB. Please note that during both time periods (2014-2016 and 2017 onwards), DUO for a JOB also received additional funding from donations, grants, and other sources. Using these funds, DUO has expanded to Antwerp and Liège.

So far, the TF intervention has shown better results in terms of job starts than the SIB (55% of mentees have found a job, compared with an average of 40% during the SIB). However, it is also important to note that lessons learned during the SIB were incorporated into the TF programme.

Traditionally financed Probation Trusts

The Transforming Rehabilitation reform, which included a PbR component, was phased in between 2014 and 2015 to change the scope and structure of community and prison-based probation and rehabilitative services. It is therefore possible to compare the intervention that was in place before the reform was implemented with the services delivered under the Transforming Rehabilitation reform implemented since 2015.

Prior to June 2014 when the Community Rehabilitation Companies (CRCs) were introduced, probation services in England and Wales were delivered by 35 self-governing Probation Trusts, working under the direction of the National Offender Management Service.⁴⁷ Probation trusts worked with offenders serving a community-based sentence and offenders who had released from custody.⁴⁸ However, the target population was somewhat different between the two programmes: Probation Trusts targeted prisoners released from custody after more than 12 months, whereas CRCs targeted those who had served less than 12 months. This is because, prior to Transforming Rehabilitation, similar services were not offered to offenders serving short custodial sentences.

The services provided by the Probation Trusts were considered effective by both Her Majesty’s (HM) Inspector for Probation and HM Prison and Probation Service.

⁴⁶ Personal communication with the service provider (2020).

⁴⁷ MoJ (2013). Transforming Rehabilitation: A revolution in the way we manage offenders, Cm 8517. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/228580/8517.pdf; House of Commons Justice Committee (2013). Crime reduction policies: a coordinated approach? Interim report on the Government’s Transforming Rehabilitation programme. Twelfth Report of Session 2013–14. Retrieved from: <https://publications.parliament.uk/pa/cm201719/cmselect/cmjust/482/48205.htm>

⁴⁸ NAO (2014). Probation: landscape review. Retrieved from: <https://www.nao.org.uk/wp-content/uploads/2014/03/Probation-landscape-review.pdf>

3. Outcomes measurement

The evaluation of outcomes is one of the key characteristics of SOC schemes (see Chapter 2). For this reason Chapter 3 of this report maps and reviews the outcomes measurement methods used in the 15 SOC schemes. To this end, we first provide an overview of the ways in which outcomes were defined and data collected in the schemes analysed. Following this, we review the outcomes measurement methods used. Our ultimate goal is to identify which methods for monitoring, measuring and evaluating outcomes in SOC schemes are more appropriate to establish the causality of impact. Our analysis is based on the data reported under each scheme analysed. More detail on each scheme is presented in Annex 1. A summary of the findings is presented at the end of this chapter.

3.1. Definition of outcomes and data management

In this section, we discuss the insights that emerge from our analysis of the measurement methods and outcomes identified in the various SOC schemes, as well as the features of the evaluation processes analysed. In particular, we focus on the definition of outcomes, including a summary of all indicators and metrics used in the SOC schemes selected for analysis. Furthermore, we detail the enabling factors and barriers to data collection and management in the SOC schemes.

3.1.1. Definition of outcomes and metrics

In the SOC schemes analysed, outcomes were operationalised using a variety of outcome indicators, depending on the issues the scheme was intended to address. A full list of outcomes and measures is included in Table 4 at the end of this section. Some general trends in the measurement of success could be identified according to service area:

- In relation to employment, the most frequently recurring outcomes measured were job entry and job sustainment. Job entry was measured either by the number of people getting a job or moving off benefits; job sustainment was measured on the basis of the number of people remaining in employment for a set amount of time. This varied from a few weeks to a year in different schemes. Both measures – job entry and job sustainment – were considered successful on the basis of the specific target set in each contract. A special employment subcategory included the unemployment of young NEETs. In those cases, outcomes related to the number of young people participating in full-time education, part-time education or full-time employment with a part-time training equivalent, and then sustained engagement for five or six months.
- In the case of drug and alcohol addiction, measures related to the number of patients who reduced their consumption by statistically significant levels or were abstinent at any two reviews within the preceding 12 months.
- Schemes dealing with offending/reoffending considered either the percentage of offenders who re-offended during a specific period of time, or a reduction in the number of offences per offender over a specific period of time. The interventions were deemed successful when there was a reduction in the two measures compared with either a historical baseline or a national/regional average. In these schemes, accompanying evaluations focused on the quality of the service assessed in terms of the helpfulness of the relationship with the keyworker, and the ability to assess and manage the risk of reoffending.

- Various schemes also included assessments of the health and wellbeing of the beneficiaries. Several specific measures were considered, such as the number of patients completing a course of vaccinations, the number of people with a specific level of 'quality of life' score, beneficiaries' perceptions of life satisfaction or health status. Two schemes focused on the wellbeing and safety of families and their children. The outcome linked to payments in these cases related to a reduction in contacts with the child protection system, and was measured using various metrics such as the number of helpline reports, safety and risk assessments and entries into out-of-home care. Other measures assessed, but not linked to payments, were changes in children's cognitive development, social competence and emotional regulation, satisfaction with parenting, the presence of secure and stable relationships within the family, the financial stability of the family, school attendance and anti-social behaviour.

The analysis revealed that, in general, reaching **agreement among stakeholders as to the definition of outcomes and metrics** was quite difficult. The misalignment of viewpoints between commissioners and providers in the choice of evaluated dimensions was identified in many schemes.

In particular, issues emerged concerning the balance between a **one-dimensional success criterion**, which was an appealing option for commissioners, and the use of **soft outcomes**, which many service providers considered more suited to their way of working. For instance, in the Work Programme, some providers felt that they had put a great deal of effort and resources into making significant improvements to the employability of their participants by building the participants' self-confidence, practising interviewing skills, etc. However, providers argued that these efforts were not adequately rewarded by a model in which payments were only issued if participants found and sustained jobs. In the Mental Health and Employment Partnership (MHEP) SIB, the measures included by the commissioner were different from the standard measures previously used by the service providers.

Importantly, **additional evaluations** were often carried out to track outcomes that **were not payment-related**. These typically focused on softer milestones for programme participants. For example, in Austria's Perspektive:Arbeit SIB, payments were only issued if women found jobs that had particular characteristics. However, the accompanying evaluation also touched on outcomes such as the increased self-esteem of programme participants, absence of violence, self-reported improvements in children's mental states or improved school achievements, better health conditions, and increased mobility in terms of passing a driving test or purchasing a car. The main two payment criteria in the PL Pathways scheme were finding a job and sustaining it for 26 weeks – but a number of accompanying evaluations also assessed softer factors such as whether participants who remained unemployed at the end of the programme were thinking about work and taking steps towards it. In both Perspektive:Arbeit SIB and PL Pathways, however, the measures assessed in the accompanying evaluations were not included in the schemes' payment structures.

According to commissioners, focusing on **hard outcomes** (such as the number of jobs found) as success criteria, rather than softer outcomes, brought greater **objectivity** to the payments issued and simplified the process of tracking outcomes. However, when outcomes were defined as **binary metrics** (yes/no indicators), service providers were often subject to **perverse incentives**, resulting in **cherry picking**. In such cases, service providers tended to focus on users who were more likely to achieve a specific outcome (not reoffending, getting a job, ceasing their substance abuse). For instance, in Transforming Rehabilitation, binary metrics created a perverse incentive for the provider to focus on those offenders who were least likely to reoffend. The commissioner later considered including soft outcomes to assess the performance of providers, such as the number of offenders who obtained access to housing and universal credit, as well as successful employment or mental health programme outcomes. Another example comes from the Work Programme.

Because the issue of perverse incentives had already been observed early on in PL Pathways, the subsequent Work Programme attempted to counteract this by providing additional payments for outcomes achieved by participants who were deemed harder-to-help. It is questionable whether this approach was very effective: some service providers noted that those who were hardest-to-help often did not manage to meet even the minimum outcome thresholds required to trigger payments, so the disincentive to work with them remained.

To avoid cherry picking and perverse incentives in general, it is important to track outcomes for different **segments of participants** (e.g. those who are easier-to-help and harder-to-help). This is helpful in understanding whether the intervention is more effective for some segments than others, and informs commissioners on how to design future programmes. Some effective examples encountered during the study include the Work Programme, JobPath and Drug and Alcohol recovery pilots. In the last of these schemes, for instance, the level of help that would be required for each beneficiary was assessed locally by LASARS (Local Area Single Assessment and Referral Systems), and outcomes were assigned based on a national complexity tool.

However, measuring outcomes for different segments of participants is not easy in practice. First, it requires a large number of participants in order to identify differences between the segments with sufficient statistical power. This might be counterintuitive in some SIBs, which are designed to test out new interventions on a small scale before rolling them out as traditionally financed models. Second, various stakeholders highlighted disagreements over how to define harder-to-help groups, while service providers emphasised the added burden of reporting various outcomes for different groups of participants.

Another issue relating to the definition of metrics is that it is difficult to define metrics that can account **for long-term impact** rather than outcomes. This could potentially be overcome by designing a dashboard of KPIs, or using a more longitudinal perspective with initial, interim and final outcomes rather than a single outcome. On the flip side, in SIBs, interim outcomes reduce the risk for investors, while increasing the risk for the commissioners.

To deal with challenges related to the definition of outcomes and metrics, some schemes implemented **a co-design phase**, which helped to align the interests and expertise of services providers and commissioners. For instance, in the Drug and Alcohol Recovery Pilots, the co-design phase allowed service providers to push for interim outcomes and in-treatment indicators. In the DUO for a JOB SIB, the timespan used for measurement (a year after the last potential exit day of the cohort) was opposed by investors, who would have preferred to find out the outcome of the intervention – and thus be repaid – sooner. Despite this initial opposition, the outcome was still measured one year after the intervention on the initiative of the service provider, DUO, in order to detect non immediate effects of the programme. While codesigning the scheme, DUO also ensured that the programme would not go against its mission by limiting services to a particular group of people.

A comparative overview of all outcomes and indicators applied in the evaluations of the 15 schemes reviewed — **including those that were linked to the payment mechanisms, and those that were not** — is presented in Table 4. In the table, we report all the effects that have been included in the schemes as ‘outcomes’ (column ‘Measured outcomes’), and we analyse the indicators used to measure these outcomes (column ‘Indicator/Metrics’). One interesting result is that some of the indicators used to assess outcomes were designed as output metrics, while others measured effects that were related to a broader impact of the intervention (these differences are highlighted in the column ‘Dimension’).

Table 4. Review of indicators and metrics

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
EMPLOYMENT				
Employment of women subject to domestic violence	Number of women from the target group placed in a job that has the following characteristics: is subject to social insurance contributions; pays a living wage (i.e. EUR 19,500 gross salary), at least 20 hours per week; for at least 12 months during the term of the project	OUTCOME	Yes	Perspektive:Arbeit
Employment of hard-to-reach disadvantaged youths (younger than 25 years old, NEET, no school-leaving qualification, no employment)	Number of young people in work or apprenticeship for at least nine months	OUTCOME	Yes	Eleven Augsburg
	Average number of days that the unemployed youth tend to receive benefits	OUTCOME	Yes	Buzinezzclub
Paid employment for the unemployed	Job sustainment:			
	Sustained employment for 13, 26, 39 and 52 weeks	OUTCOME	Yes	JobPath
	Sustained employment for 26 weeks	OUTCOME	Yes	Pathways to Work
	- Sustained employment for three or six months, depending on the target group - Various additional sustained employment thresholds for different target groups	OUTCOME	Yes	Work Programme
	- Employment measured one year after the end of the intervention. The indicator is proxied by a binary variable that is equal to 1 if the individual has been employed for at least 90 working days in the last 12 months	OUTCOME	Yes	DUO for a Job
	Job sustainment (six weeks/six months)	OUTCOME		MHEP SIB
	Job entry: - Number of benefit recipient who get a job - Number of job entries (>16 hours/week) (<16 hours/week) Movement off out-of-work benefits	OUTPUT	Yes	Pathways to Work Work Programme MHEP Troubled Families (phase II) BOAS Werkt
	Public cost savings:	IMPACT	No	BOAS Werkt

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
	<ul style="list-style-type: none"> - Number of days each person in the treatment and the control groups claimed unemployment benefits, including respective costs 			
	Level of earnings: <ul style="list-style-type: none"> - Amount of earnings from employment compared with the amount of money received in social welfare payments - Satisfaction of an earning threshold - Socio-economic classification of current work - Level of household income Household ability to keep up with bills and regular debt repayments over last two years	OUTCOME	No	JobPath Troubled Families (phase II) Pathways to Work
Participants' aspirations and motivation to find work, and effort put into finding work	Job applications and steps towards finding a job Active steps towards finding work in the last four weeks	OUTPUT	No	Work Programme Troubled Families (phase II)
	Perception of employment perspectives over the coming year	OUTCOME	No	Troubled Families (phase II)
EDUCATION & EMPLOYMENT				
Engagement of NEETS in education, training or employment	<ul style="list-style-type: none"> - Participation in full-time education or training, leading to an accredited qualification - Participation in part-time education, including re-engagement provision. Young people are required to participate in at least seven hours of directed learning per week - Participation in an apprenticeship - Participation in full-time employment (20 hours or more each week) with part-time training equivalent to at least 280 guided learning hours per year (around one day per week) - Sustained participation for at least five out of six months in full-time education or training, leading to an accredited qualification funded by the Education Funding Agency - Sustained participation for at least five out of six months in an apprenticeship - Participation for at least five out of six months in full-time employment with part-time training equivalent to at least 280 guided learning hours per year (around one day per week) 	OUTPUT	Yes	Youth Contract

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
DRUG AND ALCOHOL ADDICTION RECOVERY				
Improvement in drug and/or alcohol use Alcohol or drug addiction recovery	Number of patients who reduced their consumption by statistically significant levels for all presenting substances at any two reviews within the last 12 months	OUTCOME	Yes	Drug and Alcohol Recovery Pilots
	Number of patients who were abstinent from all presenting substances at any two reviews within the last 12 months	OUTCOME	Yes	Drug and Alcohol Recovery Pilots
	Number of patients who completed treatment	OUTPUT	Yes	Drug and Alcohol Recovery Pilots
	Number of unplanned discharges	OUTPUT	Yes	Drug and Alcohol Recovery Pilots
	Retention in treatment	OUTPUT	Yes	Drug and Alcohol Recovery Pilots
CRIME AND ANTI-SOCIAL BEHAVIOUR				
No proven offending	Number of patients with no proven offending in a six-month period, from the point of beginning a recovery intervention with a provider	OUTCOME	Yes	Drug and Alcohol Recovery Pilots
	Proportion of offenders who reoffend computed as a percentage of offenders within a quarterly cohort that are convicted of an offence within a 12-month period.	OUTCOME	Yes	Transforming Rehabilitation
	Contact(s) with the police in the last six months	OUTCOME		Troubled Families (phase II)
Reduction in average cohort offending	Percentage change in recidivism bed-days (change in the average number of days these adolescents spent in jail) following arrest on a new charge in the 12-month period following their initial release	OUTCOME	Yes	Drug and Alcohol Recovery Pilots Adolescent Behavioural Learning Experience SIB
Relationships with probation officers	Frequency of contacts with keyworkers	OUTPUT	No	Transforming Rehabilitation Troubled Families (phase II)
	Improved perception of the helpfulness and clarity of the keyworker Improved attitudes towards the help received from keyworkers	OUTCOME	No	Troubled Families (phase II)

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
Average number of re-offences per reoffender	The average number of re-offences per re-offender committed within an annual cohort over a 12-month period.	OUTCOME	Yes	Transforming Rehabilitation
Public is protected from harm	<p>Survey Questions:</p> <p>Was there sufficient assessment of the risk of harm posed to prisoners/staff/children/victims/public in general?</p> <p>Was there sufficient planning to manage and minimise the risk of harm posed to /staff/children/victims/public in general?</p> <p>Up to this point in the order/licence, has the responsible officer made sufficient progress in influencing the risk of harm posed by this service user to prisoners/staff/children/victims/public in general?</p>	IMPACT	No	Transforming Rehabilitation
Individuals abide by the sentence	<p>Survey Questions:</p> <p>Up to this point in the sentence, have probation services made sufficient progress in delivering the requirements of the order/licence?</p> <p>Has the service user sufficiently abided by the requirements in this order/licence?</p> <p>Were the contact levels sufficient for the needs of the case?</p> <p>Was the work undertaken likely to have a positive impact on reducing reoffending, public protection (when relevant) and abiding by the sentence?</p>	OUTCOME	No	Transforming Rehabilitation
Action taken to stop anti-social behaviour	Number of evictions, warning letters, possession orders	OUTPUT	No	Troubled Families (phase II)
HEALTH AND WELLBEING				
Injecting by those intravenous users at the start of treatment	Those who reported no days injecting on any two review Treatment Outcome Profiles (TOPs) within the last 12 months	OUTCOME	Yes	Drug and Alcohol Recovery Pilots
No fixed abode (NFA) / Housing problems of those NFA or with a housing problem at the start of treatment	Number of patients who no longer had any housing problems at any two reviews, where these were during the last 12 months or at their exit TOP ('no' to both housing questions).	OUTCOME	Yes	Drug and Alcohol Recovery Pilots
Hep B Vac of those eligible	Number of patients who had appropriately completed a course of Hepatitis B vaccinations within the previous 12 months.	OUTPUT	Yes	Drug and Alcohol Recovery Pilots

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
Wellbeing	Number of clients achieving a normative quality of life score in any two TOP reviews, where this was in the last 12 months Increased self-esteem Perception of life satisfaction Proportion of respondents feeling relaxed/optimistic/dealing well with problems	OUTCOME	No	Drug and Alcohol Recovery Pilots Perspektive:Arbeit Troubled Families (phase II)
Health	Mortality rate (drug-related poisoning) Mortality rate (non-drug-related poisoning)	OUTCOME/ IMPACT	No	Drug and Alcohol Recovery Pilots Troubled Families (phase II)
	Perception of health status	OUTCOME		Troubled Families (phase II)
FAMILY WELLBEING AND SAFETY				
Reduction in contact with the child protection system	Fewer helpline reports Fewer safety and risk assessments Fewer entries into out-of-home care	OUTCOME	Yes	Troubled Families (phase II) Benevolent Society SIB
Child's social-emotional wellbeing and emergent behavioural problems	Changes in children's cognitive development, social competence and emotional regulation based on indicators relying on: emotional symptoms scale, conduct problems, hyperactivity scale, peer problems scale, pro-social scale	OUTCOME	No	Benevolent Society SIB
Changes in families' protective factors	The protective factors are: knowledge of parenting, nurturing and attachment, family functioning, social support, concrete support	OUTCOME	No	Benevolent Society SIB
Parenting sense of competence	Satisfaction with parenting: parental frustration, anxiety and motivation Perceived self-efficacy: competence, problem-solving ability and capability in a parenting role	OUTCOME	No	Benevolent Society SIB
Parenting, family and relationships	Secure and stable relationship, increasing safety, increasing self-efficacy, improving coping/self-regulation	OUTCOME	No	Benevolent Society SIB
	Perception of happiness in family relationships	OUTCOME	No	Troubled Families (phase II)
	Proportion of carers not regretting marrying their partner Absence of non-sexual abuse by the partner No experience of partner abuse in the last six months	OUTCOME	No	Troubled Families (phase II)

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Result measured	Indicator/metrics	Dimension	Payment related	SOC scheme(s)
Increasing Safety	Personal Wellbeing Index: global measure of distress based on questions about anxiety and depressive symptoms Improvement in home physical environment Improvements in family resource management	OUTCOME	No	Benevolent Society SIB
	Proportion of families falling behind on rent payments Perception of family financial management Proportion of families having at least one type of credit or product loan Episodes of using force or violence in the home in the last six months Perception of feeling safe at home	OUTCOME	No	Troubled Families (phase II)
Children's behaviour at school (main carers called into school to talk about children's behaviour)	Episodes of "being in trouble" at school in the last six months	OUTCOME	No	Troubled Families (phase II)
Concerns about children attendance at school or college	Proportion of young people reporting unauthorised absence from school Perception of likelihood of being bullied	OUTCOME	No	Troubled Families (phase II)
Children's crime and anti-social behaviour	Proportion of children having an alcoholic drink Episodes of heavy drinking in the last four weeks Proportion of children smoking cigarettes Proportion of children who report having tried at least one street drug Contact with the police in the last six months Police action against children in the last six months Reported episodes of involvement in a crime in the last few months Proportion of children carrying a weapon Proportion of children receiving actions used to discourage anti-social behaviour in the last six months	OUTCOME	No	Troubled Families (phase II)

3.1.2. Data collection and management

Evaluations of most of the schemes combined the collection of primary and secondary data with both quantitative and qualitative analysis. These involved various challenges.

In all schemes in which the service provider was involved in or fully in charge of data collection and management, the whole process was perceived as a burden by the service providers. Some local authorities and providers pointed out that the development and management of data systems for PbR schemes had been a significant transaction cost. These issues concerned the volume of data required – particularly where service providers had to report to multiple stakeholders in different formats and requirements. This occurred, for example, when multiple departments were responsible for the evaluation and verification of payment-related and non-payment-related outcomes. Clear guidelines regarding the data and documents required to prove that outcomes had been achieved, as well as test runs at the beginning of the intervention in relation to data collection, appeared to help reduce the impact of some of these issues.

Regarding the data used by service providers to demonstrate payment-related outcomes, the majority of schemes relied on **paper-based systems**. For example, timesheets or signatures from employers were required to prove that participants had found jobs and remained employed. Such proofs of outcomes are burdensome for service providers to collect, for beneficiaries to provide, and for evaluators or commissioners to verify. In some instances, such data was not even available. For example, in the Perspektive:Arbeit SIB, service providers sometimes struggled to gather employment contracts because written labour contracts are not mandatory in Austria. In other cases, such as in PL Pathways, constant check-ups from service providers discouraged some potential employers from employing programme participants. Evidence emerging from the analysis shows that requests for data (during the initial phases of various programmes) may have discouraged some users from joining the initiatives. In JobPath, some programme participants refused to share details about their ongoing employment, which meant that providers could not claim payments for the services provided.

Digitalised systems to collect data could be a solution to these problems, and were used in some of the schemes (e.g. Transforming Rehabilitation). However, as highlighted in interviews with evaluators, outcomes often could not be reported using digitised administrative datasets, which would also have been less time-consuming. There are a few reasons for this. First, in some of the cases analysed, such datasets still did not exist at the time of the intervention – and might not be available in various EU Member States that are considering the implementation of SOC schemes. Second, even where administrative datasets were available, evaluators and commissioners often required data that was stored by different, siloed government departments (e.g. information on employment records and histories of benefit claims). Commissioners and evaluators therefore not only faced technical challenges in terms of how to acquire and merge that data, but also legal ones, as government agencies might not have been authorised to share such data. Finally, administrative databases were sometimes not updated with sufficient regularity to be of viable use in PbR schemes. Given that service providers required regular outcome-based payments to secure sufficient cash flow for the continuous provision of social services, it would not have been feasible for the providers to wait for an annual update of administrative data in order to receive their first payment.

Another challenge concerned the **time lag** between data collection and the payment, particularly when data needed to be checked or matched by public agencies. In Youth Contract, the providers reported difficulties linked to the time lag between an employer taking on an apprentice, and the Skills Funding Agency confirming employer eligibility and then actually providing the funds to work with the service beneficiary (e.g. attachment payment). In PL Pathways, providers complained that the outcomes measurement process hindered timely payment, since the DWP did not issue payments in cases where documents were considered incomplete, but more information was difficult to gather.

One possible solution to address issues relating to data collection and management is to allow an initial 'ramp-up' phase in the implementation of the programme. During this phase, data are not used to assess the outcomes upon which payments is conditioned, but are instead used to allow providers and commissioner to fix potential issues relating to the data collection process. This practice was suggested by service providers in the MHEP SIB, then implemented when the scheme was replicated in other locations. Initial input-based funding would be necessary for the 'ramp-up' phase.

On a related note, the **quality of data collected** by service providers was mentioned as an issue in some schemes. Often, service providers had no specific evaluation skills, making it harder to assure the quality of data collected. Nevertheless, some improvements to data collection processes were implemented over time. For example, when the Work Programme replaced PL Pathways, the DWP developed an internal system of off-benefit checks to verify all outcome claims prior to issuing payments. Interviewees noted that the most recent nationwide employment programme in the UK, the Work and Health programme, has also improved in that regard.

Despite the challenges outlined above, service providers reported that **frequent reviews** had a positive impact on service quality because issues were likely to be addressed more quickly due to the outcome payment imperative. They also added that when service providers oversaw data collection (as they did in most of the schemes analysed), it appears crucial to involve them in the design of the evaluation.

3.2. Evaluation methods used in SOC schemes

Having reviewed the definition of outcomes used in SOC schemes, as well as the challenges associated with their operationalisation, we now classify the evaluation methods adopted by the schemes studied. The classification helps us to identify in which schemes causality of impact can be attributed, and how.

The methods used are summarised according to the following typologies:

- **Non-experimental methods without comparison**, which cannot attribute the impact of an assessed intervention to a particular cause, because they do not control for various factors that might have affected the results of the intervention.
- **Non-experimental methods with comparison** (either with a control group or over time), which also cannot be used to attribute impact to a particular cause in a strict sense, because they do not have a high internal validity from a scientific point of view. Nevertheless, the comparison group, despite the risk of selection bias in the construction of the comparison, allows for the stronger validity of results than in non-experimental methods without a control group.
- **Quasi-experimental design** allows causal inference, internal validity is higher, and the possibility of selection bias in the control group is lower than in non-experimental designs.
- **Experimental design** can attribute causality of impact by eliminating selection bias.

Based on the mapping of all evaluations conducted on the SOC schemes selected, we observed that in most schemes **multiple evaluations** had been performed using different methods, in particular combining qualitative and quantitative data and quasi-experimental and non-experimental designs. These tended to assess the success of the scheme from two different perspectives. Some evaluations assessed the outcomes upon which payments were conditioned. The measurement of these outcomes aimed to establish the success of the scheme in relation to the targets that were agreed in the contract. Other evaluations were not directly related to the payment mechanisms, but instead contributed further to understanding the quality of the service, the perceptions of different stakeholders

concerning challenges and benefits related to their involvement in the SOC scheme, the various impacts of the intervention, as well as costs and benefits.

All of these evaluations, both related and unrelated to payments, are included in the mapping and review of methods presented in this section. Each type of evaluation design implemented in the SOC schemes investigated is presented in the following three subsections. Experimental methods are not covered, as these were not applied in any of the SOC schemes studied. For each evaluation approach, we aggregate the schemes using that type of design and outline their main features. Each subsection is introduced with an explanation of the type of analysis presented, and the rigour of that method in terms of drawing causal inferences. We conclude by discussing the decisions involved in choosing evaluation methods for SOC schemes.

3.2.1. Non-experimental approaches without comparison

Of the 15 schemes investigated in this study, nine featured evaluations that used a non-experimental design without comparison to assess the achievement of the outcomes upon which payment was conditioned. As mentioned above, this type of evaluation cannot determine the causality of the impact of interventions analysed. These evaluations are summarised in Table 5.

The nine schemes were characterised by a very simple definition of outcomes and related indicators, mostly counting the number of beneficiaries achieving a pre-defined objective (e.g. the number of beneficiaries who recovered from drug dependence, found and maintained a job, did not reoffend for a certain period of time, etc.).

Non-experimental design was often used to analyse qualitative data collected through interviews, focus groups or surveys with users, providers delivering the service and other stakeholders. These data collection activities often focused on the stakeholders' perceptions. The evaluation designs used did not consider any statistical techniques to determine the causality of the impacts observed. However, these qualitative analyses enabled a deep and comprehensive understanding of the ways in which different stakeholders perceived the effects of the intervention, and to investigate the process by which the intervention was delivered.

Table 5. Synthesis of non-experimental methods without comparison

Description	Quantitative and qualitative data collected from stakeholders (beneficiaries and others) involved in the intervention
Used to assess outcomes upon which payment was conditioned	Nine schemes used this method to assess the achievement of outcomes upon which payment was conditioned Nine schemes used this method for other evaluations that were not related to the payment mechanism
Causality of impact	Attribution not possible
Control group	No
Sources of data	The method uses both secondary data and primary data. <i>Secondary data</i> are self-reported data from the providers or the beneficiaries; or may be collected from existing databases. <i>Primary data</i> are collected through surveys on small samples, focus groups, informal discussions, or semi-structured interviews.
SOC schemes using the method	<i>Linked to payment:</i> - SIB : Perspektive:Arbeit, Eleven Augsburg, MHEP SIB - PbR : Drug and Alcohol Pilots, Troubled Families, Work Programme, PL Pathways, Youth Contract, JobPath <i>Not linked to payment:</i> - SIB : Eleven Augsburg, Perspektive:Arbeit, Benevolent Society, - PbR : Drug and Alcohol Pilots, Transforming Rehabilitation, Troubled Families, PL Pathways, Work Programme, Youth Contract
Geography	UK, Ireland, Australia, Austria, Germany
Time span	Often accompanies the intervention throughout its entire duration

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Type of evaluator	University, local authorities/commissioner, consultancy firms, service provider, not-for-profit research centre, law firm
Type of service provider	Charities or non-profit organisations (NPOs), local public administrations, private for-profit company
Social issues	<ul style="list-style-type: none"> - Drug and alcohol addiction - Recidivism - Safety and wellbeing of children and family - Employment (adult, youth, people with severe mental health issues) - NEETs
Strengths of the method	<ul style="list-style-type: none"> - Outcomes and related metrics are easily defined and understandable. - Having a clear definition of outcomes helps in identifying, or developing from scratch, a service that is more specifically tailored to solving the problem. - Checking of self-reported data from the provider by a third party is seen as helpful in increasing the reliability of the method. Structured and codified data analysis procedures such as thematic analysis increase the validity of results. - The method provides a comprehensive qualitative assessment of the intervention and enables feedback to be given stakeholders on how to improve service provision or the financing model. - Different rounds of data collection allow the monitoring of users at different points during the intervention.
Weaknesses of the method	<ul style="list-style-type: none"> - No possibility of fully addressing deadweight issues, ensuring the representativeness of results and determining the causality of impacts. - A method based on a one-dimensional criterion for success ignores softer success criteria/soft outcomes. - If outcomes are defined in a binary way (e.g. employed vs not), this can create perverse incentives (such as 'cherry picking' or excessive focus on those outcomes). - Frequency-based outcomes (e.g. how many months participants remain employed, avoid using drugs, etc.) are burdensome for service providers to track. - The heavy reliance on self-reported measures to assess the outcomes upon which payments are conditioned makes this method subject to the risk of fraud. - Metrics relating to the volumes of referrals to the programme can create a lot of pressure on service providers if actual volumes are different from those anticipated during programme design.

3.2.2. Non-experimental approaches with comparison

We encountered fewer schemes that employed a comparison group in their evaluation designs, compared with the number that utilised non-experimental methods without comparison. Table 6 gathers together all the evaluations using a non-experimental design with a comparison group. These evaluations make two types of comparisons:

- between **two different points in time**, e.g. the situation at the beginning or before the intervention, and the situation during or after the intervention; or

- two groups, e.g. the group that received the intervention and a **control group** of people who did not receive the intervention.

In neither of these methods was the control group robustly constructed. In evaluations that use an experimental or a quasi-experimental design, treatment and control groups are normally designated at random to ensure that no unobserved factors affect the results of the intervention. By contrast, the control group in non-experimental designs is sometimes composed of those people who chose not to participate in the intervention, or who matched similar demographic characteristics to the programme participants. In addition, the data analysis of the SOC schemes concerned did not include the required control tests to verify the variance between the treated and control groups.

Table 6. Synthesis of non-experimental methods with comparison

Description	<ul style="list-style-type: none"> - Pre-post comparison for treated only - Pre-post comparison between treated and non-treated
Used to assess outcomes on which payment was conditioned?	<p>Three schemes used this method to assess the achievement of outcomes upon which payment was conditioned</p> <p>Four schemes used this method for other evaluations that were not related to the payment mechanism</p>
Causality of impact	Causal inference is weak; internal validity is low; possibility of selection bias is high.
Control group	Not necessarily
Sources of data	<p><i>Primary data sources:</i> longitudinal surveys or interviews</p> <p><i>Secondary data sources:</i> public sector administrative data systems, e.g. prison data, police data or social security data on employment and benefit claims</p>
SOC schemes using the method	<p><i>Linked to payment:</i></p> <ul style="list-style-type: none"> - SIB: DUO for a JOB, Benevolent Society - PbR: Transforming Rehabilitation <p><i>Not linked to payment:</i></p> <ul style="list-style-type: none"> - SIB: BOAS Werkt - PbR: Troubled Families, Work Programme, PL Pathways
Geography	UK, Belgium, Netherlands, Australia
Time span	Data collected and analysed both at the beginning and end of the intervention, and sometimes even longer after the programme has ended
Type of evaluator	Public agencies, university research centre, local public administration
Type of service provider	Private for-profit companies, NPOs, local authorities
Social issues	<ul style="list-style-type: none"> - Reoffending - Immigration and unemployment - Safety and wellbeing of children and family
Strengths of the method	<ul style="list-style-type: none"> - Allows the progress of participants in the intervention to be tracked, even if that progress cannot be solely attributed to the intervention. - Use of statistical techniques such as t-tests or McNemar's tests can enhance the comparability of the treatment and control groups, even if they have not been randomly allocated.
Weaknesses of the method	<ul style="list-style-type: none"> - Inability to draw causal inference. - Long time span of measurement might entail investors having to wait for repayments and service providers engaging in new interventions before first learning about the effectiveness of the intervention in question. - Intervening factors are not addressed.

3.2.3. Quasi-experimental approaches

Somewhat surprisingly, quasi-experimental approaches – the most rigorous methods encountered during our review – were also the method least widely used to assess outcomes upon which payments were based. Table 7 presents the methods and schemes that used a quasi-experimental design in which there was a level of randomisation regarding the allocation of participants into treatment or control groups (either at individual or regional level). In particular, these schemes used differences-in-differences and propensity score matching as evaluation methodologies.

The differences-in-differences approach estimates treatment effects by comparing the pre- and post-treatment differences in the outcomes of a treatment and of a control group. Propensity score matching entails the use of statistical techniques to construct an artificial control group by matching each treated unit with a non-treated unit with similar characteristics. The resulting propensity scores in essence force an observational (non-randomised) study to mimic some of the particular characteristics of a randomised controlled trial (RCT).

Table 7. Synthesis of quasi-experimental methods

Description	<ul style="list-style-type: none"> - Differences-in-differences - Propensity score matching
Used to assess outcomes on which payment was conditioned	<p>Two schemes used this method to assess the achievement of outcomes upon which payment was conditioned</p> <p>Five schemes used this method for other evaluations that were not related to the payment mechanism</p>
Causality of impact	Causal inference is possible, although with caveats. Internal validity is high, the possibility of selection bias is low.
Control group	Yes
Sources of data	<p>Primary data sources: data collected by the local authorities, or by the service provider in the case of those participating in the programme.</p> <p>Secondary data sources: data are taken from administrative databases held by the public sector (National Drug Treatment Monitoring System dataset; Police National Computer; Hospital Episodes Statistics database, Jobseekers Longitudinal Dataset, etc.)</p>
SOC schemes using the method	<p>Linked to payment:</p> <ul style="list-style-type: none"> - SIB: Adolescent Behavioral Learning Experience, Buzinezzclub SIB <p>Not Linked to payment:</p> <ul style="list-style-type: none"> - PbR: JobPath, Drug and Alcohol Recovery Pilots, Troubled Families, PL Pathways, Youth Contract
Geography	UK, USA, Ireland, Netherlands
Time span	Data are usually tracked throughout the intervention, but analysis is carried out after the intervention
Type of evaluator	University, public agencies, non-profit research centres specialising in policy, local authorities, consultancies
Type of service provider	NPOs, local authorities, private for-profit companies
Social issues	<ul style="list-style-type: none"> - Drug and alcohol addiction - Reoffending - Unemployment (incl. specific target groups: unemployed due to health reasons, long-term unemployed, unemployed immigrants and youth) - Safety and wellbeing of family
Strengths of the method	<ul style="list-style-type: none"> - High level of internal validity, thanks to its quasi-experimental design. The validity of inferences can be increased thanks to the implementation of a sensitivity analysis of results and to

	multiple robustness checks using different layers of control variables.
Weaknesses of the method	<ul style="list-style-type: none"> - Lack of randomisation in impact evaluation: because it is not an RCT, some users in the comparison group may receive services related to the programme or may be affected by other intervening factors. - If the intervention is implemented differently (slightly different features of the intervention) at several sites (as in some of the analysed schemes), it is more difficult to design the measurement in a way that allows matching and, later on, the comparison of results. - Resource-intensive (both in terms of time and costs). - Can only be carried out when sufficient and reliable administrative data are available.

3.2.4. Discussion of overall evaluation design

None of the schemes analysed in this study implemented an **experimental design** to evaluate the outcomes of the intervention, and the use of quasi-experimental design was limited, as shown in Table 8. However, one programme – the ABLE SIB – attempted to carry out an RCT regarding a behavioural therapy programme for jailed youth offenders. The scheme eventually decided against using an RCT because clear boundaries could not be maintained between treatment and control groups when participants had to change housing units. Service providers from other programmes also felt that it would be unethical to deny services to some potential participants for the sake of establishing a control group, particularly in cases where other providers or state institutions could not provide the same services outside the SOC programme.

Only seven out of 15 schemes were evaluated using **quasi-experimental designs**. Interestingly, only in two cases did these evaluations relate to payment mechanisms. This means that rigorous evaluation methods that were able to determine causal links between interventions and outcomes triggered payments in only two schemes.

Table 8. Synthesis of evaluation design in the soc schemes identified

Evaluation design	Table header cell		Table header cell	
	SIB	PbR	SIB	PbR
Non-experimental methods without comparison	3	6	3	6
Non-experimental methods with comparison (either with a control group or over time)	2	1	1	3
Quasi-experimental design	2	0	0	5
Experimental design	0	0	0	0

The use of quasi-experimental designs to measure the achievement of **outcomes relating to payments** was implemented only in SIBs. In PbR schemes, outcomes relating to payments were assessed via non-experimental methods, mostly using data reported by service providers and metrics without comparison.

The differing designs of SIB and PbR schemes might have influenced the choice of more rigorous methods to evaluate payment-related outcomes in SIBs rather than PbR schemes.

The differences-in-differences approach or propensity score matching—the two quasi-experimental methods identified in our overview—are usually carried out after an intervention, or at least a few years into the intervention, so that the sample sizes are large enough for analysis. In SIBs, this time lag before investors are repaid on the basis of evaluation results is not an issue, because SIBs tend to attract large investors or philanthropic concerns that can maintain their activities until the SIB is complete. In the meantime, service providers receive financing from the upfront investment. By contrast, service providers in PbR schemes rely on outcomes-based payments to continue funding the services being delivered. This effectively means that service providers in PbR schemes cannot wait multiple years for evaluations to be carried out before receiving payment, particularly if that payment turns out to be lower than expected, as was the case in some of the PbR schemes studied.

The most commonly used evaluation design among the schemes – used in particular to **trigger payments** – was **non-experimental** design without any form of comparison. Many schemes employing this evaluation design tracked one-dimensional outcomes, and the indicator adopted was often a binary metric (achieved/not achieved). In these cases, neither intervening factors nor different proxies that could affect the situation of the service recipient were considered, making causal attribution quite problematic. Moreover, in most of these cases, data were reported by service providers or beneficiaries. This mechanism entails the risk of fraud (for example, see Provider-led Pathways in Annex 1).

It is important to note that even though payments were often based on the achievement of outcomes without a comparison group, commissioners monitored provider performance on the basis of historical data. Both the Work Programme and JobPath, for example, included minimum performance targets. If providers underperformed, they could be subjected to a review of their contract.

In many of the schemes studied, the independence of the evaluation was not ensured because the measurement process was implemented **by the commissioner or by public agencies related to the commissioner**. The justification for carrying out evaluations internally rather than contracting independent evaluators often included having sufficient in-house expertise, easier access to the data required for the evaluation, such as the tax records of programme participants, and the desire to learn more about the programme. In some cases, such as in Troubled Families (phase II), the risks relating to the evaluator's lack of independence were mitigated by the presence of an **advisory group** of experts which monitored and validated the design and implementation of the evaluation itself.

With regard to the **type of data** required to carry out evaluations, both primary and secondary data were used in the evaluations studied. Most frequently, data were collected by the service provider, then reported to the commissioner. In other cases, when administrative data were used, the evaluation did not require the creation of a new dataset, but instead relied on existing administrative databases. This often occurred in schemes relating to reoffending or unemployment that were commissioned at central level and with a national scope, where the commissioner itself performed the measurement. When primary data were collected, this was quite often through surveys, interviews or focus groups, directly involving either the service users or the staff of the service providers.

Overall, our analysis shows that a **good relationship between commissioners and evaluators** was a crucial enabler in effectively carrying out the measurement process. Furthermore, the involvement of evaluators in the design of the scheme and not just the evaluation also helped to ensure that the programme's effectiveness could be adequately measured. For example, in Pathways to Work, it was the evaluator's suggestion to roll the scheme out in phases, with different financing models (TF vs SOC) covering different areas of the UK and each delivery model expanded across the UK over time rather than at once. This effectively allowed the differences-in-differences method to be used for impact measurement. In JobPath, not all long-term unemployed were referred to the programme. Instead, the commissioner chose relevant benefit recipients **at random** to decide whether

they would take part in the programme. These programme design elements substantially enhanced the rigour of their evaluations. Similarly, support for evaluators by service providers, well as the transparency of data supplied by the providers, were considered very important factors in effectively monitoring outcomes.

3.3. Summary of findings on outcomes measurement

The mapping of the outcomes measurement methods conducted on the SOC schemes revealed that several types of evaluations design were adopted, particularly those combining qualitative and quantitative data and quasi-experimental and non-experimental designs. None of the schemes used an experimental design, and the use of quasi-experimental design was fairly limited, mostly in relation to outcomes not linked to payments. Thus, only in the few schemes in which experimental and quasi-experimental design were used could the **causality of impact** be attributed in a strict sense. In the remaining cases, the evidence was insufficient to claim that outcomes were achieved directly because of the interventions.

Some evidence emerged concerning the reasons why experimental designs were not used (and quasi-experimental less so). The first reason relates to the technical requirements for implementing those designs, such as the design of a control group or a misalignment between the timing of the measurement process and service implementation. Furthermore, some providers raised ethical concerns regarding control groups, as these would require denying services to some potential participants for the sake of methodological rigour.

Additional evaluations were often carried out to track outcomes that were not payment-related. These typically focused on softer milestones for programme participants. Such evaluations contributed to an understanding of the quality of the service, the perceptions of the various stakeholders concerning the challenges and benefits related to their involvement in the SOC scheme, the various impacts of the intervention, as well as its costs and benefits.

The **viewpoints** of commissioners and providers were often **mis-aligned** as to their choice of outcomes and metrics. Many service providers felt that the work they put in to delivering soft outcomes such as an increase in self-confidence or improvements in the mental state of participants, as well as supporting those who were harder-to-help, was not fully rewarded because of the use of one-dimensional success criteria based on a binary metric.

Different choices with regard to the definition of outcomes and metrics had a significant influence in encouraging or discouraging **perverse incentives** such as creaming and parking. In those schemes where a **binary metric** was used, service providers tended to focus on those users who were more likely to achieve the specified outcome. Metrics considering **long-term impacts** and the **influence of the users' characteristics** on their likelihood of achieving these outcomes were rarely included in the evaluation design. This created problems in achieving targets.

Finally, it emerges from our analysis that in the schemes in which the service provider was involved in or completely in charge of **data collection and management**, the process was perceived as a burden. The main causes were the volume of data required, especially when service providers had to report to multiple stakeholders using different formats and requirements; the reliance on paper-based systems; and a lack of specific evaluation skills among the service providers. Furthermore, the timeline for the collection and management of the data had to be carefully planned to avoid delaying payments to providers. Data supplied by service providers also had to be subjected to rigorous auditing on behalf of the commissioner, which required extra resources. Wherever possible, therefore, the designers of future SOC programmes should investigate possible ways to collect evaluation data independently – for example, through existing administrative databases.

4. Effectiveness

Little research exists that rigorously compares SOC contracts with traditionally financed programmes. As stated by the UK's National Audit Office, "Payment by results potentially offers benefits such as innovative solutions to intractable problems. If it can deliver these benefits, then the increased risk and cost may be justified, but this requires credible evidence. Without such evidence, procuring bodies may be using this mechanism in circumstances to which it is ill-suited, to the detriment of value for money."⁴⁹

In order to fill this gap research, Chapter 4 assesses the effectiveness and added value of interventions delivered under the SOC schemes considered, in comparison with the interventions delivered under traditional financing models. The standard definition of *effectiveness* is the ability of an action to accomplish specific objectives defined *ex ante*⁵⁰. Meanwhile, the *added value* literally refers to an improvement or addition to something that makes it worth more; in other words, it includes all those added features/consequences that increase value. In this specific case, we focus on the various benefits and drawbacks of SOC schemes compared with TF, in order to assess the added value of outcomes-based contracts.

It is important to mention that comparing the effectiveness of SOC and TF schemes entails certain challenges. First, not all TF schemes were evaluated or measured their results. Thus, data on the achievement of specific outcomes and impacts is often lacking. Second, for those TF schemes that were evaluated (or for which outcomes were systematically measured), the definitions of outcomes and the metrics, and indicators adopted to measure them, were often different from those adopted for SOC schemes. Lastly, the selection of TF equivalents has some limitations (specified in Section 2.2.2 as well as under each scheme in Annex 1). This is because the SOC and TF programmes were sometimes implemented at different times, in different settings, addressing slightly different targets, or delivered different types of service. Thus, all of the intervening factors should be considered when comparing their effectiveness.

That said, where impact evaluations had been conducted on both the SOC scheme and its TF equivalent, it was possible to compare the schemes based on their achievement of outcomes, taking into account the limitations outlined above. In the other cases, we could gain qualitative insights by comparing schemes on the basis of the data collected from desk research and interviews.

Based on an analysis of the data, this chapter first discusses the effectiveness of SOC schemes in achieving their expected outcomes. It then goes on to consider added value in terms of the main benefits and drawbacks of SOC schemes compared with their TF equivalents, as perceived by the stakeholders interviewed. Lastly, drawing on the information gathered, it discusses the scalability and replicability of SOC schemes. We summarise the main findings at the end of the chapter.

4.1. Achievement of outcomes

In this section, we look at whether and how the various SOC schemes achieved the targets defined in their contracts. Table 9 presents the targets and the extent to which they were achieved by each of the SOC schemes analysed. We also added two columns indicating, respectively, the policy area in which the scheme operated, to highlight relationships between programme effectiveness and policy area (if any); and the duration of the

⁴⁹ Amyas Morse, head of the National Audit Office, 19 June 2015.

⁵⁰ European Commission (2015). Better Regulation Toolbox. [SWD (2015) 111]. Retrieved from:
https://ec.europa.eu/info/sites/info/files/better-regulation-toolbox-2015_0.pdf

schemes, since this helps to contextualise the targets achieved (it is different matter to serve 20 people in a few months or in five years).

The table shows that out of 14 schemes⁵¹:

- six schemes achieved their expected targets;
- one is ongoing, and positively progressing towards achieving its target;
- seven did not achieve their targets.

Looking at the different types of schemes, four out of the eight SIBs hit their targets, as did three of the seven PbR schemes. This finding does not indicate any clear trend in the relative effectiveness of SIB vs PbR. Neither is there evidence of any clear relationship between the effectiveness of the programme and the policy area in which it operates. Nevertheless, it appears that one policy area is chosen more frequently to implement SOC schemes —labour market related interventions. Even within programmes that primarily serve clients with disabilities, mental health issues, those in crisis situations or facing social exclusion, employment targets and outcomes are dominant. This was also observed in early stages of the study, when compiling a long list of SOC programmes for consideration (see Annex 3). Therefore, it seems that the **feasibility of outcome and impact measurement is more important** when considering the design of a potential SOC scheme **than a specific policy area or target group**. In this sense, it is reasonable to assume that many SOC schemes deal with employment, since it is relatively easy to measure a simple outcome such as whether a person finds a job, especially in countries where datasets are available to collect relevant data on the labour market (e.g. registering the status of a person as unemployed or a jobseeker, and data about work benefits). Nonetheless, this reliance on employment outcomes might come at the cost of overlooking other important multi-dimensional impacts that are more difficult to operationalise and measure (see Chapter 3, Section 3.1.1 regarding the definition of outcomes).

⁵¹ One scheme—Drug and Alcohol Recovery Pilots—lacked data to assess the achievement of outcomes.

Table 9. Achievement of targets

Scheme name	Targets	Achievement of targets	Policy area	Duration
SIBs				
Eleven Augsburg	Bringing 20 difficult-to-reach juveniles into work or apprenticeships lasting at least nine months.	YES: 20 hard-to-reach juveniles were placed into a job.	Labour market-related	Two years
Perspektive:Arbeit	Placing at least 75 survivors of domestic violence in a job with at least 20 hours of employment per week that is subject to social insurance contributions and pays a living wage, for at least 12 months during the term of the project.	NO: 52 women have been successful cases according to the agreed targets. The main target has not been met because the employment acquired either did not last 12 months or remuneration was below the required threshold.	Crisis and emergency Labour market-related	Three years
DUO for a JOB	Achieving 10% higher relative rate of employment among the DUO for a JOB group relative to the control group.	YES: 322 mentee-mentor duos were established, which resulted in 133 job placements for DUO's mentees. The placement rate was 16% to 42% higher than the control group's during the programme's three years.	Labour market-related	Three years
Buzinezzclub	Median reduction of 210 days in the duration of unemployment benefit payments compared to a baseline.	YES: the median duration of unemployment benefit payments was reduced by 248 days during the first year and by 324 days in the second year, compared to a baseline each year.	Labour market-related	Two years
BOAS Werkt	Getting at least 75% of 138 individuals from the 7,000 households receiving benefits in Enschede into jobs.	NO: 58% of people trained found a job, but the evaluation still reports overall savings.	Labour market-related	Two years
ABLE Programme at Rikers Island	Reducing recidivism rates by 10% relative to a comparison group; 30% of participants completing the programme.	NO: only 11% of participants completed all 12 stages of the programme. Readmission bed days for 16 to 18-year-olds (ABLE cohort) increased 13.4% between 2006-2010 and 2013.	Social exclusion	Two years
Benevolent Society Social Benefit Bond	A minimum of 5% performance percentage (based on number of entries into out-of-home care per child; helpline reports; safety and risk assessments) was required to repay the investor.	YES: Performance Percentage was 16% in July 2018.	Caring obligation	Five years

Scheme name	Targets	Achievement of targets	Policy area	Duration
Mental Health and Employment Partnership	For each provider in Year 1: 140 people start a job; 90 people sustain the job for six weeks; No target for the number of participants engaged sustained the job for 6-month outcome.	NO: 30% of engaged people started a job; 21.5% of all participants achieved a six-week outcome.	Labour market-related Disabilities	Three years
PbRs				
JobPath	Moving 14.8% of long-term unemployed participants into full-time paid employment in the first year of the programme.	YES: 22% of participants moved into full-time paid employment in the first year.	Labour market-related	Three years
Provider-led Pathways to Work	On average across service providers, contractors were expected to move one in three clients into employment (37%).	NO: actual job rates for claimants who were required to participate in Pathways ranged between 3% and 11%.	Labour market related Disabilities	Seven years
Work Programme	The initial expectation was to move roughly one in three Jobseeker's Allowance claimants into work, and one in 10 Employment and Support Allowance claimants. However, these expectations were downgraded after providers failed to meet them. After adjustment, roughly 12% of all participants were expected to find jobs and stay employed for three of six months (depending on the group).	YES: performance levels were exceeded after target adjustment: for example, in 2016, 16.2% of all participants founds jobs and remained employed for three/six months.	Labour market-related Disabilities Social exclusion	Eight years
Youth Contract	50% sustained re-engagement	NO: 47% of the young people referred to the programme achieved a sustained outcome	Labour market-related	Four years
Drug and Alcohol Recovery Pilots	The pilot areas were free to set the price and to weigh outcomes differently	N/A	Social exclusion	Two years
Transforming Rehabilitation	Reducing reoffending by 3.7 percentage points over the life of the contracts	NO: by March 2017, there was a 2.5 percentage point reduction in the proportion of proven reoffenders since 2011; a 22% overall increase in the average number of reoffences per reoffender; just six out of 21 CRCs consistently achieved their targets to reduce reoffending.	Social exclusion	Six years
Troubled Families (phase II)	Progress for 400,000 families in terms of the Troubled Families Outcome Plan defined by the Local Authority.	ONGOING: By 8 March 2019, 380,426 families had been assisted by the TF Programme, out of the 400,000 target families.	Social exclusion Labour market-related Caring obligations	Seven years

The **main reasons why targets were not achieved** were:

- **Lower enrolment/referrals** for the programme than expected *ex ante* (Boas Werkt SIB, Provider-led Pathways to Work). For instance, in the Boas Werkt SIB, the true take-up rate was roughly half that anticipated; in Provider-led Pathways to Work, the level of enrolment on a voluntary basis was much lower than expected.
- **Contextual factors** unrelated to the design or functioning of the scheme, hampered the engagement of users or the completion of the programme (see the ABLE SIB, Boas Werkt SIB, Provider-led Pathways to Work and Transforming Rehabilitation in Annex 1). For example, in the ABLE SIB, one of the possible causes for the lower percentage of participants completing the programme is that the number of sessions participants received was directly related to the instability of the jail system and the inmates' indeterminate length of stay. Furthermore, the evaluator argued that cuts in police activities and changes in police behaviour were the main cause of the increase in the number of reoffences.
- The **profiles of users** affected the results (Perspektive:Arbeit SIB; MHEP SIB). In Perspektive:Arbeit SIB, the qualitative evaluation showed that participants faced difficulties in acquiring or maintaining the employment due to childcare commitments, rural living locations, and other factors.
- The competitive bidding process encouraged providers to set **unrealistically high targets** (Provider-led Pathways to Work; MHEP SIB).

As mentioned above, **comparison between TF and SOC schemes with regard to the achievement of outcomes** was not always possible. Therefore, for those cases where a strict comparison in relation to the mere achievement of targets was impossible, we relied on other, qualitative data collected on perspectives of various stakeholders regarding the overall success of the two schemes (i.e. meeting the needs of the beneficiaries, scoring a good level of user satisfaction, generating positive outcomes and impacts). In this regard, Table 10 illustrates not only whether each scheme achieved its targets, but also the overall success of the programme according to other qualitative information that was collected through document analysis and interviews. It is important to mention that targets were not always set in the case of TF schemes. The last column of the table briefly summarises the main issues with the comparison. These are presented more extensively for each programme in Annex 1. It is important to keep them in mind when analysing the outcomes achieved for each SOC-TF pair.

Overall, the data showed mixed results and we cannot argue that SOC schemes were more effective than TF schemes or vice versa.

Table 10. Achievement of outcomes in TF and SOC schemes

#	SOC and TF Schemes	Targets Achieved	Overall success of the programme	Issues of comparability
SIBs				
1	Eleven Augsburg (SOC)	YES	The programme was perceived as successful, but there are legal hurdles in launching SIBs as well as a risk that the commissioner will be perceived as allowing the private sector to benefit from the public sector.	Activities: slightly different service offering in the two programmes.
	AMA-Zukunft (TF)	YES	N/A	
2	Perspektive:Arbeit (SOC)	NO	Despite its failure to meet the targets, the SIB programme was subsequently funded via traditional financing (see the next section). This is because the SIB was perceived by all stakeholders to be a success.	Outcome measurement: the minimum income clause regarding outcomes was no longer in effect in the TF.
	Phase III of Perspektive:Arbeit (TF)	N/A	N/A	
3	DUO for a JOB (SOC)	YES	The scheme was overall perceived as successful by stakeholders.	Timing: Lessons learned from the SIB were taken into account in the subsequent TF phases.
	DUO for A JOB Actiris-based funding (TF)	YES	Data showed that the TF intervention achieved better results in terms of job starts than the SOC scheme (55% of mentees have found a job, compared with an average of 40% during the SIB).	
4	Buzinezzclub SIB (SOC)	YES	Overall, the SIB was considered a success and investors were paid in full.	Timing: Macro-economic conditions were more favourable during the period of the traditionally financed scheme.
	Subsequent traditionally funded scheme (TF)	N/A	The TF was positively perceived.	
5	BOAS Werkt (SOC)	NO	Participants were half as likely to claim benefits as the control group, so the programme had a positive impact despite its early discontinuation.	Activities: The TF focuses more on providing generic and customised support with red-tape compared with more active placements in BOAS Werkt SIB.
	Transfer Point (TF)	N/A	N/A	
6	ABLE Program at Rikers Island (SOC)	NO	N/A	Target population: RESTART programme targets high-risk individuals and is not specifically targeted at adolescents.
	RESTART (TF)	N/A	A high percentage of participants successfully completed the programme.	
7	Benevolent Society Social Benefit Bond (SBB) (SOC)	YES	Resilient Families proved to be a highly effective model for preventing entries into out-of-home care.	Target population: slightly different

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

#	SOC and TF Schemes	Targets Achieved	Overall success of the programme	Issues of comparability
			The evaluation of the SBB showed the added value of the scheme with regard to long-term outcomes.	
	Intensive Family Preservation (TF)	N/A	N/A	
8	Mental Health and Employment Partnership (SOC)	NO	Although the programme did not achieve its outcome, the services score more than Fair across all respondents who were asked for their overall impression of how the MHEP SIB had gone to date. Overall, the personalisation of a joint employment and mental health support was considered very positive by users.	No specific issue
	Camden Individual Placement Support (TF)	NO	Although the programme did not achieve its targets, most clients reported improved mental wellbeing and staff reported positive distance travelled with regard to mental health and employment. The users suggested the personalisation of the service as a particularly positive element.	
PbR				
9	JobPath (SOC)	YES	Overall, the scheme was perceived as successful – although the favourable economic situation did aid the achievement of its outcomes.	Target population: the target groups are not identical. Intreo mostly focuses on the short-term unemployed.
	LES (TF)	NO	N/A	
10	Provider-led Pathways to Work (SOC)	NO	Both Provider-led and Jobcentre Plus-led Pathways to Work ended in 2011 when the initiatives were replaced by the Work Programme. The replacement was partially motivated by negative reviews from the National Audit Office and the Committee of Public Accounts. Thus, overall, both programmes were evaluated negatively. However, it can be noted that PL Pathways was more effective in getting people off benefits, but was less effective for harder-to-help individuals.	Activities: the intensity of activities varied
	Jobcentre Plus Pathways (TF)	N/A		
11	Work Programme (SOC)	YES	Despite the achievement of outcomes, the intervention was not very effective for groups that were harder to help, and performance of providers varied widely. Thus, the overall perception of the programme was mixed.	Timing: the New Deal was in effect during the 2008 recession.
	NDYP/ND25PL (TF)	YES	Overall, the schemes were perceived differently by different stakeholders. Evidence showed a 'carrot and stick' effect, with some people intensifying job search to avoid joining an option or benefit sanctions. However, it should be noted that the	Target population: people were referred to the two programmes after different periods of unemployment.

#	SOC and TF Schemes	Targets Achieved	Overall success of the programme	Issues of comparability
			programme was in effect during the 2008 recession, which may have negatively impacted its performance.	
12	Youth Contract (SOC)	YES	Despite the achievement of outcomes, some stakeholders perceived the scheme negatively because a large proportion of the planned budget was not used after the scheme was discontinued.	Target population: the client group of YC included clients who were more difficult to reach, given the tight eligibility criteria. Outcome measurement: outcomes in the AA programme had to be sustained for three months, compared with five months in the Youth Contract.
	AA PILOTS (TF)	N/A	The scheme was discontinued before its end due to budget cuts.	
13	Drug and Alcohol Recovery Pilots (SOC)	N/A	Providers and users reported a positive view on the piloting because the focus on the full recovery outcome introduced by the PbR model led to a greater focus on the complete recovery of the patients on the part of the caring service provided by keyworkers.	Activities: full-recovery outcome focus, the “localist” approach of the SOC programme, and the inclusion of both alcohol and drug recovery aims.
	Lincolnshire Drug and Alcohol Programme (TF)	N/A	N/A	
14	Transforming Rehabilitation (SOC)	NO	The quality of the CRC’s probation work was found to be poor by the independent Chief Inspector of Probation. Through the Gate services, a peculiar service introduced by the TR reform, have consistently failed to meet offenders’ resettlement needs.	Target population: the reform extended the compulsory post-sentence supervision to offenders serving short custodial sentences (i.e. prison sentences of under 12 months).
	The Probation Trust (TF)	YES	Before the TR programme, the reoffending rate among adult offenders serving court orders was already falling. The services provided by the Probation Trusts have been considered effective by both HM Inspector for Probation and HM Prison and Probation Service.	
15	Troubled Families (phase II) (SOC)	Ongoing	The programme is ongoing but has been progressing gradually towards its targets. User perceptions emerging from the analysis of case studies is positive overall.	No specific issue
	Earned Autonomy Model (TF)	N/A	Data from the case study suggest that the model is positively perceived overall by the stakeholders interviewed.	

Interestingly, as illustrated in the table, among those **schemes that did not achieve their targets**, several cases were **nevertheless considered successful** by the stakeholders and were subsequently extended or replicated. One example is the MHEP SIB: although the programme did not meet the targets, users reported a high level of satisfaction. The evaluation report confirmed that all providers successfully engaged with and supported individuals in finding employment to a varying degree. Service users saw job outcomes as a wider category than simply getting a job, and perceived that the services had positively impacted their ability to find employment. Receiving employment support from staff trained in mental health meant that they could discuss their health issues with their employment specialists if they wanted to, and felt supported when they did so. Indeed, the SIB was extended in one of the boroughs in which it had already been implemented, and subsequently replicated in another three London boroughs. Learning how to engage persons with mental health issues therefore appears to be its key positive outcome.

The case of Perspektive:Arbeit is also informative in this regard. Despite the programme's failure to achieve its targets, the SIB was perceived as a success by all stakeholders, and thus the government funded the programme traditionally. After a follow-up phase of three months (September 2018 to November 2018), which was funded by the intermediary to finalise all project-related work, the social intervention was funded by the Ministry of Labour for a further 14 months in the form of public subsidies. Since February 2020, the social measure has been commissioned by the Public Employment Service Austria for another 14 months. The intermediary saw the success of being able to replicate their SIB model in Austria, after it had piloted a technically similar SIB in Germany. Even the investors – who lost their investment – saw value in supporting and showcasing impact-oriented SIBs for the other stakeholders who had not previously worked with impact investing. Furthermore, the investors saw success in the continuation of the measure after the SIB. The social service provider perceived the scheme as a success, as otherwise no such services would have been delivered to the target group. Impact was created at various levels and for a number of individuals, albeit that the targets were not technically achieved.

Another important point when comparing the effectiveness of SOC and TF schemes is that **in some cases the positive results achieved by the TF interventions were facilitated by the previous work developed by the SOC scheme**, because lessons learned during the SIB had been taken on board in the traditionally financed scheme. Stakeholders recognised that this was the case for both the DUO for a JOB SIB and the Buzinezzclub SIB, **which were used as pilots for equivalent TF schemes**. Similarly, the experience of JobPath and the Work Programme provided some insights for the programmes that followed. The Department of Employment Affairs and Social Protection (the commissioning department for JobPath) is considering the possibility of replicating the PbR model in other employment services. The Merlin Standard developed under the Work Programme was later adopted in the Youth Contract to manage quality across supply chains.

Four schemes (the ABLE SIB, Transforming Rehabilitation, BOAS Werkt and Youth Contract) were **discontinued before the expected end** of the scheme, due to the proven ineffectiveness of the intervention, inability to recruit a sufficient number of participants, or a change in the overall economic climate. As for Transforming Rehabilitation, the probation service will be transferred back into the public sector in June 2021 after the PbR model has failed to deliver the expected results. Thus, after testing out the PbR model, the Ministry of Justice (MoJ) has returned to a wholly traditionally financed probation service. In the case of Youth Contract, the programme was discontinued due to political opposition, the scheme's hefty price tag, lower take-up than expected and the fact that the labour force and education participation levels were high, making the rationale no longer relevant. One of the reasons BOAS Werkt was discontinued is also a decrease in unemployment due to improved macro-economic conditions.

In some cases, **stakeholders did not see any difference** between the effectiveness of interventions under SIB or traditionally financed models, regardless of whether the overall results were negative (Transforming Rehabilitation) or positive (MHEP SIB, DUO for a JOB,

Drug and Alcohol Recovery Pilots). For instance, the MHEP SIB and its equivalent TF programme delivered the same Individual Placement and Support (IPS) service, which in both cases was considered very helpful by the users because it was personalised. It thus appears that the financial mechanisms did not make any difference to the participants in the intervention, but the positive outcome was due to the type of social programme delivered. Similarly, in Transforming Rehabilitation, the new service introduced (providing resettlement support for short-term offenders) was considered effective by both the users and the evaluator.

4.2. Added value with regard to the benefits and drawbacks of SOC compared with TF models

Although there are certain limitations to our comparison between SOC schemes and their TF equivalents, as mentioned above, it was possible to gain some qualitative insights on the added value of SOC from the data collected for this study, by analysing the benefits and drawbacks of SOC compared to TF.

4.2.1. Benefits

We identified a number of key benefits among the SOC schemes analysed, including:

- the development of a measurement infrastructure;
- the opportunity to support evidence-based policymaking;
- the opportunity to test new interventions;
- knowledge sharing among different stakeholders;
- unlocking financial resources;
- enhanced flexibility.

Below, we provide more details about each of these benefits.

Development of a measurement infrastructure

The presence and features of the measurement infrastructure developed to support the evaluation process in SOC schemes has been extensively discussed in Chapter 3. In terms of comparing SOC and TF models, various stakeholders highlighted that the measurement process enabled them to develop new skills and data management tools that would not have been developed under a traditionally financed model. The service providers in DUO for a JOB reported that the SIB forced them to develop indicators to measure their impact beyond job placement, because they wanted to assess their service in a comprehensive way and to anticipate any negative outcomes in comparison to the control group. According to the provider of the Australian SBB, the measurement activity helped in changing the mindset across all of the organisations involved in the contract, enhancing the focus on the measurement of outcomes and the impact produced. Finally, in the Troubled Families programme, the local authorities interviewed noted that the measurement requirements helped them to improve their culture regarding data management. The central commissioner underlined that the data collection process fostered the overall development of data-based public services, since the programme required data to be shared between different agencies working on social services.

Commissioners, service providers, evaluators and investors — all noted that the measurement infrastructure put in place in SOC schemes drives organisations to focus on outcomes rather than outputs. This can contribute to enhancing programme results. Since the contract entails a precise estimation of the expected results in terms of outcomes to be achieved (not just outputs) and returns for investors, stakeholders not only measure but

also focus on outcomes when designing their actions. From this perspective, the importance of measuring and monitoring the outcome in SOC can help in defining the size and scope of the interventions in advance. In this respect, as noted by one provider of Boas Werkt, this kind of contract provides clear information about the expected size and scope of the intervention, which can help in estimating the expected costs and resources required for it.

Evidence-based policymaking

By measuring outcomes and impact, SOC schemes support the development of evidence-based practices. As one investor noted during their interview, SOC schemes can allow a new intervention with uncertain outcomes to become a proven intervention, thus providing evidence for the development of future service provision. For instance, in the Australian SBB, the independent evaluation of the intervention allowed governmental agencies to collect data that was helpful in implementing similar programmes.

Testing interventions

Some stakeholders emphasised the opportunity to use SOC schemes to test interventions that were not previously part of welfare provision. According to these stakeholders, following the positive results of an SOC, these interventions could be transferred into the regular service portfolio of the respective public agency using traditionally financed instruments. The investor of Buzinezzclub wrote in 2019 that the model of SIBs “seems most useful when it is used for yet unproven interventions”⁵², given that it allows the effectiveness of an intervention to be proven, and enables its future adoption on a broader scale. One example of this is the DUO for a JOB SIB analysed in this study. The same intervention (a mentoring programme for youth with immigrant backgrounds) successfully tested through the SIB was later funded through a traditionally financed model.

Knowledge sharing among the stakeholders

The complex design of SOC schemes entails collaboration between various stakeholders. In the case of the PbR schemes analysed, this involved (at least) a commissioner and a service provider. Evaluators were often involved as a separate party as well, but there were cases in which evaluations were carried out by the commissioning department (for an example, see JobPath in Annex 1). In addition to commissioners and service providers, investors and evaluators participated in SIBs. Furthermore, some SIBs also included intermediaries. Although the coordination of different actors may involve some drawbacks in terms of management complexity and coordination (see Sections 4.2.2 and 5.2.2), some of the stakeholders interviewed noted that their participation in an SOC scheme allowed them to develop new collaborations, to get to know new potential partners for future projects, and to share best practices. For instance, in the Work Programme, collaboration among providers was common, and providers shared information about vacancies. Similarly, Youth Contract appears to have enhanced collaboration between prime providers and local authorities.

From this perspective, SOC schemes have the potential to enable cross-sectoral collaboration through the involvement of different public and private stakeholders in the design and implementation of services. Nevertheless, it is worth noting that the practices of knowledge sharing and collaboration can also be implemented under TF models, for instance when co-creation or co-design take place.

⁵² Dekker, F. & Verhoeven, J. (2020). Vijf Jaar Social Impact Bonds: De Balans Opgemaakt. Start Foundation.

Unlocking financial resources

Particularly in the case of SIBs, various stakeholders noted that SOC schemes enabled the provision of new sources of funding that would not have been available under a traditionally financed model. This was the case in the Austrian SIB. The specific social issue addressed by the scheme (women victims of domestic violence) was outside the core mission of the investor, the ERSTE Foundation. However, the foundation decided to fund the scheme to experiment with a new financial instrument, the SIB, as a tool to test innovative social solutions. In other words, the type of financial instrument triggered the interest of the investor more than the social issues addressed. Furthermore, survivors of domestic violence were chosen as a target group for the SIB because they would not otherwise have received specific services based on the allocated annual budget.

In some cases it was noted that PbR schemes also enabled the provision of funding for previously neglected social issues. This was particularly the case where PbR schemes piloted new intervention models to address specific social needs. For instance, in the Drug and Alcohol Recovery Pilots, the local commissioner noted that under the pilot there was actually a higher budget available, which allowed the performance of the services to be enhanced.

Another way in which SOC can unlock resources is by merging different sources of funding that would not otherwise be used to fund a single intervention. This was the case in the MHEP SIB, which provided an opportunity to combine national and local outcomes-based funding (from areas in which the intervention took place) and brought together funding for sets of outcomes that were not normally aligned (funding streams from partners interested in mental health and work outcomes).

Enhanced flexibility

Some providers underscored that the design and management of interventions under SOC schemes allowed greater flexibility than under traditionally financed models. One example is the Perspektive:Arbeit SIB. In this case, it was possible to fund driving licences for women who could not get to work using public transport. Such expenditures were no longer possible under the subsequent traditionally financed model. Similarly, one service provider involved in the Eleven Augsburg SIB stated: “I had my own pot of money that I could administer myself under the SIB. It was very relaxed in terms of daily working procedures. I could choose for what I would spend the money. This was gigantic!”.

Besides flexibility in the use of financial resources, some providers also stressed that SOC contracts allowed personalising the service provision. In JobPath, providers could develop services based on individual cohorts and their needs. In Troubled Families, local authorities earned a degree of autonomy that gave providers greater flexibility in setting priorities and programming, allowing them to reset and change programme priorities on the basis of the progress of families in the programme. Flexibility of SOC contracts could be particularly beneficial when the social services available lack personalisation.

4.2.2. Drawbacks

From both the document analysis and the interviews conducted for this study, it is possible to identify some drawbacks of SOC models compared with TF schemes. Specifically, four main drawbacks were reported across the cases analysed. These were:

- perverse incentives;
- management complexity;
- lack of systemic change;
- susceptibility to the economic context.

Each of these is discussed below.

Perverse incentives

Overall, we found that some form of perverse incentives existed in most of the SOC schemes analysed. These risks were linked to both SIBs and PbR contracts, although they were encountered more frequently in PbR models. Whereas SIB providers at times creamed those participants most likely to succeed at the point of recruitment into the programme, PbR providers parked those participants who were least likely to achieve outcomes after they had already been recruited. This is probably because in many PbR programmes, participants were referred to providers by the commissioner. Due to perverse incentives, the PbR schemes analysed were less effective at serving those groups that were **'harder to help'** compared with other groups of participants.

This was particularly true in schemes tackling unemployment such as Transforming Rehabilitation, the Work Programme, Youth Contract, JobPath and Provider-led Pathways to Work (see Annex 1 for details). Around 40% of all jobs achieved in Provider-led Pathways areas were for claimants who had volunteered to participate in the programme (as opposed to those for whom participation was mandatory in order to continue receiving monetary benefits), compared with 9% of all jobs achieved through its equivalent TF programme, Jobcentre Plus Pathways. Volunteers were easier to support into employment because they were possibly more motivated to take steps towards securing a job.⁵³ In the Work Programme, results varied substantially between groups: those claiming benefits due to disability failed to reach the target outcomes more often than other jobseekers, even though lower targets were set for those with disabilities. Personal characteristics made a difference to the likelihood of participants finding and sustaining work while on the programme. In particular, multivariate statistical analysis (i.e. controlling for other intervening factors) showed that, after two years, participants' total duration of employment while on the programme was higher if they were female, young, did not have a disability or health condition, had recent work experience prior to joining the programme; and lived in a less deprived local labour market.⁵⁴

These differences in results for those who were harder to help related to the practice of creaming and/or parking of participants. Examples of these practices were encountered in many of the cases analysed (see Annex 1 for details regarding each programme). In Provider-led Pathways to Work, advisers reported strong management pressure to focus on job-ready clients, which led to less time being spent with clients who were less ready for work. Moreover, there was also evidence of parking, because those groups of customers who were harder to help continued to take part in the interventions without receiving the proper assistance, just for the sake of proving that they were attending the programme.⁵⁵ In Youth Contract, the creaming of participants was a significant issue according to the commissioner interviewed. Some providers chose young people with less disadvantaged backgrounds to ensure that re-engagement and sustainability payments could be achieved. In the Work Programme evaluations, providers openly reported that they provided a more stable support to their most job-ready participants.⁵⁶ In Drug and Alcohol Pilots, a portion of payment was based on the number of successful completions of the programme. An even greater share of income depended on whether drug users relapse (re-present themselves for treatment). The evaluator argued that such an incentive system might have conditioned providers to be risk-averse, and keep the most complex cases in treatment for a prolonged duration, to ensure they did not present themselves again for treatment. An

⁵³ NAO (2010), 9.

⁵⁴ Meager et al. (2014), 26.

⁵⁵ Hudson et al. (2010), 60.

⁵⁶ Foster et al. 2014b, 5-6.

unintended consequence of such a payment mechanism was that some service users declined treatment, which may have adversely affected their recovery.⁵⁷

Nevertheless, it is worth mentioning that SIBs can be successfully used to address specific groups of harder-to-help users and, in particular, to test innovative interventions that place them as a core target group (see Section 4.2.1 about testing interventions using SIBs). Examples are the Austria SIB and Perspektive:Arbeit, which targeted very specific groups such as women victims of domestic violence, and the DUO for a JOB SIB, which targeted young unemployed immigrants.

Management complexity

As explained in detail in Section 5.2.2, most of the SOC schemes analysed in this study had greater management costs than the comparable TF models. These relate to the need to coordinate multiple stakeholders, to collect data for measurement and evaluation (which often is not collected in TF programmes), and to the complexity of the contract and financial mechanisms (see Chapter 7 for more details on the design of the contracts). Various commissioners mentioned that the process of overseeing the SOC was quite intensive, particularly when compared with traditionally financed interventions. However, it is worth mentioning that the higher effort and complexity required to manage SOC schemes has been beneficial in designing TF models more smoothly in those cases where SIBs were used to test an intervention that was subsequently scaled or replicated using a TF model, such as DUO for a JOB, Perspektive:Arbeit, and the Benevolent Society SBB (see Section 4.3.1).

Lack of systemic change

With particular regard to the use of SIBs to test new interventions, it has been noted that if the successful intervention does not become part of regular welfare provision, the positive effects of the intervention remained limited to a restricted group of participants. For instance, the provider of the Benevolent Society SBB noted that the successful intervention remained outside regular governmental programmes and was not properly integrated into core welfare provision, thus limiting the effectiveness of the scheme to a restricted number of beneficiaries.

Susceptibility to the economic context

It is important to note that the effectiveness and success of SOC schemes can be greatly influenced by the broader economic situation of the country and region in which the SOC scheme is developed. For example, the Provider-led Pathways to Work programme took place during the 2008 recession, which partly accounts for why the programme failed to reach its targets. When payments depend on the macro-economic context, it is difficult to provide financial security for the providers. Nevertheless, we encountered some strategies to limit these dangers, discussed in Section 6.1.

4.2.3. Debated points

Besides some benefits and drawbacks recognised by different stakeholders as typical of SOC schemes in comparison to traditionally financed ones, a few topics were mentioned as benefits of SOC by some stakeholders, and as drawbacks by others.

⁵⁷ Mason et al. (2015), 2.

Financial stability for providers

The issue of financial stability is key for all providers involved in PbR schemes – it was mentioned during interviews and also underlined in some of the evaluation reports. The PbR model transfers a great deal of financial risk on to providers, thus limiting their ability to develop long-term plans, particularly in relation to hiring and staff training. In contrast, financial stability was mentioned as a benefit provided by SIBs compared with traditionally financed interventions. This was based on the fact that SIBs unlock new sources of funding and do not transfer the financial risk to providers (when providers do not invest their own resources). Furthermore, since SIBs' contracts entail an estimation of the expected results in terms of outcomes to be achieved (not just outputs) and returns for investors, some noted that they can help providers to anticipate the size and scope of the intervention, allowing them to estimate the expected costs and develop plans in the medium term. For instance, the provider of the DUO for a JOB noted that the scheme ensured financial stability in the medium term, also thanks to the additional financial resources unlocked by the SIB. For more information regarding the impact of SOC on service providers, please see Section 6.3.

User satisfaction and quality of service

SOC schemes are also often credited with increasing **user satisfaction**. SOC models are sometimes expected to improve the quality of service, thanks to the feedback collected through the measurement process, the freedom given to providers to constantly evolve the delivery of services, and the possibility of designing tailored, responsive and intensive service interventions.⁵⁸ Data were not available for all of the schemes analysed with regard to the satisfaction of users and beneficiaries. However some insights emerged during interviews. Overall, there is little evidence to support the notion that SOC schemes respond to users' needs better than TF models. In some cases, schemes surveyed user satisfaction, yielding positive results – but it is hard to say if this represented added value compared with TF, since many of these surveys do not compare user satisfaction with a SOC scheme against its TF equivalent. Such an analysis was conducted for the JobPath PbR programme. The results showed that slightly more than half of JobPath participants found the services provided by the contractor better than those of a similar, traditionally financed programme, Intreo.⁵⁹ Participants noted in particular that the provider's premises – and therefore the overall engagement – were more welcoming and comforting than the premises of the public programme, Intreo.

In other SOC schemes, issues relating to higher caseload, poor expertise on the part of staff, a tendency to overlook services that were not related to outcome measurement, and cuts to costs due to the competitive bidding process, proved detrimental to the quality of services. At interview, the providers in the Drug and Alcohol Pilots mentioned that they had to reduce the staff to patient ratio due to the PbR mechanism (they could not employ more people since they were not sure that they would achieve the expected targets and receive payment). Also, given the financial uncertainty of the PbR mechanism, they were forced to offer fixed-term contracts, which are less acceptable to experienced staff. Less experienced staff may in turn have affected the quality of the service.

In most schemes where the same type of intervention was provided under a SOC and a TF model, no relevant differences emerged concerning the satisfaction of users (e.g. Eleven Augsburg SIB). In this regard, during interviews some providers noted that user satisfaction

⁵⁸ Dimitrijevska-Markoski, T. (2016). Social impact bonds: a new tool for governance of social programs-evidence from the UK, USA and Australia. International Journal of Public Policy; Cabinet Office (2016). Guidance Social Value Act: information and resources. Retrieved from: [https://www.gov.uk/government/publications/social-value-act-information-and-resources](https://www.gov.uk/government/publications/social-value-act-information-and-resources/social-value-act-information-and-resources)

⁵⁹ DEASP 2017, 54.

was more closely linked to the ability of providers to address users' needs more than to the financial model. For instance, the provider of the Benevolent Society SBB noted: "Some people think it is better when the programme is funded with outcome-based instruments and others [that] it is going to be the same; what drives how effective a service is, it's the manager's focus on the delivery the intervention." The Transforming Rehabilitation PbR was the only case in which the quality of services provided under the SOC scheme was clearly perceived by users as worse than under the previous TF scheme.

Service innovation

Service innovation is one of the main results expected when developing a contract, particularly on the side of the commissioning bodies and investors. Indeed, in many of the schemes analysed, various stakeholders recognised that a certain degree of innovation was developed thanks to the contractual arrangements of SOC schemes. For instance, in the Troubled Families programme, all previously provided services were reconfigured to adopt a 'whole family' approach.⁶⁰

In some schemes, the **main innovation related to a focus on new outcomes** that had not previously been considered by traditionally financed interventions. The Drug and Alcohol Recovery Pilots PbR, for instance, led providers to develop new approaches by expanding the range of services offered and strengthening their focus on full recovery. Similarly, JobPath focused on helping programme beneficiaries to sustain jobs – an area overlooked by similar TF programmes such as Local Employment Services.

In some schemes, however, the expected level of innovation in terms of services delivered was not fully achieved. The investors in Buzinezzclub, for example, stated that while SIBs are useful financial instruments for testing innovative interventions, they doubted the actual contribution made by that specific SIB to innovation in social service provision. In PL Pathways, commissioners expected the PbR model to yield completely new services for programme beneficiaries, hence they gave providers the freedom to deliver whatever services they saw fit. This 'black box' approach did not yield the expected results in terms of innovation (see Section 6.1). According to PL Pathways service providers, the potential to innovate was limited by the funding available, and most prime providers were not in a position to spend funds on service innovation during the programme.⁶¹ In other words, they highlighted that the potential for innovating services in SOC schemes is linked to the actual resources made available to the providers.

4.3. Scalability and replicability

Information gathered during the study allowed us to assess under what conditions social interventions implemented through the SOC model may be replicable and scalable:

- We define **scalability** as the possibility to benefit a broader set of beneficiaries (service expansion) and to foster policy and programme development on a lasting basis (institutionalisation).⁶²

⁶⁰ Interview with the provider

⁶¹ Hudson et al. (2010). The Influence of Outcome Based Contracting in Provider-led Pathways to Work, DWP Research Report 638, HMSO, Norwich.

<https://webarchive.nationalarchives.gov.uk/20130314010800/http://research.dwp.gov.uk/asd/asd5/rports2009-2010/rrep638.pdf>, 2, p. 45.

⁶² World Health Organization (2009). Practical guidance for scaling up health service innovations. World Health Organization. Retrieved from: https://apps.who.int/iris/bitstream/handle/10665/44180/9789241598521_eng.pdf; Carter B, Joshi A, & Remme M. (2018). "Scaling Up Inclusive Approaches for Marginalised and Vulnerable People" K4D Emerging Issues Report. Brighton, UK: Institute of Development Studies; Fixsen A. (2013). "Monitoring and Evaluating Scaling up of health system interventions: Theory and Practice." Draft manuscript. Retrieved from: <https://irh.org/resource-library/theory-and-practice-monitoring-evaluating-scale-up-of-health-system-innovations/>

- **Replicability** refers to the possibility to implement the intervention in different geographical contexts.

We first outline which of the analysed cases were scaled and/or replicated. We then discuss the lessons learned from these processes and provide some illustrations of the scaled and replicated programmes. To briefly sum up the findings, four schemes (all SIBs) have been replicated in other geographical areas, and seven have been scaled or extended (four SIBs; three PbR schemes). Interestingly, four schemes have been scaled or replicated using a TF mechanism rather than a SOC design.

The analysis revealed that replication was enabled by strong endorsement from the commissioning agency and the presence of an expert intermediary able to develop a hallmark scheme in terms of activities, governance, measurement and contractual arrangement. Conversely, the main challenge to replication and scaling perceived by stakeholders was the need to adapt the design of the SOC scheme to the context in which it was being replicated, causing additional effort and costs.

4.3.1. Replicated and/or scaled SOC cases

A number of SOC schemes were scaled and/or replicated after they were implemented using an outcomes-based contract (see the table below). Four have been replicated in other geographical areas, while seven have been scaled or extended.

Table 11. SOC schemes scaled and/or replicated

Scheme	Scaled	Replicated
Perspektive:Arbeit	Not scaled	The programme was subsequently funded via traditional financing in other Austrian states.
DUO for a JOB	The same intervention provided under the SIB was scaled using a traditionally financed model. DUO also scaled its operations in Brussels.	DUO for a JOB expanded its activities to other cities.
Buzinezzclub SIB	Extended to a third year. Moreover, the service provider won a contract via public procurement to implement the same intervention.	The intervention was replicated in Utrecht and Eindhoven through a SIB.
Provider-led Pathways to Work	The Pathways to Work Programme was initially implemented in seven districts through Jobcentre Plus, and later scaled across the UK.	Not replicated
Mental Health and Employment Partnership	The SIB in Haringey was extended, and another SIB started in May 2019. It will last until April 2023.	The SIB has been replicated in three London boroughs: Barnet, Camden and Enfield.
Drug and Alcohol Recovery Pilots	In some of the pilot areas, the PbR model was maintained, although the contract and design of the schemes changed.	Not replicated
Troubled Families	The number of beneficiaries and LAs involved in the programme increased between phases I and II. The programme has been extended for one additional year.	Not replicated
Benevolent Society Social Benefit Bond	The programme was extended for two years using a traditionally financed model rather than an outcome-based one.	Not replicated

Of the eight SOC schemes that have been scaled or replicated, four (Benevolent Society Social Benefit Bond, DUO for a JOB, Buzinezzclub SIB and Perspektive:Arbeit) have been **scaled or replicated using a TF mechanism rather than an SOC design**. In some cases, such as DUO for a JOB and the TBS SBB, the SOC scheme was explicitly intended as a means to test the intervention, thus the scaling using a TF model was a natural result of the successful intervention. The commissioners interviewed also noted that SIBs entail additional costs (see Chapter 5), so it may be cheaper to replicate an intervention using traditional financing after it has been proven to work. In cases where the SIB mechanism was retained when the programme was replicated (e.g. Buzinezzclub, MHEP), the intervention was applied to different geographical locations and involved different commissioners who wanted to make sure that the intervention would work equally well in a different context.

4.3.2. Lessons learned about scaling and replication

Interviewees suggested that several **lessons learnt** have been applied in scaling or replication efforts.

The first is that **vigorous support is needed from the commissioning agency**. The support and endorsement of the public commissioner encourages the active involvement of different stakeholders and a high level of cooperation between them, allowing the intervention to be scaled/replicated more successfully. Furthermore, the involvement of the public actor allows the intervention to be integrated more effectively with existing procedures and services.

The second lesson is that it was easier to engage stakeholders to replicate the model when a hallmark scheme outlining the intervention's activities, governance structure, management process, templates for contractual agreements and the outcome evaluation infrastructure was designed during a piloting experience by an expert intermediary (as was the case in DUO for a JOB and MHEP SIB).

The main **challenge** mentioned by stakeholders in replicating the social interventions is the **highly contextual nature of social services**. The delivery of social services often needs to be adapted to the specific needs of users, which vary from place to place. This is a general issue, regardless of the funding model supporting the intervention. However, it also has a specific implication for SOC: when a social intervention is replicated, its financial model and measurement process also need to be adapted (e.g. it is not always the case that the same portion of payments can be paid using an outcomes-based approach). For instance, the defining of metrics is highly dependent on the types of service provided, the time of delivery, the types of measurement process, the availability of data and the needs and expectations of the specific stakeholders involved. Thus, it is rarely possible to simply replicate the same SOC design in different contexts.

The three most interesting examples of replication and scaling practices – the MHEP SIB, DUO for a JOB SIB and the Drug and Alcohol Recovery Pilots PbR – are summarised in the boxes below. In each of these cases, the programme was either scaled or replicated using an SOC design, traditional funding or a mix of SOC and TF funding.

Replication of the Mental Health and Employment Partnership SIB

The MHEP SIB was first applied between 2016 to 2019 in Tower Hamlets, Haringey and Staffordshire. Officially, the SIB in Haringey was extended from April 2017 to April 2019 from about 250 up to 1,000 adults (though both programmes in Haringey and Tower Hamlets were still running as of February 2020, according the report issued by the Behavioural Insights Team). Another SIB began in May 2019, which is scheduled to last until April 2023. This concerns the delivery of IPS services, but is supported by a different source of national funding, the Life Chances Fund. In addition, the SIB has been replicated in three further London boroughs: Barnet, Camden and Enfield. This

five-year SIB will support the delivery of Individual Placement Support services to assist up to 2,632 adults.

Our analysis suggests that the first three SIBs helped the commissioner gain some confidence from the fact that other outcomes-based contracts were already in operation. In addition, the intermediary in the MHEP scheme developed the specification for the Invitation to Tender, the outcomes payments structure and the performance management infrastructure, which represents a toolbox ready to be used by other commissioners. However, according to the intermediary, it was challenging to standardise the contract because the requirements of the commissioners differed between locations.

Replication and scaling of DUO for a JOB

The same intervention provided under the DUO for JOB SIB was scaled using a TF model because the SIB was seen as an instrument to test whether or not the intervention was effective. Following the SIB, Actiris allocated DUO a first 'bridging' subsidy for two years (2017-2019) on the basis of the preliminary results of the SIB. The subsidy was given because DUO had delivered good results, and the organisation had to wait two years to apply to the quinquennial call for tenders (for traditional procurements). In 2019, DUO acquired funding from Actiris for an additional five years (until 2024) after a regular tendering procedure. The TF intervention was scaled up both in terms of the funding provided by charitable organisations and the number of recipients served.

Replication of the SIB in other countries is currently under discussion. According to interviews, opportunities may exist to replicate the SIB in Paris and Rotterdam. The SIB model is seen as a way to test the same intervention and demonstrate its impact in different institutional contexts.

Scaling of the Drug and Alcohol Recovery Pilots

The Department of Health piloted a PbR scheme addressing drug and alcohol recovery in eight local administrative areas in the UK. After the pilots were completed, the PbR model was continued in some of the pilot areas, although the contract and design of the schemes changed. The PbR approach was retained because the stakeholders acknowledged the benefits of having an outcomes measurement system in place. In Lincolnshire, for example, the new programme tracks group outcomes for cohorts of participants rather than for each individual, and uses the National Drug Monitoring database to monitor the performance of substance misuse services across the country. This substantially reduces the administrative burden on both the commissioner and the providers.

4.4. Summary of findings regarding SOC effectiveness

To compare the effectiveness of SOC and TF programmes, we assessed the extent to which each programme was able to achieve the outcomes set at the start of the programme. These included the number of participants employed, reductions in re-offending rates, the number of children who avoided out-of-home care, and other outcomes. In terms of the target outcomes achieved, there is insufficient evidence to claim that outcomes-based contracts are more effective than traditional financing or vice versa. This is partly because outcome targets were not set in the majority of comparable TF programmes.

Nevertheless, our analysis shows that the achievement of targets alone was not enough to determine the overall effectiveness and success of a scheme. In fact, as illustrated in Section 4.1, some schemes that did not achieve their targets were nevertheless considered successful, since they were able to address the needs and expectations of the various stakeholders involved. Thus, they were extended or replicated.

A crucial issue concerning the effectiveness of SOC schemes was the presence of perverse incentives (i.e. creaming and parking) because they reduced the quality of services for those who were harder to help. Aside from perverse incentives, the financial uncertainty that was transferred to providers reduced the effectiveness of some of the PbR models analysed, compared with TF.

With regard to the main benefits of SOC and their potential to enhance innovation in service provision, our analysis shows that SOC schemes – particularly those in the form of SIBs – were successfully used to test innovative interventions that were not previously part of welfare provision (see Section 4.2.1). However, the ‘black-box’ approach applied in some of the PbR programmes analysed did not meet commissioners’ expectations in terms of creating completely new and better services for beneficiaries. Stakeholders highlighted that the potential to innovate a service was linked to the actual resources that were made available to the providers (see Section 4.2.3).

Other key benefits of SOC identified in the analysis include: the development of a measurement infrastructure that enabled a culture of outcomes measurement across the organisations involved and provided evidence to support policymaking; the potential of SOC to unlock financial resources; and enhanced flexibility over spending for service providers.

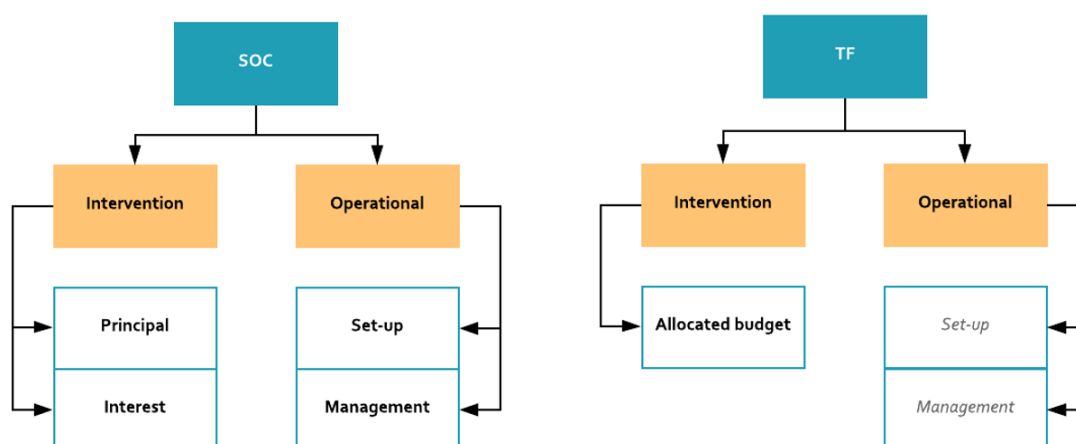
Lastly, although the need to coordinate numerous different actors may entail some drawbacks in terms of management complexity, the involvement of multiple actors in SOC proved to be beneficial in some cases, in terms of developing new collaborations, getting to know potential new partners for future projects, and sharing best practices.

5. Efficiency

As discussed in Chapter 4, SOC schemes can sometimes deliver additional benefits that TF models may be less able to bring. However, little information exists about the costs associated with running SOC schemes. To that end, wherever possible we have **assessed the level of these costs** for SOC schemes and their TF equivalents. Where information was limited, we supplemented it with broader insights on the differences in operating costs between SOC and TF schemes, gained from interviews and desk research. Please note that comparisons between individual SOC schemes and their TF counterparts are available in Annex 1. In this chapter, we present the findings that stem from the investigation of all 15 cases. A summary of the findings is presented at the end of the chapter.

In order to clarify what costs we considered, we present a side-by-side comparison of the various costs associated with SOC and TF schemes in the figure below. We mostly looked at the costs from the perspective of the **commissioner**, as the main goal was to provide information for public sector bodies on the costs to consider when choosing to finance an intervention through traditional procurement or via social outcomes contracting. Interviews with service providers did, however, also yield valuable insights. We analytically separated the **cost of delivering interventions** from the **additional operational costs** required to run a particular financing model. Each of these is discussed and compared with TF schemes in separate sections below. After completing this comparison of the costs of SOC schemes and their TF equivalents, we discuss important decisions in their design, including design features that support third-sector organisations and social economy enterprises.

Figure 4. Operationalisation of SOC and TF costs



Source: PPMI.

Note: payment of interest only applies in the case of SIBs.

5.1. Intervention costs

In this section, we focus on the intervention costs of SOC and TF schemes. As outlined in the figure above, intervention costs for SIBs include the principal investment as well as interest (if this is paid out). For PbRs, intervention costs are represented by the total amount of payments made by the commissioning department to the contracted service provider. We defined the intervention costs for TF programmes as the budgets allocated to run them. However, depending on the information available for each scheme, at times we considered actual spending instead. Details on what constitutes intervention costs are provided in Annex 1 under each SOC scheme.

In our analysis, we focus on two key parameters:

- We first estimated the **total intervention cost per beneficiary**. This cost shows which of the two programmes received greater funding in relative terms, which may have directly affected its effectiveness. This cost was calculated by taking the total funding available for the programme and dividing it by the number of beneficiaries.
- We then estimated **the cost per outcome achieved**. This variable shows which programme delivered greater value for money. The cost was calculated by taking the total programme funding and dividing it by the number of beneficiaries who achieved the programme outcomes.

Please note that comparisons were conducted only where outcomes were similarly defined and measured for both the SOC and the TF intervention. Where the programme costs, the number of beneficiaries served, or the number of successful beneficiaries were not available for the entire duration of the programme, we derived estimates from looking at the funding allocated and the participants included during a specific time period. Details on the sources of information, as well as the calculations made to arrive at each estimate presented below, are detailed in Annex 1.

These estimates are subject to a number of limitations, which are detailed under each section in Annex 1. Most notably, **we cannot claim that the cost differentials observed below are because of the different funding models** (SOC vs TF). This is because we were not able to control for participant characteristics, differences between the interventions, and other factors that might have influenced outcomes as much as the funding model. Instead, we show whether the analysed SOC programmes appear to have been more efficient compared with similar TF interventions, without claiming causal attribution.

The table below presents the summary of the cost per participant and the cost per outcome achieved for each SOC scheme and its TF equivalent. To aid readability, SOC-TF pairs have been shaded in the same colour. Please note that outcomes vary between different programme pairs. For example, Eleven Augsburg and AMA-Zukunft measure the cost per job achieved by programme participants, whereas the figures for the ABLE and RESTART programmes represent the cost per participant who completed each programme. Nevertheless, outcomes are largely similar within each SOC-TF pair (the table notes 'not comparable' where this is not the case). For details on the outcomes considered in the estimates below, please see Annex 1.

Based on the information regarding cost per outcome presented in Table 12 and Table 13, the SOC schemes fall into four groups:

1. SOC schemes whose cost per outcome achieved was **higher** than that of the TF equivalent;
2. SOC schemes whose cost per outcome achieved was **lower** than that of the TF equivalent;
3. SOC schemes whose cost per outcome achieved was **similar** to that of the TF equivalent; and
4. SOC schemes for which **information is insufficient** to make a valid comparison.

Table 12. Costs per participant and per outcome for SIB and TF schemes

#	Type	Programmes compared	Cost per Participant	Cost per Outcome	Notes
1	SIB	Eleven Augsburg (2013-2015)	EUR 3,732	EUR 12,875	According to the SIB service provider, the SIB funding did not cover project costs and the provider cross-subsidised around EUR 48,000 from other funding streams. This is not included in the estimates because providers might cross-subsidise in TF programmes as well. If it were, it would increase the cost per participant as well as the cost per outcome of the SIB relative to the TF programme. Intermediary costs in the SIB are also excluded because they are discussed separately in Section 5.2.1.
	TF	AMA-Zukunft (2015-2019)	EUR 5,505	EUR 11,981	
2	SIB	Perspektive:Arbeit (2015-2018)	EUR 2,587	EUR 15,475	Given that the SIB failed to achieve its goals, the investors were not repaid, so the cost to the taxpayer (or to the commissioner) is equivalent to zero, excluding the costs of setting up and evaluating the intervention.
	TF	Same intervention (2018-ongoing)	Unknown	Unknown	
3	SIB	DUO for a JOB (2013-2016)	EUR 5,023	EUR 9,477	The 2013-2016 period includes not only the funding received via the SIB, but also traditionally from other funding sources (e.g. charitable organisations), because outcomes and outputs were tracked for the organisation as a whole, without being separated according to funding stream.
	TF	DUO for a JOB (2017-2018)	EUR 3,450	EUR 9,400	
4	SIB	Buzinezzclub (2014-2016)	EUR 7,003	EUR 12,313	N/A
	TF	Buzinezzclub (2018-2021)	Unknown	Unknown	
5	SIB	BOAS Werkt (2016-2017)	EUR 16,667	EUR 28,947	The exact costs of the SIB are not publicly available, so the estimates are based on the assumptions outlined in Annex 1. The TF programme cost is based on staff salaries that amount to 2.5 full-time equivalents. Note that activities in the two programmes differ, which impacts on their costs.
	TF	Transfer point (2019- 2020)	Unknown	EUR 20,717	
6	SIB	ABLE (2013)	USD 1,633	USD 14,814	Given that the SIB failed to achieve its goals, investors were not repaid, so the cost of the SIB to the taxpayer (or to the commissioner) is equivalent to zero, excluding the costs of setting up and evaluating the intervention.
	TF	RESTART (2017-2018)	USD 753	USD 1,427	
7	SIB	Benevolent Society SBB (2013-2018)	AUD 30,261	Unknown	N/A
	TF	Intensive Family Preservation (2014)	AUD 40,617	Unknown	
8	SIB	MHEP (2017-2019)	GBP 1,200	GBP 4,300	

#	Type	Programmes compared	Cost per Participant	Cost per Outcome	Notes
	TF	Camden IPS (2016-2017)	GBP 1,254	GBP 4,933	The figures for the SIB are based on the funding received by providers, to make them comparable with Camden IPS' figures. The true costs of MHEP SIB are slightly higher because some providers incurred losses, meaning that they cross-subsidised from other streams.

Source: Compiled by PPMI. For details on the sources and estimates for each intervention, please see Annex 1.

Note: Figures have not been adjusted for inflation.

Table 13. Costs per participant and per outcome for PbR and TF schemes

#	Type	Programmes compared	Cost per Participant	Cost per Outcome	Notes
9	PbR	JobPath (2015-2018)	EUR 780	Not comparable	The assessment is based on JobPath's evaluation.
	TF	Local Employment Services, Intreo	Higher than JobPath	Not comparable	
10	PbR	Provider-led Pathways (2007-2011)	GBP 451 for both programmes	GBP 2,942. Costs are similar for both programmes	The assessment is based on the National Audit Office (NAO) analysis.
	TF	Jobcentre Plus Pathways (2003-2011)			
11	PbR	Work Programme (2012-2019)	GBP 1,579	Not comparable	The estimate is based on the fact that the Work Programme cost GBP 3 billion and served 1.9 million participants.
	TF	NDYP and ND25pl (1998-2009)	Unknown	Not comparable	
12	PbR	Youth Contract (2012-2016)	GBP 1,416	GBP 3,004	The YC estimates are based on assumptions outlined in Annex 1. The YC figures are likely overestimated. Sustained re-engagement outcomes were measured differently between the two programmes: YC participants had to be enrolled in school, work or an apprenticeship for five out of six months, whereas information for AA Pilots indicates how many youth were in employment or education three months after the programme.
	TF	Activity Agreement Pilots (2006-2010)	GBP 3,004	GBP 4,331	
13	PbR	Drug and Alcohol Pilots (2012-2013)	Higher than TF	Unknown	Assessment is based on programme evaluation.
	TF	Non-pilot areas (2012-2013)	Unknown	Unknown	
14	PbR	Transforming Rehabilitation (2015-2016)	GBP 5,245	Unknown	The number of participants used in the Probation Trusts estimate reflects the 2012 calendar year rather than the 2012-2013 financial year.
	TF	Probation Trusts (2012-2013)	GBP 3,794	Unknown	

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

#	Type	Programmes compared	Cost per Participant	Cost per Outcome	Notes
15	PbR	Troubled Families, phase II (2015-2021)	Unknown	Unknown	N/A
	TF	Earned Autonomy Model (2018-ongoing)	Unknown	Unknown	

Source: Compiled by PPMI. For details on the sources and estimates for each intervention, please see Annex 1.

Note: Figures have not been adjusted for inflation.

Starting with the SIBs, the two interventions with a cost per outcome higher than that of their TF counterparts were Eleven Augsburg and BOAS Werkt (see Table 14). Each of these programmes cost more than its respective pair for a different reason. According to one of Eleven Augsburg's service providers, who is also involved in the AMA-Zukunft project, the service delivered across the two interventions is very similar, which explains why the cost difference per outcome achieved is relatively small.⁶³ Eleven Augsburg's cost per outcome achieved is nevertheless higher likely because of the interest associated with paying back the investors, which amounted to 3% of the investment, as well as to the greater number of social service providers involved. It is important to note that **interviewees saw the greater cost of the SIB as acceptable because its main purpose was not to save money, but to test out an intervention among a group of people who have previously not been effectively helped.**

In contrast to Eleven Augsburg, the cost of BOAS Werkt is substantially greater than that of Transfer Point. This is in part because the two interventions are somewhat different: while both help unemployed Dutch residents to find jobs in Germany, the BOAS Werkt intervention is more intensive, involving a short-term placement in Germany; Transfer Point, meanwhile, was designed mostly to help jobseekers overcome the administrative burden of working in Germany. Nevertheless, even taking the different nature of the interventions into account, **the SIB's costs per participant and per outcome achieved were substantial because few participants overall joined the programme.** Importantly, the municipality was contractually obliged to recruit most of the participants (40 out of 55) each year. Therefore, in contrast to the other failed SIBs discussed below, the investors are likely to receive at least part of their funds back (the financial settlement is still under negotiation).

Table 14. SOC costs per outcome in comparison to TF equivalents

	SIB	PbR
SOC cost per outcome higher than TF	Eleven Augsburg BOAS Werkt	-
SOC cost per outcome lower than TF	Perspektive:Arbeit ABLE MHEP SIB	Youth Contract
SOC cost per outcome similar to TF	DUO for a JOB	Provider-led Pathways
Insufficient information to compare cost per outcome	Buzinezzclub Benevolent Society	JobPath Work Programme Drug and Alcohol Pilots Transforming Rehabilitation Troubled Families, phase II

Somewhat counterintuitively, we classified the outcome costs of the Perspektive:Arbeit and the ABLE SIBs as being lower than those of their TF pairs. Table 12 shows that the costs of Perspektive:Arbeit's TF equivalent are unknown (suggesting that the comparison was not possible), while the ABLE SIB's cost per outcome appears ten times greater than that of the RESTART programme. We nevertheless classified both SIBs as being cheaper **because they both failed** to achieve their target outcomes. In both of these cases, the investors were not issued any payments, so the programmes **cost nothing to the taxpayer** (for both of these SIBs, Table 12 illustrates the cost to the investor rather than the commissioner – see the notes next to each SIB in the table). Cases such as these illustrate the true benefits of SIBs. First, they allow the government to test out a new intervention with little risk; and second, if the programme is found to be ineffective, it can easily be shut down. According

⁶³ AMA-Zukunft probably costs more per participant because service providers in this programme also receive additional funding for administrative expenses. No such funding was provided under the SIB.

to the investor involved in the ABLE SIB, this contrasts sharply with many other government programmes that continue to run despite a lack of evidence that they are effective.

Finally, the cost per outcome for DUO for a JOB during the period 2013-2016 and in its subsequent years (which we use for TF comparison) was almost identical. This may be partly because the outcomes during the years 2013-2016 were tracked for the organisation as a whole, regardless of the funding source, and SIB funding consisted only one-fifth of the total funding the organisation received. However, it may also reflect **a theme reiterated by multiple service providers across the SIBs analysed: that their service approach did not change, whether they were funded traditionally or through an outcomes-based contract.** This was because in both cases it was the investors, rather than the service providers, that bore the risk if the intervention failed.⁶⁴ This contrasts with the way service providers acted in PbR arrangements, to which we now turn our attention.

Among the PbR schemes analysed, we did not encounter any cases in which an outcomes-based contract cost more per outcome than its TF equivalent. Importantly, this might be the case because the cost information for most PbR schemes was insufficient to allow a valid comparison. Even where the cost per outcome could be estimated for both the PbR programme and its TF pair – as in the case of JobPath – the outcomes were measured differently,⁶⁵ making the comparison invalid. Nevertheless, **a common theme across the interviews with both the commissioners and the service providers of PbR schemes was that a focus on outcomes drove organisations to eliminate inefficient practices,** because if the service providers failed to achieve the target outcomes, they would not get paid.

In some cases, ‘inefficient practices’ might mean engaging participants in activities that do not actually work—and thus, dropping such activities would be a welcome benefit of a PbR contract compared with traditionally-financed programmes. Nevertheless, multiple evaluators who were interviewed cautioned that providers might perceive working with those who were hardest-to-help as being inefficient, because these participants are the least likely to achieve the target outcomes. Such perceptions could result in parking and creaming. Therefore, **the perceived efficiency of PbR contracts might come at the cost of a lower quality of service for some groups of participants.**

The Department for Work and Pensions (DWP) in the UK – one of the most notable commissioners of PbR schemes – has attempted to **mitigate this risk by experimenting with different payment arrangements.** In the Work Programme, for example, providers received greater payments for helping those with mental or physical disabilities to find jobs compared with other groups among the unemployed. Nevertheless, **this did not completely prevent parking or creaming.** This is in part because providers had to deliver services to nine different groups of participants, which made it difficult for providers to ensure that each group could receive the specialist services it needed. Moreover, the programme was mandatory for many participants, who did not think they could work due to their disability or personal situation, so advisers might have found it more difficult to engage participants who did not want to participate in the programme.

Among the cases analysed, the overall cost of SIBs to the taxpayer was lower than that of their TF equivalents only when the programme failed and, consequently, investors were not repaid. In other cases, the cost was greater or similar – yet this greater cost was seen as justified by the stakeholders involved. By contrast, where the costs per outcome of PbR schemes could be compared to those of TF, they appear to have been similar or lower. This difference can be explained by the different reasons for choosing SIB and PbR contracts (see Chapter 2): whereas SIB commissioners often quoted the need to test out an

⁶⁴ Nevertheless, there are exceptions to this rule if service providers also invest funds of their own. For example, see the Buzinezzclub and BOAS Werkt SIBs, described in detail in Annex 1.

⁶⁵ JobPath measured employment sustained for at least 13 weeks, whereas the Irish Local Employment Services measured jobs achieved regardless of how long they were sustained for. Therefore, although the cost per outcome appears higher in JobPath, it is likely that the cost per outcome in LES would also increase if it measured the same outcome.

innovative intervention, thus justifying added cost in the form of interest, PbR commissioners more often cited the desire to achieve value for money.

5.2. Operational costs

SOC interventions are characterised by contractual complexity, multi-layer governance, and complex monitoring and evaluation activities, absent from many TF schemes. Therefore, in addition to comparing SOC intervention costs, we also considered the operational costs of SOC schemes. As shown in Figure 4, we define operational costs as:

- the cost of setting the intervention up;
- the cost of managing it. This includes work by the commissioner to supervise the work of the service provider, coordinate between different parties involved, etc.

In order to collect this information, we engaged in desk research and interviews. Across most of the schemes, we found that the working hours spent on designing and monitoring the intervention from the commissioner's side were not systematically tracked. Commissioners were therefore unable to provide exact estimates, preventing us from benchmarking these costs. Nevertheless, commissioners were able to evaluate whether the workload was higher or that that associated with the TF programmes they usually work on. In the following two sections, we draw on this information to show which programmes were considered more costly to **set up** and **manage** than similar TF programmes, and discuss why. In each section, we provide some information on the exact fees that we were able to identify, in order to illustrate what share of intervention costs they represent.

Please note, however, that SIBs and PbR programmes also entail additional evaluation costs. We do not present information on evaluation costs for two reasons. First, mostly with regard to PbR schemes, external evaluations were conducted to assess non-payment-related outcomes only, and so their cost in relation to TF programmes would depend on the methodology used to evaluate the two programmes. In other words, if both the PbR and TF programmes are evaluated using the same method, their evaluation costs will be similar because of the method used, rather than because of the programme funding model. Second, payment-related outcomes in PbR schemes were mostly verified by the commissioner, so we discuss these costs together with the management costs. Lastly, in the case of both SIBs and PbR schemes, where evaluations were outsourced, the exact costs could not be shared with the research team due to commercial sensitivity.

5.2.1. Comparison of set-up costs

In the following table, we categorise the programmes analysed according to their set-up costs in relation to similar TF programmes, as well as the type of funding (SIB vs PbR).

Table 15. SOC set-up costs in comparison to TF programmes

	SIB	PbR
SOC set-up costs higher than TF	Eleven Augsburg Perspektive:Arbeit DUO for a JOB Benevolent Society SBB	-
SOC set-up costs lower than TF	-	Youth Contract
SOC set-up costs similar to TF	MHEP SIB	Troubled Families
Insufficient information to compare set-up costs	Buzinezzclub BOAS Werkt ABLE	JobPath Provider-led Pathways Work Programme Transforming Rehabilitation

Source: Compiled by PPMI, based on commissioners' impressions stated during interviews.

Of the SIBs about which comparisons can be made, most appear to have involved greater set-up costs than the other programmes with which the commissioners interviewed typically work. This was partly because **SIBs involved a more complex legal structure** than traditional programmes, requiring legal fees. For example, in the Perspektive:Arbeit SIB, investors funded the SIB as a loan to the intermediary and not as a grant, although the latter option is more typical. This choice was linked to tax implications for the investors in the event that the intervention failed: if the investment was given as a loan, it could be written off in the event of failure; if the investment were treated as a grant and the programme failed, the investors would still be subject to tax. The investors' decision to treat the investments as loans meant that some of the interested investors could no longer participate in the SIB, because their charitable tax status did not allow them to provide loans. Please also see the box below regarding the legal challenges encountered in setting up the Eleven Augsburg SIB.

Legal issues pertaining to Eleven Augsburg

Several legal issues arose with regard to Eleven Augsburg SIB in Germany. In particular, these concerned:

- Which financing instruments provided in the German Social Code to apply. One option was to use public funding in accordance with public law. In this case, it must have been demonstrated beforehand that the multi-stakeholder arrangement involving an intermediary and an evaluator was necessary to generate impact for the intended target groups. Moreover, public bodies usually do not cover the full costs of services that go beyond the basic supply. This made it difficult to justify the full coverage of costs and the interest payment. The other model – which the Eleven Augsburg commissioner opted for – is a service agreement under civil law between the public body and the intermediary. The latter then enters into sub-contracts with the investors and social service organisations. Under this option, the contract defines the results that must be achieved, as well as duties such as information obligations and permit requirements, and the payment mechanism. For the social services provided under SIBs, however, the contract leaves a high degree of flexibility.
- The budgetary treatment of SIBs. German budgetary law requires that potential savings arising from a novel mechanism must be demonstrated *ex ante* (the principle

of economy). The Eleven Augsburg team achieved this in collaboration with the Court of Auditors.

- The German law on foundations. This law requires a risk-averse investment strategy that preserves the capital endowment. This makes mission investing in SIBs, (i.e. investments within asset management) difficult, because of the considerable risk of default in SIBs. The other option would be investments from operational resources. These, however, might conflict with restrictions on commercial business activities. In Eleven Augsburg, the foundations involved used both opportunities.

Source: Scheuerle & Nieveler 2017, p. 5-7, who drew on Fliegauf et al., 2015.

Generally, SIBs included more parties than traditional programmes, so **negotiations also took longer, in particular with intermediaries** as the task-division between the commissioner and the intermediary needed to be delineated. For example, negotiations between the Juvat/ Benckiser Foundation—which served as the intermediary in the Eleven Augsburg SIB—and the Bavarian State Ministry for Labor and Social Affairs, Family and Integration lasted between a year and a year-and-a-half, compared to a year to set up the comparable TF project AMA-Zukunft.⁶⁶ The set-up of Perspektive:Arbeit took about a year and a year-and-a-half as well,⁶⁷ and almost two years for MHEP SIB.⁶⁸ Furthermore, it took time to find the investors and agree on the terms with them as well. Some commissioners – as in the case of DUO for a JOB – opted to conduct a **feasibility study** prior to engaging service providers. In other cases such as Perspektive:Arbeit, the TF programmes that were used for comparison directly followed the SIB. Therefore, partners were already available, the programme activities were established, and the project structure was set, meaning that the set-up phase was relatively quick compared to that of the preceding SIB.

Although exact set-up costs were difficult to quantify, there was some information on the **intermediary fees**, which are relevant given that intermediaries can provide substantial help in finding investors, service providers, commissioners, and setting up the legal and outcome-tracking structure. Information was only available for three of the SIBs analysed (see the table below). In these cases, intermediary fees represented around 5% of the total investment, in line with other figures cited in the literature.⁶⁹ Please note that in the case of Buzinezzclub and Eleven Augsburg, the intermediary was involved not only in the set-up, but also in the coordination of the project's activities once the project began, whereas the MHEP SIB intermediary fees relate to activities during project initiation.

Table 16. Intermediary fees

SIB	Total investment	Intermediary fees	Share of total investment
Buzinezzclub	EUR 990,00070	EUR 40,00071	4%

⁶⁶ Jebabli, J. (2016). Social Impact Bonds: Collaboration Across Institutional Domains. Master Thesis. Vienna University of Economics and Business; and correspondence with TF commissioner.

⁶⁷ Millner, Reinhard, Üstün, Susan, Moder, Clara and Meyer, Michael (2019): Social Impact Bond Pilotprojekt "Perspektive:Arbeit". Wien: WU Wien.

⁶⁸ CBO (2016). Mental Health and Employment Partnership (MHEP) Social Impact Bond, p. 7.

⁶⁹ For example, see Gustafsson-Wright (2016). Using Impact Bonds to Achieve Early Childhood Development Outcomes in Low- and Middle-Income Countries. Brookings Institution. Retrieved from: <https://www.brookings.edu/wp-content/uploads/2016/07/Impact-Bonds-for-ECDweb.pdf>, p. 51, which show that intermediary fees account for roughly 4% of the total capital investment; and McKay, Kyle A. (2013). Evaluating Social Impact Bonds as a New Reentry Financing Mechanism: A Case Study on Reentry Programming in Maryland. Department of Legislative Services, Office of Policy Analysis, p. 6, which puts the figure at 8%.

⁷⁰ The figure represents total investment prior to interest payments. Eurocities (2016). Rotterdam Buzinezzclub Social Impact Bond. Retrieved from: http://nws.eurocities.eu/MediaShell/media/Feb16_BuzinezzclubRotterdam.pdf, p. 2, puts the figure at EUR 1,030,000, but this includes the EUR 40,000 fee. We subtracted this to make it comparable with the other schemes presented in the table.

⁷¹ Van Es (2015). Social impact bonds. Een nieuw instrument om jeugdwerkloosheid te bestrijden. Hogeschool Rotterdam.

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

SIB	Total investment	Intermediary fees	Share of total investment
MHEP SIB	GBP 2.9 million ⁷²	GBP 150,000 ⁷³	5%
Eleven Augsburg	EUR 250,000 ⁷⁴	0.75 full-time equivalent (FTE) in 2014, which decreased to 0.33 in 2015 and 2016. ⁷⁵	N/A

Source: Compiled by PPMI, based on the available information (see footnotes for sources and limitations).

Different types of arrangement were adopted for paying intermediary fees. In the Buzinezzclub SIB, for example, intermediary fees were included in the initial investment put together by the investors, who were subsequently repaid by the commissioner once the SIB turned out to be successful. In the MHEP SIB, one of the philanthropic investors – the Commissioning Better Outcomes (CBO) Fund – issued a grant to Social Finance to coordinate with local commissioners.⁷⁶ In the ABLE programme, intermediary work was funded through a separate grant from Bloomberg Philanthropies. In Eleven Augsburg, the intermediary invested its own resources to get the intervention started, in order to pilot the innovative funding model in Germany.⁷⁷

In contrast to SIBs, information about the set-up of PbR contracts was more limited. While commissioners stated that the oversight of PbR schemes was, in general, more strenuous than with TF programmes (see the next section), in terms of set-up, PbR schemes appear to follow similar design and tendering procedures to TF programmes (unless these are delivered in-house). The box below describes the example of the Work Programme's procurement.

Procurement Process for the Work Programme

"The procurement of the Work Programme took place between July 2010 and June 2011. This was a two-stage process where potential providers first bid to join DWP's Employment-Related Support Services Framework and then took part in 'mini-competitions' for Work Programme delivery within 18 contract package areas (CPAs)... The mini-competitions attracted 177 bids, with between nine and 17 bids in each CPA. Thirty of the 35 framework providers bid, 18 of which were successful. Many of these providers were successful in more than one CPA and there are two or three 'prime contractors' or 'primes' in each of the CPAs."⁷⁸

"With just six months between the Invitation to Tender and go-live, the Work Programme procurement process was substantially quicker than procurement of previous programmes. This rapid process, in particular the time between the award of contracts and go-live, was seen by providers (and DWP) as a significant

⁷² CBO (2016), p. 5.

⁷³ CBO (2016), p. 7.

⁷⁴ The figure represents total investment prior to interest payments. Juvat (2016). Abschluss erster deutscher Social Impact Bond: Das Modellprojekt in Augsburg. Fact Sheet 27.09.2016, p. 2.

⁷⁵ Interview with a consultant who was hired by the intermediary to set up and coordinate the project. Please note that the costs incurred by other staff within the intermediary are not included.

⁷⁶ CBO (2017). Midpoint in-depth review report produced as part of the CBO Fund Evaluation. Retrieved from: https://www.tnlcommunityfund.org.uk/media/research-documents/social-investment/commissioning_better_outcomes_in_depth_review_190320_122442.pdf?mtime=20190320122441

⁷⁷ Interview with the intermediary.

⁷⁸ Lane, Pippa, Rowan Foster, Laura Gardiner, Lorraine Lanceley and Ann Purvis (2013b). Work Programme Evaluation: Procurement, supply chains and implementation of the commissioning model. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/197710/rrep832.pdf, 14.

achievement, but also as a pressure on start-up. Particular issues included difficulties in securing staff and premises in areas where the prime had not delivered before. For some providers, legal issues surrounding the Transfer of Undertakings (Protection of Employment) (TUPE) was a major and complex issue that added to the pressure.”⁷⁹

Source: Lane et al. 2013a and 2013b.

Note: for more information on TUPE, please see the Work Programme in Annex 1.

One important lesson is that the set-up of PbR schemes becomes less costly as the commissioner acquires more experience in working with such schemes. Youth Contract’s commissioner, for example, stated that setting up the programme was easier than either the previous TF intervention (Activity Agreement Pilots) or the preceding Work Programme, because the commissioning authority borrowed concepts from these interventions.⁸⁰ For example, YC contractors were required to adopt the Merlin Standard developed under the Work Programme to ensure the quality of services across their supply chains.⁸¹

Arguably, SIB set up costs could also decrease over time as commissioners become more familiar with the funding mechanism. This is likely to be the case if SIBs are implemented in the same areas of social intervention, and outcomes can be defined and tracked in the same way as in the initial SIBs. In other words, set up costs would decrease because commissioners would incorporate the lessons learned from initial experiences with SIBs. Nevertheless, given that SIBs are often used to test innovative interventions, it remains to be seen whether the lessons from initial SIBs can be incorporated into subsequent SIBs and whether this reduces their set up costs. We were not able to assess this aspect in our study because most of the SIBs studied were the first to be implemented in their respective countries (see Chapter 2).

5.2.2. Comparison of management costs

According to interviewees, **most of the SIBs and PbR schemes analysed in this study demanded greater resources to implement on the part of the commissioner, compared with similar TF programmes** (see the table below). In most cases, management costs were higher because commissioners adopted **more intensive outcome tracking and verification methods** that were not previously used in similar TF programmes. Even in SOC-TF pairs that tracked exactly the same outcomes – for example, the Troubled Families programme, which was delivered through both outcomes-based contracts and TF – the reporting burden was greater for the PbR contract because outcomes were subject to a higher auditing standard, given that they triggered payments. Nevertheless, some commissioners argued that greater attention to performance monitoring also improved the results of the programmes.

Implementation costs were further increased by the need to **re-negotiate contracts** with service providers **when participant volumes were lower or higher than expected** during the planning stage (for examples, see MHEP SIB, Provider-led Pathways and the Work Programme in Annex 1). Finally, in the Eleven Augsburg SIB, the oversight process was very similar to that applied when the intervention was funded traditionally, with the exception of negotiations to determine which participants could be considered eligible after the programme went live. This was because the SIB was required to target individuals that were

⁷⁹ Lane, Pippa, Rowan Foster, Laura Gardiner, Lorraine Lanceley and Ann Purvis (2013a). Research Summary: Work Programme Evaluation: Procurement, supply chains and implementation of the commissioning model. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/197712/832summ.pdf, 2.

⁸⁰ Interview with the Commissioner (2020).

⁸¹ The Merlin Standard was designed by the Department for Work & Pensions (DWP) to help evolve successful, high performing supply chains, and champion positive behaviours and relationships in the delivery of provision and ensuring fairness within supply chains.

particularly 'difficult to reach' in order for the programme to bring added value, which in practice was difficult to prove.

Table 17. SOC management costs relative to TF programmes

	SIB	PbR
SOC management costs higher than TF	Eleven Augsburg Benevolent Society SBB MHEP SIB	Provider-led Pathways Work Programme Drug and Alcohol Recovery Pilots Troubled Families
SOC management costs lower than TF	Perspektive:Arbeit	Youth Contract
SOC management costs similar to TF	DUO for a JOB	-
Insufficient information to compare management costs	Buzinezzclub BOAS Werkt ABLE	JobPath Transforming Rehabilitation

Source: Compiled by PPMI. based on commissioners' impressions stated during interviews.

In some cases, however, the management costs of the TF programmes were greater than or similar to those of SOC schemes. According to the commissioner of Perspektive:Arbeit, their role in the traditionally funded phases of the project that followed the SIB entailed a greater workload for financial settlement because tracking multiple outputs was more cumbersome than tracking outcomes. Furthermore, the intermediary during the SIB arrangement took over some of the functions typically performed by the commissioner. The Youth Contract commissioner noted that the programme included a few large service providers, compared with the multitude of smaller organisations delivering services under the Activity Agreement Pilots (its TF equivalent). Some of these organisations were less experienced than their YC counterparts, making the management of the latter programme more resource intensive for the commissioner.⁸²

While management costs could not be quantified for the majority of the programmes, the table below provides these costs for the MHEP SIB, JobPath and the Work Programme. It appears that the share of the contract allocated to management costs decreases as the contract size increases.

Table 18. SOC management costs as a share of total cost

SOC	Total cost	Management cost	Share of total cost
MHEP SIB	GBP 1,142,264 ⁸³	319,000 ⁸⁴	28% ⁸⁵

⁸² Interview with the commissioner.

⁸³ Gadenne et al. (2020). Individual Placement Support: A Social Impact Bond Model. The Behavioural Insights Team, p. 31. Implementation includes: the cost of MHEP staff wages; audit, insurance, and board costs and exceptional use of consultants of SPV; cost of time for a specialist team member, providing operational advice and guidance to providers to improve performance; investor returns, debt interest payments and any additional project surplus; as well as evaluation costs.

⁸⁴ Gadenne et al. (2020), p. 30.

⁸⁵ Derived by PPMI based on the information presented in the table.

SOC	Total cost	Management cost	Share of total cost
JobPath	EUR 88.3 million ⁸⁶	18 FTE; of these, 5 FTE work on outcome validation alone ⁸⁷	N/A
Work Programme	GBP 3 billion ⁸⁸	“a few tens of millions” GBP ⁸⁹	N/A

Source: Compiled by PPMI, based on the available information (see footnotes for sources and limitations regarding each estimate).

5.3. Summary of findings regarding SOC efficiency

Sufficient information was available to make comparisons between intervention costs in seven out of the 15 cases selected. Given that the effectiveness of SOC or TF programmes might depend on the level of funding invested, we compared the *cost per outcome* for each SOC-TF pair. This is the cost per participant who found a job, remained in employment for a certain duration of time, or completed the entire rehabilitation programme, and so on (depending on the goals of each programme). In this way, we could assess which programme was more effective while controlling for the different levels of funds allocated to each programme.

Among the cases in which comparisons could be made, SIBs tended to cost more per outcome achieved than similar traditionally funded programmes in cases where the SIBs successfully achieved their target outcomes and investors were repaid, including the interest payment. This additional cost was nevertheless seen as justified by the commissioners, given that the SIBs were used to test interventions in different contexts or at different scales than those in which the same programmes had been implemented previously, rather than to deliver the programme more efficiently. In other cases, the commissioner issued no payments to investors because the SIB failed to reach its target outcomes, meaning that the programme cost nothing to the taxpayer (discounting the resources invested to set up and oversee the programme). Cases such as these illustrate the true benefits of SIBs when testing innovative interventions. If the programme fails to achieve outcome targets, it can be shut down easily. This contrasts with some TF programmes, which continue to run despite a lack of evidence that they are effective.

A commonly cited argument for the use of PbR is that they are more efficient than traditionally funded programmes because their focus on outcomes drives service providers to innovate and eliminate inefficient practices. In our study, the two PbR programmes whose costs per outcome could be compared against TF programmes did indeed have costs that were similar or lower than comparable TF programmes. Nevertheless, comparisons were impossible in five of the seven PbR programmes selected due a lack of information about the costs of PbR programmes, as well as the outcomes of TF programmes. More such information therefore needs to be collected to assess whether PbR programmes truly are more efficient than TF programmes. Furthermore, given the concerns about perverse incentives in PbR contracts, it is important to investigate whether quality of services is maintained at a lower cost.

⁸⁶ The cost is the total for the years 2015-2017. DEASP (2019a). Working paper: Evaluation of JobPath outcomes for Q1 2016 participants. Retrieved from: <https://assets.gov.uk/36499/ffdce98cddc34addb05cf41a70aaf4e7.pdf>.

⁸⁷ Interview with the commissioner.

⁸⁸ The figure indicates programme cost between 2011 and 2019, and was quoted by the commissioner during the interview. It is consistent with other sources of information. According to the British Parliament 2017, the cost totalled GBP 2.4 billion between 2011 and 2016.

⁸⁹ Interview with the commissioner. Please note that an impact assessment upon which this statement was based is forthcoming. The cost includes DWP staff wages. It does not include costs related to bidding and programme design.

Lastly, both SIBs and PbR programmes generally involved higher operational costs (set-up and management) than TF programmes. Insufficient information is available to assess whether these additional costs offset the savings PbR programmes potentially offer by promoting intervention efficiency. There is also little transparency as to the costs involved in the design and oversight of SIBs. It is therefore important to track and report these costs when designing new SOC programmes.

6. Designing SOC schemes

In the previous chapter, we reviewed the costs and time required to set up SOC schemes, in addition to their management costs. We now turn our attention to key design decisions and lessons learned from the 15 cases studied, to help inform the design of future SOC schemes. The effective and efficient design features of each scheme are detailed in Annex 1. In this chapter, we first provide an overview of SOC design decisions that apply to a number of the SOC programmes analysed. We then discuss design elements that affect the balance of risk between commissioners and investors in SIBs, and commissioners and service providers in PbR programmes. Following this, we summarise the extent to which design decisions supported or weakened the ability of third-sector and social economy enterprises to deliver social services. After discussing the impact of SOC on service providers, we also consider their impacts on the public sector. The summary of findings is presented at the end of the chapter.

6.1. Design elements

Key design elements observed across the 15 SOC programmes analysed are detailed in Table 19, together with their main advantages and disadvantages. These key elements fall into four categories, relating to:

- SOC stakeholders;
- Payment arrangements;
- Participant recruitment; and
- Other contractual provisions.

Please note that we also identified lessons learned regarding the outcomes measurement methods used. These are detailed in Section 3.1.

Starting with SOC stakeholders, one of the key decisions in SIB design is whether or not to engage an **intermediary**. Eleven Augsburg, Perspektive:Arbeit, Buzinezzclub, DUO for a JOB, MHEP SIB and the ABLE programme all involved an intermediary. Interviewees from these SIBs emphasised the advantages of intermediaries, including:

- help in finding and reaching agreement among the commissioners, service providers and investors;
- providing the necessary expertise; and
- at times, relieving some of the service provider's burden in relation to outcome tracking, which freed up more time for service providers to work with programme beneficiaries.

One of the service providers involved in the Eleven Augsburg SIB, however, noted that they would have preferred to deal directly with the commissioner, because they would have been able to better represent their interests than the intermediary. Therefore, it also seems important to maintain regular communication between all parties involved in a SIB even when the SIB is intermediated. Furthermore, the presence of the intermediary increases the cost of the programme (see Table 16 in Section 5.2.1), and might therefore be better used in more complex SIBs such as those involving multiple investors, commissioners or service providers.

Table 19. Advantages and disadvantages of key SOC design elements

Design element	Advantages	Disadvantages
SOC stakeholders		
Engaging multiple investors	Easier to attract investors because the risk to each investor is lower; More feedback on the service provided; More relationships formed for future cooperation.	More complex legal structure for the SOC, hence higher set up costs; Higher management costs.
Engaging multiple commissioners	Lower risk to each commissioner; Helps solve the 'wrong pockets' challenge.	More complex legal SOC structure, hence higher set up costs; Higher management costs; Might result in multiple reporting lines for the service provider.
Engaging an intermediary	Intermediary can facilitate the alignment of interests, saving time, effort and costs; Intermediary can assist the provider in gathering evidence to prove outcomes, freeing up time to work with beneficiaries; Intermediary can help find investors, commissioners and service providers.	For service providers it might entail not being able to represent their own interests in front of the commissioner; Additional costs resulting from intermediary fees.
Payment arrangements		
Binary success criterion	Easier to measure and issue payments, hence lower management costs compared to frequency criterion; Less risk to the commissioner; Helps to avoid the perception that the private sector profits from the public purse.	More difficult to attract investors to SIBs because of the greater risk compared with frequency success criteria; More difficult for small third-sector organisations to participate in PbRs, because of the risk associated with failure to achieve targets.
Different outcome targets specified for crisis and non-crisis times	Helps to take account of changes in macroeconomic conditions that affect outcomes without having to re-negotiate contracts in the event that macroeconomic conditions change drastically.	Higher start-up costs due to the need to establish two sets of targets.
Payment caps and floors	Setting payment caps and floors relative to every outcome indicator, rather than precise figures, is especially useful whenever there is uncertainty about the likely level of outcome achievement within a fixed budget. This helps control the risk for the commissioner.	The service provider does not have the opportunity to make substantial profits (but also avoids substantial losses in the event of underperformance).
Setting a low level (less than 3%) of return for investors	Helps to avoid the perception that the private sector profits from the public purse; Saves public funds.	Makes it more difficult to attract investors: only philanthropic organisations might be interested.

Design element	Advantages	Disadvantages
Tying payments to investors with savings to the municipality	Ensures that investors cannot be compensated beyond the level of savings to the commissioner.	Savings might be difficult to estimate.
Having a service provider invest own resources	Aligns the interests of service providers with those of investors, motivating service providers to be efficient and to prioritise the achievement of target outcomes; Additional revenue for the service provider in case of success	Stifles innovation due to the risk to the provider of losing their own resources; Makes it more difficult for smaller third-sector organisations to participate.
Differentiating remuneration per target group	Provides a disincentive for parking, creaming and cherry-picking.	The payment model might become too complex for providers to respond to incentives and to deliver the programme effectively, if many groups are included in the programme.
Tracking overhead costs	Transparency with regard to the overall costs of SOC's.	Higher management costs.
Regular communication between all parties involved	Helps share feedback, manage expectations and monitor performance.	Higher management costs.
Participant recruitment		
Establishing minimum required participation levels	Helps ensure that the programme can be discontinued easily if it fails to achieve the scale required to be cost-effective.	Requires it to be established which parties will be responsible for participant recruitment, compensation for each party in the event that minimum participation levels are not achieved, and conducting a pre-assessment to investigate the extent to which the legally binding level of participants is feasible.
Mandatory vs voluntary participation ⁹⁰	Mandatory participation makes it easier to predict participant volumes and plan the intervention; Mandatory participation ensures that providers work with groups that are hard-to-help.	In cases where the participants do not perceive project outcomes to be achievable or desirable (for example, persons with disability who think they are not ready for work), participants might perceive service providers as pushy, especially because of the pressure on providers to acquire outcome payments. Service providers are also likely to engage in parking.
Other contractual provisions		
'Black box' design	Provides flexibility for service providers to innovate and change the service delivered to participants according to their unique needs.	A risk that some participants will receive inadequate services; Inability to know which programme features are effective.

⁹⁰ The advantages and disadvantages of mandated social services are beyond the scope of this paper, so we limit ourselves only to aspects relevant for SOC. For an in-depth discussion of mandating, please see Dwyer, P.J. (2018). Punitive and ineffective: benefit sanctions within social security. *Journal of social security law*. Retrieved from: <http://eprints.whiterose.ac.uk/138989/>, p. 142-157; Griggs, J. & Evans, M. (2010). Sanctions within conditional benefit systems: a review of evidence. Joseph Rowntree Foundation; Oakley, M. (2014). Independent review of the operation of jobseeker's allowance sanctions validated by the Jobseekers Act 2013. Department for Work and Pensions. Retrieved from: <http://www.haringeyadvice.org.uk/uploads/ja-sanctions-independent-review-print.pdf>

STUDY ON THE BENEFITS OF USING SOCIAL OUTCOME CONTRACTING IN THE PROVISION OF SOCIAL SERVICES AND INTERVENTIONS

Design element	Advantages	Disadvantages
Establishing minimum service standards	Helps to ensure that all participants receive at least basic services, thus ensuring the commissioner's public accountability.	Limits the service provider's ability to innovate; If minimum service standards require substantial work from providers, it is possible that no budget will be left to spend on additional/customised services.
One vs multiple providers per contract area	Having one provider per contract area helps to ensure economies of scale; It renders contracts more attractive to tenderers; It is less demanding for the commissioner to manage; May be less confusing for service recipients.	Service recipients lose the freedom to choose service providers; Inability to foster competition among service providers to encourage better quality services.
Flexibility to adapt the commissioning model at the local level	Might result in a programme designed to better reflect local realities; Might yield more information on what works well.	Higher start-up and management costs.
Contract length	Longer contracts (over three years) provide greater stability to service providers in terms of hiring staff; Longer contracts provide more time to experiment, thus fostering innovation and enhancing the understanding of which elements of the intervention are effective.	More difficult for commissioners to secure funding; Greater risk of commissioners getting 'locked in' with an underperforming service provider.
Feasibility study	Helps to ensure that SIBs do not become policy goals in themselves.	Higher start-up costs.
Bidding procedure	It is relatively easy to set programme targets based on the expected performance service providers indicate in their bids; The process is transparent and fair.	Providers are incentivised to overstate their future performance in order to win the bid. If programme targets are based on these expectations, the programme might be discontinued if providers fail to achieve them.

With regard to participant recruitment, a common theme across the schemes studied was the need for **flexibility within the contractual agreement, to account for changes in volumes of participant and macroeconomic conditions**. Lower or higher participant volumes than expected during the planning stage of the intervention have hampered the delivery of a number of programmes. In the BOAS Werkt SIB, for example, the inability to recruit the contractually required minimum level of participants resulted in the discontinuation of the programme. In Provider-led Pathways, low participant volumes made it financially non-viable for some subcontractors to deliver the intervention, with two-thirds of subcontractors surveyed by the NAO expecting to make a financial loss from the programme.⁹¹ This was made worse by the 2008 recession, which made it harder for service providers to achieve outcomes and thus receive outcome-related payments. Macroeconomic fluctuations are relevant to SOC schemes because many of them deal with employment outcomes.

Some ways to build-in safeguards against participant volume and macroeconomic changes are mentioned in the table above. In the Buzinezzclub SIB, for example, different outcome targets were specified for crisis and non-crisis times. Drug and Alcohol Recovery Pilots, Transforming Rehabilitation and JobPath PbR schemes have set payment caps or floors to prevent substantial losses for service providers, but also to safeguard commissioners against having to pay out excessive profits. In other cases, such as the Work Programme, performance expectations were simply downgraded after service providers failed to meet them. This demonstrates the need for commissioners to robustly inspect providers' bids and to base performance expectations on the performance of the control group (if available) or historical data, rather than the targets set out by bidders in their bids. This is because providers are motivated to exaggerate their future performance in order to win the bid, which may later jeopardise the programme.

Another closely related issue concerns the **definition of the target group**, given that the narrower the target group, the harder it is to recruit participants. The cases analysed run the gamut from those with a very limited target group to those with wide target group specification. Each situation has its own advantages and disadvantages.

The Eleven Augsburg SIB, for example, was intended to fill the gap in public programmes by serving people who had not yet been helped effectively. It also aimed to motivate service providers to look for the most difficult cases. The operationalised definition of the target group meant that participants had to be younger than 25 years-old, NEET, with no qualifications, and **had not registered with the local job centre or youth welfare offices**.⁹² According to the service provider interviewed, such participants were "non-existent", because it is "difficult to leave no trace in Germany." Service providers therefore struggled to find participants who would satisfy the programme's eligibility criteria. Eventually, the criteria were broadened to include participants who were registered with youth welfare offices but not the local job centre (for more details on Eleven Augsburg, please see Annex 1).

By contrast, the eligibility criteria for the Work Programme were wide, as the scheme served nine different groups of participants ranging from regular jobseekers who had been unable to find work for a number of months, to prison leavers. The rationale was to streamline existing programmes so that beneficiaries could enrol in one programme instead of having to navigate their way through various services. In practice, however, service providers struggled to successfully meet the needs of such diverse participants. It was therefore argued that the programme was too broad, and the eligibility criteria were narrowed for the subsequent Work and Health Programme (for more information, see the Work Programme in Annex 1).

⁹¹ NAO (2010). Support to incapacity benefits claimants through Pathways to Work. Retrieved from: <https://www.nao.org.uk/report/support-to-incapacity-benefits-claimants-through-pathways-to-work/>, p. 11.

⁹² Interview with the intermediary and interviews with the service providers.

Striking the right balance in defining the eligibility criteria for the target group might be difficult during the planning stage, and there is no single recipe that will work for all programmes. For this reason, interviewees from multiple schemes suggested that a **start-up phase**, after the results of the tendering and the launch of service delivery, is crucial to sorting out these and other issues. The service provider involved in the Benevolent Society SBB suggested that the start-up phase should last longer than six months.

With regard to other contractual provisions – particularly for PbR schemes – the designers of the programmes we studied struggled to strike the right balance between the ‘**black box**’ approach and prescribed delivery. The black box approach does not prescribe which services providers have to deliver; rather, they can implement whatever services they see fit to achieve the specified outcomes. Such an approach, to a varying degree, was chosen in Provider-led Pathways, the Work Programme and Youth Contract. It was intended to give service providers the freedom to innovate and customise the services they delivered. **Minimum service requirements** were still stipulated to ensure that all participants received at least a basic service package. Feedback on the black box approach remains mixed. The service providers interviewed appreciated its flexibility and were able to identify a number of innovations that had been implemented because of it. One provider in the Work Programme hired physiotherapists and nutritionists and trained them to use their backgrounds to help participants with health conditions. Staff with a background in health care were able to engage more meaningfully with the participants and their doctors about what employment options may or may not be appropriate, given an individual person’s health condition (e.g. chronic back pain). Overall, however, commissioners were disappointed as they expected that the black box approach would result in completely new and better services for beneficiaries. Furthermore, with providers applying different approaches to individual participants, commissioners and evaluators were unable to assess which features of the programme were effective and which were not, echoing general criticism of the black box approach in the literature⁹³.

6.2. Risk sharing between commissioners and investors/service providers

Decisions made during the design of the SOC studied also affected how much of the risk associated with programme failure could be transferred from commissioner to investor (in SIBs) or service provider (in PbR contracts). Given that payment arrangements, and hence the transfer of risk, differs between SIBs and PbR schemes, we discuss these separately below. In both types of SOC, we found that **no model design could reduce the risks for all parties involved: if the risk was reduced for the commissioner, it increased for investors or service providers**. This contrasts with the existing rhetoric that outcomes-based contracts – SIBs in particular – are win-win situations in which all parties stand to benefit.⁹⁴ Nevertheless, higher risk for investors was also associated with the potential for greater gains in terms of interest.

6.2.1. Risk transfer in SIBs

The level of risk transferred from the commissioner to investors in the SIBs in this study typically depended on:

- the targets set;
- the definition of outcomes, and the number of different outcomes;
- guarantees on the investment;

⁹³ Albertson et al. (2018).

⁹⁴ Tan et al (2019).

- the frequency of payments;
- participant recruitment.

All SIBs analysed in this study transferred some risk to investors, but the level of risk transferred depended on how ambitious the outcome targets were, upon which payments were conditioned. The process used to arrive at the expected performance levels typically involved examining the performance of similar programmes in other contexts, estimating the potential savings for the commissioner if outcomes are achieved (which usually determined the interest payments as well), and ultimately on negotiations between the commissioner and the investors (with greater risk to investors typically being associated with higher interest payments). Given concerns about the profitisation of public services⁹⁵, we assessed whether SIBs that involved both for-profit and philanthropic investors were more likely to achieve their target outcomes than those that involved only philanthropic investors. It might be expected that for-profit investors would negotiate lower outcome targets. However, we did not find evidence to support this notion, as we encountered both successful and unsuccessful SIBs that included both types of investors.

We previously touched upon binary outcome targets in Chapter 3, with regard to outcome measurement. In relation to risk, it is important to note that binary outcomes increased the risks for the investors in the SIBs analysed, whereas frequency outcomes reduced this risk (and thereby increased the risk for commissioners). The investors involved in the Eleven Augsburg SIB – in which service providers had to help at least 20 juveniles into work for investors to receive any kind of repayment – said that they were unlikely to engage in another SIB that used a binary outcome logic, even though the target in the case of the Eleven Augsburg SIB was achieved.⁹⁶ The investors argued that binary outcome logic does not reflect the success of a programme: for example, could the programme truly be considered to have failed if 19 participants found jobs as opposed to 20? Yet, for the commissioner, it was important to have a simple and easily measurable outcome target that could be communicated to the public to avoid the perception that the private sector was profiting from the public.

Investors may carry substantial risk even in SIBs that use frequency outcomes, if the lowest payment threshold is still ambitious. As shown in the figure below, the ABLE SIB investors were to be repaid an increasing amount of money for additional reductions in the rate of re-admission to prison, as long as the programme reduced the rate by at least 8.5%. During the first year of the programme, the lowest threshold of 8.5% was not achieved, following which the SIB was discontinued and the investors were not repaid.

Figure 5. Payment arrangements in the ABLE SIB

Reduction in Re-Admission Rate	Projected Long Term City	
	Net Savings (\$)	City Payment to MDCR
≥20.0%	\$20,500,000	\$11,712,000
≥16.0%	\$11,700,000	\$10,944,000
≥13.0%	\$7,200,000	\$10,368,000
≥12.5%	\$6,400,000	\$10,272,000
≥12.0%	\$5,600,000	\$10,176,000
≥11.0%	\$1,700,000	\$10,080,000
≥10.0% (<i>breakeven</i>)	\$ ≥ 1,000,000	\$9,600,000
≥8.5%	\$ ≥ 1,000,000	\$4,800,000

Source: Berlin 2016, p. 4.

Another aspect of the ABLE SIB that is interesting from a risk perspective is that one of the investors – Bloomberg Philanthropies – guaranteed USD 7.2 million of the USD 9.6 million

⁹⁵ Erikson, Susan (2019). Faking Global Health. *Critical Public Health*, Vol. 29, Issue 4.

⁹⁶ Jebabli 2016, p. XX.

loan that another investor, Goldman Sachs, made to the intermediary, thus reducing the risk to Goldman Sachs.⁹⁷ It is important to note that there are SIB arrangements in which the commissioner, rather than a second investor, guarantees part or all of the principal to the investor, thus substantially reducing the risk associated with investing.⁹⁸ However, such arrangements were not used in any of the SIBs analysed.

To balance the risk between commissioner and investors, some SIBs included multiple outcomes as criteria for success. The Benevolent Society SBB, for example, included three different metrics for success. According to the service provider, these were used not only to assess the multiple impacts of the programme, but also to diversify the risk to investors: the chances were lower that the programme would fail to meet all three outcome targets, rather than just one. Similarly, the frequency of payments also lowered the risk to investors. In the Eleven Augsburg SIB, only one re-payment was due to be made to investors at the end of the SIB; in others, investors received partial repayments at the end of each year, depending on whether or not outcomes were achieved during that time (e.g. DUO for a JOB SIB).

Finally, certain contractual provisions can substantially increase the risk to the commissioner. In the BOAS Werkt SIB, the commissioner was responsible for recruiting a contractually agreed number of participants. The rationale was that it was necessary to operate on a certain scale in order to show that the programme was effective. Nevertheless, the programme did not generate sufficient interest and the SIB was thus discontinued early. The final settlement between the investors and the commissioners is still ongoing, but the case illustrates the point that the risk in SIBs stems not only from the ability to achieve targets, but also from other contractual obligations.

6.2.2. Risk transfer in PbR schemes

In PbR programmes, the transfer of risk from the commissioner to the service providers depends in large part on how much of the payment was based on outcomes. Among the PbR programmes analysed, we did not encounter a single case in which all payments were outcomes-based. Instead, a portion of the payments were issued in a traditional way; for example, through fees for the services delivered. Nevertheless, the schemes varied dramatically in terms of how much of the contract value was based on outcomes and hence, how much risk was transferred from commissioner to service provider in the event that the intervention failed (see the figure below). In Transforming Rehabilitation, for example, only 10% of the payments were based on outcomes and were intended to be paid as a profit share from the contract.⁹⁹ By contrast, between 89% and 100% of payments in the Work Programme were outcomes-based (depending on the year of the programme).¹⁰⁰

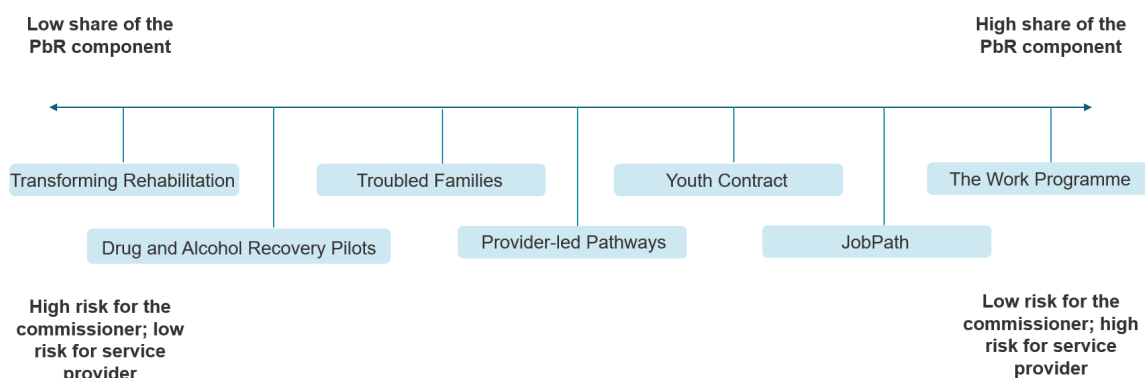
⁹⁷ Berlin, G. L. (2016). Learning from Experience: A Guide to Social Impact Bond Investing. MDRC, 5.

⁹⁸ Maier, F., Barbetta, G.P. & Godina, F. (2017). Paradoxes of Social Impact Bonds. Social Policy & Administration.

⁹⁹ Noble, J. (2015). The transforming rehabilitation tier 1 tendering process: the voluntary sector perspective. NPC, p.2-3.

¹⁰⁰ The figures are based on maximum attachment fees available in the first year of the programme, and the fact that attachment fees were phased out by the fourth year of the programme. Finn, Dan (2013). Opening up the 'Black Box': What services are Work Programme providers delivering and how are they doing it. Flexwork Research Conference, Amsterdam, 24-25 October 2013, 6-7; Foster et al. 2014a, 125.

Figure 6. Share of contracts based on outcomes in PbR schemes



Source: PPMI.

The share of the contract that was not based on outcomes was often paid for engaging with each programme participant. For example, in JobPath, providers would get EUR 311 for every participant registered on the programme.¹⁰¹ Importantly, attachment fees in some programmes were reduced over time because they were envisaged as support for the programme's start-up costs. In the Work Programme, for example, providers received no attachment fees by the fourth year of the programme.¹⁰²

It is important to note that the greater the share of payment that was based on outcomes, the more risk was transferred to service providers, and hence it became more difficult for smaller third-sector organisations and social economy enterprises to participate in PbR contracts. We now discuss this issue in further detail, in relation to both PbR schemes and SIBs.

6.3. Impact on third-sector organisations and social economy enterprises

Generally, **SIBs appeared to have more positive impacts on third-sector organisations and social economy enterprises than PbR contracts**. This was because NGOs and small-scale organisations often could not withstand the financial risk associated with PbR schemes (for example, if outcomes were not achieved), especially those in which most of the payment was outcome-based. Furthermore, across the PbR schemes observed in the study, commissioners more often gave preference to large contractors, so that they could shoulder the costs associated with programme start-up. These large contractors – often referred to as prime providers – then subcontracted third-sector organisations. The arrangement meant, however, that third-sector organisations did not have the same negotiating power as the prime providers, which at times resulted in fewer participants being referred to these organisations, making the PbR scheme not financially viable (for examples, see Provider-led Pathways, the Work Programme, the Youth Contract, Drug and Alcohol Recovery Pilots, and Transforming Rehabilitation in Annex 1). In other cases, only harder-to-help participants were referred to subcontractors. Given that it was more difficult for these participants to achieve the target outcomes, it was also more difficult for subcontractors working solely with these groups to receive payments, compared with providers that also worked with participants who were easier to help. Lastly, service providers noted that the budgetary uncertainty associated with PbR contracts made the hiring process and employee retention more difficult, in addition to the expensive data collection systems that needed to be developed to track outcomes. Table 20 below outlines

¹⁰¹ DEASP 2019, 17-18.

¹⁰² Foster et al. 2014a, 125.

the main impacts that outcome-based contracts had on third-sector organisations and social economy enterprises.

While SIBs shared some of the same negative impacts on third-sector organisations – such as insecurity in the hiring process – they presented greater opportunities for service providers than did PbR schemes. Given their novelty, the SIBs observed in this study generated media attention, which helped social economy enterprises to gain publicity, expand their networks, and ultimately acquire new funding or cooperation opportunities when their intervention was proven to work. Buzinezzclub, for example, scaled its service to other cities after the initial SIB. In SIBs, service providers are generally not required to invest their own resources in order to foster innovation, which means that the risk is shouldered by investors, making it easier for small NGOs to participate. However, there are exceptions to this rule. Service providers in the Buzinezzclub SIB and BOAS Werkt SIB, for example, were asked to invest their own resources. Furthermore, although providers did not invest their own resources in the MHEP SIB, the funds from investors were transferred to providers based on the providers' performance, hence resulting in similar risks for providers to those in PbR programmes.

Table 20. SOC impacts on third-sector organisations and social economy enterprises

	Positive impacts	Negative impacts
Social impact bonds	<p>New funding opportunities (both in the SIB itself, and the SIB can act as leverage to raise additional funds if the intervention is deemed successful);</p> <p>Networking opportunities and advice;</p> <p>Flexibility to customise social services for the target group;</p> <p>Increased visibility/publicity where SIBs are successful;</p> <p>Collection of rich data to better understand what features of the programme are effective;</p> <p>Serving participants that may not otherwise benefit from similar programmes.</p>	<p>Insecurity regarding hiring processes;</p> <p>Greater staff turnover due to job insecurity;</p> <p>Potential for mission drift;</p> <p>Bad press if the intervention is not successful;</p> <p>Time-consuming and expensive data collection systems.</p>
PbR schemes	<p>More opportunities to deliver services compared with the in-house delivery model;</p> <p>Flexibility to customise social services for the target group;</p> <p>Collection of rich data to better understand what features of a programme are effective;</p> <p>Strong emphasis on outcomes drives the organisation to focus on key priorities.</p>	<p>Inability to participate as prime providers, where priority is given to large for-profit service providers that can afford the programme's start-up costs;</p> <p>Inability to participate as prime providers due to greater risk for the organisation inherent in outcomes-based contracts compared with TF;</p> <p>Disproportionate risk carried by organisations that participated as subcontractors compared with prime providers;</p> <p>Budgetary uncertainty;</p> <p>Contracts were not financially viable for organisations that experience lower-than-expected participant volumes or delivered services during economic downturns;</p> <p>Time-consuming and expensive data collection systems.</p>

Source: compiled by PPMI, based on the information provided in Annex 1.

Nevertheless, there are certain design features that could increase the appeal of PbR contracts to third-sector organisations. The service providers interviewed noted that **contracts should be more heavily weighted toward input-based components rather than outcome-based**. For example, if 80% of the contract value were paid out as a fee-

for-service, it would enable third-sector organisations to participate in cases where they could not otherwise shoulder the up-front cost of the intervention, with 20% of payment based on outcomes. An alternative approach would be to consider outcomes-based payments only as a **premium** if performance exceeds expectations. Incorporating **milestones along the distance travelled** by the programme participants as criterion for payments could also make the contract more viable for third-sector providers. For instance, providers could be paid for drawing up an action plan with participants and helping them land job interviews in addition to the job outcome and job sustainment payments.

Lastly, in both SIBs and PbR schemes, **contract length** affected the extent to which third-sector organisations could participate. Longer contracts provided greater stability for service providers in terms of staffing and training. Longer contract duration is also needed to test out which programme features work effectively, and which do not. Interviewees generally noted that SOC contracts should ideally be longer than three years due to the steep learning curve involved in tracking outcomes and working with outcomes-based contracts more generally. Ultimately, it took time for participants to achieve the target outcomes. However, longer contracts pose the risk to commissioners that they might provide long-term funding to ineffective programmes. One way to deal with this risk is to add an **early discontinuation clause**, under which the programme can be discontinued early if it fails to achieve the minimum level of performance expected (for an example, see BOAS Werkt in Annex 1). On the flip side, the term could be extended in cases where the scheme delivers strong results (see the Work Programme in Annex 1).

6.4. Impact on the public sector

Overall, the involvement in SOC schemes was perceived positively by commissioners. This may be related to the fact that the commissioners interviewed were often those who had initiated the SOC programmes. In some cases, public sector commissioners promoted SOC schemes in order to experiment with a new form of commissioning and/or test innovative interventions, particularly in the case of SIBs such as Perspektive:Arbeit SIB, the DUO for a JOB SIB, and the Benevolent Society SBB.

One positive impact of SOC on public commissioning concerns a change of mindset and related organisational improvements in relation to **outcome measurement**. Although data collection systems were time-consuming and expensive to create, some interviewees noted that the process of setting up a measurement infrastructure was also beneficial in terms of using and integrating existing datasets and developing new ones.

As mentioned in Section 4.2.1, the measurement process in SOC schemes was helpful in terms of **providing evidence to support public decision making**. Thanks to the data collected during SOC schemes, the effectiveness of new interventions could be assessed to some extent, thus enabling public authorities to plan future service provision based on evidence. For instance, in the cases of the Benevolent Society SBB and DUO for a JOB, the interventions tested were subsequently commissioned by public authorities through TF after the interventions were shown to be effective. Nevertheless, our study also identified a number of issues in the processes used to measure of SOC schemes which affected their ability to attribute outcomes to the intervention. Therefore, measurement processes are helpful only so far as they employ a rigorous methodology to assess the impacts of SOC programmes.

As mentioned in Section 4.2.1, the benefits of SOC include enhanced collaboration among stakeholders. The potential to enable **cross-sectoral collaboration** can be particularly beneficial for the public commissioner involved in the scheme, in terms of sharing knowledge with providers and other stakeholders.

The commissioner of DUO for a JOB noted that the SIB made it **easier to acquire funding** for an intervention that had not yet been proven to work. Furthermore, it allowed access to **funding for multiple years** of the programme, since financing was not subject to annual budgets.

In terms of **efficiency**, PbR commissioners generally believed that the results-based payment structure drives service providers to deliver the best services they can, thus making the programmes more efficient. Nevertheless, as illustrated in Chapter 5, efficiency comparisons between PbR and TF programmes were not possible in half of the cases studied, because outcomes were not tracked in TF programmes, or were tracked differently from SOC programmes. Furthermore, even when comparisons were possible, they involved limitations including the inability to control for intervening factors such as participant characteristics, and thus did not allow us to claim that PbR programmes were more or less efficient specifically as a result of the funding mechanism. Thus, it was not possible to assess the belief of these commissioners empirically.

Commissioners generally perceived SIBs as costlier than comparable TF programmes, due to interest payments and management costs. Likewise, in the PbR programmes analysed, management costs were generally perceived to be greater than in TF programmes. This related in part to the outcome tracking and verification methods in place in SOC schemes. However, from the perspective of commissioners, the higher cost of SOC was acceptable in certain cases:

- **Where SOC allowed an intervention to be tested**, particularly if it provided services to a group of people that cannot count on other types of support.
- **Where higher costs in the short term could bring greater savings in the long term**. One local commissioner said that SOC can be a valuable tool to provide cost-savings in the long term, but this can only happen if, in the short term, commissioners invest more resources (e.g. in setting up the scheme, developing the measurement process, compensating the stakeholders for uncertainties related to the payment mechanisms).

In some cases, SOC schemes also received a lot of media attention. This was particularly true in countries where this new form of contracting was being implemented for the first time, such as in most of the cases analysed in continental Europe. In this sense, it is important to note that SOC can help to promote an image of commissioners as innovative. Having said that, it also increases scrutiny on the actions of the public administrations involved. Publicity can be beneficial in the event of a scheme being successful, but it also place a lot of extra pressure on stakeholders – especially public commissioners – if the intervention fails.

6.5. Summary of findings regarding SOC design

The cases analysed include various sub-types of SIB and PbR models, each of which had different implications in terms of effectiveness, efficiency, and the risk transferred from the commissioner. By ‘risk’, we mean the financial consequences in the event that the intervention does not succeed in achieving its target outcomes.

We found that an intermediary brought added value when designing large-scale SIBs with multiple commissioners, service providers and/or investors because of the expertise required to design contracts and engage stakeholders. Nevertheless, intermediaries also entailed additional costs, amounting to roughly 5% of the contract value.

In both types of SOC, we found that **no model design could reduce the risks for all parties involved: if the risk was reduced for the commissioner, it increased for investors or service providers**. This contrasts with the existing rhetoric that outcomes-based contracts – SIBs in particular – are win-win situations in which all parties stand to benefit.

The amount of risk transferred from commissioners to investors in SIBs varied, depending on a number of factors. Ambitious **programme targets** increased the risk for investors to lose their investment, yet lowered the risk for commissioners that they would have to fund ineffective interventions. Similarly, a greater **number of outcomes** assessed

was associated with a lower risk for investors that none of the outcomes would be achieved. The risk for investors was further lowered if philanthropies or commissioners **guaranteed part of the investment**. The more **frequent the payments** were, the lower the risk was for investors to lose all of their investment. On the other hand, more frequent payments implied the risk for commissioners that they would partly re-fund the intervention whose results could not be sustained. Finally, commissioners carried substantial risk when they were responsible for **recruiting** a contractually-specified number of participants, irrespective of other payment arrangements with investors.

Among the PbR programmes studied, the greater the proportion of a contract's value that was based on outcomes, the more difficult it was for small third-sector organisations and social economy enterprises to bid for the contract as prime providers (as opposed to subcontractors). Third-sector organisations lacked sufficient funds to shoulder the upfront costs of the programme, and were unable to bear the financial risks if outcomes were not achieved. It was generally easier for small third sector organisations and social economy enterprises to participate in SIBs, because such schemes posed less financial risk to service providers if the interventions failed. Nevertheless, both SIBs and PbR programmes were generally positively perceived by commissioners due to the benefits they generated, even taking into account the additional costs.

7. Conclusions and recommendations

This chapter of the Report considers the implications of the key findings of the study, and presents possible policy actions that could harness the potential benefits of social outcome contracting while minimising its drawbacks, as well as areas for further research. The chapter begins by presenting conclusions structured around four themes: effectiveness, efficiency, the design of SOC schemes, and outcome measurement. Following on from this, we present recommendations with respect to the areas and target groups for which SOC schemes are most impactful; the most effective and efficient models for SOC design, as well as outcome measurement methods. Wherever relevant, we differentiate between SIB and PbR funding models. Otherwise, our conclusions apply to both SIBs and PbR schemes. Given that the EU is considering support for innovative ways of designing public services in order to achieve the best results for the recipients of social services, these conclusions and recommendations are aimed at both commissioners and public authorities that are considering the introduction of outcomes-based contracts in the procurement of social services.

7.1. Conclusions

7.1.1. Effectiveness and added value

To compare the effectiveness of SOC and TF programmes, we assessed the extent to which each programme was able to achieve the outcomes set at the start of the programme. These included the number of participants employed, reductions in re-offending rates, the number of children who avoided out-of-home care, and other outcomes. In terms of target outcomes achieved, **there is insufficient evidence to claim that outcomes-based contracts are more effective than traditional financing or vice versa**. This is partly because outcome targets were not set in the majority of comparable TF programmes.

However, our analysis showed that the achievement of targets did not necessarily determine the overall effectiveness and success of the scheme. In fact, **some of the schemes that did not achieve their targets were considered successful, since they were able to address the needs and expectations of the various stakeholders involved**. Thus, they were extended or replicated. This was particularly true of SIBs when they were used to fill a gap in funding for a particular group of participants, or as instruments to test whether an innovative intervention is effective and can be scaled/replicated. The Perspektive:Arbeit SIB (Austria), for example, targeted survivors of domestic violence, for whom the availability of personalised social services was limited. Although the SIB technically failed to achieve its targeted outcomes, all stakeholders nevertheless perceived the programme to be successful. They argued that the programme was too short to provide the necessary training and skills to help survivors of domestic violence find jobs, which was used as justification for funding the programme traditionally in other Austrian states.

Another SIB scheme, DUO for a JOB (Belgium) addressed an emerging social problem – immigrant youth unemployment – which was not the focus of existing employment interventions. This scheme was also later funded traditionally after sufficient evidence was gathered during the SIB to show that the programme worked. It is important to note that both Perspektive:Arbeit and DUO for a JOB existed as programmes prior to the SIB. However, the SIB instrument was used to test whether the interventions could deliver the results expected.

Furthermore, interviewees noted additional benefits of SOC programmes, including:

- the development of a measurement infrastructure;
- evidence-based policy making;
- knowledge sharing among the stakeholders;

- the unlocking of financial resources;
- enhanced flexibility for service providers.

Although SOC's appear to focus more frequently on employment for specific disadvantaged categories of beneficiaries (such as people affected by mental health problems, migrants, and former prisoners), no specific target group proved to be more positively impacted by the SIB or PbR models than others. **The effectiveness of SOC's related more to the context of the intervention and its design than to specific types of beneficiaries or areas of social services.**

In line with the existing literature, **we found that PbR contracts encouraged episodes of creaming¹⁰³ and parking¹⁰⁴ of harder-to-help participants**, due to the financial pressures on service providers to achieve performance targets. The evidence gathered shows this problem stemmed from two causes. First, the PbR payment mechanism pushed providers to prioritise easier-to-help participants, especially when the programme funding was limited and targets were difficult to attain, because otherwise the programme would not have been financially viable for providers. Second, in programmes in which participation was mandatory, some participants were unwilling to participate or did not think that programme's goals (e.g. employment) were within their reach due to the multiple barriers (e.g. childcare responsibilities, disability, lack of transportation, etc.). Service providers therefore found it appropriate to park such participants. The study identified fewer cases of parking in SIBs, although cases of creaming at the point of participant recruitment were reported in SIBs as well.

Furthermore, in cases such as Provider-led Pathways, the Work Programme, the Mental Health and Employment Partnership (MHEP) SIB and Transforming Rehabilitation, contracts were formally re-negotiated or contract managers informally applied different performance criteria when providers failed to meet target expectations due to changing macroeconomic conditions. Such practices were not only costly, but they also undermined the public accountability of the commissioner. The added pressure for providers to achieve unattainable targets also encouraged perverse incentives and added to the complexity of contract management.

Although some of the PbR schemes analysed were not very effective in relation to the hardest-to-help groups, they were fairly effective for other groups. For example, the Work Programme performed worse among people claiming unemployment benefits due to disability, but it exceeded its targets for most other jobseekers. As a result, the programme's net benefits overall exceeded net costs. The JobPath (Ireland) commissioner also observed a similar trend. This would suggest that PbR programmes would perhaps be best utilised for target groups that are relatively easy to help, but not so easy that they could achieve the outcomes without the intervention (e.g. the long-term unemployed).

Nevertheless, when we compared the performance of PbR programmes with respect to these groups against the performance of similar TF programmes, we found mixed results. For example, JobPath users were slightly more satisfied with the JobPath programme than with the traditionally funded Intreo service, but the Provider-led Pathways PbR scheme performed similarly to the parallel Pathways programme delivered in-house through Jobcentre Plus. Furthermore, direct comparisons could not be made as to the effectiveness of a number of programmes because outcomes were not consistently tracked in TF programmes. More information on the outcomes of TF programmes is therefore needed in order to compare them to the outcomes achieved in PbR programmes.

¹⁰³ In this study, creaming refers to the process of selecting those participants into the programme that are most likely to achieve the programme's outcomes, in order for service providers or investors to receive payments.

¹⁰⁴ Parking refers to a process by which providers try to keep costs down by doing little to serve those with the poorest anticipated outcomes, while instead focusing resources on more able clients with better employment prospects.

7.1.2. Efficiency

We compared the cost per outcome in each matched pair of SOC and TF programmes to assess the efficiency of SOC and TF programmes. This represents the cost per participant who found a job; who remained in employment for a certain duration of time; who completed the entire rehabilitation programme, etc., depending on the goals of each programme. Sufficient information was available to draw comparisons in the intervention costs for 7 out of the 15 selected cases. Importantly, our comparisons were limited by the fact that we could not control for intervening factors such as participant characteristics, so **we cannot claim that SOC programmes were more or less efficient specifically due to the funding mechanism used.**

Among those cases in which comparisons could be made, **SIBs tended to cost more per outcome achieved than similar traditionally funded programmes**, in cases where the SIB successfully achieved its target outcomes and investors were repaid, including the interest payment. **The commissioners nevertheless saw these additional costs as justified:** SIBs were used to test interventions in different contexts or at different scales to those in which the same programmes has previously been implemented, rather than to deliver the programme more efficiently. Furthermore, we also encountered cases in which SIBs failed to reach their target outcomes, and thus the commissioners issued no payments to the investors, meaning that the programme cost nothing to the taxpayer (discounting the resources invested in setting up and overseeing the programme). **SIBs were thus useful in that when a programme failed to achieve its outcome targets, it could be easily shut down.** This contrasts with some TF programmes, which continue to run despite a lack of evidence that they are effective.

With regard to PbR schemes, a commonly cited argument for their use is that they are more efficient than traditionally funded programmes because their focus on outcomes drives service providers to innovate and eliminate inefficient practices. Indeed, the two PbR programmes in our study whose costs per outcome could be compared against those of TF programmes showed either a similar or lower cost than equivalent TF programmes. However, it is impossible to draw comparisons with regard to the other selected PbR programmes, due **insufficient information about both the costs of PbR programmes and the outcomes of the TF programmes.** Furthermore, given concerns about the perverse incentives that may arise in PbR contracts, it is important to investigate whether the quality of services is maintained at a lower cost.

Lastly, **both SIBs and PbR programmes generally entailed greater operational (set-up and management) costs than TF programmes**, for example, **sometimes taking between a year and two years to set up.** Available information is insufficient to assess whether these additional costs offset the savings that PbR programmes potentially achieve by promoting the efficiency of interventions. **There is also little transparency as to the costs involved in the design and oversight of SIBs.**

7.1.3. Design of social outcomes contracts

The cases analysed include various sub-types of SIB and PbR models, with each model having different implications in terms of the level of risk transferred from commissioner to investors or service providers. By 'risk', we mean the financial consequences in the event that the intervention is not successful in achieving its target outcomes. In both types of SOC, we found that **no model design could reduce the risks for all parties involved: if the risk was reduced for the commissioner, it increased for investors or service providers.** This contrasts with the existing rhetoric that outcomes-based contracts – SIBs in particular – are win-win situations in which all parties stand to benefit.

The amount of risk transferred from commissioners to investors in SIBs varied, depending on a number of factors. Ambitious **programme targets** increased the risk for investors to lose their investment, yet lowered the risk for commissioners that they would have to fund ineffective interventions. Similarly, a greater **number of outcomes** assessed

was associated with a lower risk for investors that none of the outcomes would be achieved. The risk for investors was further lowered if philanthropies or commissioners **guaranteed part of the investment**. The more **frequent the payments** were, the lower the risk was for investors to lose all of their investment. On the other hand, more frequent payments implied the risk for commissioners that they would partly re-fund the intervention whose results could not be sustained. Finally, commissioners carried substantial risk when they were responsible for **recruiting** a contractually-specified number of participants, irrespective of other payment arrangements with investors.

In the PbR programmes studied, the greater the proportion of the contract value that was based on outcomes, the more difficult it was for small third-sector organisations and social economy enterprises to bid for the contract as prime providers (as opposed to subcontractors). Third-sector organisations lacked sufficient funds to shoulder the programme's upfront costs, and could not bear the financial risks if outcomes were not achieved. SIBs were generally easier for small third-sector organisations and social economy enterprises to participate in, because they posed less financial risk to service providers if interventions failed. Nevertheless, both SIBs and PbR programmes were generally perceived positively by commissioners due to the benefits they generated, even taking into account their additional costs.

7.1.4. Outcome measurement

Outcomes evaluation is one of the key characteristics of SOC schemes, which ultimately determines the payments made to investors or service providers. Therefore, we investigated how such outcomes were defined and measured in the SOC schemes selected, and what impact that had on the quality of the programme.

Among the cases analysed, **the way in which outcomes were defined had a significant influence in terms of encouraging or discouraging perverse incentives such as creaming and parking**. Outcomes were often defined using either a **binary or a frequency approach**. When binary outcomes were chosen, providers had to achieve an absolute target, and no payment was granted for achieving lesser results. An example of a binary outcome would be whether or not a person found a job. In frequency schemes, rewards were staggered according to an agreed frequency of results, with payments increasing as results increase. For example, providers or investors would receive additional payments for every three months that a programme participant remained employed. Of these two models, **binary outcomes were more often associated with creaming and parking** because the target outcome (e.g. employment) was often out of reach for some programme participants, motivating providers to focus on other clients.

Furthermore, **investors argued that frequency measures better reflected the true success or failure of the programme**. Many of the investors interviewed who worked on programmes in which success payments were conditioned on binary outcomes questioned whether a programme can truly be considered to have failed if the target outcome is reached, for example, for 19 out of 20 participants. The downsides of frequency measures were that they were more difficult to track, they were less intuitive to communicate to the public, and they limited the risk that could be transferred from commissioners to investors or service providers. To accommodate the differing interests of the stakeholders involved, in a third of the cases analysed, both types of measure – frequency and binary – were used to track outcomes.

Importantly, **we encountered instances in which involving the service provider in the choice of outcomes helped to prevent mission drift and ensured that medium-term impacts were considered in addition to short-term impacts** (for examples, see DUO for a JOB and the Drug and Alcohol Recovery Pilots in Annex 1). Involving service providers also meant that their expertise could be drawn upon in determining which outcomes would be the best to ensure the well-being of the beneficiaries.

In a number of programmes studied, service providers were responsible for gathering the evidence used to assess whether target outcomes had been achieved. This process involved a number of challenges, including:

- providers lacking the expertise to gather the necessary information;
- the information that was supposed to be used as evidence (e.g. employment contracts; education records; proof of participant eligibility) was not available, meaning that providers could not claim payments for some of the outcomes they achieved;
- the process was more time-consuming than the providers anticipated, so they could dedicate less time than expected to working with beneficiaries;
- conflicts of interest emerged regarding the data self-reported by service providers in PbR programmes, because these data determined the payments made to providers.

Furthermore, even though randomised controlled trials (RCTs) are considered the gold standard in positivist evaluations, **none of the 15 analysed cases employed an RCT**. One programme – the ABLE SIB – attempted an RCT in relation to a behavioural therapy programme for jailed young offenders, but eventually decided against this approach because clear boundaries could not be maintained between treatment and control groups when participants had to change housing units. Service providers from other programmes also felt that it would be unethical to deny services to some potential participants for the sake of forming a control group, particularly in cases where other providers or state institutions could not provide the same services outside the SOC programme. Lastly, RCTs also entail large costs.

As a result, the most rigorous methods encountered among the cases analysed were quasi-experimental, yet these were only used in seven cases. Out of these cases, only two schemes used the quasi-experimental design to evaluate the outcomes related to the payment mechanisms. This means that rigorous evaluation methods, which causally link the intervention to its effects, triggered the payment only in two schemes. **In most cases, the absence of a control group meant that the outcomes achieved in SOC programmes could not be definitively attributed to the intervention.**

7.2. Recommendations

Recommendation 1: SIBs should be commissioned in areas in which there is a gap in funding, and for the purpose of testing whether an innovative intervention is effective and/or scalable.

Based on the findings outlined above, we recommend commissioning SIBs in areas in which there is a gap in funding, and for the purpose of testing whether an innovative intervention is effective and/or scalable. Such SIBs were perceived as the most successful by the stakeholders interviewed, even when the SIBs failed to achieve their target outcomes. However, it is difficult to justify the use of SIBs when they fall outside these criteria due to their additional management costs as well as interest payments to investors. To ensure that taxpayer money is used efficiently, feasibility studies should be conducted prior to launching a SIB, so that that the funding instrument brings added value in return for the additional costs it entails.

Recommendation 2: The commissioning of PbR schemes should be avoided when services are being contracted for groups that face multiple barriers to achieving the target outcomes, particularly when participation in the programme is mandatory.

Given the perverse incentives identified in the PbR programmes studied, we suggest that PbR contracts should not be used when contracting services for groups that face multiple barriers in reaching target outcomes, particularly where participation in the programme is mandatory. Rather, we recommend that services for groups with complex needs should be funded in a traditional way.

It is important to note that SIBs, as opposed to PbR programmes, could be an effective way to test interventions targeting beneficiaries with multiple, complex needs. For example, the creation of the Mental Health and Employment Partnership SIB in the UK was partly motivated by the desire to find alternative financing mechanisms to help those with mental and other health issues who were not effectively helped by the heavily outcomes-based PbR scheme, the Work Programme. As outlined in Recommendation 1, however, SIBs entail additional costs and programmes should therefore not continue to be funded through SIBs once there is sufficient evidence of their effectiveness in a particular context.

Recommendation 3: Assess the effectiveness and efficiency of PbR programmes for easier-to-help groups by tracking outcomes in TF programmes and utilising control groups.

Given mixed results observed in terms of the effectiveness of PbR and TF models, coupled with a lack of information on their respective efficiency, we argue that more data is required to determine whether PbR programmes are more effective and efficient than TF programmes for target groups that do not face multiple barriers to achieving their target outcomes. These data include the outcomes of TF interventions. Below, we suggest two approaches to tracking outcomes in TF programmes.

First, we recommend tracking the outcomes achieved in traditionally financed social service programmes for which it is feasible to collect data through administrative databases. We acknowledge that the collection of outcome data can impose a costly burden on all programme stakeholders, which is why we recommend starting with programmes whose outcomes could be assessed using or linking data that is already being collected (e.g. datasets from tax authorities, social security agencies, child support agencies and others). As an added benefit, the service providers interviewed during the research programme argued that data collected on outcomes that were achieved (or not achieved) helped them improve their service offering and to learn from more effective competitors. Collecting outcomes data could therefore help to improve both SOC and TF programmes, in addition to providing more information with which to compare SOC and TF models.

Second, we suggest including TF control groups in the design of PbR schemes. The most robust assessment of the effectiveness of SOC and TF would require the same intervention to be funded using both models, ensuring that factors such as the target group, services provided, duration, location and others are identical across the two funding schemes. One way to implement this would be to use a TF model as a control group for the PbR model, as was done in the Pathways to Work programme and is currently being done in the DWP's Work and Health Programme in the UK.

Recommendation 4: Track and report operational costs to ensure that public procurement is transparent.

We found it particularly challenging to assess the efficiency of both PbR and SIB models because limited information is available on the costs incurred in designing these types of contract, setting the programmes up, monitoring performance, and evaluating results. These costs should be included in the total reported programme cost, even though this is not currently standard practice. Furthermore, they should be made publicly available to ensure that public procurement is transparent, and governments can be held accountable.

Recommendation 5: To ensure that small third sector organisations can participate in outcomes-based contracts, use the PbR mechanism as a premium for results accomplished beyond the expectation, base a small portion (e.g. up to 20%) of the PbR contract on outcomes, or implement SIBs.

Among the PbR cases analysed, third-sector organisations lacked sufficient funds to shoulder the upfront costs of the programme, and could not bear the financial risks if outcomes were not achieved. This was particularly true in the case of programmes in which a large portion of the payments was based on outcomes. Therefore, to ensure that small third-sector organisations are able to participate in outcomes-based contracts (as prime providers as opposed to subcontractors), we recommend using the PbR mechanism as a premium for results achieved beyond the expectations of the targets set by programme commissioners. An alternative would be to base a small portion (e.g. up to 20%) of the PbR contract on outcomes. This would mean that most of the payments in the contract would be based on outputs. Another alternative would be to use SIBs instead. These may be easier for smaller third-sector organisations to participate in, compared with PbR programmes. This is because investors, rather than service providers, bear the financial risks. Furthermore, under SIBs, investors provide the funds up-front (or in batches over time), so even small organisations that have little of their own capital might find it viable to participate. For both SIBs and PbR models, setting the length of the contract to at least three years (and ideally longer) would also give more stability to service providers in terms of hiring and project management.

Recommendation 6: Contracts should provide some flexibility with regard to performance targets, to ensure that targets remain achievable under changing conditions. This could be achieved by tying performance expectations to the performance of the control group, setting different targets reflecting different macroeconomic environments, setting payment caps and floors, and/or including early discontinuation clauses.

To avoid the risk of having to renegotiate agreements, SOC contracts should build in flexibility with regard to performance targets, so that these targets can alter depending on changing macroeconomic conditions. Performance expectations could be tied to the performance of a control group, which would receive services at the same time as participants in the SOC programme – for example, its performance target could be specified as 10% better performance compared with the control group (see DUO for a JOB in Annex 1). Alternatively, as was the case in the Buzinezzclub SIB, different performance targets could be specified prior to the start of the programme, according to whether or not the economy is experiencing recession. Payment caps would safeguard the commissioner against having to pay out excessive profits in the event that the programme performs better than expected (see JobPath in Annex 1). On the flip side, payment floors could be implemented to ensure that SOC contracts are financially viable to providers even when participant volumes are low. Finally, early discontinuation clauses would limit the risk to both commissioners and investors in cases where the programme's early performance fails to meet minimum thresholds (see the ABLE SIB and BOAS Werkt in Annex 1).

Recommendation 7: Combine both binary¹⁰⁵ and frequency outcome measures, and consider the distance travelled by participants as well as customer satisfaction to prevent perverse incentives.

To discourage perverse incentives and accommodate the interests of different stakeholder, we recommend combining frequency outcome measures with binary when designing SOC schemes. However, even when frequency outcome measures are combined with binary, some programme participants may still be far from achieving even the lowest outcome threshold, which often results in creaming and parking. For example, if a programme aims to place participants into employment and sustain that employment for a number of months, some participants will not be able to find a job. Therefore, we further suggest including 'soft' or 'distance travelled' outcomes into PbR contracts. These could include outcomes such as agreement on an action plan, CV preparation, interview attendance and others, which would make it more financially viable for providers to work with less job-ready clients.

Another way to limit creaming and parking is to monitor performance among different groups of individuals (i.e. those closer to and further from achieving the outcomes) while the programme is ongoing. In addition, surveying participant satisfaction while the programme is ongoing might help to ensure that all participants are receiving high-quality services. This step could ensure that participants' employment (or other outcomes) are sustainable, and that they do not get pushed into jobs that they do not intend keeping, just so providers can claim outcome payments. A good example would be the JobPath programme, in which customer satisfaction was considered among the minimum service standards.

Recommendation 8: Involve service providers in defining outcome measures and evaluation design.

Given that the involvement of service providers helped to ensure that the most relevant outcomes were chosen for a number of the SOC programmes analysed, we recommend that service providers should be included in the process of defining outcomes. Furthermore, given the burden that evaluations impose on service providers, they should also be consulted regarding evaluation design. For example, during the initial phases of the programme, a test-run of data collection techniques could be implemented that would clarify what information will be used to gather evidence for payment-related outcomes. If necessary, service providers should be trained in how such data should be handled.

¹⁰⁵ With binary outcomes, providers have to achieve an absolute target, and no payment is granted for achieving lesser results. An example of a binary outcome would be whether or not a person found a job. In frequency schemes, rewards are staggered according to the agreed frequency of results, with payments increasing as results increase. For example, providers or investors would receive additional payments for every three months programme that participants remain employed.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at:

https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: <https://op.europa.eu/en/publications>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: <http://eur-lex.europa.eu>

Open data from the EU

The EU Open Data Portal (<http://data.europa.eu/euodp/en>) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

