IDS Bulletin
Transforming Development Knowledge
Volume **53** | Number **1** | February 2022

THEORY-BASED EVALUATION OF INCLUSIVE BUSINESS PROGRAMMES

Issue Editors **Giel Ton and Sietze Vellema**

Assessing the Contribution to Market System Change of the Private Enterprise Programme Ethiopia*

Giel Ton,[1] Ben Taylor[2] and Andrew Koleros[3]

Abstract An impact evaluation of a pro-poor market system development programme, 'Making Markets Work for the Poor' (M4P), poses several methodological issues for evaluators. M4P interventions intend to change the contextual conditions in which stakeholders take business decisions so that it triggers change processes in the wider social system and ultimately benefits poor people. An impact evaluation design for such a programme thus needs to be robust enough to adequately capture these systemic outcomes, acknowledging dynamic changes in intervention delivery as well as in market conditions over time. Theory-based evaluation can provide learning and accountability when it incorporates methods that allow a critical reflection on the key causal steps in an intervention's theory of change. We present our learnings about indicators and methods to reflect on the importance of the contributions to market system change of a large M4P programme in Ethiopia.

Keywords market systems, value chain development, mixed methods, job creation, theory-based evaluation, monitoring and evaluation, logframes.

1 Background

In interventions aimed at catalysing change in a complex system – that is, systems where multiple actors act and interact with each other and the wider environment to bring about change, such as economic sectors in a country, transformation of a political system, or mitigation of climate change – it becomes unreasonable and methodologically challenging to assess the relative effectiveness of the support of one of these actors to wider systems changes, let alone credit this to a single contributor (Earl, Carden and Smutylo 2001). The systemic effects that are measured at the system-wide level are well outside the sphere of direct influence of any one actor group, and hence any direct

The *IDS Bulletin* is published by Institute of Development Studies, Library Road, Brighton BN1 9RE, UK
This article is part of *IDS Bulletin* Vol. 53 No. 1 February 2022 **'Theory-Based Evaluation of Inclusive Business Programmes'**; the Introduction is also recommended reading.

intervention with that actor group. For example, if wider systemic outcomes were to decline over time, it would be unreasonable to blame an intervention for this negative change.

Moreover, an implementer would come up with countless reasons as to why this negative change could not be attributed to this intervention. The reverse logic is also true; if wider systemic outcomes are positively changing, it is unreasonable for the intervention to claim the credit for an improvement. Nevertheless, commissioners of evaluations often want to have an idea about the size or importance of a contribution to change at this systems level for multiple reasons, including the need to present this information at an aggregate level; for example, to account for their commitments to the Sustainable Development Goals. This creates the paradox that commissioners of impact evaluations pose legitimate but unanswerable questions about the precise size of their contribution to wider systems changes. As such, impact evaluators are often faced with identifying ways to reconcile the impossible with the possible (Ton *et al*. 2019).

In this article, we present learning from an attempt to do so in the area of market system development. The Private Enterprise Programme Ethiopia (PEPE) was an ambitious, £69m programme, funded by the UK's Department for International Development (DFID, now Foreign, Commonwealth & Development Office – FCDO), designed following the Making Markets Work for the Poor (M4P) approach. M4P is an approach to developing market systems so that they function more effectively, sustainably, and beneficially for poor people, building their capacities and offering them the opportunity to enhance their lives (Elliott, Gibson and Hitchins 2008; Nippard, Hitchins and Elliott 2014). By addressing underlying causes (rather than symptoms) of weak performance of the market system, M4P aims to unleash large-scale change. Interventions may be small in themselves but are expected to leverage the actions of key market players to bring about extensive and deep-seated systemic change (Tschumi and Hagan 2008).

The M4P activities of PEPE were implemented from 2013 to 2020 by a consortium led by the global consultancy firm DAI through a team based in Addis Ababa, named Enterprise Partners (EP). EP supported M4P innovations in three priority sectors (leather, textiles, and horticulture) and provided technical assistance to the financial sector to improve access to finance by micro, small and medium enterprises (MSMEs). Additionally, DFID contracted an external evaluation team, led by the consultancy firm Palladium. DFID had ambitious objectives both with PEPE and its external evaluation. It was intended to be the first evaluation to include an *ex post* analysis five years after programme completion, to capture the scale and sustainability of the innovations developed through the M4P approach within the wider market systems over a longer time horizon.

The total budget for the external evaluation was £2m and covered the costs of ten process evaluations to assess the outputs and progress of EP on an annual basis (referred to as annual reviews by DFID), and four impact evaluation milestone reporting events, at baseline (2016), mid-term (2018), endline (2020), and *ex post* (2024), to assess the outcomes and impact of EP's support to wider market systems changes in the three priority sectors: leather, textiles, and horticulture. In addition to the data-gathering exercises conducted by the evaluation team, EP had its own monitoring and results measurement (MRM) system following best practice in MRM system design (DCED 2017), which consisted of intervention logics for each support component along with progress indicators that were reviewed quarterly by programme teams to drive intervention adaptation and pivots (Enterprise Partners 2020; Yohannes 2020).

In this article, we present lessons learnt from this ambitious monitoring and impact evaluation effort. It starts with a brief introduction to the main features of M4P programmes. This includes a discussion of the adaptive management that is required for these programmes to work effectively, and the challenges this presents to rigorous impact evaluation, particularly when the desired impact is a wider sector-level change, such as job creation. Section 2 then discusses two major challenges faced by evaluators: defining and capturing early signs of systemic change; and designing an appropriate mix of methods to reflect on the importance of the contribution of the programme to these changes. Next, Section 3 illustrates how our impact evaluation addressed these challenges. Finally, Section 4 reflects on the results and Section 5 draws conclusions with advice for commissioners of impact evaluations in these types of programmes.

2 Challenges in the evaluation of M4P programmes

The ambition of M4P programmes is to find leverage points in market systems that change the dynamics in these systems in a way that more poor people benefit. That is, an M4P programme wants to trigger the motivation of firms and other stakeholders to innovate existing production, service delivery, or transaction practices in order to improve the functioning of the market system while including more poor people within markets. Consistent with wider principles around how change happens in complex adaptive systems such as economic systems (Beinhocker 2006), these changes rarely follow a dose-response relationship: the amount of effort or investment in an activity is not proportional to the size of the outcome; a small change in one actor or institution can cause a dramatic shift in the overall systems performance. These innovation processes often involve many stakeholder groups, and each will have a different perception of the related risks and rewards.

M4P programmes try to find leverage points by multiple activities, such as organising brokering events, elaborating proposals for

a policy change, or peer-learning activities around experiments and innovation. However, market systems change continuously and M4P programmes might find that the relevance of these activities shift or fade over time; what promised to be a leverage point may cease to be one once there is a shift in the market constellation. Due to this uncertainty inherent to complex systems (Snowden and Boone 2007), M4P programmes need to adapt and improvise, trying multiple activities while making sense of the ripple effects – experimentation and innovation is inherent to M4P programmes.

Moreover, the effects of support activities can manifest themselves much later. For example, an event where various sector stakeholders meet for the first time, such as producers and processing companies that discuss the strategies to improve the quality of raw material inputs, may appear to be fruitless in the short term, when they do not reach any common agreement on the ways to tackle the issue. However, this 'fruitless' activity may have resulted in personal networks between persons and organisations that lead to important systemic effects several years later, when the same persons contact each other for a rapid response to a policy proposal in the sector. What first could appear as being an insignificant event may prove a key event in the explanation of significant outcomes some years later. Evaluators of M4P programmes need to find ways to capture these unpredictable outcomes as a result of multiple, adaptively managed activities.

Implementers of M4P programmes, of course, have strong economic and reputational incentives to attribute results to themselves. They are often international consultancy companies that rely in their business model on successful projects or, at least, satisfied commissioners. They will have a tendency to overestimate their contribution. This requires evaluators thus to critically scrutinise both the rationale of the support activities and the evidence that supports the contribution claims (Stern *et al*. 2012; Mayne 2019). In M4P programmes – due to the multitude of activities – there is almost always a contribution to changes at the direct beneficiary level, often through business service providers or beneficiaries of innovation grants. The more interesting, but also more contestable, claims are usually related with its contribution to the performance of firms in the sector, such as increased trade, employment, or value addition in the sector, among firms that are not directly supported by the intervention and lie outside the sphere of direct influence.

Evaluators must thus find ways to verify whether the support can indeed be considered as a contributing factor in the wider configuration of changes among actors and other external factors over time that produced the observed outcome in performance. This implies a structured process of critical counterfactual thinking about alternative explanations of the

change process (Spellman and Mandel 1999; Stern *et al*. 2012; Yin 2013), and answering the question whether it is plausible – as assumed – that the support has been a non-redundant component in the configuration of causes that resulted in the outcome (Mackie 1974; Shadish, Cook and Campbell 2002; Mahoney 2008).

Even when the evidence suggests that an M4P programme has contributed to a systemic change in the market, this does not answer the question about the importance and effectiveness of this public investment in for-profit private actors. Donors need to aggregate and compare programmes at a higher, portfolio level, for example when deciding on new programme priorities in the region (ICAI 2015). John Mayne (2019: 4) indicates various ways for reflecting on the relative importance of the intervention's efforts in bringing about change in comparison to other factors. However, understanding the relative contribution of a support programme in one particular complex change process (= a causal configuration), is not enough for this portfolio analysis; there is still a need for some sort of ranking of various, alternative programmes (= multiple configurations) according to the size or importance of the outcome that resulted. Commissioners legitimately ask evaluators to give them an idea of the size of the impact to make priorities in future programming and budget allocation decisions.

This outlines two big challenges in the impact evaluation of M4P programmes. First, it is difficult to pinpoint what a systemic change in markets is, and how to 'capture' and monitor the early signs of it with sensible indicators. Second, impact evaluators need a research design that not only verifies whether a programme contributed to this change, but also helps to reflect on the importance of this contribution to judge the relevance for future funding of similar programmes.

3 Impact evaluation of PEPE

In this section, we describe how we attempted to address these two main challenges through the design of the impact evaluation of PEPE. As mentioned above, PEPE intervened in three priority sectors of the Ethiopian economy: leather, textiles, and horticulture; as well as interventions in the cross-cutting financial sector. DFID had selected these sectors in 2013 because of their potential for sector-wide transformation and poverty reduction, and followed the priorities defined in Ethiopia's Growth Transformation Plans (Diriba and Man 2019). For example, the focus on horticulture linked to the ambition to support the large number of newly commercialising smallholder farmers who could increase their incomes. With the focus on labour sourcing for industrial parks and the development of the leather manufacturing value chain, DFID expected to provide employment to part of the growing youth population.

3.1 Theory of change and logframe indicators

The overarching theory of change of PEPE and its accompanying logical framework (DFID 2018) were developed over time through a participatory process between the programme commissioners, the implementers, and evaluators, and provided the framework for the logframe indicators; i.e. how the programme was intended to report on progress and outcomes over time to DFID. Figure 1 is a simplified version of PEPE's theory of change; the full logframe differentiates between agro-industrial sectors and financial services, and included two non-M4P output areas, related to work with the International Labour Office and the Ethiopian Competitiveness Facility (a grant fund), and the external evaluation, led by Palladium. The backbone theory of change depicted in Figure 1 has a linear dimension that shows the intention to create higher level outcomes related with poverty alleviation through sector-level outcomes by multiple activities and outputs that are intended to find the leverage point in the market system.

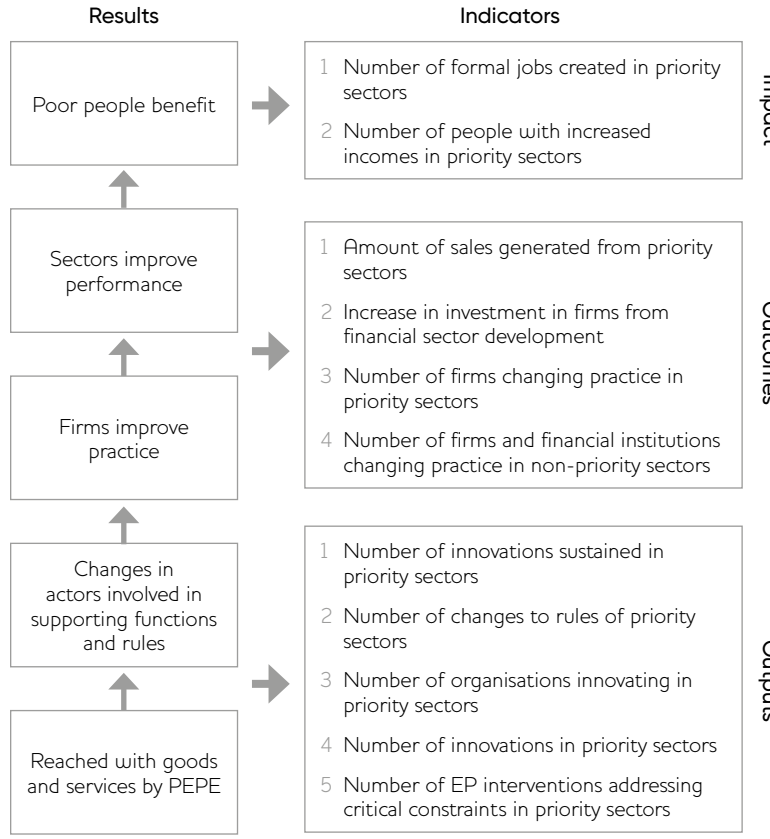
The logframe indicators align with the principles of M4P and allow outputs even when these do not directly lead to outcomes and impact. Changes in market systems result from non-linear processes and multiple activities need to be trialled to find leverage points that shift the system to a higher level of performance.

Following the M4P logic, all of these interventions were designed to trigger an innovation, but it was acknowledged that only part of these innovations would be successful in doing so and result in a significant and sustained change in practices of core market firms. EP describes its M4P approach as:

> a process of 'facilitated muddling through' [that] can only take place if a programme is set up as a learning organisation, able to adapt and respond to the context it engages with. Core to adaptive management in an MSD [market system development] programme is an ability to design innovative partnerships, map out how they are expected to work by means of a theory of change, test whether this theory holds true by means of continuous monitoring, and use the insight generated to engage stakeholders.
> (Bekkers 2020: 47)

The PEPE theory of change served as the backbone of EP's monitoring and result measurement system (Enterprise Partners 2020) but was much more granular in practice. Before starting with an activity, EP mapped the market system and developed more granular, actor-based theories of change (Koleros *et al*. 2020) with indicators and targets related to the specific subsector. Across the lifespan of EP, there were a total of around 100 interventions, each with their own intervention logic and activity plan.

Figure 1 PEPE intervention logic and logframe indicators



Source Authors' own, simplified from the PEPE logframe (DFID 2018).

EP developed a programme-level monitoring and results measurement (MRM) system that followed best practice guidance in the field (DCED 2017; Posthumus *et al*. 2020). The MRM team in EP generated annual, bi-annual, and quarterly data on the results of each of the programme's interventions. Each sector team had a person responsible for monitoring the results: the MRM person generated real-time data and analysis to support intervention managers to make decisions, and managers in turn provided qualitative input to data-gathering activities. Quarterly half-day workshops for each sector team created an opportunity for staff to provide input into each other's decision-making (Enterprise Partners 2020; Yohannes 2020). The MRM system was third-party audited by the Donor Committee for Enterprise Development (DCED).

Below we provide a more detailed description of the logframe indicators designed to capture the systemic nature of M4P (see Figure 1), and how these were operationalised by EP and the external evaluation team. Progress on the indicators was

self-reported by EP, using its MRM system, and reviewed by the independent impact evaluators during the annual reviews before being submitted to DFID.

3.1.1 Output indicators

As described above, output indicators were designed to report progress around how EP's interventions supported the development and sustainability of an innovation or rule change within the priority sectors. In the logframe (DFID 2018: Output 1), 'innovation' was defined as 'a change in the way a supporting function works in response to a critical constraint identified in the sector strategy'. Changes in rules included policies enacted, standards revised, regulations released, strategies validated, and directives or other rules enabled that address critical constraints in the relevant sectors. The term 'sustained' meant that the innovation continued for a minimum of 12 months after the end of direct support to the intervention. It was assumed that 50–70 per cent of innovations would be sustained, recognising that it would take two years after the start of the innovation before they could be reported as sustained.

For all output-level changes, EP developed a results chain showing how it contributed to the new innovation or rule change. The evaluation of the outputs sometimes implied expert judgements about what was considered as being an innovation – where to draw the boundary? Generally, the differences in judgement between EP and the evaluation team were small at output level, reflecting an unwritten rule that allowed flexibility in outputs as long as at least some of these delivered outcomes.

3.1.2 Outcome indicators

Much more discrepancy between the assessments of EP and the evaluators was present when the evaluation reflected on outcomes that resulted from these outputs. The logframe acknowledges that the indicators around investment posed particular problems for attribution. This is because investment is often a significant decision for a company, which is made based on many factors – not just EP. The guidance provided in the logframe, therefore, suggested that EP would seek to assess attributable investment where possible. Where it reports contribution, it isolates the specific investment that it has contributed to, and explains how this contribution was made, rather than reporting the whole investment. This obviously opened up a discussion between EP and the independent evaluation team, and indicated the methodological rigour required from the latter when verifying the reported outcomes.

3.1.3 Impact indicators

In spite of a consensus between implementers and evaluators that net effects of M4P programmes are only meaningful when applied on outcomes that are still in the sphere of influence of the intervention – firms that change their business practices in

response to the support – DFID maintained an interpretation of impact as being 'additional jobs created' and the 'number of smallholders that increased their incomes by a minimum of 20 per cent' (DFID 2018: Impact 1) as this was the basic metric on which the programme was awarded to DAI through a competitive procurement process. The logframe clearly asked for an estimate of the plausible size of the impact that resulted from their £69m investment in PEPE. The evaluation team, therefore, had to come up with a research design to do the impossible with the possible, and decided to give well-reasoned plausible range for this impact, instead of point estimates.

3.2 Impact evaluation design

Throughout the programme period, but particularly in the first half, the external evaluation team functioned as technical back-stoppers to the programme, with the annual reviews as the key moments of interaction. The accountability question became more dominant in the second half of the programme, from 2017 onwards, not least because the logframe targets became partly linked to a 'payment for results' element (DFID 2015). The methodological design for the impact evaluation was a learning process that can be divided into three phases, each associated with a different core method to assess and quantify the outcomes and impacts along the theory of change. Each phase had a methodological design that was revised and approved by DFID's Evaluation Quality Assurance and Learning Service (EQuALS).

3.2.1 Phase 1 – baseline

At baseline (2015/16), the core method proposed was to measure net effect in sector performance through changes in constraints in a firm's business environment. This followed the logic used to assess the value for money in the DFID Business Case by estimating the induced growth of value added at sector level. The impact evaluation team planned to use different data set analytical methods (econometric methods, social network analysis, and qualitative comparative analysis) that could show that the sector performance was associated with EP-induced changes in the perceived severity of constraints. The data needed included questions that asked for business performance (sales, employment, exports), and modules to identify and rank a long list of constraints and incentives that affected a firm's decision-making to invest, and asked respondents to rank their importance, similar to the World Bank Enterprise Surveys.

The results of the baseline survey, however, proved somewhat unsatisfactory. While the response rate of the survey was good and covered most formally registered small and medium-sized enterprises (SMEs) in each sector, not all the surveyed firms answered all questions related to the business constraints which made it difficult to aggregate results. Moreover, the team concluded that the baseline constraints prioritisation exercise was too imprecise to be used at mid-term as the core method

for inferences about the impact of EP support, for three main reasons. First, the persons that would respond at mid-term and endline about the firm's status and constraints could well have changed. Second, the prioritisation of constraints was unlikely to capture longer-term, more structural changes in the market system over time. Third, the data collected about the firm's economic performance were highly incomplete, as firm owners did not always want to give the exact figures – which often resulted in lacking or unreliable data (e.g. often, the enumerator was told to come back to interview other staff, even when this later proved to be unfeasible).

3.2.2 Phase 2 – mid-term

In 2017, based on this baseline survey experience, the methodology was revised. Compared with phase 1, the qualitative and quantitative research were much more interlinked in a mixed-method design that allowed triangulation of findings. Additional to the survey, the evaluation introduced process tracing case studies (see Figure 2), in the subsectors where EP claimed that the most significant outcomes had occurred. These process tracing case studies assessed the strength of the evidence for the claim that EP's outputs were a necessary, non-redundant component in the configuration of factors and actors that caused the change at outcome and impact level – the (early signs of) systemic change. The evidence for the claim was provided by EP and complemented by additional interviews and analysis by the evaluation team.

The case studies verified the logic of the result chain and, for example, probed whether the service providers had indeed improved their services due to EP support or would have provided these services to the firms in any case. To assess the strength of the contribution claim, we used four leading questions in a process of logical reasoning about the counterfactual situation (Ton and Glover 2019). These questions were adapted to the specific case at hand:

1 Did the change occur?

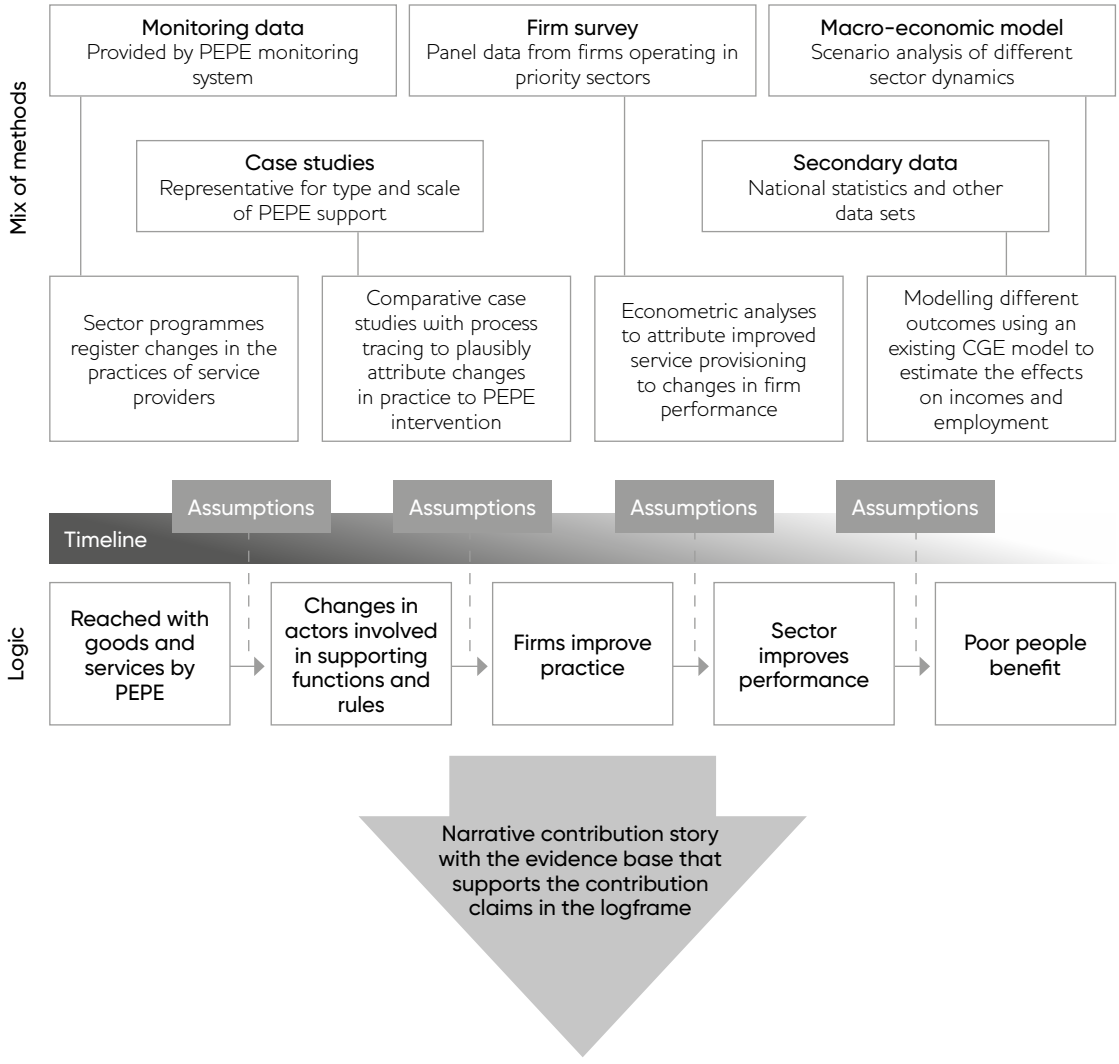
2 Did it result from a process in which PEPE-supported services were used?

3 Can this support be considered as a necessary (non-redundant) causal factor for that process to have taken place? And, if not, was it a necessary causal factor in accelerating or scaling of the outcomes?

4 Are there any other institutions or programmes that may have provided similar support to the change process, if the PEPE-supported services had not been present?

Each case study took around 25 days of research. At mid-term, this included a week of interviews in Ethiopia, and at endline, it included a series of online interviews. Most time, however, was

Figure 2 Methodological design at mid-term



Source Authors' own, based on Koleros, Taylor and Ton (2018).

spent on reviewing the documentation provided by EP in the 'evidence pack' distilled from their MRM system, and additional information provided by EP at the request of the evaluators. The case studies, especially at mid-term, explored the sustainability and importance of the effects of EP's activities at output and outcome level. At endline, the case studies focused primarily on the claims related to investment being mobilised and the impact level, the jobs created, and the number of smallholders with at least 20 per cent improvement in income.

Another adaptation to the methodology between baseline and mid-term in order to more accurately measure firm-level

changes was a complementary way to assess the firms' business performance. In order to address the real-world problem of incomplete survey data on the financial performance of firms that often need to operate in the grey area between the formal and informal in relation to the payment of taxes, the team decided to include a second, complementary way to ask for the performance change, using a less threatening way of asking about business performance.

Instead of relying only on the formal reported figures for before-after estimates of impact, the survey introduced questions that asked for perceptions of change in these performance indicators over the last three years. The scale had four intervals to indicate an increase or decrease in sales, exports, and employment: 0–25 per cent; 25–50 per cent; 50–100 per cent; more than 100 per cent. The information was based on the firm manager's perception of the change over the last three years, without requiring the exact figures of this change. This resulted in complete data on these percentual estimates and, again, a high number of missing values for the formal, absolute numbers from the financial statements of the firms. The survey resulted in data on 335 firms.

Moreover, the baseline modules for the prioritisation of constraints were substituted by modules that asked, for each of the 23 constraints, two perception questions that could be used to compute 'contribution scores' (Waarts *et al*. 2017; van Rijn *et al*. 2018). The perceived change in the severity of the constraint (using a five-point Likert scale) was combined ('multiplied') with the information about the perceived influence of the EP-supported service providers on this improvement (also a five-point Likert scale) into a contribution score. This was a sector- and constraint-specific list of business service providers and government institutions provided by EP's MRM team. These contribution scores can be interpreted as the 'perceived impact of EP-supported service providers on the constraint/outcome'. The average contribution score, considering all relevant constraints for which the perception questions were asked, was converted in percentage points and could fluctuate between 0 (no change or no influence) and 100 per cent (large change with a large perceived influence). The contribution scores allowed comparative analysis and subgroup analysis to detect meaningful differences in outcome pattern between types of firms and between sectors in statistical analyses.

3.2.3 Phase 3 – endline

The Covid-19 pandemic and related budget and logistical constraints forced us to make a change in the survey method. In 2020, when the endline survey was held, it was decided to limit the sample to only firms that were likely to have been within the influence of the EP-supported service providers. The 2020 endline survey covered 74 firms that had been in contact with one or

more EP-supported service providers. The Covid-19 pandemic and lockdowns affected the firms. The perceived change in performance, therefore, used the question 'Imagine the situation that the Covid-19 pandemic had not affected your firm. Can you give an estimate of the percentage change [in sales/exports/profits] that you would have had, without Covid-19, compared with three years ago?'

While the mid-term analysis of the contribution scores was largely restricted to map the differential impact of PEPE between different types of firms and between different sectors, for the endline evaluation, we went a step further and used them to assess the outcomes in sales, exports, and profits. To estimate a plausible treatment effect, we converted the 23 contribution scores for each of the 74 firms into seven support components using principal-component analysis. These components were used in regressions to test their association with the outcomes. For those sectors where a component proved significant (and only when the case studies confirmed the contribution claim), the coefficient in the regression was used as a scenario in the macro-economic CGE model of the Ethiopian economy (Tebekew *et al*. 2015) to estimate the lower and upper bounds of the EP-induced employment effects in the economy.

During the annual reviews in 2019 and 2020, it became evident that EP's M4P interventions managed to meet the output targets in the logframe but that these outputs did not (yet) result in the outcomes and impacts that were expected at the start. Most of the job creation was due to the support to a labour-sourcing innovation in Hawassa Industrial Park and the financing of women entrepreneurs and SMEs in two programmes funded by the World Bank that had received the direct technical assistance of EP. The challenge for the PEPE evaluation team was to come up with a reasonable way to estimate the EP-attributable effects within the total effects of these large multi-donor programmes. Therefore, two of the three process tracing case studies at endline verified the contribution claims related to the work in the financial sector.

3.3 Results

The impact evaluation resulted in an endline report that synthesised and combined the information from the methods depicted in Figure 2 (Koleros *et al.* 2018; Ton *et al*. 2021). The report shows that PEPE managed to reach their output targets but that this did not result in the expected level of outcomes and impact. Six case studies estimate, for each case, the causal effect of the activities and outputs on these outcomes and explain how this is backed up by the evidence and data generated in the MRM and the critical verification by the external evaluation team. For reasons of space, we illustrate the results by zooming in on only three of the six markets where PEPE claimed to have made a contribution to outcomes and impact.

These three cases best exemplify how the evaluation methods helped to critically verify the contribution claim of the implementer. The cases concern the support to labour sourcing in industrial parks, the development of agent-based seedling propagation models for smallholders, and the activities to develop a private capital investment advisory market. The texts in these three cases are taken from a much more comprehensive analysis of the cases in the endline report (Ton *et al*. 2021) and illustrate the way inferences were made.

3.3.1 Seed and seedling market

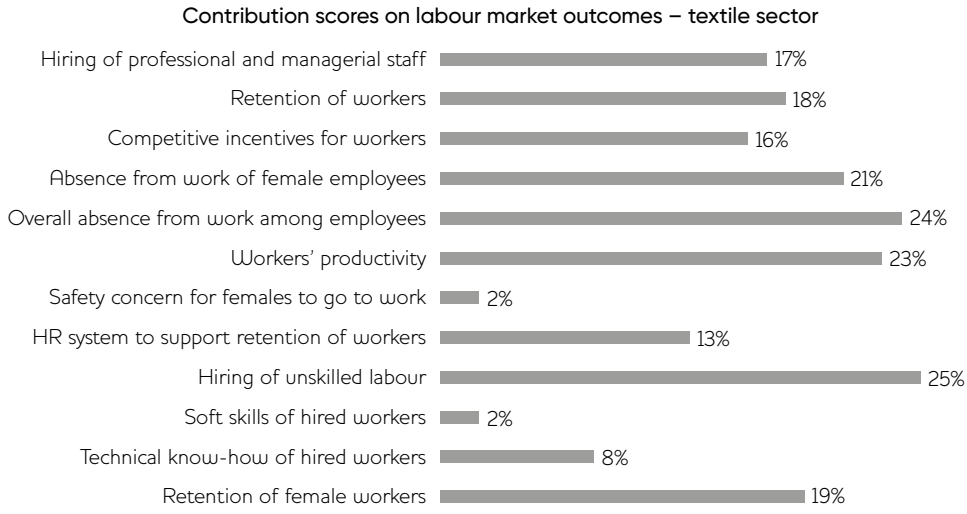
In the vegetable and fruit sector, PEPE identified that smallholder farmers struggled to consistently procure good-quality seeds and seedlings. With PEPE support, 14 propagators set up satellite nurseries in remote locations, involving around 400 agents who provided on-farm extension support to farmers that bought the seedlings. The propagators reached the vegetable farmers mainly through model farmers. Each of these was assumed to reach around five fellow farmers, who learn through demonstration effects. For the seed distribution model, the MRM system included a comparison between participating and non-participating farmers that suggest income increases in vegetable producers of more than 30 per cent.

However, the sample included mainly model farmers who are more likely to receive better training and technical support, and to have established stronger market linkages. In the seedling distribution model, farmers adopted improved fruit tree seedlings. However, it is not yet certain whether these farmers will have an income increase of 20 per cent in the future, because the income rise is still uncertain and contingent upon the continuation of care of these trees and future harvests. Taking both considerations into account, the endline report estimates that the vegetable seedling programme improved the income of a minimum of 3,416 farmers and a maximum of 17,082. The difference is due to this uncertainty in spread of the innovation beyond the model farmer.

3.3.2 Labour sourcing in industrial parks

An important theme of discussion in several annual reviews related to the way that jobs were created by the innovative labour-sourcing system in Hawassa Industrial Park (HIPSTER[4]), where EP had helped to establish a system of sourcing and grading of labourers to meet labour demand by the textile manufacturers that started operating there. Hawassa is the first and largest industrial park in Ethiopia located in a region where the potential workers are primarily located in rural villages. Consequently, Hawassa Industrial Park had unanticipated problems in attracting sufficient workers for the (textile) factories. The case study concluded that EP had effectively become part of the problem-solving task force to address issues with labour in Hawassa.

Figure 3 Perceived contribution of EP support to labour market outcomes in the textile sector



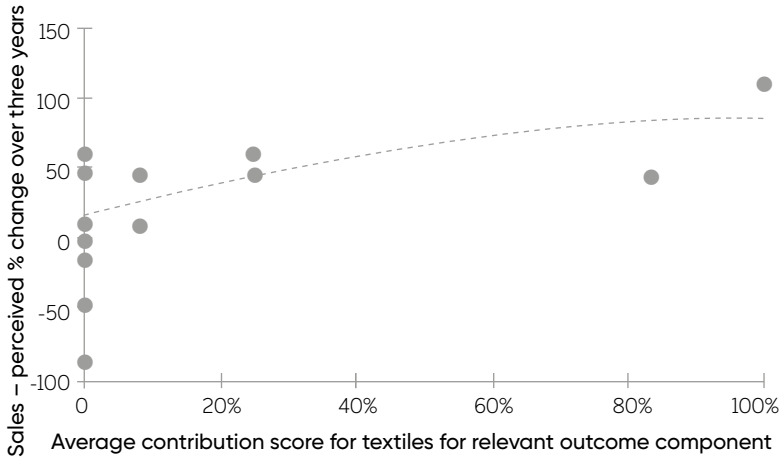
Source Authors' own, using endline data (Ton *et al*. 2021).

Since its start, more than 70,000 workers in Hawassa were screened and 60 per cent were also graded based on their skills. The problem for the impact evaluation was that the labour-sourcing system was mandatory for all firms in the industrial parks that required workers. The screening and grading component in HIPSTER is where PEPE has put most efforts. The screening and grading component was, however, also the component that did not appear to have worked particularly well in Hawassa and has not been replicated in other, more recently established industrial parks, which carry out combined sourcing and screening through local government departments.

The contribution scores (Figure 3) suggest that there is almost no perceived effect of EP support on soft skill but a small to fair effect on absentee reduction and workers' productivity. To assess the impact, the evaluators used the definition of job creation in the intervention logic that envisioned job creation through increased productivity. This means that it is inappropriate to count the number of workers that went through the sourcing system as jobs created by EP. However, HIPSTER has undoubtedly increased the effectiveness of the system. Without having a point estimate, the evaluation estimated that this efficiency is most likely not less than 4 per cent and not more than 10 per cent of the jobs that were created in Hawassa Industrial Park.

The contribution scores also helped to reflect on the performance-enhancing effect of EP, and an econometric regression showed that, for textile firms, the support was associated with an increase in sales of 2 per cent per year (Figure 4). Both elements were used to estimate the total

Figure 4 Association between contribution scores and performance in the textile sector



Note The curved line, instead of a linear one, being the most approximate to the regression results.
Source Authors' own, using endline data (Ton *et al*. 2021).

additional induced job creation with the CGE model in the range of 1,184 and 2,963 jobs in the priority sectors and an induced job creation in the national economy in the interval of 5,672 and 13,413 jobs.

3.3.3 Private capital advisory market

The case of the private investment advisory market was important for the outcome target related to investments in firms mobilised by new financial instruments, and illustrates the forensic approach used to assess the non-redundancy of the support in the complex process of systemic change. With a Private Capital Advisory Fund (PCAF), PEPE created a fund (initially intended as a revolving fund) for companies to hire investment advisors, who would help facilitate transactions by ensuring they meet the requirements of private equity investors in areas such as business plans, international financial reporting requirements, and valuations of assets. More than 30 grants were made available in 2018–19. In 2020, four companies were able to create improved business propositions and also attract a business partner, and generated a total investment of US$25m.

However, interviews by the evaluation team with the firms involved showed that in three of these cases, the advisors already had existing working relationships with companies prior to PEPE's intervention. While PCAF certainly helped to cover some of the costs associated with these deals, the deals themselves did not rely on PCAF. Moreover, PCAF saw uptake from companies only when it offered funds in the form of a grant rather than a loan, implying that there was no appetite for taking on the risk of hiring

an investment advisor. It was too early to ascertain any significant change in the market system. The case also highlighted the issue that the outcome indicators were not defined as additional (net effects) unlike the impact indicators. Therefore, even the small contribution of PCAF could be registered as investment mobilised in the logframe agreed upon with DFID.

4 Reflection

We found that the indicators used at output level in the PEPE logframe reflected the programme's theory of change and implementation approach reasonably well. However, they did not provide a particularly good accountability framework. The flexibility afforded at the output level – designed to allow evidence-based and adaptive programming – meant that it became possible to achieve the output targets by developing interventions that would never have impact on job creation or smallholder incomes at scale. And in some activities for which outcomes were reported, as in the case of PCAF, the importance of the contribution could be called into question. The disaggregate analysis of the logframe indicators helped the reflection on the importance of the different M4P components but the aggregates will hide nuances and lead to opportunistic, direct, programme-funded support activities with farmers and firms instead of the indirect support that characterises the M4P approach.

Monitoring of the ultimate outcomes and impact indicators is important for reflection on the relevance of the £69m investments of UK public development funds. However, estimating the net effects of changes in market systems that are well beyond the sphere of influence of an intervention is trivial and methodologically problematic (Ton, Vellema and Ge 2014). DFID insisted in its requirement for the evaluation team to quantify the net effects of PEPE on job creation. The sophisticated method developed to do so showed that PEPE's impact was far below targets, even when considering the higher bounds of the confidence interval.

Other donors, such as the Netherlands, decided to shift the focus in evaluating the importance of private sector development programmes away from this net-effect perspective and to ask for monitoring data about the aggregate sales and employment of all firms that were reached by a private sector intervention with a 'significant contribution' (DGIS-RVO 2017: 4). Monitoring the reach of a programme in relation to the number of firms, farmers, or jobs supported is far easier than computing net effects and still results in rough, indicative numbers that help to compare between programmes and interventions. Instead of requiring precise baseline–endline data with counterfactual designs, this requires research methods that evaluate the significance or importance of a contribution made by an intervention but without the need to quantify it, which appears, similar to what is argued by other

scholars (Goertz 2006; Mayne 2019), a better and more workable approach for evaluating the development impact of private sector development support, such as M4P programmes.

Especially in M4P programmes, there is a need to redefine what a rigorous impact evaluation design implies. Our experience showed that when the 'treatment' is highly variable, as inherent to M4P programmes, a treatment-comparison design to assess changes in outcomes or impact indicators is extremely vulnerable to changes in intervention modalities. We show that a careful analysis of change trajectories, and within the group of supported firms only, can yield a plausible estimate of impact, without the use of a comparison group. Estimates of the relative change over the last two years in sales, profits, and exports, combined with the perception questions used to compute contribution scores, proved sufficient to roughly estimate the impact of the support provided. We argue that asking directly for the perceptions of contributions or impact is a useful add-on to any survey that wants to capture M4P effects. Perception questions allow cross-sectional analyses and real-time reporting, can capture a wide range of outcomes, and help to build the resiliency of an impact evaluation design to changes in interventions, sample attrition, and evaluation conditions.

We learnt that rough measures of performance with high response rates are preferable over precise measures but with many missing data points. The competitive nature of firms makes it difficult to collect precise performance data. Therefore, even with relations of trust between the respondent and the enumerator and with well-crafted confidentiality agreements, missing data on sales, profits, and investments is notorious in firm surveys. Less precise but easier-to-collect data, for example asking for rough percentual changes in business performance indicators, as we did, helped to get a full data set that allows statistical pattern detection.

We maximised the potential to capture evidence/responses that could support the contribution claim but at the same time made it possible to critically assess the effects. The two core methods used, the firm survey, and the process tracing case studies, had features that allowed falsification of the claims. The survey did so quite straightforwardly, by asking the firm managers directly whether they used the improved services or regulations that addressed each constraint (see Figure 2) and, if so, how they rated the influence of these services in their business development. The contribution scores showed that only a few firms perceived a positive effect on these outcomes that they attributed to some degree to EP support. The average contribution scores in each sector rarely exceeded 16 per cent, which reads as 'a slight improvement and a slight influence' (Ton *et al*. 2021: 52).

The major drawback in the evaluation of PEPE related to the division of labour between the leading implementing company (DAI) and the leading impact evaluation company (Palladium). The contract stated that the evaluators were not allowed to influence the field activities and detailed interventions of the implementer, except by reflecting on the theory of change and the M4P approach. The evaluation's main task was to help DFID reflect on the effectiveness of the M4P approach and assist in fact-checking the reported progress according to the logframe targets. As inherent to *ex post* impact evaluations, the learning from the impact evaluation often comes too late to have a short-term follow-up. Also, in this case of PEPE, the decision to follow up was taken long before the results of the endline evaluation findings about the importance and size of the contribution to employment and smallholder incomes were available.

This delink between the endline evaluation outputs echoes the warning of the ICAI who warned that 'the more that evaluations are seen as prompts to evaluative and strategic thinking by programme teams, rather than products in their own right, the more useful they are likely to be' (ICAI 2015: 23). We think that as external evaluators, we could have done better in creating and feeding this strategic learning, continuing the more developmental evaluation process that characterised the annual reviews in the early stages of the PEPE programme. The logframe targets and external accountability became more important in the last years and, logically, created more sensitivities around the way the contributions to outcomes and impact were assessed and quantified. Together with personnel changes in EP, DFID, and Palladium, the decision to design a non-M4P programme as a follow-up, and the logistical challenges due to the Covid-19 pandemic, this translated into a more distant relationship between the stakeholders involved in the evaluation.

5 Conclusion

In private sector development programmes, where government funds are used to support private profiteering, the impact on development and public goods needs critical scrutiny; the risks of market distortion and corruption are simply too big to ignore. Therefore, we argue that it is legitimate to ask for an assessment of the size and importance of a contribution claim. However, computing a precise quantitative estimate of the size of a contribution is not possible. Nevertheless, as shown in the PEPE example, it might be possible to give a rough idea of the plausible range of effects that result from the support.

In the end, the quality of any evaluation design depends on the room for and quality of critical scrutiny (Patton 2012; Pawson 2013; Yin 2013). We argue that the critical scrutiny of contribution claims, articulated by the implementing stakeholder, and based on a reflection on the theory of change or intervention logic

provides a good starting and endpoint for an impact evaluation in M4P programmes. The theory of change provides the grammar for the contribution claims, and proper logframe indicators help to pinpoint the expected size and importance of the impact that is being pursued. However, due to contractual obligations and donor dependency, the targets specified in a logframe often take on a life of their own and activities are geared towards meeting the logframe's targets without concern for quality of the outputs and outcomes, and the nature of the impact.

We found that *ex post* process tracing of the most significant outcomes reported by the implementers is a method of critical inquiry and counterfactual reasoning that helps to balance the overreporting bias. *Ex post* process tracing is inherently resilient to changes in interventions, and economic and policy dynamics, including changes in the expectations and evaluation questions of the commissioners. It is especially useful when the contribution claim includes an outcome at the boundary of the sphere of influence (e.g. the ultimate outcomes) where the causal arrow is important but contested (or uncertain).

We argue that the commissioners could do better in prioritising methods and sense-making events that can inform the discussion around impact at mid-term in the terms of reference, instead of the current emphasis on rigorous impact evaluation designs that only produce evaluative insights at endline. Our advice for future impact evaluation in M4P programmes is threefold: verify the logic of the contribution claims with critical, forensic research methods; take perceptions of firm managers seriously; and refrain from point estimates of outcomes and impact but use minimum and maximum bounds of plausible effects. In sum, combine 'good-enough methods' with critical, evaluative, and counterfactual reasoning, to feed iterative learning cycles, involving the implementers, commissioners, and evaluators together in reflecting on the importance and logic of the evolving intervention logics of a programme.

Notes

* The authors were contracted by Palladium. Andrew Koleros was team leader between 2013 and 2017 before he moved to Mathematica. Ben Taylor of Agora Global was a member of the evaluation team between 2013 and 2021. Giel Ton of the Institute of Development Studies (IDS) joined the evaluation team in 2016. The authors acknowledge the support of Seife Ayele, Marinella Leone, Dirk Willenbockel, Ayako Ebata, and Keir Macdonald at IDS and Deepti Sastry and Ciana-Marie Pegus at Palladium, and Adam Kessler and Shumete Belete for designing and implementing EP's monitoring and results management system. We also acknowledge the constructive peer review of Sietze Vellema at Wageningen University that helped us to shape and sharpen the arguments.

1 Giel Ton, Research Fellow, Institute of Development Studies, UK. Corresponding author: **g.ton@ids.ac.uk**.
2 Ben Taylor, CEO, Agora Global, UK.
3 Andrew Koleros, Senior Researcher, Mathematica Policy Research, USA.
4 Hawassa Industrial Park Sourcing and Training Employees in the Region (HIPSTER).

References

Beinhocker, E.D. (2006) *The Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics*, Boston MA: Harvard Business School Press

Bekkers, H. (2020) *Enterprise Partners in Support of Industrial Transformation: Building an Industrial Labour Services Market in Ethiopia*, Enterprise Partners Case Study Series 2, Addis Ababa: Enterprise Partners

DCED (2017) *The DCED Standard for Measuring Achievements in Private Sector Development: Control Points and Compliance Criteria*, London: Donor Committee for Enterprise Development

DFID (2018) *Private Enterprise Programme Ethiopia Logframe 2018*, London: Department for International Development

DFID (2015) *Private Enterprise Programme Ethiopia Annual Review 2015*, London: Department for International Development

DGIS-RVO (2017) *15 Methodological Notes – Instructions for Calculation, Validation and Reporting of Performance Indicators*, The Hague: Dutch Ministry of Foreign Affairs

Diriba, G. and Man, C. (2019) *Building a Big Tent for Agricultural Transformation in Ethiopia*, CSIS Global Food Security Project, Washington DC: Center for Strategic and International Studies

Earl, S.; Carden, F. and Smutylo, T. (2001) *Outcome Mapping: Building Learning and Reflection into Development Programs*, Ottawa: International Development Research Centre

Elliott, D.; Gibson, A. and Hitchins, R. (2008) 'Making Markets Work for the Poor: Rationale and Practice', *Enterprise Development and Microfinance* 19.2: 101–19

Enterprise Partners (2020) *Enterprise Partners' Monitoring and Results Measurement System and DCED Experience*, Addis Ababa: Enterprise Partners

Goertz, G. (2006) 'Assessing the Trivialness, Relevance, and Relative Importance of Necessary or Sufficient Conditions in Social Science', *Studies in Comparative International Development* 41.2: 88–109

ICAI (2015) *DFID's Approach to Delivering Impact*, London: Independent Commission for Aid Impact

Koleros, A.; Taylor, B. and Ton, G. (2018) *Independent Evaluation of the Private Enterprise Programme Ethiopia (PEPE): Midterm Evaluation Final Report*, London: Palladium

Koleros, A.; Mulkerne, S.; Oldenbeuving, M. and Stein, D. (2020) 'The Actor-Based Change Framework: A Pragmatic Approach to Developing Program Theory for Interventions in Complex Systems', *American Journal of Evaluation* 41.1: 34–53

Mackie, J.L. (1974) *The Cement of the Universe: A Study of Causation*, Oxford: Oxford University Press

Mahoney, J. (2008) 'Toward a Unified Theory of Causality', *Comparative Political Studies* 41.4–5: 412–36

Mayne, J. (2019) ***Assessing the Relative Importance of Causal Factors***, CDI Practice Paper 21, Brighton: Institute of Development Studies (accessed 13 July 2021)

Nippard, D.; Hitchins, R. and Elliott, D. (2014) 'Adopt-Adapt-Expand-Respond: A Framework for Managing and Measuring Systemic Change Processes', *Briefing Paper*, Durham: Springfield Centre

Patton, M.Q. (2012) 'A Utilization-Focused Approach to Contribution Analysis', *Evaluation* 18.3: 364–77

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*, London: SAGE

Posthumus, H.; Shah, R.; Miehlbradt, A. and Kessler, A. (2020) *A Pragmatic Approach to Assessing System Change*, Boekel: Hans Posthumus Consultancy, Springfield Centre, Miehlbradt Consulting, and Donor Committee for Enterprise Development

Shadish, W.R.; Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston MA: Houghton Mifflin

Snowden, D. and Boone, M. (2007) 'A Leader's Framework for Decision Making', *Harvard Business Review* 85.11: 68

Spellman, B.A. and Mandel, D.R. (1999) 'When Possibility Informs Reality: Counterfactual Thinking as a Cue to Causality', *Current Directions in Psychological Science* 8.4: 120–23

Stern, E. *et al.* (2012) *Broadening the Range of Designs and Methods for Impact Evaluations: Report of a Study Commissioned by the Department for International Development*, Working Paper 38, London: Department for International Development

Tebekew, T. *et al.* (2015) *Ethiopia: Input Output Table and Social Accounting Matrix*, Addis Ababa and Washington DC: Ethiopian Development Research Institute and International Food Policy Research Institute

Ton, G. and Glover, D. (2019) ***Improving Knowledge, Inputs and Markets for Legume Expansion: A Contribution Analysis of N2Africa in Ghana and Ethiopia***, IDS Practice Paper 10, Brighton: Institute of Development Studies (accessed 13 July 2021)

Ton, G.; Vellema, S. and Ge, L. (2014) '**The Triviality of Measuring Ultimate Outcomes: Acknowledging the Span of Direct Influence**', *IDS Bulletin* 45.6: 37–48 (accessed 13 July 2021)

Ton, G. *et al.* (2021) *PEPE Evaluation: Endline Impact Evaluation*, London: Palladium

Ton, G. *et al.* (2019) ***Contribution Analysis and Estimating the Size of Effects: Can We Reconcile the Possible with the Impossible?***, CDI Practice Paper 20, Brighton: Institute of Development Studies (accessed 13 July 2021)

Tschumi, P. and Hagan, H. (2008) *A Synthesis of the Making Markets Work for the Poor (M4P) Approach*, Bern: Swiss Agency for Development and Cooperation, Federal Department of Foreign Affairs

van Rijn, F. *et al.* (2018) *Verification of PUM's Intervention Logic: Insights from the PRIME Toolbox*, The Hague: Wageningen Economic Research

Waarts, Y. *et al.* (2017) *Assessing IDH's Contribution to Public Good Impacts at Scale*, Wageningen: Wageningen University & Research and KPMG

Yin, R.K. (2013) 'Validity and Generalization in Future Case Study Evaluations', *Evaluation* 19.3: 321–32

Yohannes, L. (2020) *Adaptive Management: From the Inside Looking Out. Managing the Enterprise Partners Market Systems Development Programme in Ethiopia*, Addis Ababa: Enterprise Partners

This page is intentionally left blank