REVIEW OF M4P EVALUATION METHODS AND APPROACHES

Date: **April 2013**

Commissioned by the UK Department for International Development (DFID)

Prepared by Tim Ruffer & Elise Wach

Results in development

Table of contents

Executive Summary

This report presents the findings of a study undertaken by Itad for DFID to critically review the methods used to evaluate M4P programmes and thereby help guide the design and implementation of future evaluations.

Scope of this review

The review compiled information about the scope and purpose of evaluations of M4P programmes and analysed the trends, strengths, and weaknesses of these. The reviewers analysed 32 M4P programme reviews and evaluations, monitoring and evaluation (M&E) frameworks and programme reports, including 14 evaluation reports, and consulted with 14 individuals and organisations in the field.

M4P programmes are defined as playing a facilitative, adaptive role in order to contribute to systemic, large scale and sustainable market changes that positively affect the poor. The nature of the approach and the complexity of the markets within which it operates present a number of challenges for evaluation. Evaluation approaches that address these challenges are needed to avoid inaccurate estimations of impact.

Key findings

The M4P evaluations reviewed here were generally weak in terms of:

- consideration of **systemic, sustainable changes** in market systems;
- **data quality** (small sample sizes with little consideration of sampling frames, statistical significance or bias);
- **triangulation practices** (particularly with regard to qualitative data collection);
- the use of **theories of change** (those used were often linear, not externally vetted, with assumptions not adequately tested);
- **consistency in units** for facilitating accurate aggregation; and
- consideration of **unintended negative effects**.

Recommendations for practice

1. Evaluation method (how)

Evaluations that assess the extent to which M4P interventions result in market changes which are systemic, large scale and sustainable can serve to both 'improve' M4P programmes (through facilitating adaptive management) and 'prove' results for accountability. The majority of M4P evaluations reviewed here did not adequately assess impact in terms of whether it was systemic, large scale and sustainable. Those that were most successful at doing so were based on a theory of change that explicitly incorporated systemic change and evaluated results through a mixed-methods approach.

A **theory of change-based approach** to evaluation can help establish whether the linkages between interventions and intended impacts are plausible, account for other contributory factors, and capture unintended effects. Theories of change should be revisited frequently and vetted by stakeholders external to the project. While the majority of M4P evaluations were based on a theory of change, most of these evaluations did not adequately test the linkages contained in the theory.

The use of **mixed methods** in evaluation mitigates the risk of over-relying on one or two sources of evidence in the face of unpredictability. It is important for mixed methods approaches to be conducted with the same rigour and attention normally given to experimental approaches in order to minimise bias and ensure credibility. The majority of M4P evaluations were weak in terms of their qualitative data collection practices, reducing the reliability and robustness of their findings. Increased attention to rigorous qualitative data collection approaches is recommended for M4P evaluations.

Quasi-experimental approaches can be useful for measuring specific stages in the results chain or assessing discreet interventions at the pilot stage (before effects multiply) but face a number of challenges in terms of timing and location due to the adaptable, nonlinear

nature of M4P approaches. They are not suited to assessing the extent to which market changes are systemic, large scale or sustainable

The complex nature of market systems and the systemic nature of M4P interventions mean that adequately assessing **unintended effects** (both positive and negative) is crucial in M4P evaluations. There was considerable scope for improvement here in all of the evaluations reviewed.

Finally, evaluations need to examine more closely the impact and effectiveness of the facilitative and adaptive approach to M4P programmes – this is often held to be the key to the success of M4P programmes and yet has not been effectively measured or evaluated to date.

Evaluation timing (when)

In order to both estimate contribution in the context of other contributory factors and assess long-term changes, evaluation needs to happen both (a) during the programme, to ensure contributions are measurable and help facilitate an adaptive approach, and (b) later on (at the end or post-project), when systemic, long-term change has had time to unfold.

Evaluation responsibilities (who)

Institutional arrangements that ensure both objectivity in the evaluation and in-depth understanding of interventions and the context are important considerations for M4P evaluations. Approaches to achieve this balance include longitudinal evaluations through which the evaluator and evaluand build a collaborative relationship and / or internal data collection with external audits (i.e. the approach advocated by the DCED Standard).

Evaluation level (what)

Programmes can be evaluated at an intervention level or programme wide level. Where the intention is to demonstrate impact of the programme as a whole, a combination of top-down (programme-wide) and bottom-up (intervention specific) measurement is likely to address the inherent drawbacks of each approach to results measurement.

1. Introduction

A market systems approach to international development, often referred to as the 'Making Markets Work for the Poor' (M4P) approach, is being increasingly applied by many international development agencies. At the same time, agencies are placing greater emphasis on the need for evidence on the effectiveness of their investments. This has led to increasing demands for M4P programmes to better demonstrate results.

Significant efforts have been made to develop improved approaches to results measurement of M4P programmes, including through the development of the DCED Standard for Results Measurement[1] (a framework for private enterprise programmes to measure, manage, and demonstrate results) and the commissioning of a growing number of independent evaluations.

In this context, DFID commissioned Itad to critically review the methods used to evaluate M4P programmes and provide recommendations for good practice to help guide the design and implementation of future evaluations. The findings of the review are presented in this report. The review compiled information about the scope and purpose of evaluations of M4P programmes and analysed the trends, strengths, and weaknesses of these. The reviewers analysed 32 M4P programme reviews and evaluations, monitoring and evaluation (M&E) frameworks and programme reports in terms of evaluation scope, quality of evidence and key findings. The team also consulted with 14 individuals and organisations about the strengths, weaknesses and trends in M4P evaluation[2].

The report is structured as follows:

- Section 2 lays out the definitions of evaluation and of M4P which guided this review.
- Section 3 explores the characteristics of M4P evaluations in terms of why evaluations are performed, who performs the evaluation, when the evaluation take place and what is evaluated
- Section 4 presents an analysis of M4P evaluation methods, their strengths and weaknesses of current approaches to M4P evaluation.
- Section 5 summarises the conclusions and recommendations of the review.

[1] http://www.enterprise-development.org/page/measuring-and-reporting-results

[2] Annex 3 provides a summary of the 14 evaluations of M4P programmes that were considered in the review. In addition to the documents listed in Annex 3, a large number of programme reviews, case studies, M&E frameworks and guidelines were reviewed and have fed into our overall findings. These are listed in Annex 1. However the content of many of these documents was not suitable for summary in the template used for Annex 3.

2. Defining evaluation and M4P

2.1. Defining 'evaluation'

The OECD DAC defines evaluation as: "…an assessment, as systematic and objective as possible, of an on-going or completed project, programme or policy, its design, implementation and results. The aim is to determine the relevance and fulfilment of objectives, developmental efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors" (OECD DAC 1991).

For the purpose of the review, we have taken a wide interpretation of the definition of evaluation to ensure that we pick up lessons from a range of external process and impact evaluations, as well as programme reviews and internal results measurement processes (including monitoring guidelines and internal monitoring reports). The review has considered the following sub-categories of a broad definition of 'evaluation', each of which implies a different institutional model for the relationship between the evaluator and the evaluand (see Table 1).

- Internal results measurement.
- One-off independent evaluations or reviews.
- Independent impact evaluations.

Table 1: Alternative institutional arrangements for evaluation

Approach	Strengths	Weaknesses	Documents reviewed
Internal results measurement The majority of results measurement and impact assessment is undertaken internally (sometimes with support from external consultants). This increasingly follows the DCED Standard under which the results measurement system is periodically audited externally to ensure credibility of the results reported. In the case of Katalyst, which currently favours this approach, the programme occasionally commissions discreet intervention-specific impact assessments from external parties.	External engagement in results measurement inputs more targeted. Internal adaptive management more easily facilitated. In-depth of knowledge of programme and context because of involvement of implementers in results measurement. Timing more tailored to specific interventions (but no reason why longitudinal evaluations can't do this if managed well).	Risks of bias in results, for example due to: • self importance bias[3] • incentives of implementers to inflate success	The review considered the internal results measurement practices of a wide range of donors and implementing agencies, through a variety of guidelines, monitoring reports and other similar documents.
Once-off independent evaluations or reviews These are generally undertaken at the mid-point or end of the programme. While some may entail independent data collection, this type of evaluation typically relies on project monitoring and secondary data. The evaluations often apply a 'process' approach, focusing on the process of implementation, i.e. the way in which the interventions work, rather than concentrating on the achievement or non-achievement of objectives.	High degree of objectivity, although rigour depends on data collection and analysis processes.	Normally rely on secondary data and often limited verification of the quality of the data. Often superficial and not quantitative. Once-off nature means that they are limited in their ability to track longitudinal change with rigour. Risk that external and short term involvement will lead to a lack of ownership of the findings within the programme which may compromise the extent to which findings are internalised.	The review analysed five external process evaluations which reviewed the way in which the programme was implemented as well as the achievement of outputs, outcomes and impacts.
Independent impact evaluations These generally aim to look beyond the immediate results of programmes to identify longer-term effects. This is a model increasingly adopted by DFID for M4P programmes).	High degree of objectivity. Generally apply quantitative rigour. Long term engagement of evaluator can provide opportunity for them to develop familiarity with the programme.	Too much distance and short inputs mean evaluators can miss context. Risk of lack of internal ownership of evaluation findings. Limited application for adaptive management.	The review included analysis of five external impact evaluations which we define as impact evaluations that were conducted by a party independent of the donor or implementation agency.

[3] White & Phillips (2012).

2.2. Defining M4P

The M4P approach is based on recognition that economic poverty is the result of the structure of market systems in which poor participate. When markets work efficiently and produce equitable outcomes for the poor, they are a powerful vehicle for delivering growth and poverty reduction. The M4P approach aims to sustainably improve the lives of the poor by analysing and influencing market systems that affect them as business people (in terms of higher margins, increased volumes and improved market access), consumers (in the form of better access to products and services, lower prices and wider choice) and employees (in the form of higher wages and improved working conditions). It works to identify the underlying causes, instead of symptoms, of why markets do not work for the poor. M4P activities aim to facilitate change to the behaviour, capabilities, incentives and relationships of market actors in order to:

- improve target market systems, and
- create the conditions for markets to be continuously strengthened after the M4P 'intervention' is completed.

M4P is a flexible approach to development rather than a defined instrument. It has application in both economic and social fields. Building on a wide range of experience and learning, it recognises both the achievements and limitations of many conventional (i.e. more direct delivery) approaches and the growing number of diverse, successful applications of M4P[4]. There are therefore no textbook M4P projects or blueprints for intervention, since intervention choices should emerge based on needs and context.

The M4P literature is broadly consistent in specifying the key attributes that define the approach. In terms of the methods of implementation, M4P programmes play a **facilitative**, **adaptive role**. In terms of the impacts they seek to achieve, M4P programmes aim to contribute to **systemic**, **large scale** and **sustainable** changes that positively affect the poor. Each of these attributes is described below[5].

Implementation Approach:

Facilitative role: M4P programmes aim to adopt a facilitative role, acting as a catalyst to stimulate, but not displace, market functions or players, thereby 'crowding in' market players and activity. Achieving this requires a rigorous analysis of complex social, political or economic systems to ensure that programme designers think about the incentives and interests that encourage individuals to undertake particular roles or functions in systems. Transforming complex systems sustainably is often about finding subtle, innovative and enduring ways to respond to and change incentives or challenge particular interests, rather than directly orchestrating shifts in behaviour en masse.

Adaptive in nature: The dynamic and unpredictable nature of market systems means that programmes need to be flexible and presents a strong case for an experimental and adaptive approach.

Desired Impacts:

Systemic change is defined as transformations in the structure or dynamics of a system that leads to impacts on the material conditions or behaviours of large numbers of people. M4P focuses on

[4] Springfield Centre (undated).

[5] Whilst these attributes are relatively easy to describe, the degree to which a programme's interventions are consistent with them is difficult to assess objectively, particularly in the context of a desk-based exercise where limited information was available to the reviewers. The review applied a generous interpretation of compliance with these attributes in the selection of "M4P programmes" that were considered to ensure that it was comprehensive and drew lessons from a wide range of relevant programmes and evaluations.

systemic action: understanding where market systems are failing to serve the needs of the poor, and acting to correct those failings. The approach takes account of interrelationships in the market system and targets interventions at critical weaknesses in the system.

Sustainability: M4P seeks sustainable change from the outset - delivering sustainable outcomes by better aligning key market functions and players with the incentives and capacities to work more effectively. Sustainability is not just about maintaining the *status quo* achieved by a project intervention without continued external support. It is also about the long-term integrity of dynamic processes, the resilience of the system to shocks and stresses, and the capacity to evolve or innovate in response to an inevitably changing external environment. This dynamic dimension to sustainability is very important because it suggests that underpinning the outward or superficial performance of any 'sustainable' system are a variety of critical but often less visible institutions and functions.

Large-scale: M4P programmes are designed to achieve large-scale change, benefitting large numbers of poor people beyond the programme's direct sphere of interaction. Interventions explicitly envisage mechanisms for replicating, extending or multiplying results so that, at least potentially, they could reach very large numbers of beneficiaries. It is not that every intervention has to directly reach the large scale, but rather that the envisaged route to large-scale impact is credible. Whatever scaling up logic is envisaged should be explicit in the design of programmes and interventions.

These attributes of M4P programmes have been used to structure the analysis of evaluation methods provided in this report in Section 4.

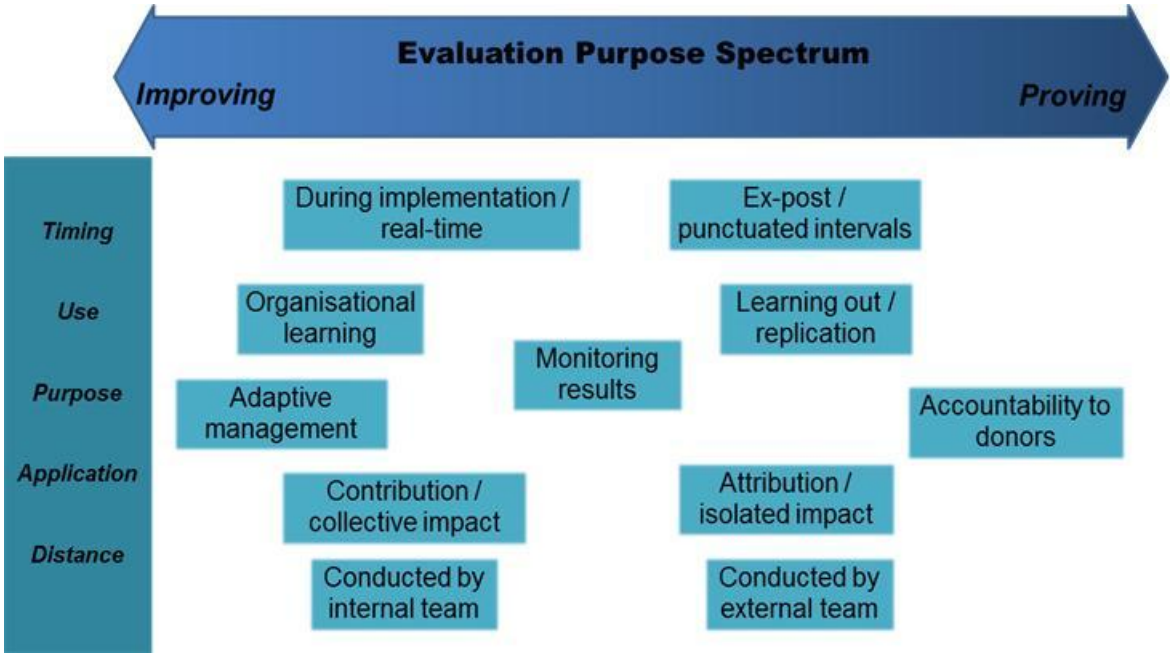
3. Characteristics of M4P evaluations

This section considers the characteristics of M4P evaluations by looking at questions of **why** evaluations are performed, **who** performs the evaluation, **when** the evaluation takes place and **what** is evaluated.

3.1. Why: evaluation purpose

Evaluations are often seen to have two primary purposes: *improving* programme implementation or *proving* programme effectiveness. There can potentially be tensions between these purposes, including balancing degrees of usefulness and credibility (Creevey et al 2010), which relates to issues of independence versus an in-depth understanding of the programme and the context (Tarsilla 2010), the indicators used for measurement, the timing of the evaluation, and other concerns. Given that many evaluations do not fit neatly into one category or another it is useful to think of them along a spectrum, as shown in Figure 1.

Figure 1: Evaluation purpose spectrum



Improving: Evaluation used for 'improving' practice is typically focused on how and why change has happened and how the programme can adapt and improve (i.e. organisational or programmatic learning) in order to maximise its effectiveness (Osorio-Cortes and Jenal 2013). This requires significant focus on the processes and changes in the system (e.g. rules, incentives, relationships, capacity, and practice) rather than the final impacts at the household or enterprise level (e.g. income or poverty). Evaluations focused on 'improving' are often carried out by teams that include members of the implementing organisation, though can also be facilitated externally.

Due to the experimental nature of many M4P programmes, evaluation that provides real-time information to facilitate adaptive management can help contribute to improved programme performance. Impact assessments (either internal or external) conducted partway through implementation fall into this category. Many programmes have an adaptive and responsive internal research programme (e.g. Katalyst), that includes impact assessment, which is intended to direct programme planning.

Proving: Evaluations focused on 'proving' results are typically used for accountability and cost-effectiveness reasons: they seek to understand the overall impacts of the programme. These tend to be *ex post* evaluations which often seek to ascertain attribution: the changes that have occurred as a

result of the intervention (Kandhker et al 2010). Information from these evaluations might inform decisions for continuing, scaling up, or replicating programmes. They are typically conducted by teams external to the intervention for the sake of objectivity, though the need for an in-depth understanding of both the intervention and the context is often required to ensure accurate capture of effects (Tarsilla 2010).

Recommendations for good practice

From stakeholder consultations and a review of existing evaluation practice, in the case of M4P programmes, it appears that the best evaluation approaches need to serve both 'proving' and 'improving' functions. Improving inevitably cuts across the responsibilities of both programme managers and evaluators and therefore collaboration between these two parties is required. The improving focus of mid-term evaluations should be on strengthening the efficiency and effectiveness of programme delivery. The improving focus of final evaluations should be on generating learning regarding the M4P delivery model and therefore improving its future application.

The implications of this on the institutional arrangements, timing and selection of interventions that are included in the evaluation are considered in turn below.

3.2. Who: Responsibilities for evaluation

Many M4P programmes simultaneously apply a variety of models for 'evaluation' (i.e. internal evaluation, external review, and independent impact assessment, as detailed in Table 1 above). Interestingly, none of the evaluations involved a long-term, sustained collaboration between evaluator and evaluand: all of the 14 evaluations reviewed were either one-off externally conducted evaluations or were largely conducted internally.

An alternative way of categorising evaluations is according to whether data is collected and analysed internally or externally. Of the 14 evaluations reviewed, eight were based primarily or exclusively on internally collected data, one was based on secondary data, and five used externally collected data. This is detailed in Table 2 below. Unsurprisingly, independent impact evaluations all rely on externally collected data, whereas internal results measurement uses internally generated data. External reviews apply a variety of models for data collection and analysis, and a reliance on internally collected data is not uncommon.

Whilst independence on the part of an evaluator is desirable for objectivity, for M4P evaluation, it is equally important that the evaluator has in-depth knowledge of interventions and context, given the complex nature of the programmes and markets. Internal data collection can bring this familiarity with the programme and also maximise the chances that evaluation findings are internalised. However, internal data collection is subject to risks of bias.

Table 2: Responsibilities for data collection and analysis in M4P evaluations reviewed

Evaluation Report[6]		Internal	External	Evaluation type
1. Katalyst Impact Assessment	Data collection		X	Independent impact evaluation
	Data analysis		X	
2. Cambodia MSME final M&E report	Data collection		X	Independent impact evaluation
	Data analysis		X	
3. AgLink Egypt final report	Data collection	X Supplemented by stakeholder interviews		External review
	Data analysis	X		
4. PrOpCom Project Completion Report	Data collection	X		Internal results measurement
	Data analysis	X		
5. SECO Cooperation in Business Environment Reform,	Data collection	X		External review
	Data analysis		X	
6. Impacts of the KBDS & KHDP projects in the tree fruit value chain in Kenya	Data collection		X	Independent impact evaluation
	Data analysis		X	
7. Effectiveness assessment of the GMED India project	Data collection	X		Independent impact evaluation
			X	
8. Second Thanh Hoa bamboo survey	Data collection	X		Independent impact evaluation
	Data analysis		X	
9. PrOpCom tractor leasing case study report	Data collection	X		Internal results measurement
	Data analysis	X		
10. Enter-Growth Project Sri Lanka, Final Evaluation	Data collection	X Supplemented by external stakeholder consultations		External review
	Data analysis		X	
11. PROFIT Zambia Impact Assessment Final Report	Data collection		X	Independent impact evaluation
	Data analysis		X	
12. Joint SDC – Irish Aid Review of the Mekong Market Development Portfolio Programme (MMDPP)	Data collection	Mixed team		Internal results measurement
	Data analysis	X		
13. Enterprise Challenge Fund Mid-term review	Data collection	X Supplemented by external stakeholder consultations, field visits		External review
	Data analysis		X	
14. Cross-section of independent evaluations in PSD 2007 GTZ	Data collection		X	External review
	Data analysis	X		

From stakeholder consultations, there appear to be three different institutional arrangements that can provide the necessary balance of objectivity and distance:

[6] See Annex 3 for full titles of the evaluation reports.

Internal data collection verified through an independent DCED Standard audit and occasionally supplemented by discreet impact evaluations. This is currently favoured by Katalyst. DCED audit provides assurance regarding monitoring processes and includes analysis of the way in which impact is attributed to the programme. DCED Standard audits supplement the external evaluation process, but the audits do not necessarily address donor-specific evaluation questions, or make use of donor-preferred evaluation methodologies.

A longitudinal collaboration between evaluator and evaluand. This model is being increasingly adopted by DFID and can help ensure a desirable combination of independence, relevance and utility of M4P evaluations. However, there are no historical examples of this arrangement from which to draw lessons.

The use of a **pool of external evaluators** who have in-depth knowledge of the M4P approach, but who are sufficiently distanced from programmes and implementing teams to avoid bias. However, stakeholders pointed to the scarcity of such a pool, which creates challenges in the application of this approach.

The appropriate institutional model is partly dependent on the scale of the programme and the resources available for evaluation. Clearly a larger programme is more likely to have available the resources required for external longitudinal evaluation than smaller programmes. The review found that the M&E budget for an M4P programme is typically in the region of 6-9%, of which around 50% is normally spent on internal monitoring and 50% on external evaluation or impact assessment.

Recommendations for good practice

Evaluators need to be both (a) sufficiently familiar with M4P approaches to design appropriate evaluations, and (b) sufficiently distanced from the programme in order to provide the required degree of independence. This was stressed repeatedly by stakeholders in our consultations as a key factor in ensuring relevance and accuracy of M4P evaluations, and is also advocated by Tarsilla (2010). Thus, internal results data collection and analysis with external audits and / or longitudinal collaborations is preferred over purely internal or purely external evaluation arrangements.

The appropriate model is partly dependent on the context of the programme and the resources available. For larger M4P programmes (e.g. valued at over £10 million), a combination of internal monitoring with external audits *and* longitudinal external evaluation is in many cases likely to be optimal. Where this is the case, the approach should ideally include:

The appointment of the evaluator at the beginning of the project;

Collaboration between the project team and evaluator on reviewing the theory of change and the monitoring data that will be gathered for the project;

Agreement on the programme of “evaluative learning”/research that will be conducted and the division of responsibilities for this work between the project team and evaluator;

Agreement on the application of the DCED standard and its impact on the scope of the evaluator's work.

3.3. When: Timing of the evaluation

The timing of the evaluation and data collection determine the type of information provided and the ways in which it can be used. Data can be collected *ex-ante*, mid-project, end-project and post-project.

Around half of the evaluations reviewed here were conducted during project implementation, implying that the balance of purpose was more towards ‘improving’. The other half of the evaluations analysed in this review were conducted at the end of the project. For these, the

implication is that the emphasis was on 'proving' and providing information for wider lesson-learning external to the project.

The review did not uncover any evaluations that had been conducted post-project (i.e. two or more years following project completion). As much of the impact of M4P interventions is achieved after programme intervention has completed, evaluations conducted during implementation or at completion are likely to only identify full change at certain levels in the results chain. There is therefore a risk that they will underestimate impact and be unable to effectively assess whether changes have been systemic and sustainable. However, clearly, the degree of attributable change will reduce over time as the influence of other contextual factors builds.

The timing of different levels of effect is difficult to predict and there is a trade-off in selecting the timeframe for the follow-up data collection and analysis. With too short a period between the before and after surveys, the impacts may not have had time to take hold and may not be measurable. But with too long a time period between the before and after surveys, the amount of attrition in the panel of respondents may distort the results.

For example, an evaluation of Katalyst interventions[7] identified that impacts will occur at the wholesale service provider level first, then at the retail service provider level, and only after the retail-level service providers have adopted and promoted the innovations could we expect to see change at the end-user beneficiary level. Furthermore, the type of impact we could expect to observe would itself reflect the behavioural change pattern. We would first expect to see increases in knowledge about a new technique or practice, then changes in attitudes about adopting or using the new technique or practice, then changes in terms of actual behaviour in implementing the practice. Only after the practice is implemented would changes in production, productivity, income, employment and standard of living occur.

Similarly, in a meta-evaluation of business training and entrepreneurship interventions[8], it was highlighted that whilst one might expect firms to make some changes relatively quickly after training, the full impact of training may take some time. However, firms could start some practices and then drop them, so that surveys which measure what is taking place in the business only several years after training may miss the period of experimentation. Ideally, then studies should trace the trajectories of impacts, measuring both short and longer-term effects. However the majority of studies take a single follow-up survey, providing a snapshot of information on the training impact, but no details on the trajectory of impacts.

Recommendations for good practice

For M4P programmes, evaluation needs to happen **both** (a) during the programme, to ensure contributions are measurable and help facilitate an adaptive approach, **and** (b) later on (at the end or post-project), when systemic, long-term change has had time to unfold. This justifies the application of a longitudinal approach to evaluation.

The theory of change and logframe should clearly indicate when the expected impacts are likely to occur and define the anticipated "trajectory of change"[9]. These assumptions should then be used to determine the evaluation strategy and timing at the start of the project. However it is important to recognise that unexpected change is almost inevitable due to the unpredictability of market dynamics; and that as such, deviations from the anticipated trajectory of change should not necessarily be considered negatively.

[7] Magill, JH & G Woller (2010).

[8] McKenzie D & C Woodruff (2012).

[9] See Woolcock (2009).

3.4. What: Selection of interventions for evaluation

Many M4P programmes apply a 'portfolio approach' where multiple interventions are implemented under one programme. A testing and piloting approach is applied where successful interventions are pursued further and/or replicated, and unsuccessful interventions are dropped. In evaluating such multi-stranded programmes, there is an option of either evaluating the entire programme's impact (i.e. 'top-down' measurement, which assesses the key changes in indicators at the impact level and then analyses the factors at the market level that have driven this change) or focusing on a handful of interventions (i.e. 'bottom-up' measurement, whereby intervention-based monitoring is applied to assess whether and how interventions have achieved market level change and the extent to which this has led to an improvement in the performance and inclusiveness of the market system). There are advantages and disadvantages of each approach (see Box 1).

Box 1: Bottom-up and top-down measurement challenges

Bottom-up measurement (intervention-based)	Top-down measurement (programme-based)
Risk of double counting impact across interventions. Risk that measurement will ignore deadweight loss and displacement. Difficult to account for impact of synergies across different components of a programme. Risk of 'self-importance bias' in estimating attribution.	Very challenging / costly to make surveys & quantitative analysis representative. Large attribution challenges – big jump between micro interventions and macro economy-wide impacts.

Source: Itad (2012)

The pros and cons of both the bottom-up and top-down approaches to evaluation have resulted in suggestions that the two should be combined (e.g. ITAD 2012).

Bottom-Up (Intervention specific) Approach

Where only a selection of interventions can be evaluated, a variety of approaches to intervention sampling can be applied:

Random sampling: Some argue for random sampling to reduce selection bias. For example, DCED audits (e.g. in the case of Katalyst) take a square root of the total number of interventions and select a random sample. Others argue for 'hand-picking' of interventions – deliberately selecting either success stories or failures.

Evaluating success: The justification for focusing on successful interventions is that the experimental and 'portfolio approach' of an M4P programme often means that the majority of overall impact from a programme is likely to come from a small number of interventions. The implication is that a random selection of interventions could miss the major areas of impact that a programme achieves. However a challenge in 'hand picking' successful interventions in longitudinal evaluations is that an evaluator will not know in advance which interventions will be a success and therefore for which interventions a baseline should be developed.

Evaluating failure: An opposing view, which has recently been discussed in the wider context of evaluation in international development suggests that "*the role of evidence is not to prove that things work, but to prove they don't, forcing us to challenge received wisdom and standard approaches*" [10]. This suggests that learning is most likely to be effective from failure, in a context where there are strong incentives for programme implementers and donors to not be open about failure, which results in systemic bias (Sarewitz 2012, Ioannidis 2005). This has been a driver behind

[10] http://www.oxfamblogs.org/fp2p/?p=13590

the "Admitting Failure" initiative[11] of Engineers Without Borders, which includes documentation of failure in at least one M4P programme (PROFIT in Zambia[12]).

Box 2: The importance of capturing failure

"It's only when we reached the stage when we were admitting that it was not working that we started to learn...Failure is painful to admit. And a lot of people that are in development work push an intervention just because they need to keep the numbers on and they don't admit it in a report on the table".

Carity Ngoma, discussing learning from failures in the PROFIT Zambia programme

Of the evaluations reviewed, seven analysed the full set of interventions implemented by the programme, whilst seven selected a sample. The basis for the selection of a sample of interventions is not always made clear in the evaluation report. Examples of the approach to intervention sample selection where it is explicitly explained are provided below:

- The **PROFIT Zambia impact assessment** selected interventions based on an evaluability assessment at the start of the evaluation. However, because the location for one of the interventions shifted over time, and the evaluation had employed a longitudinal target and control group approach, the evaluation of one of the interventions was not effective.
- The **Cambodia MSME M&E** report took a random sample for the before/after analysis that was undertaken.
- The **PrOpCom PCR** focused primarily on the interventions that enjoyed the greatest success through the long life of the project (and whose implementation was therefore maintained through to the latter years of its implementation).
- The **Katalyst impact assessment** examined four of the approximately forty activities carried out under the Katalyst/DBSM programme. The evaluation report states that "these activities were selected for practical reasons – their timing, duration, and magnitude"[13].

None of the evaluations reviewed here explicitly evaluated failure. Given the opportunities for learning from failure and the biased information provided by only reporting on successes, it is recommended that more M4P programmes incorporate case studies of failure in their evaluations.

When combining the results of bottom-up evaluations of a number of interventions care must be taken because[14]:

- Many indicators may be defined in different ways in different contexts. This may result in the aggregation of inconsistent units.
- There is a risk of double counting between interventions and components – e.g. at the impact level between employment and income effects

[11] http://www.admittingfailure.com/

[12] http://www.admittingfailure.com/failure/charity-ngoma/

[13] Timing was important because the study needed to assess projects that were just beginning or that had not progressed to a stage at which meaningful baseline data could not be collected. Magnitude was important because the activity had to have the potential of causing a measurable impact in the targeted beneficiary population. And duration was important because there needed to be a sufficient period of activity to project measurable impacts. Many of Katalyst's ongoing activities were mature programmes that offered no possibility of collecting needed a priori information or baseline data. Others were too limited in scope and time to serve in a longitudinal study.

[14] Itad (2012).

- By aggregating results from individual interventions or components, there is a risk that the impact of synergies between programme parts is missed.

Many of the evaluations reviewed were not fully clear on whether these factors were taken into account in aggregation processes. An example of good practice is PrOpCom:

"Overlaps due to PrOpCom's interventions are likely since PrOpCom works in both value chains and cross cutting sectors, moreover these interventions in many cases occur in the same geographical locations and hence are likely to reach the same beneficiaries. Therefore, when aggregating impact the M&E team will identify the overlapping interventions and the areas where overlaps are likely and take steps to account for those".

Top-Down (Programme-Based) Approach

Programme-wide evaluations can capture the effects of the overall programme, which can help in capturing synergistic effects between multiple interventions and also prove useful for accountability purposes. However, the breadth of interventions means that programme-wide evaluation can be challenging and resource intensive, whilst evaluating the effects of discreet interventions can provide opportunities for in-depth learning about what is working, not working, and why.

None of the programmes reviewed as part of this study applied a top-down programme based approach to evaluation.

Combination (Top-down and Bottom-up) Approach

A third approach is to bring together the intervention based and programme based evaluation approaches. This approach can enable the triangulation of evidence of change, a reasonably robust approach to measuring attribution and minimise self-importance bias. The approach recognises the difficulty of assessing the influence of interventions at outcome and impact levels without a broader understanding of sectoral social and economic performance.

The approach entails synthesising and cross-checking the results of top-down and bottom-up measurement by focusing on the levels where the two steps come together at the output and outcome levels. It assesses the extent to which the outputs and outcomes achieved by programme interventions are consistent with the market-level factors that have driven changes in impact level indicators. This allows summarising, synthesising and double-checking of results.

Of the evaluations that applied rigorous quantitative methods to their impact assessment, there was very limited triangulation of evidence of wider sector change through a top-down approach. In stakeholder consultations, it was generally agreed that combining a top-down with a bottom-up approach could add depth and rigour to evaluation findings, although caution was raised with regard to the additional resources and information requirements that would be required for this in many cases.

Recommendations for good practice

A combination of 'top-down' and 'bottom-up' approaches to measurement is most likely to address the inherent drawbacks of applying either approach on its own.

The selection of interventions for evaluation is important in determining its efficiency and effectiveness. A number of factors need to be taken into account in selecting interventions for evaluation, including:

The strength of the assumptions and causal links in the programme's theory of change, and the need to generate further evidence to support them.

The evaluability of the interventions.

The resources available for the evaluation and a consideration of the optimal balance between breadth and depth.

The need to avoid selection bias and the tendency to only select success stories.

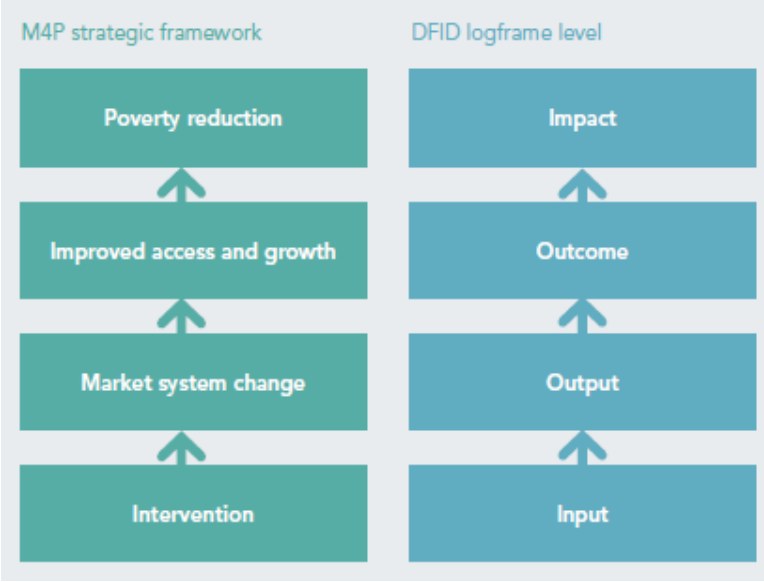
For longitudinal impact evaluations, there is significant risk associated with selecting a small number of interventions at the baseline of an evaluation, given that some may be dropped, or adjusted so significantly that quantitative baselines become obsolete.

Care must be taken when aggregating results to ensure the consistency and accuracy of conclusions made.

3.5. What: Levels of measurement

Our review of the results frameworks used in M4P evaluations has shown that they are generally very linear in nature, mirroring the levels espoused in the M4P strategic framework and DFID logframe (see Figure 3).

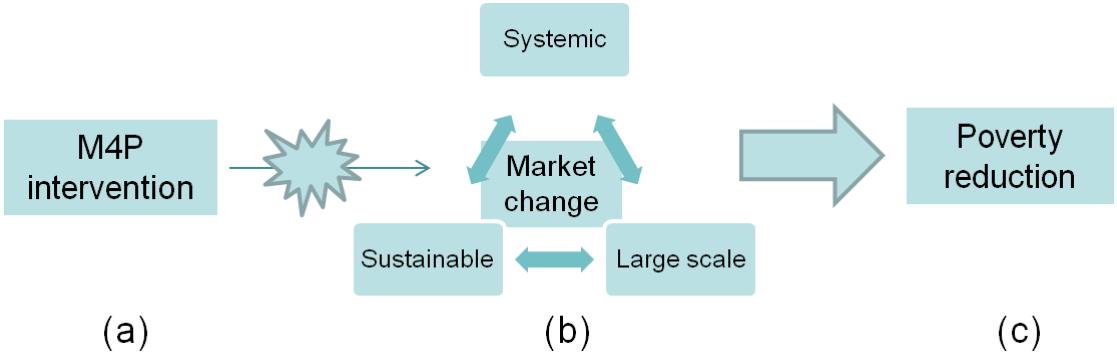
Figure 2: A linear logic model is often applied but can be unhelpful for M4P evaluation



In the selection of indicators for the measurement of impact, four evaluations considered poverty impact, eight considered household income, nine considered effects on jobs, and eleven considered impacts on enterprise income (note that some evaluations considered more than one of these indicators). In most cases, these impact indicators were only measured in relation to the *direct* effects of programme interventions.

The use of linear results frameworks often fails to capture the complex interactions and feedback loops that exist within the results chains of M4P interventions. Indicators set at each level in the results chain generally ignore indirect, systemic effects, which is the channel through which M4P programmes seek to achieve significant impact (see Figure 4). Therefore applying a direct, linear approach to results measurement presents a risk that evaluations will miss many of the results which M4P programmes are inherently designed to achieve. This approach can equally risk overestimating impacts through inaccurate assumptions about linkages between various levels of results.

Figure 3: A systemic logical model for M4P enables better capture of M4P impact



Evaluations need to place more emphasis on assessing the extent to which systemic market change, sustainability and large-scale impact has been achieved (i.e. point b in Figure 4) to help assess the degree to which the underlying objectives of M4P programmes have been achieved. Sections 4.6 to 4.8 consider in detail the extent to which the evaluations reviewed have done this.

Box 3: Complexity and systems

Concepts from complexity science are increasingly being applied to the implementation of market interventions and arguably should equally be applied to their evaluation.

The systems that M4P programmes seek to alter can be considered to be 'complex' - they are nonlinear, dynamic, and relatively unpredictable. They consist of multiple interconnected and interdependent pieces. One small shift could catalyse a substantive change in the entire system, and it can be difficult to know with certainty the nature and magnitude of that change in advance.

Complex initiatives require evaluation strategies that account for their unpredictable, adaptive and non-linear nature. This typically involves the use of a theory of change (in which the assumptions about the linkages between the intervention and the desired outcome are articulated, tested, and revisited frequently), mixed methods and triangulation, and approaches that allow for the capture of unanticipated effects (also called 'emergence')[15]. Osorio-Cortes and Jenal (2013) also detail some of the implications for applying complexity science to monitoring and evaluating market development programmes.

[15] Patton 2011, Rogers 2008, Funnel and Rogers 2011, Forss et al 2011.

Recommendations for good practice

Linear results frameworks run the risk of mis-estimating impact and are not suitable for evaluating M4P programmes. Instead, M4P evaluations should explicitly focus on assessing the extent to which systemic market change, sustainability and large-scale impact have been achieved. This can be achieved through an evaluation approach that is based on and frequently revisits and tests the linkages and assumptions in the project's theory of change.

4. M4P evaluation methods

This section provides an analysis of ***how*** M4P programmes have been evaluated. It begins with a discussion of the issues associated with assessing attribution and contribution and then reviews the following considerations in the evaluation of M4P programmes :

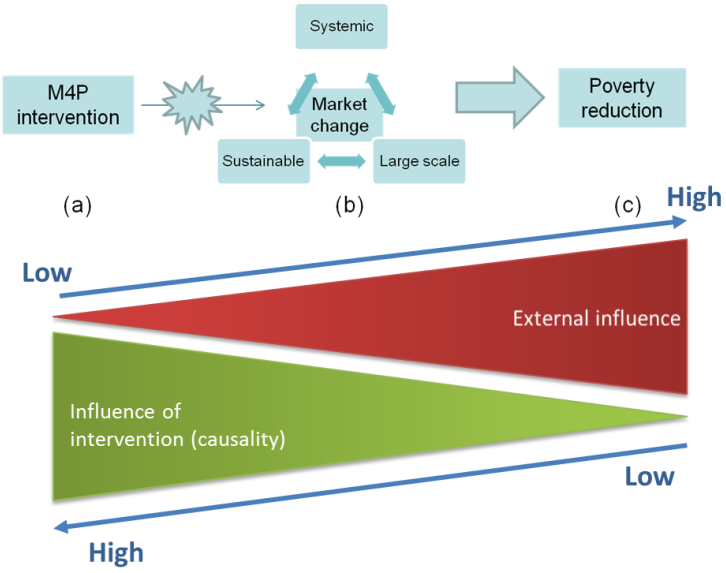
- Methods.
- M4P facilitative and adaptive approach.
- Systemic change in markets.
- Sustainability.
- Large-scale impact.
- Unintended effects.

4.1. Assessing attribution and contribution

Attribution, as defined by the OECD DAC (2010), is the ascription of a causal link between observed (or expected to be observed) changes and a specific intervention. In other words, attribution is "*the extent of change that can be claimed by a project/intervention out of total change that takes place*" (Sen 2013).

M4P interventions operate as part of a wider system where, in nearly all cases, they interact in some way with other public and private activities to achieve their intended results. Interventions aim to catalyse change, inducing spill-over effects to indirectly scale up change. Most interventions can therefore be seen as a 'contributory' cause – i.e. the intervention is a vital part of a 'package' of causal factors that are together sufficient to produce the intended effect. On its own, however, the intervention may not be sufficient, nor even necessary, to achieve a desired result. It is important, therefore, to appreciate the position of the intervention within a wide range of causal factors, and to understand that external factors have an important influence on the scale and nature of change. Moreover, the importance of external factors and complexity of the problem becomes greater as we move further away from the initial intervention to look at the final poverty-reduction impacts (see Figure 4).

Figure 4: Attribution – causality and external influence



Our review of the literature points to two broad schools of thought on addressing attribution for M4P programmes (including in external evaluations and internal impact assessments):

The complexity of the contribution chain makes attribution impossible. For example, USAID's guidelines on assessing the effectiveness of economic growth programmes indicate that it is "impossible to prove attribution" when the intervention aims to generate changes in the business enabling environment, such as government regulations and strategies (Creevey et al 2010). David Snowden argues: *"One of the things you have to start to show donors is how their funding […] had a positive impact on the area but you can't show actual attribution because it's dependent on other things that are going on at the same time*"[16]. Such arguments are made particularly strongly for the 'higher' levels in an intervention's results chain or theory of change, such as assigning attribution of an intervention to increased household income or poverty reduction.

Efforts should be made to assign attribution by isolating and accurately measuring the particular contribution of an intervention and ensuring that causality runs from the intervention to the result. For example, the DCED Standard requires "*a clear and appropriate system for estimating attributable changes in all key indicators*", although it accepts that "*some programmes (for example improving the business environment) are creating pre-conditions for development outcomes, rather than stimulating actual change. Attribution (and measurement of impact) may be more difficult in such cases*"[17]. In assigning attribution, purely measuring 'direct impacts' can often under-estimate the effectiveness of an M4P intervention, especially when evaluations are conducted before the most significant effects of systemic changes (copying or crowding in, for example) have fully emerged. Conversely, evaluations can over-estimate the effects of the programme when other contributory factors are not adequately considered.

Recommendations for good practice

Determining attribution or contribution of M4P programmes towards observed results presents a number of challenges. Many of the changes M4P programmes seek to achieve are long-term in nature and embedded within extremely complex systems. As such, many other contributory factors also tend to affect the results, thereby making attribution to any one intervention very difficult to ascribe. Evaluations should therefore seek to ascertain the extent to which changes are systemic, carefully consider other contributory factors, and additionally collect data over an appropriate period of time following an intervention to capture long-term changes and measure sustainability.

4.2. Evaluation methods

M4P evaluations apply a variety of methods. Below we review some of the most popular methods, including quasi-experimental approaches, multiplier calculations, other quantitative methods, mixed-methods, and theory-based approaches.

Quasi-experimental methods aim to estimate attribution by comparing the 'treatment' with counterfactuals based on comparisons with a control group. Quasi-experimental difference-in-difference analysis is most commonly applied: five evaluations applied quasi-experimental approaches (all difference in difference) to measuring impact. All of these evaluations undertook their own primary data collection activities.

Quasi-experimental approaches have the potential to provide 'hard data' for measuring the discreet impacts of specific interventions, which can be useful at pilot stages or for 'proving' purposes. Some mature projects (e.g. Katalyst), incorporate quasi-experimental work as part of the programme's

[16] Osorio-Cortes & Jenal (2013).

[17] It is worth noting that the DCED Standard has been developed for private sector interventions in general, and is not limited to M4P approaches, which focus on achieving systemic change.

ongoing monitoring and learning processes. However, they are not suited to capturing whether changes are systemic, sustainable or large-scale. In addition they can be difficult to implement well due to the changing and adaptive nature of M4P programmes. For example, the PROFIT programme in Zambia experienced a shift in target location for its interventions in the cotton sector, which meant that ex-ante baseline data for evaluation in that sector was of no utility[18].

Before-and-after approaches: Most of the other evaluations that applied quantitative approaches to the evaluation used simple before/after analysis and relied on internal monitoring data. Whilst often preferred as a quantitative method because of its simplicity, the approach risks overestimating attributable results because there is no consideration of the counterfactual.

Statistical rigour: Much of the data used to provide quantitative evidence in M4P evaluations relies on small sample sizes with little consideration of sampling frames, statistical significance or selection bias. Exceptions include the Katalyst[19] and PROFIT Zambia[20] evaluations, in which the strengths and weaknesses of data sources are extensively discussed by the evaluation team. In general, M4P evaluations need to be far more transparent about the confidence levels of quantitative evaluations, accepting that achieving high levels of statistical significance are challenging to achieve for M4P programmes.

Multiplier calculations: The use of multipliers to calculate programme impacts (e.g. Keynesian multipliers) is common in the estimation of private sector programme impacts. The 2007 GTZ cross-section of PSD evaluations reviewed the multiplier calculations and assumptions made in PSD evaluations and concluded that "an automatic link between growth and employment and thus with poverty reduction is usually presumed... this automatic connection does not always exist, however". It is therefore important to be explicit about and test the assumed linkages and pathways for impact.

Katalyst also uses multiplier calculations in its monitoring system to provide quantitative estimates of impact: for example, the extent to which innovations were expanded to other players, areas and sectors. Consistent with the DCED Standard, these calculations are based on the assumptions made explicit in the results chain, and these are recorded for transparency. This enables the calculations and the assumptions on which they are based to be revisited and tested throughout the programme and by external stakeholders (including via auditing processes).

Assumptions that are critical to claims of impact should ideally be subject to external validation, but also through consultation with relevant stakeholders external to the project implementation and to the evaluation team.

Mixed Methods: Where statistical confidence levels are low, quantitative evidence can still have value, so long as it is triangulated with alternative evidence using mixed methods. The PROFIT Zambia evaluation provides an example of where mixed methods provide value in making up for unanticipated failure of a data source: one of the strongest sources of data, the household surveys, was rendered obsolete due to a change in location of project interventions. This particular evaluation was fortunately able to rely on other sources of information, but effective triangulation was not possible. Thus the adaptive nature of M4P programmes means that they cannot rely too heavily on data sets (e.g. baseline and control groups) identified *ex ante*.

Mixing qualitative with quantitative methods also helps in combining statistical proof of a relationship from quantitative methods with evidence of 'why' from qualitative methods.

[18] The evaluation report states, '*Because of the specific limitations of this study and general shortcomings of quasi-experimental approaches, most of the findings cited in this report should be regarded as suggestive rather than definitively proven*' (DAI 2010:3).

[19] Magill & Woller (2010).

[20] DAI (2010).

The majority of evaluations reviewed included qualitative consultations with stakeholder as a form of validation or triangulation for quantitative data. However triangulation of evidence was weak. This was largely due to either (a) weak triangulation and/or data collection practices, or (b) unanticipated failure of one or more of the evaluation methods. The evaluations did not include a detailed account of the methodologies used for qualitative information gathering. The common risks of bias in qualitative evaluation – on the part of both respondents and evaluators – are detailed by White and Phillips (2013), who provide suggestions for increased rigour in 'small n' evaluations[21]. Given the importance of qualitative information for M4P evaluations, and the significant risk of various types of bias, higher standards are needed for the collection and analysis of qualitative information in M4P evaluations.

Theory-based approaches: A theory of change or results chain can be used in both attribution and contribution approaches. In a results chain, calculations are made based on assumptions that one change (e.g. increased employment) will lead to another (e.g. poverty reduction). In a theory of change, the intervention is described as one of many contributory causes to the outcome, and the evaluation attempts to establish if there are plausible causal links between each step and to document and account for other contributory factors.

The use of a theory of change is recognised as good practice in evaluating complex initiatives (Hivos 2009, Funnel and Rogers 2011, Rogers 2008, Patton 2010) to both improve programme management for results and track changes to which the programme might have contributed. However, in order to be useful, theories of change must include (a) an explicit articulation of assumptions, (b) rationale on which they are based, (c) identification of alternate pathways of change, (d) identification of other factors in the system: in-links (contributing to change) and out-links (changed by the programme). They should be (e) validated externally (e.g. by stakeholders outside of the project or external evaluators), and (f) revisited frequently as the system, and/or the programme team's understanding of it, evolves over the course of programme implementation.

While the majority of the programme evaluations reviewed included a theory of change (often labelled as a results chain), they generally did not meet the above basic criteria and instead resembled more of a simple logic model. The Joint SDC-Irish Aid Review of MMDPP, for example, focused only on intended, positive effects and did not test the assumptions in the theory of change. The PROFIT Zambia impact assessment included a discussion of the causal chain but this did not entail consideration of possible negative or unintended effects. In the SECO Business Environment Reform Evaluation, the evaluation team designed a generic results chain to facilitate the evaluation, though this was a single, linear chain focused only on intended impacts. The links in the chain were partially examined with a theoretical lens (as opposed to being tested through data collection). Again, this was also identified as a problem by the GTZ evaluation review team.

For programmes engaging in complex adaptive systems, a theory-based approach presents a number of risks in relation to the integrity of the evaluation, which must be carefully taken into account in the application of the approach:

- Overstatement of the causal contribution of the intervention.
- 'Goal displacement', 'where original targets are met even though this undercuts the actual goals of the intervention' (Patton 2008:34), due to 'premature selection' of solutions based primarily on the point of view of the NGO or donor (Osorio-Cortes and Jenal 2013:3).
- Lack of anticipation and capture of unintended consequences (positive and negative).

[21] Drawing on work of Pawson and Tilley (2004), Chambers (2006), Davies and Dart (2005), Neubert (2010) and others.

More thorough articulations of theories of change can help improve the use of theories of change for M4P evaluations by: (i) embracing complexity; (ii) encompassing external points of view; and (iii) frequently revisiting them as the project evolves.

The Enterprise Challenge Fund Mid-Term Review presents an example of good practice in evaluations based on theory of change. The evaluation team explicitly examined the programme's theory of change and found it to be lacking causal logic and supporting evidence, leading to recommendations for revised programme logic for more coherence.

Recommendations for good practice

M4P evaluations should apply a theory-based, mixed methods approach which uses a range of evaluation methods to test the causal links contained in a programme theory of change. This approach relies on the development of a robust theory of change and should (i) explicitly embrace complexity; (ii) encompass external points of view; and (iii) be frequently revisited and tested. Where engaged, an independent evaluator should ideally be engaged in reviewing, and sometimes working with the project in developing, the theory of change.

Quasi-experimental methods can provide rigorous quantitative evidence of impact. However they are only able to measure relatively direct impacts and are therefore most effective when testing discrete components of the programme, or at the pilot stages of an intervention before significant contamination becomes a factor in the measurement process. Quasi-experimental approaches are risky for M4P interventions as the adaptive nature of the approach risks making baseline design obsolete as an intervention evolves.

Quantitative evaluations should be explicit about the statistical confidence level in their findings and the strength of the assumptions used to support their calculations. This is rarely done in current practice, particularly for internally conducted impact assessments.

Qualitative evidence is important for triangulating and providing explanations for quantitative findings. There is significant scope for increased rigour and improved documentation in the qualitative evaluation methods applied to address the risk of bias in evaluation findings.

4.3. Evaluation of the M4P facilitative and adaptive approach

The facilitative role and adaptive nature of the M4P implementation approach is described by advocates of the approach as a desirable response to the drawbacks of programmes that apply a more rigid and 'direct delivery' approach.

Several evaluations have considered the relevance of this approach:

The final evaluation of the Enter-Growth project in Sri Lanka reflects positively on the fact that the project had acted as a catalyst for change and operated in an innovative and risk taking manner. It does this by comparing the project's "market focus" approach with a more "conventional enterprise focus" to enterprise development.

The Katalyst external impact assessment is more critical of the way in which the programme approached its facilitative role. It concludes that many of the programme's interventions were too "hands-off" and included inadequate follow up, reinforcement or monitoring. It also concludes that there was an absence of programme synergies and that the DBSM programme was implemented in a very compartmentalised way, rather than fostering a cohesive or comprehensive approach to solving the situations faced in the different sectors. The issues were too narrowly defined and too isolated, with the result that some opportunities to relieve bottlenecks or address identifiable constraints were not implemented.

The USAID 'Understanding Facilitation' Briefing Paper[22] points to a number of benefits of using a facilitative approach and provides concrete examples of how it has: (i) built the capacity of existing actors and institutions; (ii) fostered stronger relationships among them to create incentives for upgrading; and (iii) increased the probability of reaching greater scale by targeting interventions at leverage points.

It is important to recognise that a facilitative and adaptive approach to programme implementation in itself creates challenges for the evaluator:

A **facilitative role** presents challenges in defining of 'treatment' and 'control' groups. Programme implementers often don't have control over what the treatment is: "*The problem with widespread changes that are owned and sustained by large numbers of market actors is that the more success the project has, the harder it becomes to establish causality!*"[23]. Interventions are less direct in nature and therefore linkages between interventions and results at the level of households is nonlinear and indirect.

An **adaptive approach** creates problems for longitudinal impact evaluations because target populations, and the nature of the intended change, may change as the programme evolves. A strong example of where this challenge was faced is in the PROFIT impact assessment.

None of the evaluations considered in this study explicitly compare the benefit of the facilitative approach to that of alternative approaches.

Recommendations for good practice

Evaluations should not take for granted that an adaptive and facilitative approach is optimal. It is desirable for evaluations to consider the relevance of the M4P approach alongside a 'counterfactual' direct delivery approach to achieving desired impacts in a similar context.

The theory of change supporting the intervention should explicitly identify assumptions relating to the benefits of the facilitative approach (e.g. sustainability and large-scale impact) and the risks inherent within it (e.g. more unintended consequences). The evaluation should test the assumptions which are less well supported by evidence and of most concern to the evaluator and project team.

4.4. Evaluation of systemic change in markets

Definition of systemic change

Systemic change is defined as transformation in the structure or dynamics of a system that lead to impacts on large numbers of people, either in their material conditions or in their behaviour. Systemic approaches aim to catalyse change, inducing spill-over effects to indirectly drive and scale up change.

Evaluation challenges

Systemic changes will normally only be observed through indirect beneficiaries, whose context, relationships and developmental status are affected by the system itself, not by the intervention. If evaluations do not look beyond direct beneficiaries to broader considerations of changes in the structures and dynamics of the market system with indirect effects on the target populations, they will be superficial.

[22] USAID (undated).

[23] Osorio-Cortes & Jenal (2013).

This poses a challenge in attributing changes to an intervention because:

- External factors have an increasing influence on the scale and nature of change further up the results chain.
- A facilitation approach often makes the distinction between groups that are "treated" and "untreated" unclear.
- It is difficult to distinguish which target groups are actually driving change.

Systemic change indicators and measurement methods

Systemic change in markets is frequently slow, involves multiple people or businesses, and often relates to attitudes or social norms that are difficult to observe. It is challenging to define generically *ex-ante*, and is highly context-specific. The DCED suggests the following possible indicators of systemic change[24], which are variously considered in some of the evaluations that we have reviewed:

Crowding in: The programme helps targeted enterprises provide a new service, by supplying training or improving the market environment. Other enterprises see that this service can be profitable, and start supplying it as well. For example, a programme helps agricultural suppliers start up pesticide spraying services. Other agricultural input suppliers, who did not receive any direct input from the programme, might then start up a similar service.

Copying: The programme improves the practices of targeted enterprises, to improve the quality or efficiency of production. Other entrepreneurs can see the positive impact of these new practices, and adopts them in their own business. For example, if a shoe making entrepreneur sees that his rival has improved the quality of his shoes; he copies the quality improvements and so also gets higher prices for his goods.

Sector growth: Programme activities cause the targeted sectors to grow. Consequently, existing enterprises expand their businesses and new entrants come into the market.

Backward and forward linkages: Changes in the market can trigger changes at other points along the value chain. For example, a programme increases the amount of maize cultivated. This benefits not just farmers, but others in the value chain, such as van drivers who transport maize. They receive more business as there is a greater amount of maize to transport.

Other indirect impact: As a result of programme activities, other indirect impacts may occur in completely different sectors. For example, if a programme increases the income of pig producers, they can spend more on consumer goods, benefiting other shops in the local area.

There is currently little guidance on how to measure systemic change. The DCED suggests that the methods chosen to assess systemic change link back to the results chains, are appropriate to the programme context, take attribution into account and conform to good research practices. It also suggests that it is useful to keep the direct and indirect impact channel separate from each other all the way up to the goal-level impacts, as it helps programme to add up impact separately if desired. Moreover, since there is often a time lag between when direct beneficiaries feel an impact and when indirect beneficiaries subsequently copy their performance, it is helpful to record their changes in separate channels.

Our review has found that only five of 14 evaluations considered systemic change to a satisfactory degree. Examples of where it was done relatively well include:

[24] Kessler & Sen (2013).

- The PrOpCom tractor leasing case study report, where systemic change was measured through the extent of copying of the business model by enterprises and service providers.
- The Enter-Growth final evaluation, which considered the extent to which business environment reform achievements helped build momentum and support for other project interventions.
- The Enterprise Challenge Fund mid-term review raises as a concern the fact that the ECF had a lack of conceptual clarity about what systemic change might mean, and how the ECF can support such a vision practically. The review suggests that systemic change can be demonstrated in two ways: firstly, the project itself is so large that in itself it can be considered as systemic; and secondly, it is through innovation, resulting in clear demonstration which then leads to wider take up and adoption which in turn delivers wider change.
- Katalyst measures systemic impact using a framework consisting of the stages of adoption, adaptation, expansion and response.

Recommendations for good practice

Systemic change should be explicitly included in the theory of change of M4P programmes and carefully considered in evaluations. The theory of change must clearly define the system to be changed (e.g. the markets, geographies, common processes or organisations) and the assumptions regarding the ways in which this systemic change will be realised. This is essential in order to enable the evaluation to test these assumptions and find out if and how systemic change took place in practice. The indicators selected to assess systemic change inevitably need be context-specific, but should generally relate to signs of replication, crowding in, and wider market change.

Systemic changes are difficult to measure quantitatively and a more 'journalistic' approach is normally required through dialogue with project partners and other market stakeholders. However, once signs of systemic change have been identified, it should in many cases be possible to assign numerical estimates to the level of outreach, or the degree of sector growth or income increase that the systemic change has achieved.

4.5. Evaluation of sustainability

Definition of sustainability

M4P evaluations and associated M&E guidance varies in the extent to which sustainability is considered in a static or dynamic sense. Arguably, only dynamic considerations of sustainability will take full account of the extent to which systemic changes become embedded within the market:

- **Static sustainability** is defined as the extent to which the *status quo* (in terms of the results achieved through an intervention) will be maintained after external support is withdrawn. In other words, the extent to which the project's legacy is maintained.
- **Dynamic sustainability** is defined as the achievement of structural change to the market system which enhances its resilience to shocks and stresses through evolution or innovation in response to changing external factors. Achieving this dimension of sustainability is arguably integral to the M4P approach in that it moves beyond outward or superficial market performance to look more deeply at the critical underlying, but often less visible, institutions and functions that determine systemic functioning. Dynamic sustainability is inherently linked with significant systems change

Evaluation challenges

If successful, the impact of an M4P intervention is likely to grow after it ends. In many cases, the opposite is true for 'direct delivery' programmes. This means that evaluations can only assess the full

impact of M4P programmes if they are undertaken some time after the programme ends, or if they at least provide projections of future impact at the time of measurement. However, as time progresses, it becomes more difficult to disentangle the effects that can be attributed to the M4P intervention.

Sustainability indicators and measurement methods

Our review has found that consideration of sustainability in M4P evaluations is surprisingly weak. Where it is considered, a range of indicators are used, including the following:

- Behaviours maintained after external support / intervention concludes.
- The commercial viability of a new business model after external support is withdrawn - e.g. the Enterprise Challenge Fund mid-term review.
- Private and public sector investment catalysed (as proxies for partners' buy-in to change, and hence the sustainability of change) – e.g. PrOpCom (2011).
- Institutional or structural change, for example through:
  - ✓ New or better relationships: firms in new or modified vertical or horizontal relationships have experienced win-win outcomes that lead to greater trust and continued incentives to cooperate.
  - ✓ Aligned incentives: when incentives for change are positive for all actors in the new market model, the likelihood of sustainability is high.
- The extent to which the market system is becoming more adaptable (dynamic sustainability). For example, as mentioned above, Katalyst's internal results measurement system defines the stages of adoption, adaptation, expansion and response. The concept of sustainability is considered to be linked with expansion and response. This implies that fundamental market change cannot be considered to be sustainable if a small number of market players directly engaged by a project's intervention are adopting a change. An indicator of dynamic sustainability is therefore where programme partners or service providers begin innovating new products (beyond the scope of the original intervention) to exploit the gaps in the service market, where the likelihood of sustainability is high.

The PrOpCom PCR brings together many of these indicators to assess the sustainability and points to evidence of the following as signs of sustainability:

- **Well-aligned incentives:** When incentives for change are positive for all actors in the new market model, the likelihood of sustainability is high.
- **Capacity development of service providers:** when partners show greater organisational capacity, market changes are more likely to be sustained.
- **Creation of space for innovation within markets:** when programme partners or service providers begin innovating new products (beyond the scope of the original intervention) to exploit the gaps in the service market, the likelihood of sustainability is high.
- **Leveraged investment:** this measure of investment indicates the level of commitment of our private sector partners to the change introduced; commitment signals continuity.

Recommendations for good practice

Sustainability should be explicitly considered in evaluations and defined in terms of both static and dynamic sustainability. Indicators of sustainability should relate to evidence of lasting change to institutions and behaviours of market participants, as well as the commercial viability of new business models or other innovations that are introduced. Evaluation data should be collected several years after an intervention has concluded to assess the extent to which impacts achieved are sustained after project activities have ended.

4.6. Evaluation of large-scale impact

Definition of large-scale impact

M4P programmes are designed to achieve large-scale change, benefitting large numbers of poor people beyond the programme's direct sphere of interaction. Interventions explicitly envisage mechanisms for replicating, extending or multiplying results so that, at least potentially, they could reach very large numbers of beneficiaries. While 'large' is often relative, we can take it to mean that the scale of the impact is many times larger than what could be achieved directly. The achievement of large-scale impact is closely linked with market system change that is systemic and sustainable.

Evaluation challenges

A successful M4P programme will have widespread effects. This presents challenges in establishing control and treatment groups and in attributing wider market change beyond the direct scope of an intervention.

Large-scale impact indicators and measurement methods

A key indicator of large-scale impact is that results achieved affect 'indirect beneficiaries' – i.e. people or enterprises who are not specifically targeted by the intervention, or who fall outside of its direct sphere of influence. Large-scale effects are best measured by assessing the impact (in terms of increased incomes, jobs or reduced poverty) of systemic change achieved through a programme.

As highlighted above, consideration of systemic change in the evaluations reviewed was weak. Where it was considered, little effort was made to measure the impact of the systemic change achieved.

Recommendations for good practice

The achievement of large-scale impact is closely linked with market system change that is systemic and sustainable. M4P evaluations should seek to assess the indirect (positive and negative) impacts of the programme and the extent to which project facilitation has played a role in the scaling up of successful elements of interventions. Estimates can relatively plausibly be made by: (i) clearly defining the systemic changes achieved; (ii) estimating the reach and depth of these changes; (iii) applying a mixed methods approach to assigning attribution.

4.7. Evaluation of unintended consequences

The Development Assistance Committee (DAC) defines impact as 'the positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended' (OECD 2010). Capturing all effects, rather than just focusing on the positive or intended effects, is imperative for all programming, but is especially relevant to M4P programmes due to factors of systemic, large-scale change and complexity. Because of these factors, the implications of

unintended effects (both positive and negative) are potentially great and therefore need to be monitored and evaluated effectively to allow for any necessary course correction.

Measuring unintended effects requires open-ended and flexible approaches to evaluation (e.g. semi-structured interviews) that allow for 'emergence' of the unanticipated. The construction of a 'negative program theory' (Funnel and Rogers 2011) can also facilitate the identification of possible unintended impacts.

The evaluations reviewed were extremely weak in considering and capturing unintended and negative effects, which may relate to the lack of incentive to do so as well as theory of change models that often do not consider risk explicitly or in sufficient depth. Common unintended effects such as displacement (benefitting some enterprises or populations at the expense of others) and environmental impacts were largely overlooked by M4P evaluations. Gendered assessments of M4P programme impacts were largely limited to disaggregation of direct effects by gender and did not capture changes in social relations, roles and power, which can lead to superficial conclusions about the gendered effects of a programme (Sahan and Fischer-Mackey 2011). This was corroborated in stakeholder consultations.

Recommendations for good practice

Significant consideration of possible unintended consequences is a must for M4P evaluation given the unpredictable nature of M4P and potential scales of impact. Theory of change models should explicitly consider risk and be revisited frequently. As well as considering economic effects on non-target populations, evaluations should aim to capture changes in social relations, roles and power. Environmental impacts and displacement are also common unintended effects and should be considered in evaluations.

5. Summary and conclusions

M4P programmes are characterised by playing a facilitative, adaptive role in order to contribute to systemic, large-scale and sustainable market changes that positively affect the poor. The nature of the approach and the complexity of the markets within which it operates present a number of challenges for evaluation.

Because of these challenges and the relatively recent development of the M4P approach, there are, unsurprisingly, very few examples of good practice in M4P evaluation to date.

The M4P evaluations reviewed were often weak in terms of:

- consideration of systemic, sustainable changes in market systems;
- the rigour of data collection practices;
- triangulation approaches;
- the development and use of theories of change;
- consistency in units for facilitating accurate aggregation; and
- consideration of unintended negative effects.

While there has been learning in this area, and significant improvements in practices and approaches, there remains considerable room for improvement. The following recommendations are based on our review of existing practices and consultations with stakeholders in this field.

Institutional arrangements that achieve an optimal balance between objectivity and in-depth knowledge of the intervention and context should include: internal data collection with external audits and/or longitudinal collaborations between the evaluator and evaluand.

Evaluation needs to happen **both** (a) during the programme, to ensure contributions are measurable and help facilitate an adaptive approach, **and** (b) later on (at the end or post-project), when systemic, long-term change has had time to unfold. This justifies the application of a longitudinal approach to evaluation.

A combination of **top-down and bottom-up** measurement is likely to address the inherent drawbacks of each approach. In selecting interventions for evaluation, one needs to consider the **evaluability** of the interventions, resources available, balance between breadth and depth, and bias in selecting only success stories.

Determining attribution or contribution of M4P programmes towards results presents a number of challenges in that many of the systemic changes M4P programmes seek to achieve are long-term in nature, and many other contributory causes of a given result exist. Evaluations should therefore seek to ascertain the extent to which changes are systemic, carefully consider other contributory factors, and additionally collect data an appropriate amount of time following an intervention.

A **mixed methods** approach combining qualitative and quantitative approaches and based on a **theory of change** is well-suited to evaluating M4P programmes. The theory of change should (i) explicitly embrace complexity; (ii) encompass external points of view; and (iii) be frequently revisited and tested throughout the programme.

Quasi experimental approaches are risky for M4P interventions as the adaptive nature of the approach risks making baseline design obsolete as an intervention evolves. Where successful, these methods can provide rigorous quantitative evidence of impact. However they are only able to measure relatively direct impacts and are therefore most effective at the pilot stage of an

intervention before significant contamination becomes a factor. **Quantitative** evaluations in general need to be more explicit about the statistical confidence level in their findings.

Qualitative evidence is often necessary for triangulating and providing explanations for quantitative findings. There is significant scope for increased rigour and improved documentation in the qualitative evaluation methods applied to address the risk of bias in evaluation findings.

The M4P approach was specifically developed to increase effectiveness and efficiency in comparison with direct-delivery approaches. However, evaluations should consider the **relevance of the facilitative and adaptive** M4P approach in comparison with a possible direct delivery approach to achieving desired impacts in a similar context.

Linear results frameworks present significant risks of both over and under estimating impact and are not suitable for evaluating M4P programmes. Instead, M4P evaluations should explicitly focus on assessing the extent to which systemic and sustainable market change and large-scale impact have been achieved.

The indicators selected to assess **systemic change** inevitably need be context-specific, but should generally relate to signs of replication, crowding in, and wider market change. These changes are very difficult to measure quantitatively and qualitative approaches are normally required. However, once signs of systemic change have been identified, it is often possible to estimate quantitative impacts.

Sustainability of changes should be defined in terms of both static and dynamic sustainability. Indicators of sustainability should relate to evidence of lasting change to institutions and behaviours of market participants, as well as the commercial viability of new business models or other innovations that are introduced. Evaluation data should be collected several years after an intervention has concluded to assess the extent to which impacts continue in the long-term.

The achievement of **large scale** impact is closely linked with market system change that is systemic and sustainable, and can be plausibly estimated by: (i) clearly defining the systemic changes achieved; (ii) estimating the reach and depth of these changes; (iii) applying a mixed methods approach to assigning attribution.

Significant consideration of **unintended consequences** is a must for M4P evaluation given the probability and scale of possible unintended impacts and should include economic effects on non-target populations, displacement, environmental impacts, and changes in social relations, roles and power.

Annex 1: Evaluations & guidelines reviewed

ACDI/VOCA (2004), *AgLink Final Report*, July.

Barlow, S., Kothalawala, J., Van Der Ree, K. (2009) Final evaluation of Enter-Growth Project Sri Lanka for the ILO Sri Lanka' ILO April 2009

Baulch B, J Marsh, N Bùi Linh, N Hoàng Trung & V Hoàng Linh (2009), *Key findings from the second Thanh Hoa bamboo survey, Second Prosperity Initiative Impact Assessment Report*, Hanoi, November.

Creevey, L. (2006) 'Collecting and using data for impact assessment' Impact Assessment Primer Series Publication #3, USAID

Creevey, L. (2006) 'Methodological issues in conducting impact assessments of private sector development programs' Impact Assessment Primer Series Publication #2, USAID

Creevey, L. (2007) 'Common problems in impact assessment research' Impact Assessment Primer Series Publication #7, USAID

Creevey, L., J. Downing, E. Dunn, Z. Northrip, D. Snodgrass, and A.C. Wares (2010) 'Assessing the effectiveness of economic growth programs' USAID

DAI (2010) PROFIT Zambia Impact Assessment Final Report, USAID Office of Microenterprise Development

DCED - The Donor Committee for Enterprise Development (2010) 'The DCED Standard for Measuring Achievements in Private Sector Development, Control Points and Compliance Criteria,' Version V, 13 January 2010

DCED (2010) Measuring achievements in Private Sector Development, Implementation Guidelines, Version 1g, 5th March 2010

DCED (2010) Case study in using the DCED Standard Maize production in Bangladesh with Katalyst (Available from www.enterprise-development.org, Accessed March 2013)

DCED (2013) The DCED Standard for Measuring Results in Private Sector Development, Changes made between Version V and Version VI (January 2013)

Devfin Advisors (2011), *The role and effectiveness of SECO cooperation in business environment reform*, Independent Evaluation, Economic Cooperation and Development Division, July.

Dunn, E, H Schiff and L Creevey (2011), *Linking small-scale vegetable farmers to supermarkets: effectiveness assessment of the GMED India project*, USAID microREPORT #166, February.

Elliott, D., Barlow, S., Bekkers, H. (2009) 'Enterprise Challenge Fund - Mid Term Review' Springfield Centre, Durham UK

Indochina Research Limited (2008), *Cambodia MSME project final monitoring and evaluation report*, USAID, October.

ITAD (2012), *GEMS Results Measurement Handbook*, Version 1.0, December.

Keppel, U, LD Binh & J Spatz (2006), *Streamlining Business Registration and Licensing Procedures: Experiences from the Philippines and Vietnam*, paper presented at the Asia Regional Consultative Conference, Donor Committee for Enterprise Development, Bangkok, 29 November to 1 December 2006, GTZ.

Kessler, A. and Sen, N. (2013a) 'Guideline to the DCED Standard for Results Measurement: Articulating the Results Chain', DCED, January 2013

Kessler, A. and N. Sen (2013b) 'Guideline to the DCED Standard for Results Measurement: Capturing Wider Changes in the System or Market' DCED

Kluve J (2009), *Measuring employment effects of technical cooperation interventions: some methodological guidelines*, Report Commissioned by GTZ Sector Project Employment-Oriented Development Strategies and Projects.

Magill, JH & G Woller (2010), *Impact Assessment of Four Business Service Market Development Activities in Bangladesh*, Final Report, DAI, February.

McKenzie D & C Woodruff (2012), *What are we learning from business training and entrepreneurship evaluations around the developing world*, Policy Research Working Paper 6202, The World Bank, Development Research Group, Finance and Private Sector Development Team, September.

Menocal, AR, D Booth, M Geere, L Phillips, B Sharma and E Mendizaba (2008), *Punching above its weight: An evaluation of DFID's PSPS, LAMIT and ENLACE programmes in Latin America,* ODI, December.

PrOpCom (undated), *Making tractor markets work for the poor.*

PrOpCom (2011), *Nigeria PrOpCom Project Completion Report*, October.

Ramm, H. Long, B. N., Hung, K. Q. (2011) 'Joint SDC - Irish Aid Review of the Mekong Market Development Portfolio Programme (MMDPP), SDC, Irish Aid

Sebstad J & D Snodgrass (2008), *Impacts of the KBDS and KHDP projects on the tree fruit value chain in Kenya*, microREPORT#129, USAID.

Sen, N. (2013) 'Guideline to the DCED Standard for Results Measurement: Estimating Attributable Changes' DCED

Snodgrass, D. (2006) 'Assessing the impact of new generation private sector development programs' Impact Assessment Primer Series Publication #1, USAID

Spath, B. (2007) Cross-section evaluation of independent evaluations in 2007 in the thematic priority area Private Sector Development (PSD) GTZ, Eschborn

Tomecko, J. (2010) Impact Assessment Manual for the SC Development Programme, Draft 3, June 16 2010, Swisscontact

Woller, G. (2007) 'Causal models as a useful program management tool: Case study of PROFIT Zambia' Impact Assessment Primer Series Publication #5, USAID

Woller, G. (2007) 'Developing a causal model for private sector development programs' Impact Assessment Primer Series Publication #4, USAID

Woolcock, M. (2009) *Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy*, Journal of Development Effectiveness Vol. 1, No. 1, March 2009, 1–14

Annex 2: Other references

Albu, M (2008), Making Markets Work for Poor: International development cooperation: seeking common principles that underpin a coherent approach to poverty reduction, paper prepared for the Swiss Agency for Development and Cooperation, June.

Chambers, R. (2006) Poverty Unperceived: Traps, Biases and Agenda. Institute of Development Studies, Working Paper 270, University of Sussex, UK. Available at http://www.ids.ac.uk/files/Wp270.pdf (Accessed February 2013)

Creevey, L., J. Downing, E. Dunn, Z. Northrip, D. Snodgrass, and A.C. Wares (2010) 'Assessing the effectiveness of economic growth programs' USAID

Davies, R. and Dart, J. (2005) The "Most Significant Change" (MSC) Technique: A Guide to Its Use, April 2005. Online at http://mande.co.uk/docs/MSCGuide.pdf

Forss, K., Marra, M., and Schwartz, R. (Eds.) (2011) Evaluating the Complex: Attribution, Contribution and Beyond : Comparative Policy Evaluation, Volume 18, Transaction Publishers

Funnel, S. and P. Rogers (2011) Purposeful Program Theory: effective use of theories of change and logic models John Wiley and Sons, San Francisco California#

Hivos (2009) 'Working with a theory of change in complex change processes', Hivos, April 2009

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124

Kandhker, S.; Koolwal, G. And Samad, H. (2010) Handbook on Impact Evaluation: Quantitative Methods and Practices, Washington DC: The International Bank for Reconstruction and Development / The World Bank

Neubert, S. (2010) 'Description and Examples of MAPP Method for Impact Assessment of Programmes and Projects.' German Development Institute (GDI)

Online at www.ngo-ideas.net/mediaCache/MAPP/

OECD (Organisation of Economic Co-operation and Development) (2013) 'DAC Criteria for Evaluating Development Assistance' Available at

http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm (Accessed February 2013)

OECD (2010) Glossary of Evaluation and Results Based Management (RBM) Terms, OECD

Osorio-Cortes, L & M Jenal (2013), Monitoring and measuring change in market systems – rethinking the current paradigm, Synthesis Paper, January.

OECD DAC (1991). Principles for Evaluation of Development Assistance, Paris.

Patton, M.Q. (2010) Developmental Evaluation: applying complexity concepts to enhance innovation and use The Guilford Press: New York

Pawson, R. and Tilley, N. (2004) Realist Evaluation. London: SAGE Publications

Rogers, P.J. (2008) 'Using programme theory to evaluate complicated and complex aspects of interventions.' Evaluation 14:1

Sahan, E. and Fischer-Mackey, J. (2011) 'Making markets empower the poor' Oxfam Discussion Paper

Sarewitz, D (2012) 'Beware the creeping cracks of bias' Nature Column: World View, 9 May 2012 Available at http://www.nature.com/news/beware-the-creeping-cracks-of-bias-1.10600 (Accessed February 2013)

Springfield Centre (undated), A synthesis of the Making Markets Work for the Poor (M4P) Approach.

Tanburn, J. (2008) 'The 2008 Reader on Private Sector Development: Measuring and Reporting Results' International training Centre of the International Labour Organization

Tarsilla, M. (2010) 'Being blind in a world of multiple perspectives: The evaluator's dilemma between the hope of becoming a team player and the fear of becoming a critical friend with no friends' Journal of MultiDisciplinary Evaluation 6:13

USAID (undated), Understanding Facilitation, Briefing Paper.

White, H. & Phillips, D. (2012) 'Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework' International Initiative for Impact Evaluation, Working Paper 15

Annex 3: Summary of evaluations reviewed

Ranking of Evidence Framework		
Evidence score	**Ranking of evidence**	**Description of ranking**
1	*Strong*	The finding is consistently supported by a full range of evidence sources, including documentary sources, quantitative analysis and qualitative evidence (i.e. there is very good triangulation); or the evidence sources, while not comprehensive, are of high quality and reliable to draw a conclusion (e.g. strong quantitative evidence with adequate sample sizes and no major data quality or reliability issues; or a wide range of reliable qualitative sources, across which there is good triangulation).
2	*More than satisfactory*	There are at least two different sources of evidence with good triangulation across evidence, but the coverage of the evidence is not complete.
3	*Indicative but not conclusive*	There is only one evidence source of good quality, and no triangulation with other sources of evidence.
4	*Weak*	There is no triangulation and/ or evidence is limited to a single source and is relatively weak.

Summary of Evaluations Reviewed													
Evaluation reviewed	**Evaluation institutional arrangement**	**Timing of evaluation**	**Methods**		**Primary data collection**	**Use of programme theory**	**Evaluation of M4P attributes – ranking of evidence***			**Impact indicators measured**			
			Quantitative	**Qualitative**			**Large scale impact**	**Systemic change**	**Sustainability**	**Poverty**	**HH Income**	**Jobs**	**Enterprise Income**
1. Katalyst: Impact assessment of four business service market development activities	Independent impact evaluation	During	Longitudinal quasi experimental (difference in difference)	FGDs, interviews, secondary data	By evaluation team	✓	3 Considered but not effectively measured	3 Measured qualitatively but superficially	2 Considered but unable to assess well due to timing	✓	✓	✓	✓
2. Cambodia MSME project final monitoring and evaluation report	Independent impact evaluation	End of project	Before / after. Randomised sample selected		By evaluation team	✖	3 Some assessment of spill over – spread of technologies to non-direct beneficiaries	4 Not considered	4 Not considered	✖	✓	✖	✓

Summary of Evaluations Reviewed												
Evaluation reviewed	Evaluation institutional arrangement	Timing of evaluation	Methods		Primary data collection	Use of programme theory	Evaluation of M4P attributes – ranking of evidence*			Impact indicators measured			
			Quantitative	Qualitative			Large scale impact	Systemic change	Sustainability	Poverty	HH Income	Jobs	Enterprise Income
3. AgLink Egypt final report	External review	End of project	None	Process evaluation based on project monitoring information	Interviews undertaken by evaluation team. Quantitative data sourced from monitoring information	✖	4 Not explicitly considered	3 Consideration of "institutionalisation" of AgLink activities	3 Consideration of "institutionalisation" of AgLink activities	✖	✖	✖	✖
4. PrOpCom Project Completion Report	Internal results measurement	End of project	Before / after. No counterfactuals. Methodologies used to measure impacts not clearly specified.	Project narrative. Programme monitoring based on DCED standard (mock audit conducted Sept 2010).	Internal monitoring data	✓	3 Consideration of signs of crowding in for some interventions	3 Some consideration of wider market change	2 Well considered framework for assessing sustainability of interventions	✖	✖	✓	✓
5. Independent evaluation of **SECO Cooperation in Business Environment Reform**	External review	Meta evaluation – during / end of project	None	Literature review Case study in Serbia	Relied on secondary data in the form of results reporting from selected projects	✓ Generic results chain developed	4 Evaluation found very limited and weak evidence of impact	4 Not considered	2 Sustainability ranking provided for each project reviewed	✖	✖	✓	✓
6. Impacts of the KBDS & KHDP projects in the **tree fruit value chain in Kenya**	Independent impact evaluation	End of project	Panel survey. Longitudinal quasi experimental (difference in difference)	FGDs & individual interviews with sub-sample of value chain actors	By evaluation team	✓	4 Not considered	4 Not considered	2 "Sustainability impacts" explicitly considered	✖	✓	✓	✓
7. Effectiveness assessment of the **GMED India project**	Independent impact evaluation	During	Randomised sample design. Longitudinal quasi experimental (difference in difference)	Process evaluation Qualitative field study	Local research partner	✓	4 Not considered	4 Not considered	2 Creation of sustainable vertical linkages explicitly considered	✓	✓	✓	✓

Summary of Evaluations Reviewed												
Evaluation reviewed	**Evaluation institutional arrangement**	**Timing of evaluation**	**Methods**		**Primary data collection**	**Use of programme theory**	**Evaluation of M4P attributes – ranking of evidence***			**Impact indicators measured**			
			Quantitative	**Qualitative**			**Large scale impact**	**Systemic change**	**Sustainability**	**Poverty**	**HH Income**	**Jobs**	**Enterprise Income**
8. Second Prosperity Initiative Impact Assessment Report, second **Thanh Hoa bamboo survey**	Independent impact evaluation	During	Matched difference in difference. Hedonic pricing analysis.	None	By evaluation team	✖	4 Not considered	4 Not considered	4 Not considered	✓	✓	✖	✓
9. PrOpCom tractor leasing case study report	Internal results measurement	During	Before / after based on survey of tractor service providers & farmers. Control groups included in survey	Interviews & monitoring data	Internal monitoring data	✓	2 Evidence of copying & crowding in provided	2 Measured through extent of copying by enterprises & service providers	2 Strong criteria developed to measure sustainability	✖	✖	✖	✓
10. **Enter-Growth Project Sri Lanka**, Final Evaluation	External review	End of project	Internally conducted Impact Assessment reports (documents unavailable at time of writing)	Cultural assessment (open-ended interviews, FGDs)	Primary data collected by implementing partner rather than evaluation team Evaluation team conducted stakeholder interviews, workshop	✓	2 Considered in the evaluation	2 Considered in the evaluation	2 There is a plan to measure poverty reduction 2 to 3 years after project close	✖	✓	✓	✓

Summary of Evaluations Reviewed												
Evaluation reviewed	**Evaluation institutional arrangement**	**Timing of evaluation**	**Methods**		**Primary data collection**	**Use of programme theory**	**Evaluation of M4P attributes – ranking of evidence***			**Impact indicators measured**			
			Quantitative	**Qualitative**			**Large scale impact**	**Systemic change**	**Sustainability**	**Poverty**	**HH Income**	**Jobs**	**Enterprise Income**
11. **PROFIT Zambia** Impact Assessment Final Report	Independent impact evaluation	End of project *though most of data is from 2 years into the 5 year project	Longitudinal, quasi-experimental mixed methods	Interviews, FGDs	By evaluation team	✓ Focus on intended, positive effects	3 Large number of HHs included but only in target areas which caused 'severe problems' when location shifted	2 Agent model being replicated	3 Considered in evaluation but not assessed	✓	✓	✖	✓
12. Joint SDC – Irish Aid Review of the **Mekong Market Development Portfolio** Programme (MMDPP)	Internal results measurement	End of project Complements previous evaluation	Reports from firms, processors and pre-processors	HH income surveys in selected bamboo areas	Primary data collected by firm - PI (Prosperity Initiative Community Interest Company)	✓ Focus on intended, positive effects, assumptions not tested	3 Considered but not effectively measured	2 Business environment reform evaluated	3 Considered through logic chain	✖	✓	✓	✓
13. **Enterprise Challenge Fund** Mid-term review	External review	Mid-project		Consultations and project visits	Primary data collected by project team (including baselines), reviewed by external team, complemented with visits and consultations	✓ Assessed in the evaluation (found to be lacking causal logic)	3 Considered in the evaluation (recommendations provided for better data collection to differentiate between outreach and impact)	2 Considered in the evaluation (recommendations provided for better data collection)	2	✖	✖	✖	✖

Summary of Evaluations Reviewed												
Evaluation reviewed	**Evaluation institutional arrangement**	**Timing of evaluation**	**Methods**		**Primary data collection**	**Use of programme theory**	**Evaluation of M4P attributes – ranking of evidence***			**Impact indicators measured**			
			Quantitative	**Qualitative**			**Large scale impact**	**Systemic change**	**Sustainability**	**Poverty**	**HH Income**	**Jobs**	**Enterprise Income**
14. Cross-section of independent **evaluations in PSD 2007 GTZ**	External review	7 Mid-project, 6 End-project, 4 Post-Project	Evaluations were based on Keynesian employment and income multiplier; assumed link between growth and employment and thus with poverty reduction	N/A	Review of evaluation reports	✖	Difficult to assess	2	1 Reviewed and rated for all projects ** this reflects the review and not the programme evaluations	✖	✖	✓	✖

Annex 4: List of individuals and organisations consulted

Name	Position	Organisation
Markus Kupper	Director, Monitoring and Results Measurement for Katalyst	Swisscontact
Goetz Ebbecke	General Manager, Katalyst	Swisscontact
Manish Pandey	Regional Director, South Asia	Swisscontact
Liz Kirk	Global Advisor - Private Sector Programme Policy Team	Oxfam GB
Gareth Davies	Senior Manager	Adam Smith International
Bill Grant	Senior Principal Development Specialist, Economic Growth	DAI
David Elliot	Director	Springfield Centre
Alan Gibson	Director	Springfield Centre
Simon Calvert	Evaluation Adviser – Private Sector, Growth and Trade	DFID
Adrian Stone	Evaluation Adviser – Private Sector, Growth and Trade	DFID
Catherine Martin	Principal Strategy Officer, East Asia and Pacific Department	IFC
Jeanne Downing	Senior Business Development Services (BDS) Advisor	USAID
Jim Tomecko	Senior Adviser AIPD-Rural, AusAID	Independent consultant
Alopi Latukefu	Director, Food Security Policy / Food Security, Mining, Infrastructure and Trade	AusAID
Jim Tanburn	Coordinator	Donor Committee for Enterprise Development
Aly Miehlbradt	Consultant	Independent
Marcus Jenal	Consultant	SDC