Improved Learning Initiative for the design of a
Participatory Impact Assessment & Learning Approach (PIALA)

Methodological reflections following the second PIALA pilot in Ghana

Adinda Van Hemelrijck

January 2016

Project:	Improved Learning Initiative (ILI)
Funding:	IFAD – Innovations Mainstreaming Initiative (IMI) BMGF – Measurement, Learning & Evaluation
Managers:	Edward Heinemann (e.heinemann@ifad.org) Adinda Van Hemelrijck (a.vanhemelrijck@ids.ac.uk) Richard Caldwell (richard.caldwell@gatesfoundation.org)
PIALA Design:	Adinda Van Hemelrijck (project and team leader) Irene Guijt (iguijt@learningbydesign.org) Jeremy Holland (jemholland@googlemail.com) Andre Proctor (andre@keystoneaccountability.org)

Table of Contents

TABLE OF CONTENTS ..... **I**

GLOSSARY **2**

TEXTBOX AND FIGURES **2**

INTRODUCTION **3**

1 PARTICIPATORY IMPACT ASSESSMENT & LEARNING APPROACH (PIALA) **4**

2 FOCUSING AND FRAMING THE EVALUATION **7**

2.1 CRITICAL REFLECTIONS 7

2.1.1 Reconstructing and visualizing the ToC *7*

2.1.2 Considering different design options *8*

2.2 COST-BENEFITS 9

3 COLLECTING AND LINKING THE DATA **10**

3.1 CRITICAL REFLECTIONS 10

3.1.1 Developing the sample frame centred on "open systems" *10*

3.1.2 Deciding on depth and breadth of inquiry *11*

3.1.3 Maintaining independence in field mobilization *12*

3.1.4 Contextualising poverty analysis *13*

3.1.5 Dealing with power and bias in participatory research *14*

3.1.6 Methodological complementarity *15*

3.1.7 Data linking and quality monitoring *17*

3.2 COST-BENEFITS 18

4 SYNTHESIZING THE EVIDENCE FOR ANALYSING AND DEBATING CONTRIBUTIONS **18**

4.1 CRITICAL REFLECTIONS 19

4.1.1 Synthesizing the evidence *19*

4.1.2 Participatory sensemaking *19*

4.1.3 Analysing with/without configurations supported by the evidence *21*

4.2 COST-BENEFITS 23

5 KEY OUTCOMES AND LEARNING POINTS OF THE PIALA PILOTS **24**

REFERENCES **27**

ANNEX I. PIALA METHODS & PROCESSES (PURPOSES, COSTS AND PARTICIPANTS) **30**

ANNEX II. REVISED PIALA QUALITY ASSURANCE FRAMEWORK (QAF) **32**

Glossary

BMGF	Bill & Melinda Gates Foundation
CF	Constituent Feedback
CDT	Core Design Team
DBRP	Doing Business with the Rural Poor (name of the IFAD-funded project in Vietnam)
DSG	Design Support Group
ERG	External Reference Group
FGD	Focus Group Discussion
IFAD	International Funds for Agricultural Development
HH	Household
ILI	Improved Learning Initiative
IMI	Innovation Mainstreaming Initiative
KIIs	Key Informant Interviews
PIALA	Participatory Impact Assessment & Learning Approach
PRA	Participatory Rural Appraisal
QAF	Quality Assurance Framework
RIMS	Results & Impact Management System
SME	Small & Medium Enterprise
ToC	Theory of Change

Textbox and figures

Figure 1: Sampling hierarchy of the two PIALA pilots……..... 9

Figure 2: PIALA scoring……..... 19

Figure 3: Participatory sensemaking sessions…… 21

Textbox 1: Part of the configurational scoring table of the RTIMP evaluation…… 20

Introduction

1. IFAD has to report to its Members States on the total number of rural people lifted out of poverty[1]. The government programmes it funds, however, are implemented in complex ways and environments that challenge mainstream evaluation practice. The challenge for IFAD and its co-implementing and co-funding partners, moreover, is not just to rigorously assess impact but also to understand the processes generating impact in order to realize its ambitious targets (IFAD, 2011). Albeit a strong emphasis on quantitative measurement, there is a need for impact evaluation that fosters learning and responsibility.

2. In response to this need, a three-year Improved Learning Initiative (ILI)[2] was by IFAD[3] and the BMGF commissioned in October 2012. The ILI sought to realize the following three objectives (IFAD & BMGF, 2013b):

a) To assess the nature and extent of the impact of two selected IFAD-financed programmes on rural poverty;
b) To design and pre-pilot in these two programmes a cost-effective ***Participatory Impact Assessment and Learning Approach (PIALA)*** that can:
    - generate rigorous qualitative and quantitative evidence of rural poverty impact for global reporting and advocacy;
    - facilitate inclusive analysis and reflection on evidence of impact for collaborative learning and collective responsibility within the countries;
    - produce a feasible model that shows potential scalability for strengthening IFAD's self-evaluation system;
c) To facilitate internal and external stakeholder reflections at grassroots, country and global levels on the validity, feasibility and utility of PIALA.

3. PIALA was first piloted in the *Developing Business with the Rural Poor Programme* (DBRP) in one province in Vietnam (IFAD & BMGF, 2014), and subsequently at national scale in the *Root and Tuber Improvement and Marketing Programme* (RTIMP) in Ghana (MOFA/GOG, IFAD & BMGF, 2015). Both programmes aimed at improving livelihoods and increasing incomes as part of sustainable and equitable poverty reduction through enhancing smallholders' capacity to commercialise and linking local businesses to markets and industries. To realise this, the DBRP focused on developing diversified short value chain systems, while the RTIMP sought to develop longer commodity chains linked to national and export markets and industries. Both involved the systemic change requiring policy and institution building, private sector engagement, and empowerment of smallholders.

[1] Under its 9th Replenishment (2012-2015), when the PIALA initiative was launched, IFAD committed to moving 80 million rural people out of poverty cumulative from 2010 onwards to 2015, and conducting 30 rigorous impact assessments.

[2] The ILI was managed by Edward Heinemann (Senior Policy Advisor, IFAD), Adinda Van Hemelrijck (evaluation consultant, IFAD) and Richard Caldwell (head of evaluation for agricultural development, BMGF).

[3] The IFAD funds for the ILI came from the DFID-supported Innovation Mainstreaming Initiative (IMI).

4. The DBRP was implemented from 2008 to 2014 in two provinces (Cao Bằng and Bến Tre) with a budget of US$ 51 million, including US$ 36 million from IFAD. The evaluation was conducted in 2013 at a cost of US$ 90,000 and in Bến Tre only, where the programme was implemented in 50 of 164 communes in eight of nine districts (IFAD & BMGF, 2014). The RTIMP was implemented from 2007 to 2015 in 106 of Ghana's 216 districts across all ten provinces with a total budget of 24 million, of which US$ 19 million was from IFAD. The evaluation was conducted in 2015 countrywide at a cost of USD $ 233,000 (MOFA/GOG, IFAD & BMGF, 2015). Both evaluations covered the past five years of programme activity.

5. Methodological reflections were held in both pilots with researchers and programme stakeholders. Feedback was also solicited from staff and managers at IFAD headquarters in Rome, and from a select group of international experts external to IFAD and BMGF[4]. Insights from the reflections and feedback on the first pilot in Vietnam (cf. IFAD & BMGF, 2013a) enabled better addressing challenges in the second pilot in Ghana. This report presents a synthesis of the reflections following the second pilot in Ghana.

1 Participatory Impact Assessment & Learning Approach (PIALA)

6. PIALA is designed to produce rigorous quantitative and qualitative evidence for collaboratively *assessing, explaining* and *debating* programme contributions to rural poverty impact together with partners and key stakeholders. Its purpose is to foster learning and influence policy, strategy and targeting for generating greater and more sustainable impact (IFAD & BMGF, 2013b). Different from process and performance evaluation is the focus on contributions to impact *broader* than intended outcomes and performance against pre-set targets.

7. Impact in this approach is viewed from a systemic perspective, as a system of interactions between various actors and influences, rather than the direct relationship between intervention and effect. A systemic approach seeks to move beyond assessing "what works" to also answer the more difficult "why" and "how" questions and investigate the likely sustainability of changes observed. It looks at intended and unintended, positive and negative effects of a programme in relation to other influences that directly or indirectly affected impact on rural poverty (Befani et al, 2015; Burns, 2014). Generally, the questions PIALA seeks to answer are: *"what has changed (or not) for whom and why"; "how sustainable are these changes likely to be"; "what are the impacts of these changes and what has caused them"; "what has been the program's contributions to these changes in relation to other causes or influences";* and *"what are the implications for policy and strategy"* (IFAD & BMGF, 2013b).

[4] A Design Support Group (DSG) was established to provide critical inputs and feedback on the piloting processes and products, and ensure maximum alignment with IFAD's self-evaluation and knowledge management frameworks, systems, methods/tools and initiatives. The DSG included managers and advisors of IFAD's Strategy & Knowledge Management and Program Management Departments, as well as the Independent Office of Evaluation. In addition, also an External Reference Group (ERG) was established for providing strategic guidance and feedback, and advocating for PIALA in the wider development and evaluation community. The ERG comprised of four experts: Robert Chambers (IDS); Marie Gaarder (IEG. World Bank); Carlos Barahona (Statistical Service Centre, University of Reading); and Kent Glenzer (Monterrey Institute of International Studies, CA).

8. To answer these questions and address the challenge of assessing and learning about program contributions to impact from a *systemic* perspective, PIALA draws on five important design elements: (i) a systemic Theory of Change (ToC) approach; (ii) multi-stage random sampling centred on "market-bounded systems"; (iii) participatory mixed-methods; (iv) participatory sensemaking; and (v) configurational analysis (MOFA/GOG, IFAD & BMGF, 2015). Together, these five elements make it possible to arrive at rigorous causal inference in programme contexts where conventional counterfactual-based approaches are not feasible –e.g. where credible "control groups" cannot be identified, such as in programmes where institutional and policy work have affected entire populations, other donors and influences have augmented the "causal density", and innovation and self-targeting mechanisms have rendered programme treatments erratic (Befani, 201 2; Woolcock, 2013). The table in Annex 1 shows how the five PIALA elements were translated in concrete processes and methods in the three main phases of evaluation:

a) focusing and framing the evaluation;
b) collecting and linking the data; and
c) synthesising the findings for analysing and debating programme contributions to impact.

9. The ToC approach provides the structure for the entire evaluation. It involves a reconstruction and visualisation of the programme logic and the broader trends and influences, based on a desk review and discussions with stakeholders. The ToC shows the systemic links and feedback loops between different programme components and mechanisms and other influences, and their collective outcomes and impact (IFAD & BMGF, 2013b; Van Hemelrijck, 2013). The methodology for data collection is a blend of KIIs, household surveys and participatory methods investigating the different causal claims and links in the ToC. They complement and build on each other analytically. A brief household survey and a generic change analysis in gender-specific focus groups inquire changes in household food & income and the influences of changes in livelihoods on these. A livelihood analysis in gender-specific focus groups is triangulated with the generic change analysis to further investigate causes and effects of the observed livelihoods changes. Finally, a Constituent Feedback method in mixed groups is triangulated with the livelihood analysis to investigate the reach and effects of specific programme mechanisms on the changes and causes in the various intervention areas that affect livelihoods. Additionally, Key Informant Interviews (KIIs) are conducted with local, regional and national programme stakeholders to crosscheck the emerging evidence on all the causal links (MOFA/GOG, IFAD & BMGF, 2015).

10. The household survey uses a simplified RIMS3-type of questionnaire adapted to the programme context. The generic change analysis consists of two to four PRA-based tools: a social mapping with timeline (if necessary, depending on the main analysis unit), and a change ranking with a causal flow mapping of changes in wealth & wellbeing. The livelihood analysis combines two PRA-inspired tools –change matrix and causal flow mapping. In the RTIMP evaluation, the livelihood analysis also included a brief SenseMaker study. Constituent Feedback combines group discussion and anomymous scoring tools drawn from custumor satisfaction study methods used in the private sector. To permit triangulation, the KIIs used similar questionnaires as those used for the anonymous scoring in the Constituent Feedback method. Tools and guidance for systematic data collation and quality monitoring during fieldwork makes it possible to cross-check and link the data. Local sense-making workshops are organized immediately after collecting the data in each locality to facilitate debate with

local stakeholders and ensure the evidence is sufficient and consistent to assess the programme's causal claims (Ibid).

11. Multistage random sampling enables a nesting of methods for inquiring the interplay between various programme mechanisms and other influences within market-bounded systems (in the case of the two pilots: value chain systems) across a large population. These 'systems' form the principle unit of analysis and thus also the principle sample population from which to subsample the populations that are expected to be affected by and have participated in these systems. For the quantitative poverty analysis, the total size of subsamples of households must be large enough to arrive at 95 % statistical precision. For the comparative analysis of the systemic interactions affecting poverty status, the subsamples of programme beneficiaries together must sufficiently cover the variations in and configurations of conditions, treatments and outcomes across the entire programme area. A configurational analysis method is employed that starts with the clustering and comparing of evidence across all sampled 'systems' to surface the configurations supported by the evidence. This forms the basis for identifying programme contributions to impact. In participatory sensemaking workshops, key stakeholders debate and value/refute programme contributions to impact among other influences, by comparing the changes revealed by the evidence with the changes presumed in the ToC (Ibid).

12. Finally, PIALA is designed and piloted around standards of *rigour*, *inclusiveness* and *feasibility* considered essential for delivering high quality and high value evaluations. A Quality Assurance Framework (QAF) was developed and used to guide the stakeholder reflections on these three standards in the two pilots. *Rigour* in this framework refers to the quality of thought put into the methodological design and conduct of the evaluation, in order to ensure consistency and responsiveness (Rogers, 2009; Stern et al, 2012). *Inclusiveness* involves meaningful engagement of stakeholders with diverse perspectives, which has an intrinsic empowering value while also enhancing credibility of the evaluation through triangulation and cross-validation of evidence (Chambers, 2015; Pawson, 2013). *Feasibility* concerns the budget and capacity needed to meet the expectations of rigour and inclusiveness and enhance learning. The QAF is presented in Annex 2.

13. The following sections 2, 3 and 4 present a synthesis of these reflections for each of the three main evaluation phases: (a) focusing and framing the evaluation; (b) collecting and linking data; and (c) synthesising findings for analysing and debating contributions. Each starts with the key leading question from the QAF and closes with a synthesis of costs and benefits of using the PIALA methods and processes for that phase. The report concludes with a summary of results and key learning points for future pilots.

2 Focusing and framing the evaluation

14. Focusing and framing the evaluation involves determining what needs to be evaluated and for what purposes. Clarifying purposes helps to define the scope and scale of the evaluation. Scale, scope and focus are best decided with commissioners and key stakeholders based on: (a) a projection of the potential cost-benefits of different design options; and (b) a reconstruction and visualisation of the programme ToC showing the causal claims and assumptions. The ToC process helps to decide on the focus and questions and specify the criteria and standards for the evaluation (BetterEvaluation, 2014). The key question from the QAF for this phase is the following:

How and to what extent did PIALA help adequately identify the causal claims that the evaluation should focus on, and determine its purposes, questions and standards in an inclusive and rigorous manner?

2.1 Critical reflections

2.1.1 Reconstructing and visualizing the ToC

15. Every programme has an implicit or explicit ToC. Most ToC approaches are more linear and focused on the performance of a single intervention, rather than the collective impact of multiple actors and multiple interventions (Funnell & Rogers, 2011). The ToC approach used in PIALA is more systemic, showing the various linkages and feedback-loops between multiple programme components and influences and their collective outcomes and impacts. A systemic ToC approach is most effectively used as a dynamic and adaptive framework for evaluating and managing complex multi-actor programmes, for three important reasons. First, as the pilot in Ghana has shown, it allows for a rigorous assessment of multiple interacting causal links using appropriate methods for data collection and analysis that are not necessarily counterfactual-based. Second, both the pilots in Vietnam and in Ghana have demonstrated how it enables different stakeholders to engage with the evidence collected on these links, probe their assumptions, and critically analyse and debate their roles and contributions to impact on rural poverty (IFAD & BMGF, 2013a; MOFA/GOG, IFAD & BMGF, 2015). Third, the two pilots suggest that if used from the beginning of a programme, the approach will permit timely corrections or adaptations based on the learnings from the analyses and debates, while also contributing to the capacity-building of stakeholders to think and operate more systemically and evaluative (Van Hemelrijck, 2013).

16. The reconstruction and visualisation of the ToC, and the framing and focusing of the evaluation based on this, therefore, needs to be rigorous and inclusive. In Vietnam, insufficient time and budget was spent on reconstructing and visualising the ToC, affecting the rigour and inclusiveness of the approach during the entire evaluation. In Ghana, this was corrected by making the ToC a priority and key deliverable of the evaluation. In Vietnam, a brief workshop was organised *prior* to reconstructing the ToC and only with the programme steering committee and managers to discuss the programme logic. The process of reconstructing and visualising the ToC then happened after the workshop and almost independent of the evaluation design. The desk review and stakeholder interviews were conducted by one of the international consultants. The researchers meanwhile were occupied with

testing and refining the field methodology and barely involved in the ToC process. The design of the evaluation was guided by the generic framework outlined in the PIALA strategy paper (IFAD & BMGF, 2013b), rather then the ToC process. As a result, the scope of the evaluation remained too wide and the focus unclear, making it difficult for the researchers to relate the evidence to the ToC and arrive at greater precision in the causal analysis. Moreover, less ownership of the ToC on the part of the stakeholders also hindered their critical engagement in valuing programme contributions and identifying areas requiring more attention in the next programme (IFAD & BMGF, 2013a).

17. Learning from this first pilot, more time and budget was granted in Ghana for the lead researchers to conduct a thorough desk review and engage in the reconstruction and visualization of the ToC. This was crucial for them to fully grasp the programme in its broader context and identify the programme's causal claims and assumptions. A design workshop halfway the ToC process convened national stakeholders to build a shared understanding of the ToC and select the mechanisms and assumptions to be evaluated and agree on the standards for valuing programme contributions. Evaluation questions were formulated around these mechanisms and assumptions, and methods were selected specifically in relation to the causal links in which they occurred. This more rigorous and collaborative process laid the foundation for the entire evaluation.

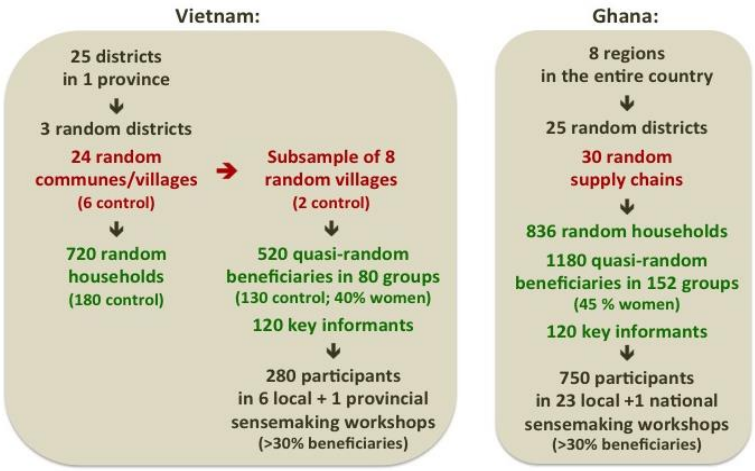
2.1.2 Considering different design options

18. Evaluations are often commissioned and conducted on a shoestring budget and under time constraints, which limits the possibility to contract research teams with sufficient capacity. In such situations, either scope or scale of the evaluation may need to be downsized. In an evaluation using PIALA, scale refers to the size of the principal sample of 'open systems' (cf. Section 1), from which households for survey and intended beneficiaries for participatory research are subsampled. Scope refers to coverage of programme components and mechanisms reflected in the ToC.

19. When choosing for a '*full scope - limited scale'* design, the emphasis is on *learning* about the programme's total contribution to impact in *select cases* under *specific conditions.* The ToC approach is useful here for obtaining a systemic understanding. Fieldwork and analysis are less resource-intensive as the samples are smaller. Evaluation findings are not generalizable, however, and thus insufficient to report on contributions to impact for the entire population (unless the programme itself is case-based, thus implemented at a limited scale) (MOFA/GOG, IFAD & BMGF, 2015). When choosing for a '*limited scope - full scale'* design, the purpose is to *learn* about the effects of one or two particular aspects or mechanisms of the programme. The ToC approach is not mandatory, which saves time and budget but runs the risk of arriving at flawed conclusions and limiting stakeholders' systemic learning. Components are studied in isolation, thus not permitting conclusions about their systemic interactions. For example, a cost-effectiveness study of Farmer Field Forums (FFFs) in Ghana recommended a scaling up because of the high adoption of new technologies (MoFA 2014), yet the PIALA evaluation showed that in a downward conjuncture the success of the FFFs contributed to market saturation negatively affecting livelihoods across the entire country (MOFA/GOG, IFAD & BMGF, 2015).

20. In Vietnam this trade-off and its consequences were not yet well understood. The assumption was that participatory research in a subsample of the already *small scale* sample of villages where household surveys were conducted, would be sufficient to conduct a *full scope* inquiry of programme contribution to impact for the entire province. This caused problems for the linking of *not case-specific* household-level findings to *case-specific* participatory research findings, and for generalising participatory research findings regarding the effects of programme mechanisms on livelihoods, hindering rigorous causal inference (IFAD & BMGF, 2013a). In Ghana, a conscious choice was made therefore to employ *all* methods and processes (including participatory) in the same sample, thus *at the same scale* (cf. Figure 1).

Figure 1: Sampling hierarchy of the two PIALA pilots



Three major design options were discussed with clients and commissioners (e.g. *'full scope - full scale'*, *'limited scope - full scale'*, or *'full scope - limited scale')* before any procurement or design work was started, as to give them a good understanding of the value-for-money they could expect. They chose the most expensive option: a *'full scope – full scale'* design (MOFA/GOG, IFAD & BMGF, 2015).

2.2 Cost-benefits

21. The total cost of this first phase of evaluation in Ghana was a little less than **USD $ 18,000**. This also included a two-day PIALA design training for less than USD 2,000. The training was crucial to afford the research team leaders with the basics of PIALA and help them understand what type of impact evaluation was aimed for, what type of questions were to be answered, the components to be adapted to the context as part of the in-country design, and the standards and principles to adhere to. It essentially sought to prepare them for a detailed design of an impact evaluation using PIALA. The product that came out of this first phase was the Evaluation Design Paper presenting the ToC, the evaluation focus and frame, the sampling, the methodology and the rating system (MOFA/GOG, IFAD & BMGF, 2015).

22. The two pilots have shown three major benefits of using a systemic ToC approach for focusing and framing an impact evaluation:

a) the creation of shared understanding among stakeholders of the programme theory and broader influences affecting rural livelihoods, enabling them to meaningfully engage in the analysis;
b) the formulation of evaluation questions around the programme's causal claims and assumptions, focusing data collection on the causal links of these claims and thus permitting a more rigorous causal inference;
c) the selection of methods specifically in relation to the causal links to be investigated, making it possible to employ the methods concurrently and independently, thus leaving more flexibility in the organisation of fieldwork.

3 Collecting and linking the data

23. This phase involves important decisions about the multi-stage sampling, the selection of methods, and the use of tools for collecting, monitoring and linking the data. The decisions must ensure data sufficiency and quality while also enabling meaningful participation (BetterEvaluation, 2014). The main QAF question for this is:

How and to what extent does PIALA help rigorously collect, triangulate and link the data needed to produce the evidence for confidently answering the evaluation questions in a participatory manner?

3.1 Critical reflections

3.1.1 Developing the sample frame centred on "open systems"

24. Market-bounded systems such as value chains have open boundaries, meaning they interact with their environment (Burns, 2014b; Humphrey, 2014). Such open systems are difficult to discern and sample. If there is no shared understanding of these systems, using proxy sample units may be necessary, which will affect the rigour in the analysis.

25. In Vietnam, for instance, villages were considered an adequate proxy for the short value chains that the programme sought to develop at the local level, close to the farmers and involving few local actors. However, not having clearly identified the value chain systems, and subsampled the households and research participants within these, made it difficult to inquire the systemic interactions and link data on changes in business environment and local capacity of service providers to changes in livelihoods within the value chains, thus compromising on analytical rigour (IFAD & BMGF, 2013a). Learning from this, much more work was put into the identification and sampling of the value chains systems in Ghana. The RTIMP evaluation focused on four commodity chains developed by the programme, each consisting of many supply chains around the country. These supply chains formed the principal unit of analysis. Hence the evaluation had four different supply chain populations from which to sample. The supply chains though formed loose geographic areas where smallholders supply raw products to a small enterprise or industrial off-taker (called "supply chain leaders") manufacturing higher-value products for bigger markets. The chains were not homogeneously defined and consistently registered. They interacted and overlapped geographically and administratively. Hence they were hard to discern and in practice often differed from what was sampled on paper. In some cases, suppliers produced different crops supplying different buyers and markets; in other cases

different supply chains were mixed up and could not be discerned; and in other cases there were no suppliers but only off-takers (MOFA/GOG, IFAD and BMGF, 2015). Ensuring the evidence collected on these systems remained comparable required much creativity and coordination.

26. Moreover, while in Vietnam demographic data for the entire sample of villages from which to subsample households was readily available from provincial government agencies, no such population lists were available in Ghana that could be matched to the sample of supply chains. Budget limits did not permit extra fieldwork to construct the lists based on house counting. This would have required 1-2 extra days of work in every district, at a total extra cost of USD 20,000-40,000, which was no option within the approved budget. Instead, a systematic sampling technique was applied in which every fifth or tenth household was selected, following a straight or zigzag line from a central point of the main community in each supply chain area. A major challenge encountered in this was that the vast majority of towns and villages in Ghana are not laid out in an orderly fashion. The settlements are mostly scattered and houses are not built in a structured layout. In some cases, it was almost impossible to count every 5^{th} or 10^{th} house in a straight line. Some settlements in the North were widely scattered with very few houses, which made the use of the 10^{th} /5^{th} rule arduous. Also, in more urban areas there were shops, sheds and other structures in between houses, which made the exact counting more challenging. Generally the teams stuck to the rule but in some communities they had to use their discretion while consulting their supervisors (Ibid).

27. These two sampling issues show the importance of 'responsiveness' to the constraints occurring during the evaluation, complementary to 'consistency' in employing methods and processes, in order to warrant rigour.

3.1.2 Deciding on depth and breadth of inquiry

28. PIALA's mixed-methods approach pursues depth through focused participatory inquiry of 'open systems' and breadth through a representative household survey in the sample of these 'systems'. The sample size of 'open systems' determines the scale (cf. Section 2.1.2), from which households for survey and intended beneficiaries for participatory research are subsampled. The larger the scale, the more challenging systematic data collection and collation using this mixed methods approach will be. Particular participatory research becomes more onerous when done at a larger scale with limited resources.

29. The choice to conduct a *'full scope - full scale'* evaluation in Ghana certainly affected quality, as research fatigue is higher when conducting six weeks of uninterrupted participatory research at a national scale. Extending the period of fieldwork would have permitted more breaks but likely affected researchers' motivation and commitment and obviously increased the cost. Limiting the period by working with more teams in parallel, on the other hand, would have led to greater variation in the quality and depth of the work and equally required more budget. So there is no clear-cut solution for this dilemma. Yet quality can be upheld if working with highly competent and motivated research team leaders and a research coordinator who takes pride in quality. In Vietnam, researchers spent only two weeks in the field and had two days of rest, yet didn't produce higher quality than in Ghana. They struggled with qualitative data capturing and data linking under time pressure, due to their limited understanding of the ToC and limited experience with participatory or qualitative

research. The large amount of qualitative data was overwhelming, which demotivated them to systematically reflect on quality and crosschecking and linking the data on a daily basis. Reports were compiled long after fieldwork was finalised, affecting the quality of the evidence and in turn challenging the aggregated analysis (IFAD & BMGF, 2013a).

30. Learning from Vietnam, a research firm capable of handling large-scale mixed-methods research, with substantial experience in participatory research, was selected in Ghana, and a senior statistician was hired and put on the team to lead on the sampling and the quantitative analysis. The research team leaders had a solid grasp of the ToC, methods were tightly geared to inquire the causal links in the ToC, and adequate guidance was provided for every single method of tool to be used. This enabled the researchers to rigorously triangulate and link the data and closely monitor quality. Data gaps and weaknesses were timely addressed and sufficient evidence was produced to adequately answer all evaluation and learning questions (MOFA/GOG, IFAD and BMGF, 2015).

31. This convincingly demonstrates an important win-win as opposed to what on first sight appeared as a trade-off –namely, that an evaluation does not have to sacrifice depth of analysis for breadth of coverage (or vice versa) but can achieve both[5], if sufficient capacity and motivation to deliver quality is present. In contexts where local research capacity is weaker, scale and ambitions need to be lowered and/or more resources have to be put into training, coaching and supervision.

3.1.3 Maintaining independence in field mobilization

32. Field mobilisation of research participants is best undertaken independently from the programme to avoid positive bias. This requires sufficient logistical capacity, time and budget for organising the mobilisation and dealing with unexpected obstructions and problems (e.g. resistance from local officials). In countries like Vietnam, however, the government controls all fieldwork. Hence complete independence is impossible. In cases where independent mobilisation is not feasible or possible, strong facilitation skills are needed to minimise undue influence or interference.

33. The challenges encountered in Ghana were quite different from those in Vietnam. In Vietnam, the evaluation took place in one province with good road infrastructure, hence travel distances were manageable and communities were easily accessible. Focus group discussions were organised within the villages, so participants did not have to travel. Local transportation and mobilisation was organised by local officials and programme staff, saving time and making the research more efficient, but also making it more difficult to maintain independence. On the other hand, staff and officials were more engaged as the evaluation unfolded (IFAD & BMGF, 2013a).

34. In Ghana, the researchers took responsibility for transportation and mobilisation to warrant greater independence. Staff and officials were only engaged in interviews and workshops. They were challenged though by the national scale, very poor infrastructure, remoteness of communities, and spread of communities in the supply chain areas. Researchers and participants had to travel far on poor

[5] In some cases more depth might be desirable in relation to some of the learning questions as to obtain more fine-grained explanations relevant for policy, targeting and scaling up of certain programme mechanisms. But for this, a separate inquiry in a select subsample would be most appropriate.

roads to conduct the focus group discussions. Locations for convening focus groups with participants from different communities (known and trusted by all) were difficult to find. The budgetted compensation of 4 GHS per person for public transportation was often insufficient particularly for those living in very remote areas with very poor access roads. Hence in these cases the research teams had to make their own vehicle and driver available, which in these remote areas was also insufficient to keep up with the 4-6 FGDs scheduled a day. In some districts, officials were unwilling to collaborate and hand over the documents requested since some thought their work was being evaluated. This was partly due to the fact that the field research was organized independently without much help of RTIMP. Hence a tension occurred between maintaining independence and engaging officials in the organization and conduct of the evaluation (MOFA/GOG, IFAD and BMGF, 2015).

3.1.4 Contextualising poverty analysis

35. To make it possible to say something about programme influences on poverty, data on these influences and on poverty need to be linkable. Hence poverty has to be defined in a way that is relevant to the context and conditions of the villagers in the programme area. In Ghana and Vietnam, this was done in a participatory manner by employing a PRA tool for wealth and wellbeing ranking as part of the generic change analysis method. This method is generally used for identifying locally relevant indicators on wealth and wellbeing and analysing changes in relative poverty status. The assumption is that by using such locally defined indicators, this helps overcome cognitive bias.

36. The participatory characteristics of wealth and wellbeing obtained from the ranking exercise, however, weren't used as indicators for collecting data on poverty and assessing changes in poverty status through the household survey. This would have implied participatory data collection prior to the evaluation as an input for designing the household survey, which in Ghana, given the large travel distances, would have required substantially more time and budget without producing remarkably more rigorous or more credible conclusions with regard to the programme's influences on poverty. After all, the pilot in Ghana didn't show major differences between the most pertinent characteristics of wealth and wellbeing identified by the most of the women and men who participated in the ranking exercise, and the ones inquired by the household survey and used as the basis for statistical poverty analysis (IFAD, GOG & BMGF, 2015).

37. In Vietnam, IFAD's general poverty impact indicators were found largely irrelevant to the Vietnamese context and the government's categories of income-based poverty too narrow to assess project influences on changes in the distribution of poverty from a multi-dimensional perspective (IFAD & BMGF, 2013a). In Ghana, greater attention was paid, therefore, to ensure that the household survey questionnaire and analysis were better adapted to the context. Poverty status was defined in a statistically relevant manner by first applying a proxy means test for assessing each household on wealth and wellbeing characteristics identified by the Ghana Statistical Service (GSS) as relevant proxies in the Ghanaian context, and then using the data collected on these characteristics through the household survey for computing the categories of poverty status applying a Principal Components Analysis (PCA). Also here greater rigor could have been obtained in computing the poverty categories if a more detailed questionnaire would have been used to inquire more household characteristics in greater detail. But again this would have added a serious cost, while also enhancing the risk of

research fatigue on both the part of the respondents and the researchers. While the survey questionnaire used in Ghana was a bit longer than the one used in Vietnam, it was still kept within the limit of a 20-minute interview to avoid this risk.

38. Thus instead of spending more resources on collecting and analysing participatory poverty characteristics prior to the evaluation for designing the household survey, or on collecting more fine-grained data on more household characteristics to identify poverty categories, the choice was made in Ghana to keep the poverty analysis short and enable participants to engage meaningfully in the causal analysis (MOFA/GOG, IFAD and BMGF, 2015).

3.1.5 Dealing with power and bias in participatory research

39. Participatory methods are often considered not suitable for impact evaluation because of the perceived risk of bias and/or power-blindness (Copestake, 2013). In-depth case-based participatory research makes it impossible to draw generalizable conclusions about the difference a project or programme has made for a large population, thus limiting the ability to influence policy and decisions. Large-scale participatory research on the other hand tends to instrumentalize and thus de-politicize the research context and relationships, similar to what large-scale surveys tend to do, thereby enhancing the risk of power-blindness and thus gaming and bias (Gaventa & Cornwall, 2006; Mohan & Hickey, 2004; Mosse, 2001).

40. A suitable way to address this challenge is to facilitate the participatory processes in a power-sensitive way, requiring strong facilitation skills. Instead of trying to gain control over the data by avoiding or ignoring power dynamics (which in real-world evaluation is impossible), researchers recognize their own position and influence, take a step back and share the control over the data with the research participants by carefully facilitating a process in which equal voice levels out power and bias. Rigour then emanates not from controlling the data and avoiding potential influence or bias, but from the sharp observation of motivation and interaction and from ensuring a continuous cross-checking during the process. This is what Chambers (2015) calls *"inclusive rigor".* Using visual tools such as causal flow mapping and matrix scoring is particularly useful in this respect, as it enables people to *see* how the reconstruction takes shape and indicate where things are mostly relevant or flawed. There is always the danger of one with more knowledge and power trying to override others with less knowledge and power, yet good facilitators notice this fairly quickly and know how to derail these attempts. They also know how to arrange groups and employ methods in ways that enable equal voice and thus ensure that the voice of the majority is stronger than the one of a single power-holder.

41. In Vietnam, this was quite challenging for the researchers. Most of them came from a purely quantitative research background and struggled with the principle of triangulation (or cross-checking). Triangulation in their view was more a means of verification of 'truth' than a way to compile a multi-perspective systemic picture forestalling the dominance of a single truth or bias. They also were less aware of their own positionality and power reflected in their value judgements about the data. This didn't pose a challenge in the pilot in Ghana. The researchers there were well versed in participatory research and skilled facilitators with a good understanding of the process of triangulation. Also a major advantage compared to the pilot in Vietnam was that the research assistants were from around

the areas where they conducted the field research, so they knew the local cultures and political economies very well and spoke multiple local languages.

42. In Ghana, there was no budget for allowances to compensate for participants' time. In Vietnam, small allowances were paid in accordance with the project's protocol. Although this wasn't people's first motivation to participate since the allowance was too small and much less than what they would earn at work, it may have generated a courtesy bias in some cases where it was associated with the project. To avoid this, no allowances were granted in the pilot in Ghana. The principle of voluntary and non-paid participation was applied very strictly, but the researchers experienced great difficulty in getting this understood and accepted by the participants, as most of them who had participated in the programme were used to receive generous allowances. Hence people were generally less keen on participating in something that did not give them a direct and tangible profit or benefit. Moreover, people also showed more reluctance as the programme had promised them markets if they would participate in the programme, but these markets had never come. In Vietnam, people were more eager to participate, with or without allowances or fulfilled promises, just for the purpose of debate and learning.

3.1.6 Methodological complementarity

43. Mixed-methods imply the use of both quantitative *and* qualitative methods. Unique is their intentional attempt to integrate frameworks, standards, methods and procedures from different social science disciplines to serve mixed purposes or uses. The intention is to draw on the strength of each of these disciplines and overcome the limitations and biases that they show when applied in impact evaluation as a single method or approach. There is a growing consensus in the international evaluation community that mixed-methods is in most cases a better choice for the design and conduct of impact evaluation than a single quantitative or quantitative method, in particular when it comes to evaluating more complex programmes (Bamberger, 2012; Stern et al., 2012; White, 2014).

44. One of the well-known weaknesses of a classic household survey, for instance, is its length. As mentioned earlier (in Section 3.1.4), long interviews increase the risk of gaming and bias due to respondent and enumerator fatigue. Another important weakness of the classic survey is that questions tend to reflect the assumptions of the designers. As Sarah White (2015: 138) illustrates in her recent analysis of a large-n RCT impact evaluation of girls' empowerment in Bangladesh: *"hypothetical questions are susceptible to 'desirability bias' in which respondents give the answer they believe the researchers wish to hear."*[6] In the PIALA pilots, the household survey was limited to no longer than a 10 to 20 minutes interview, and triangulated with other methods.

45. Methods however can only be triangulated if they inquire the same causal links. Mostly this is done in a *sequential* design in which qualitative methods serve to design and support the quantitative ones. Sometimes quantitative methods are used to collect generalizable data needed to identify the outliers and the mainstream cases, from which then a small sample can be taken to conduct more in-depth case studies. If the methods have to be used *concurrently*, however, to inquire *different* links in the same sample in order to enable data linking and arrive at causal inference, in principle they cannot

[6] This type of bias is also often called courtesy or politically correctness bias.

be crosschecked, *unless there is some overlap*. The methods and tools in the pilot in Ghana were selected and used in such a way that they inquired different causal links, while also probing for explanations or causes of causes and thus partly overlapping with each other. Where the methods overlapped, it was possible to triangulate the data and as such compensate for weaknesses and possible biases.

46. Apart from the participatory methods and the household survey that are part of the PIALA standard package, two new methods were tried out in the evaluation of RTIMP in Ghana: Constituent Feedback[7] and the SenseMaker[8]. Both were methodological experiments and thus limited in size and ambition. They served the purpose to pilot-test and demonstrate the potential added value of these methods and identify the methodological challenges when used in impact evaluation (MOFA/GOG, IFAD and BMGF, 2015). In terms of added value, both methods generate quantified perceptual data that is quite different from the statistical data produced by a classic survey. Both use software to analyse distribution patterns and correlations[9], which similarly to a classic survey gain statistical power when employed in a large sample. But instead of letting respondents answer pre-determined and pre-quantified questions, these two new methods permit the participants to quantify or score *their own views and perceptions* in response to a set of guiding questions, and they do this anonymously right after a group discussion and/or telling a personal story to which they relate their scorings. Consequently the analysis of the data produces patterns of people's perceptions that are much more recognizable afterwards when shown in a collective stakeholder debate involving the participants.

47. While Constituent Feedback focuses on performance of specific service delivery mechanisms, SenseMaker® investigates broader trends or patterns of *emergent* change and seeks to produce evidence of unknown or unexpected influences and effects that conventional methods are unlikely to detect (Deprez et al, 2012; Jenal, 2014; Van Hemelrijck, 2015). Moreover, SenseMaker is designed in a way that removes intermediary levels of structuring, interpreting and analyzing data, thus permitting rapid analysis and response, while avoiding researcher bias at each of these levels. It does so by collecting a large amount of fragmented –thus *unstructured*– experiences or narratives and letting respondents self-signify them, while enabling managers and decision-makers to directly access the raw narratives for making sense of the statistics produced by the SenseMaker® software. Tools such as *triads*, *diads* and *stones* enable people to self-signify how their story *proportionally* relate to sets of competing characteristics and variables reflecting plausible trade-offs or choices. This also helps to avoid "gaming" that occur in traditional scoring with scales (Ibid, Snowden, 2002), such as the one used in Constituent Feedback. The purpose of the latter however is rather to generate debate and learning among key stakeholders around feedback on performance, rather than to avoid bias and gaming. Rigour in Constituent Feedback is realized through triangulation and cross-validation, rather than through the method-specific procedure itself such as in SenseMaker or traditional statistical methods (Jacobs et al, 2010; Van Hemelrijck et al, 2011).

[7] Cf. http://www.keystoneaccountability.org/analysis/constituency.

[8] Cf. http://cognitive-edge.com/sensemaker.

[9] The household survey and the Constituent Feedback simply used SPSS, while the SenseMaker required special software that is designed specifically for analysing and visualising more complex landscape patterns and allows to go back to the respondents' original stories behind the patterns.

48. In the RTIMP evaluation in Ghana, the livelihood analysis was combined with a brief SenseMaker study, making the focus group discussions too long. Across the entire sample, paticipants had shown the symptoms of fatigue. Particularly for women with small children, the length of the method was troublesome, as they needed to return to their homes for preparing the food. Moreover, the abstract SenseMaker tools (triads and dyads) were challenging for farmers (particularly those illiterate) as well local translators to grasp[10]. The researchers also observed that many of the illiterate participants felt intimidated when they were given a marker to place their dote on a paper with text they couldn't read. Some didn't even want to try; others asked the researchers or other participants to place the dot for them anywhere they wanted, which created certain power dynamics around those who understood the tools and those who did not. All this may affect the quality of data, and thus the credibility of findings (IFAD & BMGF, 2015).

3.1.7 Data linking and quality monitoring

49. Quality monitoring of data and processes involves daily research team reflections using sets of questions that enables the researchers to assess the robustness of the emerging evidence and timely discover data gaps/weaknesses during fieldwork. This goes hand in hand with data collation (or data linking) in each locality, which is the process by which all data collected on the 'open system' inquired is cross-checked and linked to assess the causal claims in the ToC. It involves a reconstruction of the actual causal flow with the data that mirrors (and thus validates or refutes) the ToC for the 'open system' inquired in a particular geographic location.[11]

50. While in Vietnam the researchers experienced great difficulty with linking and processing the large amount of data, in Ghana this problem did not occur. There are three important reasons for this. First, the researchers profoundly studied the programme and understood the ToC. Second, questions and methods were well focused on the causal claims and links in the ToC to be inquired. Third, the researchers were very well versed in more complex and participatory research while also having more detailed guidance for data capturing, linking and quality monitoring. Learning from the pilot in Vietnam, much more attention was paid in Ghana to accurate and systematic data capturing and collation. To avoid data loss or contamination, clear note-taking instructions were followed, and data were sent to the head research office in Accra on a daily basis.[12] To ensure consistency, standardized data capturing templates and spread sheets were used. For data collation and quality monitoring, similar but improved tools were used as in Vietnam but with more detailed guidance (MOFA/GOG, IFAD and BMGF, 2015). Because the researchers in Ghana didn't experience difficulties with data processing and linking, they felt more confident and had more time to do their team reflections, resulting in higher quality evidence.

[10] In Ghana, there are more than 80 different languages. The researchers spoke between 5 and 12 of these. In some areas, they had to hire local translators. These were village people, thus often not highly educated.

[11] The sampled 'open systems' were formed by the commodity supply chains in Ghana and the short value chains in Vietnam; the localities concerned the districts that administered the supply chains in Ghana and the villages and communes that supported the development of the value chains in Vietnam.

[12] When there was no internet available, data was sent per mobile phone.

3.2 Cost-benefits

51. The total cost of the data collection and processing in the *'full scale –full scope'* evaluation in Ghana was about **USD 152,000**. This also included the sampling, the introduction of new methods (such as SenseMaker and Constituent Feedback), training, tool development and field-testing at a total cost of USD 22,000. Methods and tools used in Vietnam were revised, new methods and tools added, and concepts translated in local languages. Although all researchers in Ghana were well versed in participatory research, none of them had been involved in impact evaluations combining participatory methods with statistics. Hence the importance of the methods training and field-testing. The main products that came out of this phase were: the researchers' field handbook (incl. guidance notes, questions, samples, note taking formats, research schedules and contacts), and the district data collation tables and field notes.

52. The major benefits of using PIALA's sampling and participatory mixed-methods approach as demonstrated by the pilots are:

a) the possibility of achieving both 'breath of coverage' and 'depth of analysis', thus of conducting a rigorous and in-depth impact inquiry of *multiple-interacting* or *systemic* change processes in a relatively large sample of 'open systems' covering a large geographic area (e.g. country-wide in the case of Ghana);
b) the possibility of doing this in a way that enables local stakeholders to meaningfully engage in causal flow analysis and debate about change (without needing to ask them directly about "attribution");
c) the robustness of the evidence achieved by: (a) using participatory mixed-methods tailored to inquire specific causal links in the ToC; (b) systematic and extensive triangulation of methods; and (c) applying 'mixed-rigor' standards comprised of both *inclusive* and *statistical* rigor.

4 Synthesizing the evidence for analysing and debating contributions

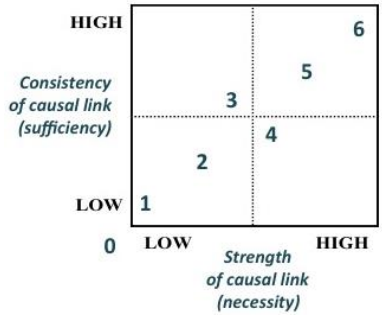
53. This phase involves the process of synthesizing the evidence to facilitate participatory sensemaking with stakeholders, and analysing the evidence and findings (incl. from the sensemaking) to draw conclusions about programme contribution to impact on poverty. The process essentially leads to answering the evaluation and learning questions agreed with stakeholders at the start of the evaluation. The leading QAF question for this phase is:

How and to what extent does PIALA help synthesize and present the evidence to facilitate rigorous analysis and solid debate of programme contributions in a participatory manner?

4.1 Critical reflections

4.1.1 Synthesizing the evidence

Figure 2: PIALA scoring



54. Data collation and quality monitoring was documented and synthesized in a standard data collation table for each locality. In Vietnam, due to time constraints and inadequate guidance, the researchers did not manage to do this systematically on a daily basis. In Ghana, greater attention was paid to training the researchers and providing them with sufficient guidance, while also important improvements were made to the tools.

55. Particularly the data collation table was key to the analysis and worked very well in Ghana. Filling in the table involved a scoring of the robustness of the emerging evidence (based on the quality monitoring) and of the *strength* and *consistency* of each of the causal link in each contribution claim in the ToC (based on the emerging evidence) (cf. Figure 2). After completion of fieldwork, the district data collation tables were further synthesized in an aggregated collation table as the basis for causal inference. Scores and explanations entered into the aggregated table were thoroughly cross-checked with original field notes and district data collation tables (MOFA/GOG, IFAD and BMGF, 2015).

4.1.2 Participatory sensemaking

56. Engaging beneficiaries, service providers and decision makers in collectively making sense of emerging evidence before turning to final analysis and reporting has both instrumental and empowering value (MOFA/GOG, IFAD and BMGF, 2015). Doing this in each researched locality (before ending fieldwork) and subsequently at programme level (right after fieldwork) helps improve and strengthen the evidence, overcome bias and create ownership of evaluation findings among stakeholders. A participatory sensemaking workshop model was developed that was first pilot-tested in Vietnam and then further expanded and improved in Ghana.[13]

57. Rigorous facilitation[14] in this model is considered essential for enhancing credibility and confidence of evaluation findings and generating solid debate and systemic learning among key stakeholders. Group compositions and plenary processes are designed with great care (using the concept of "patches & nodes") to enable participants to critically engage and express their views in the presence of power holders. Meaning and depth of engagement is acquired by going through several iterations, each adding different dimensions of evidence to the construction of the causal flow mirroring the ToC.

[13] In Vietnam, we organised six village-level workshops with 180 and one provincial workshop with 100 participants; in Ghana there twenty-three district workshops with 650 and one national workshop with 100 participants. Participants were purposively sampled from the research participants for the diverse perspectives they will bring to the analysis. Beneficiaries comprised more than 70 % of those attending the local workshops, and more than 30 % at the provincial/national workshops.

[14] Following the definition of 'rigour' used in PIALA and its QAF, 'rigorous facilitation' is considered as a well thought-through set of methods, tools and principles applied for designing and facilitating the various participatory processes, which is done in a way that is both *consistent* across all localities and *responsive* to the constraints encountered in these localities.

Textbox 1: Participatory sensemaking sessions

The sensemaking workshops essentially involve four main types of sessions[15]:

a) *Focus groups debating prepared evidence statements that are based on collated quantitative and qualitative data.* The groups are organized around the different causal claims in the ToC, in ways that convene those most knowledgeable of the topics and issues discussed. The groups represent the home "patches" in which the participants feel comfortable and safe to express critical views with their peers.

b) *Mixed groups constructing a causal flow with the evidence statements to validate/refute the ToC.* The groups are organised around geographic areas and comprise a majority of beneficiaries to make them feel more empowered in the presence of power holders. To enable equal voice in the plenary sharing and discussion, beneficiaries always present their group findings first, before any other group.

c) *Plenary discussion around areas of less success.* After minimum two iterations of small group discussions, a plenary fish bowl exercise is organised to discuss critical areas of less success requiring more investment. Chairs are placed in a circle: three for each stakeholder group (such as farmers, officials, bankers). Only two members from each can take a seat. When a new participant takes one of the empty seats, another member of his/her group has to leave. Intended beneficiaries form at least half of the discussants at all times.

d) *Focus groups valuing programme contributions* by placing the evidence statements along two scales: from *"strong contribution"* to *"contribution overrun by other influences"*, and *"positive"* to *"negative"* impact. This is done back in the patch groups. The groups then share their ratings in plenary and discuss the differences in their judgments. Beneficiaries present their ratings first, before the other groups.

58. The workshops were very successful and appreciated by all the participants in both pilots. This was overwhelmingly shown by the micro-surveys and reflections held at the end of every sensemaking workshop both at local and programme levels. Participants confirmed they gained a better understanding and more complete picture of the development processes and a better sense of the learning purpose of the evaluation. There were lively debates about programme contribution to impact and beneficiaries critically engaged in discussing and identifying priority areas for future investment. Essential to this success were the time and resources invested in organizing the workshops, and the capacity to rigorously design and facilitate them. When operating on a shoestring, the amount workshops and participants may need to be more limited. But this undoubtedly has implications for rigour and inclusiveness.

[15] In the local sensemaking workshops, the b-, c- and d-sessions are merged and simplified, and happen only in plenary.

4.1.3 Analysing with/without configurations supported by the evidence

59. Mainstream impact evaluation assumes that comparative analysis of evidence from treated and non-treated locations is both accessible (thus feasible) and necessary (thus rigorous) to reach generalizable conclusions about impact on rural household poverty. In most 'real world' evaluation contexts (Bamberger, Rugh & Mabri, 2012), however, it is not feasible to assign communities to specific interventions and identify credible and matching control groups, due to the high causal density (Woolcock, 2013). In such contexts, different ways for arriving at rigorous causal inference need to be employed (Befani, 2012).

60. In the Vietnam pilot, causal density and heterogeneity in programme treatment and conditions hampered the matching and sampling of control villages, which made us hesitant to generalize certain findings regarding program contributions to impact. Attaining the desired level of rigour in counterfactual analysis appeared not feasible. Our implicit assumption there was, however, that comparative analysis of data from treated and non-treated sites was possible and necessary to reach generalizable conclusions with regard to impact valid for *all* treated sites. If not, the programme's evaluability had to be questioned (IFAD & BMGF, 2013a).

61. In Ghana, this assumption was challenged by moving away from the classic counterfactual-based approach, and using *systemic heterogeneity* as the basis for identifying and analysing programme contributions. Instead of a classic counterfactual inquiry of household-level impact, a configurational methodology was developed (see Figure 3 below), integrated with a "generative perspective" (Punton & Welle, 2015; Stern et al, 2012). A similar sampling approach was used as in Vietnam, but without sampling control sites (or sites not covered by the programme). Instead, we sampled and compared areas with strong and areas with dysfunctional or weak program treatments in various combinations and gradations, while also taking into account other influences. Hence we looked at different blends or configurations of treatments, conditions, influences and outcomes supported by the evidence, in other to identify and assess programme contributions to impact.

Figure 3: Part of the configurational scoring table of the RTIMP evaluation

							Contribution Claim of RTIMP Component 3 ↓ Enhanced Processing (O3)			Contribution Claim of RTIMP Component 2 ↓ Enhanced Production (O2)			Contribution Claim of RTIMP Component 1 ↓ Enhanced Market-Linking (O3)			Contributions of RTIMP Components 1, 2 & 3 ↓ Improved Livelihoods (I2)	
		DSF	FFF	GPC	MEF	MEF (M3c)+C1a+M3b →C3c	GPC (M3b)+C3c →C3b →O3	Evidence Strength	FFF M2a+M2b+ (M2c) →C2a	C2a+C2b →O2	Evidence Strength	M1c+M1b+ O3+O2 +O1→C1b	DSF C1a+(M1) →O1	Evidence Strength	O1+O2+O3 →I2	Evidence Strength
Tano North (Apesika)	(CZ)	1	1	1	1	3	6	5	5	5	5	4	4	5	5	5
Techiman	(CZ)	1	1	1	1	4	5	5	5	5	5	4	4	5	5	5
Gomoa East	(SZ)	1	1	1	0	2	5	3	5	5	5	4	4	5	5	6
Assin South	(SZ)	1	1	1	1	3	4	4	6	5	4	3	3	4	4	4
Birim Central	(CZ)	1	1	1	1	3	3	4	5	5	4	3	4	4	4	5
Nkwanta South	(NZ)	1	1	1	0	3	4	5	5	4	5	3	3	5	4	5
Upper West Akim	(CZ)	1	1	1	1	2	4	4	5	5	4	3	3	5	4	5
Ashanti Mampong	(CZ)	1	1	1	1	3	4	5	5	5	5	3	3	5	4	5
West Gonja (Damongo)	(NZ)	1	1	1	0	3	4	5	5	4	5	3	3	5	4	5
Abura Asebu Kwamankese	(SZ)	1	1	1	1	3	3	5	5	5	6	3	3	5	4	4
Nanumba North	(NZ)	1	1	N/A		N/A			5	5	5	3	3	5	4	5
East Gonja	(NZ)	1	1	N/A		N/A			4	3	5	3	3	5	4	5
Central Gonja	(NZ)	1	1	N/A		2	3	5	5	4	5	2	2	5	4	5

62. 'Treatment' was looked at in terms of the presence or non-presence of functional or non-functional programme mechanisms in various gradations; 'conditions' included socio-economic, infrastructural, agro-ecological and demographic differences; and 'outcomes' were the observable changes (both positive and negative, intended and unintended) in livelihoods and household poverty.

For each supply chain (cf. first column in the Figure 3), the formal presence of programme mechanisms such as FFF or MEF[16] was inputted as a *binary code* (cf. next four columns in Figure 3). Starting with the impact claim, the districts were first clustered according to their scores on improvements in livelihoods. Differences and similarities were examined *within* and *between* these clusters. Next, findings from the analysis of impact-related changes and causes were then used to examine the changes and causes in each of the contributions claims separately, using a similar procedure as the one for the impact claim. Differences and similarities were then examined *within* and *between* the different clusters for each of the three claims (MOFA/GOG, IFAD and BMGF, 2015).

63. Rigorous configurational analysis requires high quality data and analytical skills, which in "real world" evaluation contexts are often not available. The approach can be simplified and brought within reach of evaluations with less favourable conditions, which is particularly relevant to contexts where classical counterfactual approaches are not possible. Yet this does not address the challenge of lacking research capacity. In both pilots, researchers were not experienced with rigorous aggregated multi-causal analysis. Because of the methodological innovation, the international consultants leading on the PIALA pilots took responsibility for the final analytical product. As a standard procedure, however, bringing in international evaluation experts is not the most cost-effective solution and does not foster in-country responsibility for high-quality impact evaluation. Investing in research partnerships with in-country research firms, while integrating PIALA in programme designs as part of an impact M&E process, might be a better option. Presumably this would help build greater ownership of the knowledge-generation for policymaking and over time create greater democratic space to influence decisions (Van Hemelrijck, 2013).[17]

64. In Ghana, we shifted towards this second option by ensuring those involved in the tendering process fully understood the requirements and including far more days for in-country supportive supervision and capacity-building by the international consultant who was leading on the pilot. Crucial was the intensive engagement of the coordinator of the national research team that conducted the evaluation, resulting in a much greater ownership and responsibility compared to Vietnam. Moreover, the strong learning orientation and active participation of IFAD's country programme manager and his team also fostered greater interest and engagement on the part of the partners. The Ghana initiative was experienced more as a joint learning journey, a partnership rather than a technical consultancy.

[16] FFF stands for Farmer Field Forum; MEF for Micro-Enterprise Fund.

[17] Many research firms have become donor-driven and merely implement research contracts without taking responsibility for the outcomes of the research. This is largely the result of how most research contracts are framed and research is commissioned and used by donors. Our assumption is that by framing it differently and building more of a partnership, research firms will become more critical of the quality of their research and take greater responsibility for the added value and the use of the knowledge it generates.

4.2 Cost-benefits

65. The total cost of the last phase of the evaluation in Ghana came at **USD 64,000**. This included all sensemaking workshops as well as the analysis and final reporting. The key product that came out of this phase is the evaluation report.

66. Major benefits demonstrated by the pilots of using the PIALA methods and processes for synthesizing evidence, participatory sensemaking and aggregated causal analysis are:

a) the robustness of the evidence achieved by: (a) the rigorous scoring of confidence in the evidence and of strength and consistency of causal links based on the evidence; and (b) the extensive cross-validation of evidence through participatory sensemaking;
b) the systemic learning and ownership of evaluation findings among key stakeholders created through the participatory sensemaking, which enhances the likelihood that findings will be used for policy and investment decisions;
c) the empowerment-related value of engaging stakeholders (particularly beneficiaries) in the sensemaking and contribution analysis creating space for real voice and dialogue between decision-makers, service providers and constituents.

5 Key outcomes and learning points of the PIALA pilots

67. In this last section I present a summary of the main outcomes and learning points from the two pilots, which I hope will be useful for the design and conduct of future evaluations using PIALA. I try to answer two important questions:

1. To what extent did PIALA help produce rigorous and inclusive evidence adequately answering the evaluation questions?
2. What were the main processes and decisions that spawned or thwarted rigor and inclusiveness?
3. What tensions or trade-offs occurred between the different standards of the different purposes[18] that PIALA sought to serve? What context factors influenced these?

68. The first pilot in Vietnam experienced several limitations from which much was learned in the adjusted approach employed in the second pilot in Ghana. In Vietnam, we encountered issues with data collation, translation and note taking on participatory processes that affected the quality and transparency of the evidence. Differences in language use and communication style between researchers and villagers created tensions and noticeably influenced the group dynamics (IFAD & BMGF, 2013a). Also government control on planning and mobilization made it more difficult for the researchers to maintain independence. Moreover, coverage of participatory research findings and confidence in linking *distributed survey data* on impact to *case-specific participatory data* on contributions remained limited. Although the researchers worked hard to improve the logic and quality of their analysis, their reports displayed the issues –e.g. insufficient fine-grained explanations and quant-qual integration, absence of gender differentiation, limited accurate multi-causal analysis[19], and limited indication of confidence or strength of evidence for specific assertions of causes and explanations. This affected our ability to answer the evaluation question. Despite the issues, the evidence produced was substantially crosschecked and validated by hundreds of people and therefore could not be invalidated. Yet it needed reprocessing to arrive at conclusions. A other Vietnamese team was hired to help reviewing all quantitative, qualitative and secondary data (most in Vietnamese). This delayed[20] the reporting process but enhanced the robustness of the final report.

69. The main processes and decisions that affected the quality of the evidence and the credibility of the conclusions in the pilots, were:

- The selection of a Hanoi-based research firm with a merely quantitative background, very limited experience in participatory research, limited knowledge of the local context and dialect in the Mekong delta, and limited logistical capacity to manage complex evaluations at scale.
- Insufficient involvement of the local researchers in the desk review and the reconstruction of the ToC, resulting in their limited understanding of the programme and limited ability to relate

[18] E.g. instrumental and empowering; reporting and learning; replicable and transformational…

[19] Incl. referencing and geographic specificity related to focus vs. non-focus villages.

[20] It can be safely assumed that the time lag in the reporting did not imply any difference in the validity of the conclusions, since impact in the DBRP had a slow rate of incubation and change, and thus very unlikely would have shifted in five months' time. Furthermore, compared to mainstream impact assessments that measure T and T+1 with time lags of up to 5 years, this impact assessment can be considered a real-time evaluation (IFAD & BMGF, 2014).

data to the program's causal claims and mechanisms, triangulate the data, identify complementary or competing mechanisms or explanations.

- Insufficient engagement of stakeholders in the process of reconstructing the ToC, identifying causal claims and assumptions, and selecting standards and criteria for contribution rating, which resulted in a limited ownership of the ToC and the evaluation findings[21].
- Evaluation design guided by a generic impact framework developed with global IFAD and BMGF stakeholders, rather than a program-specific ToC process with program stakeholders, resulting in a lack of clarity and focus of evaluation questions[22].
- Employment of participatory and statistical survey-based methods in different samples, based on the *assumption* that qualitative and/or participatory methods are only needed to *explain things* (thus at *limited scale*) while statistical methods must measure the real impact in large-enough representative samples (thus at *full scale*). If methods are selected to investigate different causal links, however, downgrading the scale for any of the methods jeopardizes the ability to arrive at rigorous causal inference (IFAD & BMGF, 2013a).
- Sampling of villages as a proxy for the value chain systems to be inquired, and insufficient effort put into identifying and agreeing with stakeholders on a shared definition of these systems before starting data collection.
- Sampling of non-project villages that couldn't serve as a credible control group due to the high causal density and heterogeneity in program distribution and treatment.

70. The second pilot in Ghana successfully addressed all the limitations encountered in the first pilot. Sufficient time was invested in the procurement and collaborative framing and focusing of the evaluation together with stakeholders. Local researchers were selected for capacity in mixed methods and participatory research, and were well versed in handling tensions between participatory process and rigorous data requirements, maintaining independence in mobilisation and coding and collating large amounts of qualitative data. Research assistants were from the areas where field research was conducted, and thus knew about the local political economy and spoke several of the local languages. The research team led the ToC process and engaged national stakeholders to build shared understanding of presumed change pathways and contributions. Agreement was reached with stakeholders upfront about contribution claims and mechanisms to be evaluated and questions and standards to be used. Methods for data collection and collation were organised around the contribution claims and tightly focused to inquire the causal links in these claims. Having a more solid grasp of the ToC, this helped the researchers to rigorously document participatory processes and triangulate methods, and gave them more bandwidth for data collation and quality monitoring. Also more time

[21] Here it is important to note that greater ownership does not necessarily result in more critical engagement and learning on the part of program managers and implementers. In Ghana for instance, it surprised everybody in the national sensemaking workshop how after two days of cross-validating evidence and ToC, the program managers and implementing partners still rated all contributions entirely positive and attributable to the program, whereas the ratings of the farmers and processors and other local actors were much more critical and honest. The lack of critical reflection and learning on the part of the national government officials and service-providers surprised all other participants and was received with an earsplitting and painstaking silence. This clearly shows that more than just a one-time off participatory evaluation is needed to develop a critical learning attitude.

[22] We assumed that the broader evaluation questions outlined in the PIALA strategy paper (IFAD & BMGF, 2013b) would be sufficient to guide fieldwork and focused analysis. This was clearly not the case. The researchers experienced great difficulty to relate the data to the questions and the causal links and mechanisms.

was spent upfront on defining the sample frame and the value chain system that formed its principle unit, and on discussing with stakeholders different options of counterfactual analysis. All methods and processes (including participatory) were used at the same scale to permit systemic analysis in all sampled value chains and comparison across the entire sample. Interference or influence by programme staff and local leaders was avoided by mobilising participants independently. A configurational method integrating a generative perspective was used for conducting a counterfactual analysis using systemic heterogeneity as the basis for assessing contributions among other (competing or reinforcing) influences. All these decisions and improvements contributed substantially to enhancing the quality and relevance of the evidence and our ability to answer the evaluation questions (MOFA/GOG, IFAD and BMGF, 2015).

71. Alongside these improvements, there were also some challenges encountered by the team in Ghana, which merit greater attention in future PIALA applications. These include: (a) the sampling of value chains that are open, interacting and overlapping systems difficult to discern; (b) the time and capacities required from (particularly illiterate) people to participate in some of the methods using abstract concepts (matrices, triangles, etc.) and flipcharts and markers; and (c) the rigid nature of the methodology to be applied in a systematic manner across all locations, which sometimes clashed with local village schedules and habits, and the limited time spent in each district (MOFA/GOG, IFAD and BMGF, 2015). The main take-away for future PIALA applications is that methods and tools need to be adapted to the participants' conditions as much as possible, and sufficient time is needed in the field to accommodate cultural habits and events and address unexpected challenges with regard to sampling and mobilisation. Another major constraint encountered in both pilots is the limited capacity to undertake a rigorous aggregated multi-causal analysis. Investment in building in-country analytical capacity through a research partnership with a local firm is probably the most cost-effective and sustainable solution in the long run. If PIALA would be integrated with impact M&E from the start of a programme, then arguably this would help overcome most of these challenges and contribute to creating more democratic space for stakeholders to influence decisions (Guijt, 2014; Peersman et al., 2015).

References

Bamberger, M. (2012). *Introduction to Mixed Methods in Impact Evaluation* (Guidance Note No. 3). InterAction, Rockefeller Foundation.

Bamberger, M., Rugh, J., & Mabri, L. (2012). *RealWorld Evaluation*. SAGE.

Bawden, R. (2010). *Messy issues, worldviews and systemic competencies.* In *Social learning systems and communities of practice*. Springer.

Befani, B. (2012). *Models of Causality and Causal Inference* (Review repared as part of the DFID study: Broadening the range of designs and methods for impact evaluation). DFID.

Befani, B., Barnett, C., & Stern, E. (2014). *Introduction – Rethinking Impact Evaluation for Development.* IDS Bulletin, 45(6).

Befani, B., Ramalingam, B., & Stern, E. (2015). *Introduction – Towards Systemic Approaches to Evaluation and Impact.* IDS Bulletin, 46(1).

BetterEvaluation. (2014). *BetterEvaluation Rainbow Framework and Planning Tool.* Retrieved from www.betterevaluation.org.

Burns, D. (2014a). *Assessing Impact in Dynamic and Complex Environments: Systemic Action Research and Participatory Systemic Inquiry.* CDI Practice Paper (8).

Burns, D. (2014b). *Systemic action research: Changing system dynamics to support sustainable change.* Action Research, 12(1).

Chambers, R. (2015). *Inclusive rigour for complexity.* Journal of Development Effectiveness, 7(3).

Copestake, J. (2013). *Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches* (Draft 18 Oct 2013). Centre for Development Studies, University of Bath.

Deprez, S., Huyghe, C., & Van Gool Maldonado, C. (2012). *The use of Sensemaker for measuring, learning and communicating about smallholder farmer inclusion* (Case Report). VECO.

Eyben, R. (2008). *Power, Mutual Accountability and Responsibility in the Practice of International Aid A Relational Approach* (Working Paper No. 305). IDS.

Gaventa, J., & Cornwall, A. (2006). *Challenging the Boundaries of the Possible: Participation, Knowledge and Power.* IDS Bulletin, 37(6).

Guijt, I., & Roche, C. (2014). *Does Impact Evaluation in Development Matter? Well, It Depends What It's For!* European Journal of Development Research, 26(1).

Humphrey, J. (2014). *Market systems approaches. A Literature Review.* (Beam Exchange). IDS.

IFAD. (2011). *IFAD Strategic Framework 2011-2015. Enabling poor rural people to improve their food security and nutrition, raise their incomes and strengthen their resilience.* IFAD.

IFAD, & BMGF. (2013a). *Improved Learning Initiative for the design of a Participatory Impact Assessment & Learning Approach (PIALA): Insights and lessons learned from the reflections on the PIALA piloting in Vietnam*. IFAD.

IFAD, & BMGF. (2013b). *PIALA Research Strategy.* Improved Learning Initiative (Internal Document). IFAD.

IFAD, & BMGF. (2014). *Impact Assessment of the "Developing Business with the Rural Poor" Programme in Bến Tre, Viet Nam (2008- 2013) in Ben Tre, Vietnam. Report of the Pilot Application of a Participatory Impact Assessment & Learning Approach (PIALA)*. IFAD.

IFAD, & BMGF. (2015). *Improved Learning Initiative for the design of a Participatory Impact Assessment & Learning Approach (PIALA): Methodological reflections following the second PIALA pilot in Ghana*. IFAD.

Jacobs, A., Barnett, C., & Ponsford, R. (2010). *Three Approaches to Monitoring: Feedback Systems, Participatory Monitoring and Evaluation and Logical Frameworks.* IDS Bulletin, 41(6).

Jenal, M. (2014). *Monitoring in complex systems: from spotlight and indicators to sensor networks and landscapes.* Retrieved from https://marcusjenal.wordpress.com/2014/04/16/monitoring-in-complex-systems-from-spotlight-to-sensor-networks/

MOFA/GOG, IFAD, & BMGF. (2015). *Final Report on the participatory impact evaluation of the Root & Tuber Improvement & Marketing Program (RTIMP). Pilot Application of a Participatory Impact Assessment & Learning Approach (PIALA)*. IFAD.

Mohan, G., & Hickey, S. (2004). *Relocating participation within a radical politics of development: critical modernism and citizenship.* In: *Participation : from Tyranny to Transformation?: Exploring New Approaches to Participation in Development*. Zed Books.

Mosse, D. (2001). *"People's knowledge", participation and patronage: Operations and representations in rural development.* In: *Participation: the New Tyranny?* Zed Books.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. SAGE.

Punton, M., & Welle, K. (2015). *Straws-in-the-wind, Hoops and Smoking Guns: What can Process Tracing Offer to Impact Evaluation?* CDI Practice Paper (10).

Rogers, P. (2009). *Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation.* The Journal of Development Effectiveness, 1(3).

Snowden, D. (2002). *Complex acts of knowing: paradox and descriptive self-awareness.* Journal of Knowledge Management, 6(2).

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations.* (Working Paper No. 38). DFID.

Van Hemelrijck, A. (2013). *Powerful Beyond Measure? Measuring complex systemic change in collaborative settings.* In: *Sustainable Participation and Culture in Communication: Theory and Praxis.* Intellect Ltd.

Van Hemelrijck, A. (2015). *Methodological strengths and limitations of the SenseMaker® pilot in AMD*. IFAD.

Van Hemelrijck, A., Dirirsa, B., & Ling, A. (2011). *Joint MEL framework for Strengthening the Ethiopian Agricultural Extension System (SEAES). Framework paper approved for funding by the BMGF*. (p. 33). Oxfam America, Sasakawa Africa Association and the Ministry of Agriculture in collaboration with ALINe and Keystone Accountability.

White, H. (2009). *Some Reflections On Current Debates In Impact Evaluation* (Working paper No. 2009-1). International Initiative for Impact Evaluation (3ie).

White, H. (2014). *Current Challenges in Impact Evaluation.* European Journal of Development Research, 26.

Woolcock, M. (2013). *Using case studies to explore the external validity of "complex" development interventions.* Evaluation, 19(3).

Annex I. PIALA methods & processes (purposes, costs and participants)

Methods & Processes	Purposes	Costs & Participants
Focusing and framing the evaluation		
1. Outlining of **design options** and budget implications (*full scale–full scope*; *full scale–limited scope*; or *limited scale–full scope*)	• Enable clients and commissioners to decide on scope and scale of evaluation	***Participants:*** • Evaluation clients and commissioners (8 participants) • Stakeholders who were involved in the design, management and implementation of the programme (32 participants) • Local research team (5 senior researchers) ***Total cost:*** USD 18,000 (incl. training & design workshop)
2. Reconstruction and visualisation of programme **Theory of Change (ToC)**	• Identify causal claims and assumptions • Formulate evaluation questions • Create shared understanding among stakeholders of programme theory and broader influences	
Collecting and linking data		
3. **Multi-stage sampling** with 'open systems' as principle sample unit (e.g. value chain systems) 4. Selection of methods for data collection, and drafting of 'how-to' guidance and templates for each method and for quality monitoring	• Enable systemic inquiry and comparative analysis of impact on livelihoods and household poverty • Enable rigorous use of methods and facilitation of processes • Enable systematic data quality monitoring and reflective practice	***Participants:*** • Local research team and consultants (17) • Randomly sampled households (837 households) • Quasi-randomly selected beneficiaries (439 in the generic change analysis with 51 % women; 400 in the livelihood analysis with 47 % women; 341 in the CF with 53 % women) • Officials and service providers (75 district-level and 25 regional/national participants) ***Total cost:*** USD 152,000 (incl. training)
5. Data collection on changes and causes in household food and income through • **household survey** • **generic change analysis** in gender-specific groups *(social mapping, timeline, wealth & wellbeing ranking, causal flow mapping)*	• Collect and triangulate data on impacts • With intended beneficiaries visually reconstruct and discuss causal flow of changes in livelihoods affecting household wealth and wellbeing	
6. Data collection on livelihood changes and causes through • **generic change analysis** (see above) • **livelihood analysis** in gender-specific groups *(livelihood change matrix, causal flow mapping, SenseMaker)*	• Collect and triangulate data on effects of livelihood changes on household food and income • Visualise and discuss with intended beneficiaries causal flow of changes and causes in different areas affecting their livelihood	
7. Data collection on reach and effects of selected programme mechanisms through • **livelihood analysis** (see above) • **Constituent Feedback (CF)** in mixed groups *(questionnaire for discussion & anonymous scoring)* • **semi-structured interviews** with service providers and officials *(CF-linked questionnaire)*	• Collect and triangulate data on effects of programme mechanisms on changes and causes in various areas affecting livelihoods • With intended beneficiaries discuss and anonymously score reach, benefits, outcomes of mechanisms	
8. **Data linking & quality monitoring** using a standard data collation tool and questionnaire for	• Enable instant data processing and cross-checking to identify gaps and weaknesses	

team reflections on quality of methods, processes and evidence using a standard questionnaire	• Ensure robust evidence (inclusive, sufficient consistent, rigorous)	
Synthesizing evidence and analysing and debating programme contributions		
9. Local and national **participatory sensemaking** using a workshop model consisting of design principles and methods for enabling voice and facilitating cross validation & contribution scoring 10. **Configurational analysis** using standardized data collation and scoring tools	• Probe to fill remaining data gaps • Enable stakeholders to understand impact systemically • Engage stakeholders in valuing programme contributions and identifying priority investment areas • Give voice to beneficiaries and create space for dialogue with decision makers • Arrive at rigorous causal inference and conclusions about programme contributions	***Participants:*** • Local stakeholders who participated in the research (640 in 23 district sensemaking workshops, of which 80 % beneficiaries and 48 % women) • Local and national stakeholders (106 in national sensemaking workshop, of which 40 % beneficiaries, 45 % officials and 15 % private sector) ***Total cost:*** USD 64,000 (incl. local and national workshops)

Annex II. Revised PIALA Quality Assurance Framework (QAF)

MAIN EVALUATION PHASES	RIGOUR (methodological consistency and responsiveness)	INCLUSIVENESS (credibility and meaningful participation)	FEASIBILITY (budget and capacity for creating added-value)

Focusing and framing the evaluation
How and to what extent did PIALA help adequately identify the causal claims that the evaluation should focus on, and determine its purposes, questions and standards in an inclusive and rigorous manner?
	• *Has the focus of the impact assessment been clearly defined against the programme/project's Theory of Change?*
	• *Is the Theory of Change (ToC) well articulated and visualised?*
	  ○ *Does it clearly show the presumed change pathways towards impact, and the causal links in these pathways?*
	  ○ *Does it adequately reveal the change mechanisms and the project/programme's influences on these that need to be evaluated?*
	  ○ *Does it make the assumptions to be evaluated with regard to the change pathways and the project/programme's influences explicit?*
	  ○ *Does it sufficiently reflect other influences and contextual factors interacting with the project/programme and affecting rural poverty?*
	• *Have the purposes and questions for conducting the evaluation been clearly defined?*
	• *Have standards for valuing project/programme contributions been clearly defined?*
		• *Does the theory of change sufficiently reflect the different views of those who were involved in the design of the project/programme?*
		• *Does the theory of change give room for probing different perspectives of change towards rural poverty impact?*
		• *Have key stakeholders who were involved in the design endorsed the focus, purposes, questions and standards of the evaluation?*
		• *Do other participants sufficiently understand the purposes and questions in order to meaningfully engage in the evaluation?*
			• *Were budget and capacity sufficient to employ PIALA's ToC approach for focusing and framing the evaluation in a rigorous and collaborative way?*
			• *What were the benefits of using this ToC approach?*
			• *What minimum conditions need to be in place to make this approach cost-effective?*

Collecting and linking the data
How and to what extent does PIALA help rigorously collect, triangulate and link the data needed to produce the evidence for confidently answering the evaluation questions in a
	• *Has sample units and populations been adequately defined in a way that permits rigorous causal inference?*
	  ○ *Have contextual variability as well as heterogeneity and time lags in programme/project treatment sufficiently been taken into account in the sampling structure?*
	  ○ *Have possible **biases** in sampling been adequately mitigated?*
	• *Have appropriate methods been selected to collect sufficient data on the causal links in the ToC that the evaluation is supposed to focus on?*
	  ○ *Have possible biases on both researchers' and participants' side been adequately mitigated by the selected methods and*
		• *Have methods and processes been sufficiently field-tested to ensure their appropriateness for the research participants?*
		• *Have methods and processes enabled key stakeholders (particularly intended beneficiaries, both women and men) to equally and meaningfully engage in the research?*
		• *Have methods and processes been employed in a gender/power-sensitive way*
			• *Were budget and capacity sufficient to employ PIALA's sampling and participatory mixed-methods approach?*
			• *What were the benefits of using this sampling and participatory mixed-methods approach?*
			• *What minimum conditions need to be in place to*

participatory manner?	*processes? (e.g. courtesy, availability and perception biases, or the types of biases that are due to the limitations of people's knowledge, perceptions and memories)* ○ *Have methods been employed in a way that produces reliable data and permits systematic crosschecking and elimination of alternative causes or explanations?* ○ *Have data been consistently triangulated, collated and linked in a way that permits a reliable reconstruction of the actual change pathway and a valid judgement of relative strength and consistency of its causal links?* ○ *Have methods been employed consistently across the sample in a way that permits rigorous causal inference?* • *Did the researchers have sufficient autonomy in selecting the participants, facilitating the group discussions, conducting the surveys, and processing the data?* • *Did the researchers have full access to all participants and secondary sources?*	*that is open to participants' feedback?* • *Have data been collected in an ethical manner, respectful of participants' right to prior consent, confidentiality and protection?* • *Have data been collated and linked in a transparent way that is respectful of differences in views and perspectives?*	*make this sampling and mixed-methods approach approach cost-effective?*
Synthesising the evidence and analysing and debating programme contributions *How and to what extent does PIALA help synthesize and present the evidence to facilitate rigorous analysis and solid debate of programme contributions in a participatory manner?*	• *Have the evidence strands been consistently synthesized and presented across the sample in a way that permits rigorous causal analysis and assessment of project/programme contributions in various configurations?* • *Have different configurations of project/programme's influences or contributions to rural poverty impact been valued in a reliable manner along the agreed standards?*	• *Have the evidence strands been adequately synthesised and presented in a way that permits meaningful engagement of stakeholders in sensemaking and contribution analysis?* • *Have the evidence strands been presented in a way that is sensitive enough to culture and context in order to facilitate meaningful dialogue?* • *Have the feedback from those who were supposed to benefit from the project/programme been adequately taken into account in the valuing of programme contributions?* • *Have beneficiaries, implementers, service-providers and other actors been given the opportunity to engage in a meaningful dialogue around programme contributions based on evidence?* • *Has the process of collectively analysing*	• *Were budget and capacity sufficient to employ PIALA's aggregated configuration analysis method and participatory sensemaking model?* • *What were the benefits of using this particular configuration analysis method and sensemaking model?* • *What minimum conditions need to be in place to make the method and model cost-effective?*

		the evidence and valuing project/programme contributions helped stakeholders obtain a better understanding of the systemic interactions impacting rural poverty?	
Assessing the evaluation outcomes *To what extent did PIALA help produce rigorous, contested and debated qualitative and quantitative evidence sufficient to withstand scrutiny among all key stakeholders and answer the evaluation questions?*	• *Did the evaluation produce sufficient evidence of rural poverty impact and project/programme contributions to such impact in order to answer the evaluation questions?* • *Was the evidence rigorous enough to withstand scrutiny?* • *Have useful insights about the likely sustainability of the impacts been generated based on a thorough analysis of the systemic interactions between the project/programme and other influences impacting rural poverty?* • *Have issues requiring more effort, innovative thinking, evaluative input and policy change been validly identified?*	• *Has the evidence produced been sufficiently contested and debated with key stakeholders to be found valid and credible?* • *Have stakeholders learned from this evaluation about the issues requiring greater effort and innovation?* • *Did stakeholders found their participation in the evaluation meaningful and worth their time?* • *Were stakeholders satisfied with how the evaluation was conducted and what has come out of it?*	• *Is the combination of scientific and participatory methods and processes doable and replicable?* • *Were budget and capacity sufficient to meet the expectations with regard to rigor and inclusiveness, and serve the intended evaluation purposes?* • *What were the benefits of using PIALA as compared to other approaches?* • *What minimum conditions need to be in place to make PIALA cost-effective?*