Chapter 19

Rising to the Challenges of Impact Evaluation – Insights from Piloting a Systemic and Participatory Approach

Adinda Van Hemelrijck

Abstract. *This chapter reflects on the use and value of a systemic theory-based and participatory mixed-methods approach for addressing the challenges of impact evaluation in complex development contexts. A Participatory Impact Assessment and Learning Approach (PIALA) was developed and piloted with the International Fund for Agricultural Development in Vietnam in 2013, and in Ghana in 2015, that engages partners and stakeholders in assessing, explaining, and debating systemic impacts on rural poverty. An action research was conducted around the pilots to learn about the processes and mechanisms that make impact evaluations using PIALA rigorous and inclusive but also feasible. The study concluded that inclusiveness and rigor can reinforce each other, even more so at scale, with sufficient capacity. Methodological complementarity and consistency, extensive and robust triangulation, and cross-validation are important attributes. Investing in research capacity may help to reduce costs over time, while enhancing the value of impact evaluation and the uptake of its findings.*

Adinda Van Hemelrijck, Independent Consultant, adindavanh@gmail.com.

Development today takes place in globalized contexts of growing inequality, uncertainty, and instability, with new rising powers and an infinite number of conflicting issues and interests. The 2030 Agenda for Sustainable Development calls for fundamental systemic changes, and adds demands for inclusiveness and sustainability to those of effectiveness, in order to eradicate poverty and inequality and protect our planet. Interventions, consequently, are becoming ever more complex, with stakes and stakeholders getting more diverse, influences more dense, problems more systemic, and outcomes more unpredictable. This complexity challenges the field of impact evaluation.

Traditional counterfactual-based approaches are generally found to be too costly and difficult to pursue in complex environments, due to high causal density, spillover, time lags, and the unpredictability of events (Befani et al. 2014; Picciotto 2014). They focus too narrowly on specific intervention components, thus "leaving many evaluation questions unanswered" (White 2014, 3). They also do not explain impact or assess its sustainability, given their focus on specific and isolated cause-effect relationships: therefore they cannot tell if, how, or why similar relations would or would not work elsewhere (Picciotto 2014; Ravallion 2012; Woolcock 2013). Finally, engagement of and learning with partners and stakeholders is inhibited by scientific procedures, raising questions about inclusiveness and democratic value (Van Hemelrijck 2013a, 2017a).

Alternative theory-based and complex systems approaches, on the other hand, tend to be time-intensive and to produce evidence that is not comparable across many cases;[1] therefore, they are not suitable for evaluations with larger populations (a larger *n*) that require estimates of impact distribution (Beach and Pedersen 2013). In addition, those studies that allow for participation generally do not set out to rigorously assess causality and to address concerns of bias and rigor (Copestake 2014; White and Phillips 2012). Chambers calls this "a strange omission, perhaps even a blind spot," and refers to the Participatory Impact Assessment and Learning Approach (PIALA) in this respect as "part of what should be a wave of the future" (Chambers 2017, 108).

PIALA was developed with the International Fund for Agricultural Development (IFAD) between 2012 and 2015 in an attempt to address these challenges. IFAD is a United Nations (UN) agency that provides loans and support to governments for agricultural and rural development programs that aim at reducing rural poverty by changing smallholder production and market systems (IFAD 2016). These are generally medium to large-scale programs that aspire to create sustainable systemic or transformative change, and are implemented by public and private partners in often quite complex political environments. The PIALA initiative sought not to reinvent the wheel, but to develop a model that creatively combines existing designs and methodologies (both quantitative and qualitative) in novel ways to rigorously assess such complex programs, and to bring participation in impact evaluation

[1] This is mostly because the cases themselves are not comparable.

to life (Guijt et al. 2013). Inspiration was drawn mostly from the theory-based (in particular, realist) and transformative (including rights-based) traditions (Holland 2013; Mertens 2009; Pawson 2013; Van Hemelrijck 2013a).

This chapter first describes what PIALA is and briefly presents the two IFAD pilots. It then discusses the main insights from the action research that was conducted around the pilots on how an impact evaluation using PIALA can be rigorous and inclusive. The chapter concludes with some reflections on the value-for-money of the approach, how rigor and inclusiveness may reinforce each other and generate greater value, and the key attributes and conditions for achieving this.

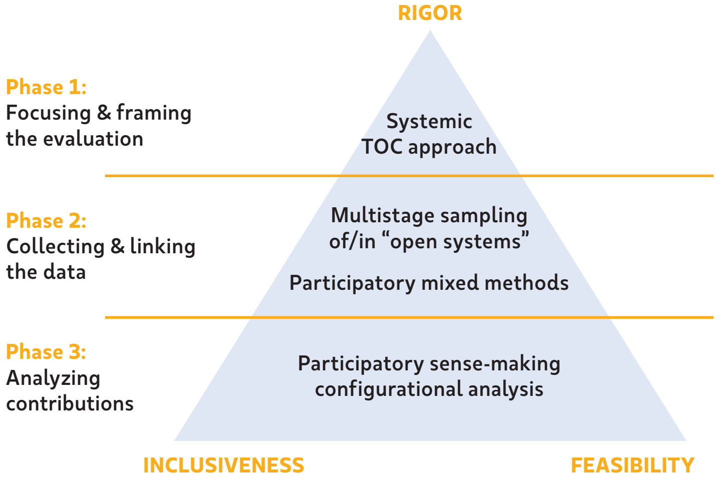
THE PARTICIPATORY IMPACT ASSESSMENT AND LEARNING APPROACH

PIALA is a theory-based, mixed-methods approach that is essentially participatory. It aims to enable stakeholders to see and learn about impact collectively and systemically, in order to bring about transformative change. It is most suitable for assessing the impact of medium to large-scale projects or programs that are targeting relatively large populations, in contexts where a conventional counterfactual approach is insufficient, difficult, or impossible to pursue. PIALA is not a specific research or evaluation methodology, but an approach that can embed any method and allows for a creative "mixed design" (Stern 2015) combining different evaluation traditions and methodologies, as long as its two overarching design principles—*evaluating systemically* and *enabling meaningful participation*—are maintained (Van Hemelrijck 2016b). These two principles allow for a flexible design, and make it possible for evaluators to adapt PIALA's five methodological elements to the specific evaluation context and purposes. The five elements follow:

- A **systemic theory of change (TOC)** for visualizing the project's causal claims, and engaging stakeholders in framing the evaluation and debating the evidence
- **Multistage sampling of/in "open systems"** for enabling systemic inquiry across medium to large-size populations
- Standardized **participatory mixed methods** for collecting, linking, and cross-checking the data in all sampled systems, in a systematic and comparable way
- A two-stage **participatory sense-making** model for engaging stakeholders at local and aggregated levels in debating the emerging evidence
- A **configurational analysis** method for assessing systemic change patterns and drawing conclusions about the distribution and magnitude of their impact across medium to large samples

As shown in figure 19.1, these five elements are designed and put to use in three consecutive phases: framing and focusing the evaluation; collecting and linking the data; and analyzing and debating contributions. To further uphold the desired quality in the design and conduct of an evaluation for

FIGURE 19.1 **PIALA elements and standards**


RIGOR
Phase 1:
Focusing & framing the evaluation
Systemic TOC approach
Phase 2:
Collecting & linking the data
Multistage sampling of/in "open systems"
Participatory mixed methods
Phase 3:
Analyzing contributions
Participatory sense-making configurational analysis
INCLUSIVENESS
FEASIBILITY

achieving optimal value within the constraints of available resources, PIALA combines standards of *rigor*, *inclusiveness*, and *feasibility*.

Methodological Elements

The **systemic TOC approach** forms the backbone for the entire evaluation. It is designed in the first phase of the evaluation process, by means of which the evaluation is focused and framed. It involves a process of reconstructing and visualizing a program's impact pathways and change hypotheses, and the broader trends and influences, based on a thorough desk review and discussions with key stakeholders. Unlike a classic program/project theory,[2] this approach uses an evaluative lens, assessing the hypotheses by looking backward, from the envisioned impact back to the interactions and mechanisms that presumably have caused or influenced the impact (Funnell and Rogers 2011; van Es, Guijt, and Vogel 2015). Moreover, it views impact from a systemic perspective, resulting from changes in systems of interactions, rather than the direct and isolated relationship between intervention and effect. A systemic TOC approach is most useful for evaluating the changes caused by many different interventions, implementers, contributors, and funders, because it helps to create a shared understanding of complex pathways, and enables different stakeholders to critically engage in parts of the analysis (Van Hemelrijck 2013a).

[2] A program/project theory is constructed from a management perspective, and is focused on strategy and performance looking forward, toward the delivery of planned results.

Multistage cluster sampling of/in open systems happens right after the evaluation focus and framing is agreed upon with the stakeholders, as part of the design for the second phase. Its purpose is to ensure sufficient representation of the various populations, in order to enable the comparison and generalization of findings about systemic impact at the medium-to-large scale. If we want to learn about systemic impact, then the system should be the main level of analysis, and thus also the main sample unit (Lain 2016). In the case of government policies and programs, the system is generally too monolithic for a classic counterfactual comparison. However, by focusing on the lowest embedded open system at the local level (e.g., the local supply-chain system) that is entrenched in and affected by the larger system (e.g., the larger commodity value chain, and national policy framework), it is often possible to have a sample that is large enough to cover systemic heterogeneity, and to have large enough subsamples for statistical comparison. Contrasting evidence can then be obtained from the areas where program mechanisms are found absent, dysfunctional, or ineffective, rather than from predetermined control areas that are sampled external to the program (Van Hemelrijck 2017b). Multistage cluster sampling of these local systems, and of populations within these systems, is the most cost-effective method, as it substantially reduces costs and logistics compared to other random sampling strategies (Levy and Barahona 2002).[3]

The appropriate **selection and mixing of methods** to collect qualitative and quantitative data on the different causal claims in the TOC is also part of the design of the second phase. The IFAD pilot in Ghana combined conventional household surveys for statistical poverty analysis; participatory methods for generic change analysis, livelihood analysis, and constituent feedback; SenseMaker for quantitative pattern analysis of perceptions; and key informant semistructured interviews for inquiring the larger system.[4] Methods are selected specific to the causal links in the TOC, and are used on an equal basis. They complement and build on each other analytically to enable the construction of the actual causal paths with the data for each locality or sampled

[3] Random sampling is needed for statistical analysis. This depends on the evaluation focus. In an impact evaluation for Oxfam GB in Myanmar, for instance, PIALA's sampling protocol was adapted to fit the specific evaluation focus and requirements, which did not require statistical analysis and thus also not a random sampling (Van Hemelrijck 2017a).

[4] Constituent Feedback (also called Constituent Voice) is a methodology developed by Keystone Accountability (http://www.keystoneaccountability.org) for collecting quantified feedback and engaging in dialogue with key constituents or beneficiaries, using standardized metrics similar to the customer satisfaction surveys developed in the private sector, and descriptive statistics to produce visual data reports. SenseMaker is a software-based methodology developed by Cognitive Edge (http://www.sensemaker-suite.com) that facilitates mass ethnography and provides a way of nearly real-time mapping of social interactions and individual perceptions and motivations to inform adaptive management and policy formulation. It collects large amounts of self-signified micro-stories that capture people's experiences and perceptions of past and future change in ways that enable us to identify emerging patterns of actions and decisions. The software permits statistical analysis at a very large scale.

system, mirroring the envisioned paths in the TOC. They also partly overlap, to permit triangulation. Systematic data collation and quality monitoring makes it possible to cross-check and link the data for building the causal paths during fieldwork in every locality, and to timely identify data gaps or weaknesses that need further inquiry before moving to the next locality. To enable comparison across the sample of systems as the basis for aggregating findings, the methods are more or less standardized. Yet they also remain sufficiently open-ended by including sensing tools such as causal flow mapping that can capture unintended effects and influences, and uncover broader dynamics that are interacting with the program (Van Hemelrijck 2015).

Participatory sense-making occurs in the third phase of the process: analyzing and debating contributions. It involves half-day local workshops with 30–50 participants (of whom 60–70 percent are intended beneficiaries) during the fieldwork in each locality, and a one or two-day program-level workshop with 100–130 participants (of whom over 30 percent are intended beneficiaries), shortly after finishing the fieldwork and before turning to the final analysis and reporting. The stakeholders participating in the workshops typically include decision makers, service providers, and intended beneficiaries. They proportionally represent all the different perspectives necessary to cross-validate the evidence and inform the final analysis. They discuss the evidence together and assign value to observed contributions (among other influences) by comparing the actual causal paths revealed by the data with those hypothesized in the TOC. Participatory sense-making in all researched localities and at the aggregated level serves to not only cross-check and strengthen the evidence, but also to create ownership, enable equal voice, and stimulate systemic learning. In essence, it makes an evaluation more democratic (Van Hemelrijck 2016b).

Finally, the **configurational analysis** compares systemic change and impact across the sample of systems to reach conclusions about the distribution and magnitude of impact. Its tools are designed and put to use in the third phase of the evaluation process, after the sense-making. It employs elements of process tracing, contribution scoring, and cross tabulation, and involves four major steps. The first is the aggregated data collation in a standard Excel matrix format, in which all evidence from the field collation matrices as well as secondary sources is synthesized and tabulated alongside the TOC. The next step involves the clustering of the evidence across all the sampled systems to surface patterns or configurations of systemic changes and causal attributes. The third step involves the comparative analysis of similarities and differences in configurations for the specific mechanisms or parts of the system of interest (including cases with and cases without functioning mechanisms).[5] The final step involves integration of the findings for

[5] Software such as EvalC3 can be applied to assess the conjuncture of different mechanisms and causal processes. This novel software, developed by Rick Davies, was piloted in an impact evaluation using PIALA for Oxfam GB in Myanmar (Van Hemelrijck 2017a). The software helped to identify sets of causal attributes that are necessary and/or sufficient for specific sets of outcome attributes to occur, and to compare and evaluate the performance of these causal models to find those with the greatest predictive power.

the different parts and mechanisms as the basis for validating (or refuting) the hypotheses in the TOC; zipping up the findings alongside the TOC; and drawing conclusions about the distribution and magnitude of the program's contributions to impact (Van Hemelrijck 2016a, 2016b, 2017b).

Quality Standards

To enable an evaluation to achieve optimal value with limited budgets while remaining true to the two core principles of systemic thinking and meaningful engagement, PIALA also features a quality framework that includes standards of *rigor*, *inclusiveness,* and *feasibility*.

Rigor is defined as the quality of thought put into the methodological design and conduct of the evaluation in a way that enables robust triangulation of different methods and perspectives in order to defeat bias or dominance of a single truth; and to ensure both consistency and responsiveness to local contexts and conditions (Van Hemelrijck 2016b). Whereas mainstream evaluation practice defines rigor as the controlled avoidance of bias through statistical procedure, PIALA builds on the premise that bias cannot be avoided by a single method or procedure, but can be mitigated through systematic triangulation of different methods and perspectives (Camfield, Duvendack, and Palmer-Jones 2014; Carugi 2016; Mertens 2010).

Inclusiveness refers to the legitimacy of the ways in which people are engaged in the evaluation, and to the level of impartiality or inclusion of all stakeholder views and perspectives. This has intrinsic empowerment value but also contributes to the robustness and credibility of the evidence and thus to the validity of the findings (Chambers 2015; Pawson 2013). Validity is understood as the extent to which findings are well founded, based on robust evidence, and correspond with the reality of all stakeholders, in particular the populations affected by the project or program being evaluated. By embracing a wide range of stakeholder perspectives and ensuring their equal weight in examining the evaluation questions, the evaluation builds a more accurate systemic picture of impact. Meaningful engagement in constructing, analyzing, and debating this picture, on the other hand, enables equal voice, and contributes to empowerment (Chambers 2017).

Feasibility concerns the budget and the capacity needed to meet the expectations of rigor and inclusiveness, and to enhance stakeholders' systemic and collaborative learning (Van Hemelrijck 2016b). The investment in building the capacity of in-country researchers, and in experimentation with novel mixed designs that stretch the limits of conventional evaluation practice, is critical for doing this at a larger scale. Considering feasibility as an explicit and intent-driven (rather than constraint-driven) quality helps to think of this investment in a positive way. Much remains to be learned about how to do it well. Excessive focus on limiting costs starves the evaluation of the oxygen it needs in order to deliver on rigor and inclusiveness and to maximize its value (Chambers 2017).

TWO IFAD EVALUATIONS THAT PILOTED PIALA

PIALA was piloted in the evaluation of two IFAD-funded programs: the Doing Business with the Rural Poor (DBRP) program conducted in one province in southern Vietnam,[6] and the Roots and Tubers Improvement and Marketing Program (RTIMP), which was conducted countrywide in Ghana.[7] Both were aimed at improving livelihoods and increasing food and income security by enhancing smallholders' capacities to commercialize, and by linking local businesses to markets and industries. DBRP focused on developing diversified short-value chain systems; RTIMP was concerned with developing much longer commodity chains, linked to national and export markets and industries (Guijt et al. 2014; Van Hemelrijck and Kyei-Mensah 2015). Both programs essentially sought to create the mechanisms needed to facilitate rural peoples' access to services, resources, and markets.

Despite some important differences in the context and quality of the two evaluations,[8] both produced quite convincing evidence of program contributions to the improvement of livelihoods as a result of the increased access to services, resources, and markets generated through these mechanisms. The RTIMP evaluation in Ghana, for instance, showed significant improvements in roots and tubers-based livelihoods, with 15 percent of households increasing their incomes above $2 a day. Very weak or no improvement was found in supply chain areas where the RTIMP mechanisms were dysfunctional or absent. Although positive, the evidence also showed that these improvements were rather limited, fragile, and susceptible to climate and market shocks, particularly for poor and vulnerable households, and in remote and marginalized areas. The improvements in roots and tubers-based livelihoods in Ghana occurred merely between 2009 and 2013, and in about 52 percent of the supply chain areas, or about half of the country. Moreover, no households gained profits above $4/day from roots and tubers, even though 61 percent

[6] The DBRP was implemented from 2008 to 2014 in two provinces (Cao Bang and Ben Tre), with a total budget of $51 million, including a $36 million loan from IFAD. The evaluation Guijt et al. 2014) was conducted in 2013 at a cost of $90,000 in Ben Tre province only, where the project was implemented in 50 of 164 communes in eight of nine districts.

[7] The RTIMP was implemented from 2007 to 2015 as a national program in 106 of 216 districts, spread across all 10 regions countrywide, with a total budget of $24 million, of which $19 million was financed under an IFAD loan. The evaluation (Van Hemelrijck and Kyei-Mensah 2015) was conducted countrywide after project completion in 2015, at a cost of $233,000, and covered the post-midterm review period from 2010.

[8] Although the evidence suggested strong connections between all observed changes, confidence in causal inference remained relatively weak in the Vietnam evaluation. In this first PIALA pilot, data collation, cross-checking, and quality monitoring was not yet done systematically with the TOC as a backbone structure. Confidence in inference and generalizability was much stronger in the second pilot in Ghana because of its systematic and multilayered triangulation and cross-validation procedure (Van Hemelrijck 2015). This is further discussed in the next section of this chapter.

of the households had invested in roots and tubers businesses. Access to new seeds and farming technologies had initially led to a boom in production across the country, triggering a spillover into processing. Adoption of new processing technologies, though, remained limited in 83 percent of the cases, partly due to limited investment capital. By and large, the finance mechanism put into place by the program proved inaccessible, as it required pre-investment without short-term capital return, posing high risks for smallholders (Van Hemelrijck and Kyei-Mensah 2015). Both in Ghana and in Vietnam, poor and vulnerable households ran considerable risks by engaging in value chains and accessing markets (Guijt et al. 2014; Van Hemelrijck and Kyei-Mensah 2015). These risks were left largely unmitigated due to inadequate market linking and forecast that otherwise could have helped avoiding the observed local market saturation and monopolization; and inadequate poverty targeting that should have made the support services and mechanisms more inclusive and sustainable (Van Hemelrijck 2016a). Recommendations for how to address these issues in similar IFAD-funded programs and projects were made by these two evaluations.

KEY INSIGHTS FROM THE PILOTING

As mentioned earlier, the PIALA initiative was conceived as an action research to inquire into the conditions, processes, and decisions affecting the rigor and inclusiveness of the two pilots. The action research combined multisited ethnography with cooperative inquiry, and involved extensive reflections with researchers and participants in the two pilot countries, as well as feedback sessions with global experts at IFAD headquarters (Van Hemelrijck 2016b). Insights from the first pilot (in Vietnam) helped to better address the challenges in the second pilot (in Ghana) (Van Hemelrijck 2015). This section summarizes some of the key lessons learned.

Creating Ownership of the Evaluation

In order to create ownership, key stakeholders need to be sufficiently engaged in the framing and focusing of the evaluation. Ownership implies that the evaluation is wanted, legitimized, and enabled by a shared sense of responsibility for its success. Ownership also enables participation in the analysis, and facilitates learning and greater uptake of evaluation findings and recommendations (Burns and Worsley 2015; Patton 2011). In the case of PIALA, stakeholders are engaged in the framing and focusing of the evaluation through a process of reconstructing and visualizing the TOC (Van Hemelrijck 2015).

In Vietnam, insufficient time and budget was spent on this process, which affected the rigor and inclusiveness of the approach during the entire evaluation. A brief workshop was organized with the program steering committee and managers to discuss program logic and expectations. The process of reconstructing and visualizing the TOC, however, happened *after* the workshop, and independently of the evaluation design. Evaluation questions did not focus on the causal links and assumptions in the TOC, which made it difficult for the researchers to relate the evidence back to the TOC and arrive at greater precision in causal analysis. Furthermore, limited ownership of the TOC by the stakeholders hindered their critical engagement in sense-making and contribution analysis (Van Hemelrijck 2013b).

Learning from the Vietnam pilot, the TOC process was made a priority and a key deliverable in Ghana. The researchers organized a design workshop to discuss the TOC and design options, and to determine the focus of the evaluation together with key stakeholders. The investment in a more robust and collaborative TOC process bore fruit and laid the foundation for attaining greater quality throughout the entire evaluation, resulting in stronger evidence and ownership of findings (Van Hemelrijck 2015, 2016b).

Deciding on the Scope and Scale of the Evaluation

Scale refers to the size of the sample of the primary sample unit. In the case of PIALA, this was the lowest embedded "open system" that the program sought to change, to generate impact. *Scope* refers to the various components and mechanisms of the system that the evaluation should cover. Generally speaking, the larger the scale, the more relevant the findings for reporting and advocacy will be. Using participatory mixed methods at scale, however, is challenging, and requires sufficient capacity and resources. When research capacity is weak, more resources are needed for training, coaching, and supervision in order to uphold quality (Van Hemelrijck 2015).

Three relevant design options are available for designing an impact evaluation: full scope–limited scale, limited scope–full scale, and full scope–full scale.[9] When choosing a **full scope–limited scale design**, the emphasis is on learning about the project's total contribution to impact in select cases, under specific conditions. Fieldwork and analysis are less resource-intensive, given the relatively small sample sizes. Yet evaluation findings will not be generalizable for the entire population: therefore they are less useful for influencing policy decisions.[10] With a **limited scope–full scale design**, the purpose is to assess the effects of one or two particular aspects or mechanisms of the project. The TOC is not strictly necessary in order to conduct such a narrow study, but skipping the TOC process may risk missing out on systemic understanding, leading to flawed conclusions. Components are studied in isolation, which does not permit analysis of systemic interactions. For example, a cost-effectiveness study of Farmer Field Forums (FFF) in Ghana recommended a scaling-up, as the adoption of new technologies had proven the success of this mechanism. The PIALA evaluation, however, showed that in a

[9] A limited scope–limited scale option is not really relevant for impact evaluation, as it limits the possibility of causal analysis through classic counterfactual comparison, frequency statistics, and/or triangulation and cross-validation of sources and methods.

[10] This does not hold true if the project/program itself is implemented at a limited scale (small *n*), in which case larger within-samples and more stringent triangulation and cross-validation procedures will take up the resources needed to attain the required level of rigor for generalization.

period of weak economic growth, this success in fact contributed to market saturation, which negatively affected livelihoods across the entire country (Van Hemelrijck 2015; Van Hemelrijck and Kyei-Mensah 2015).

In Vietnam, a choice was made for a full scope–limited scale design, but with disparate scales for the different methods. To save resources, participatory methods were employed only in a subsample drawn from the sample of villages where the statistical household surveys were conducted. The assumption was that this would be sufficient to conduct a full scope inquiry of contributions to impacts on rural poverty for the entire program area. However, it generated a disparity in the data sets that caused problems for their subsequent linking. While participatory data on causes and contributions came from only a few villages or cases, survey data on household impact were more widely distributed and not related to the specific cases or villages covered by the participatory methods. This hindered causal inference (Van Hemelrijck 2013b).

In Ghana, by contrast, a conscious choice was made to employ all methods in the same sample and at the same scale. The three design options were discussed with clients and commissioners before any procurement or design work was started, giving them a basic understanding of the cost and value of each.[11] As the future Ghana Agriculture Sector Investment Program (GASIP) was expected to scale up most of the RTIMP mechanisms, the evaluation was found necessary for both reporting and learning. The commissioners therefore chose the most comprehensive design: **full scope–full scale**. This implied six weeks of uninterrupted fieldwork—much longer and far more intensive than the pilot in Vietnam, where fieldwork took only two weeks. The budget was tighter than in Vietnam because of the larger scale and scope, but quality was upheld by a competent research team (Van Hemelrijck 2015; Van Hemelrijck and Kyei-Mensah 2015).

Deciding on the Counterfactual

Mainstream impact evaluation assumes that comparative analysis of evidence from both treated and nontreated locations is feasible and necessary in order to assess causality and reach generalizable conclusions about impact on rural household poverty. However, in most "real-world" evaluation contexts (Bamberger, Rugh, and Mabri 2012), it is very difficult and costly to arrive at an accurate assignment of locations to specific interventions and identify credible control groups. The challenge occurs, for instance, in cases of unexpected or uncontrolled project expansion and/or spillover, combined with high causal density of other interventions and influences. In such contexts, it is difficult to discern project from nonproject localities, and to find the right matches (Woolcock 2009). In addition, the open systems that form the principal sample unit in PIALA generally do not have clear boundaries such as villages or other administrative units have. Hence the identification and

[11] Including Ghana's Ministry of Food and Agriculture and the IFAD Country Office in Ghana.

matching of control units for these systems and subsampling of various populations from these systems, if even possible, requires fieldwork prior to the evaluation that substantially increases both costs and risks (Chambers 2017).

In the Vietnam pilot, comparative analysis of treated and nontreated units was considered both possible and necessary for assessing household-level impacts, and the village was thought to be the best proxy unit for investigating the short value chains developed by the program. These assumptions were flawed and compromised in terms of analytical rigor, making it difficult to generalize the findings, for three important reasons. First, without a clear definition and identification of the value-chain systems, and thus without having sampled proper proxies based on such a definition, it was difficult to relate the data on changes in capacities, institutions, and livelihoods to the specific value chains and to assess the causal links. Second, the matching of treated and nontreated villages was based on variables that applied to the village as a unit, not to the value chains, again making it difficult to relate the difference revealed by the data to the interventions. Finally, the high heterogeneity in program delivery and incoherence in its value-chain linking efforts, further conflated by the high causal density of other programs and influences in the villages, made it impossible within available budgets to obtain credible control data (Van Hemelrijck 2013b).

Learning from this experience, in Ghana much more work was done to understand and define the principle sample unit. In the evaluation design workshop, the decision was made not to waste resources on identifying and inquiring control groups of households, but instead invest all in the systemic inquiry of the four main commodity supply chains developed by the program (gari, plywood cassava flour, high-quality cassava flour, and fresh export yam). These commodity chains comprised medium to large amounts of supply chains spread over the entire country. The supply chains are loose catchment areas comprising clusters of communities of smallholders supplying the raw produce, and small enterprises or off-takers acting as “supply chain leaders” and manufacturing higher-value products for bigger markets. The supply chains were not entirely homogeneous, as they interacted and overlapped. Hence they often differed in reality from what was sampled on paper. Ensuring that the data collected on these systems remained comparable required much creativity and coordination. Furthermore, no reliable lists of households and beneficiaries were available for the subsampling of farmers, processors, and households within the catchment areas of the sampled supply chains. Identification and matching of control units and sampling of households thus would have required extensive pre-evaluation fieldwork, and the sponsors and other participants in the design workshop voted against this. Instead, a configurational analysis method, which uses heterogeneity in the sample of systems as the basis for counterfactual analysis, was developed. Supply chains with different systemic configurations of treatments and causal attributes were randomly sampled (with probability proportional to size) from the four commodities’ supply chain populations. The samples were large enough to include supply chains with dysfunctional or absent program mechanisms that could serve as a “natural” counterfactual (Van Hemelrijck and Kyei-Mensah 2015).

Maintaining Independence

In order to avoid positive bias, field mobilization of research participants is best undertaken independently from project management.[12] When research participants suspect that the research is not independent, they are more likely to over- or underreport. On the other hand, they are unlikely to trust outsiders who are not authorized and formally introduced by their leaders. Thus, for the researchers to organize fieldwork at scale and mobilize participants without any help from the program, they need to be good at logistics; know the areas and local customs; and be able to obtain authorization and introduction in ways that do not affect their independence. In contexts where this is not possible, strong facilitation skills are needed to minimize undue influence or interference (Van Hemelrijck 2015; Van Hemelrijck and Kyei-Mensah 2015).

The challenges encountered in Ghana were quite different from those in Vietnam. In both pilots, though, participants trusted the researchers' authorization and independence, which made them feel safe about expressing their views and critically engaging in the group inquiries. In Vietnam, field research cannot be conducted without government permission and interference. Hence, in the DBRP evaluation, local transportation and mobilization was organized by local officials and program staff, which was highly efficient, but challenging in terms of independence. Local leaders and program staff were quite collaborative during fieldwork but omnipresent. The researchers artfully managed to maintain sufficient distance, though, and to safeguard the privacy of the focus groups (Van Hemelrijck 2013b).

In Ghana, the researchers took care of the transportation and mobilization entirely by themselves but without prior notification or engagement of the local officials, allowing for much greater independence. Staff and officials were present only at the discussions to which they were invited. This, however, made them more suspicious of and resistant to the evaluation. Also, the scale of the fieldwork, the remoteness and spread of the communities, the large distances to travel over poor roads, and the difficulty of finding safe and trusted locations for convening people from different communities, made the field inquiries quite onerous. Independence thus came at a substantial effort and cost in Ghana, but compromise in both rigor and inclusiveness was avoided (Van Hemelrijck 2015; Van Hemelrijck and Kyei-Mensah 2015).

Contextualizing Poverty Analysis

To make it possible to say something about a program's influences on poverty, data on those influences, and on poverty, need to be linkable. Also, poverty has to be defined in ways that are relevant to the specific context and conditions of the villagers in the program area. In both of the IFAD pilots, a Participatory Rural Appraisal (PRA) ranking tool was used for identifying locally relevant

[12] According to the Organisation for Economic Co-operation and Development, "independence" implies an evaluation process that is transparent, independent from project management, and free from political influence or organizational pressure (OECD DAC 2010, 25).

characteristics of wealth and well-being, and assessing changes in relative poverty status (Van Hemelrijck 2015). This tool helped create a shared understanding among the participants of their wealth and well-being as the basis for a causal flow-mapping exercise of the changes they had experienced. It also enabled cross-checking and linking of the participatory data with the household survey data on poverty.

The characteristics of wealth and well-being that were obtained from the participatory ranking exercise, however, were not used to design the household survey. For this, the participatory data collection should have happened prior to the evaluation, which would have increased the cost of the evaluation. It is unclear if, and to what extent, this might have generated more rigorous findings on poverty impact, and therefore have justified the extra investment.

In Vietnam, the survey focused only on IFAD's generic poverty indicators, and used the purely income-based, absolute poverty categories of the Vietnamese government. These proved inadequate for assessing changes in poverty status related to the program. Learning from this, greater efforts were made in Ghana to ensure that the household survey questionnaire was sufficiently attuned to the program context and to on-the ground realities. Poverty characteristics corresponding with IFAD's poverty indicators were selected from the Ghana Living Standards Survey 2009–14 for assessing the households and computing the categories of poverty status (applying a proxy means test and principal components analysis). And here, no major differences were found between the characteristics obtained from the participatory ranking exercise and those used by the household survey (Van Hemelrijck 2015; Van Hemelrijck and Kyei-Mensah 2015).

Arguably, greater rigor could have been obtained in the findings on poverty distribution and impact in Ghana if the questionnaire had asked about household characteristics in much greater detail. Yet more lengthy surveys cost more, and also increase the risk of fatigue and gaming on the part of both respondents and researchers (Chambers 2008; White 2015). Therefore, in both Vietnam and Ghana, the duration of the household survey was kept to a maximum of no longer than 20 minutes. Also the time length of the participatory group discussions were limited to a maximum of two hours (Van Hemelrijck 2015).

Thus, instead of spending more resources on collecting and analyzing participatory poverty characteristics, or more fine-grained quantitative data on household characteristics to identify poverty categories prior to the evaluation, in Ghana the choice was made to keep the poverty analysis short and instead create room for participation that was more meaningful to the participants. This is what Chambers (2017) calls "appropriate imprecision." The group-based causal flow mapping exercises were found particularly useful by the participants, as it helped them to recall and understand the changes from a systems perspective, and enabled them to engage in collective sense-making with other stakeholders. The assumption is that this contributes to the ability of people to understand and navigate the system within which they are operating, and thus to their empowerment (Burns and Worsley 2015; Merrifield 2002).

Dealing with Power and Bias

All methods are susceptible to bias, and biases may occur in every phase of the evaluation, from the design to the analysis. Participatory methods, however, are considered more vulnerable than traditional survey-based methods, as they collect perceptions, meanings, and interpretations instead of hard numbers. Yet surveys generating hard numbers are also designed and conducted by human beings with value judgments (Camfield, Duvendack, and Palmer-Jones 2014; Copestake 2014; White and Phillips 2012). Survey questionnaires tend to reflect the assumptions of the designers, while qualitative interviews and participatory inquiries make room for the assumptions of the researched. Both may result in desirability or courtesy bias: the researched tell the researchers what they believe is expected from them (Chambers 2017; White 2015). To overcome such bias, the PIALA evaluations employed participatory mixed methods in a way that enabled extensive and systematic triangulation of different methods and perspectives, and cross-validation of the findings at scale.[13]

Scale, however, can also create bias, as it requires standardization that tends to reduce participation to power-blind procedure (Gaventa and Cornwall 2006; Mosse 2001). In both pilots, attempts were made to avoid this by carefully thinking through how to arrange and facilitate the group processes in order to equal out power imbalances, and by using tools that inherently empower people (Van Hemelrijck 2015). Visual tools were used, such as causal flow and relationship mapping, that enable participants to *see how* data is constructed, and to flag where things are flawed. Appropriate group composition further helped outnumber those who tend to dominate and empower those with less-heard voices, by means of majority or what Chambers (2017, 102) calls “the democracy of the ground.” Of course there is always a danger that more powerful and influential participants will dominate the discussions. Additionally, there is internalized power: the social norms and values that make certain groups believe in and accept their subordination and “voicelessness” (Kabeer 1999). Good facilitators know how to overcome this, and how to “empower through behaviour and attitudes” through careful listening and sharp observation of motives and interactions (Chambers 2017, 122). The researchers in Vietnam and Ghana were trained in this, and

[13] *Triangulation* is a principle social science technique that involves the use of more than one type of information or data source, method, and even theory and researcher, for the purpose of crosschecking in order to overcome weaknesses and biases and thus obtain greater credibility of and confidence in findings (cf. http://www.betterevaluation.org/en/evaluation-options/triangulation). In PIALA, this goes beyond merely verifying findings: it values different views and perspectives and crosschecks them to build a rich and comprehensive picture of the change processes as the basis for identifying and checking all plausible explanations for causality. *Cross-validation* in the case of PIALA is understood in “realist evaluation” terms as the practice of (dis)confirming findings across multiple independent inquiry cases to strengthen the explanatory power and the confidence in the conclusions about causality and contribution (Pawson 2013).

provided with detailed guidance on facilitation for each of the participatory methods and tools (Van Hemelrijck 2016b).

Rigor then emanates from the combination of good facilitation, and systematic triangulation and cross-validation (Chambers 2017). For the latter, a strict procedure was developed, involving six essential steps:

1. **Quasi-standardized data collection**, using at least two different methods per type or area of change in the TOC, and each with a minimum of two different sources or groups of people
2. **Daily reflections on the quality of the processes** by which each of the methods was used, and of the quality of the data generated by these processes
3. **Triangulation of the data** from the different methods and sources, and identification of data gaps and weaknesses, through systematic data collation alongside the TOC, and scoring of confidence in the emerging evidence, followed by additional data collection, if and where needed, to address the gaps and weaknesses
4. **Participatory sense-making of the evidence in each locality** together with the local stakeholders, to address remaining gaps and contradictions, and to cross-validate initial findings
5. **Aggregated collation and crosschecking** of all of the evidence and scorings obtained from all the localities
6. **Participatory sense-making of the evidence at the aggregated level** together with local and program-level stakeholders, to cross-validate initial findings for the entire program area, and to value program contributions to impact

In Vietnam, this procedure was not yet fully developed, and was found challenging by the researchers. The researchers, most of whom came from a quantitative research background, struggled with triangulation as a way to compile a multiperspective picture, and they were unable to uphold a daily practice of critical reflection on quality and process. Moreover, the tools and guidance for the data collation and quality monitoring proved insufficient (Van Hemelrijck 2013b).

In Ghana, by contrast, the researchers had a mixed background and substantial experience in participatory research. Data collation and quality monitoring were undertaken daily and systematically. A standard set of questions guided the daily reflections on inclusiveness of the processes, and the quality and sufficiency of the data, and a standard table structured around the causal claims and links in the TOC was used for data collation and triangulation (table 19.1). Methods were tightly focused on the causal links, making the triangulation much more straightforward and systematic. A Likert scale rubric was used to score the relative strength of emerging evidence for each of the causal links in the TOC. Since they were better equipped to handle large amounts of data, these researchers were able to finish the data collation, and to identify data gaps and weaknesses in each locality in good time, and to be well prepared for the sense-making workshops (Van Hemelrijck 2015).

Deciding on the Scale and Level of Engagement in the Sense-Making

Participation in evaluation is purely extractive if the findings are not returned to the participants and there is no opportunity for them to contest and debate them (Gaventa 2004; Mohan and Hickey 2004). Using PIALA's sense-making model, six village-level workshops with 180 participants, and one provincial-level workshop with 100 participants were organized in Vietnam. In Ghana, there were 23 district workshops with 650 participants, and one national workshop with over 100 participants. The participants in the workshops were sampled purposively from the research participants (Van Hemelrijck 2015).

The outcomes of the participatory sense-making were twofold. First, an additional layer of detail and confirmation of evidence was obtained from the cross-validations in the local and aggregated workshops, adding to the rigor of the evidence, and thus to the validity of the findings. Second, shared ownership was created of both the evidence and the findings, which contributed to the evaluation's inclusiveness and empowerment value. Having participated merely in data collection, people walked into the workshops knowing and owning little: but they left the workshops with a comprehensive picture of the systemic changes and the issues that the evidence had revealed, as well as of stakeholders' various perspectives on these.[14] Critical to the success of the participatory sense-making was both the scale of the workshops, and the way in which they were designed and facilitated. Special competencies are required particularly for doing this at scale. When operating with low capacity and on a shoestring budget, the amount and size of the workshops may need to be trimmed, at the expense of both rigor and inclusiveness (Van Hemelrijck and Guijt 2016).

A truly participatory sense-making process implies the equal and active engagement of all stakeholders. Dynamic environments were created, long presentations by experts were banned, and the various types of evidence were made available in accessible (including visual) formats. Small-group discussions were held, ensuring that people felt "listened to" rather than just "talked at" (Newman 2015). Beneficiaries constituted more than 30 percent of the participants in the provincial and national workshops, and 60–70 percent in the local workshops, giving them sufficient weight in debates with decision makers and service providers.[15]

[14] The survey and reflections held at the end of each workshop revealed a high degree of satisfaction among the participants, and of the knowledge and insights gained by them that they found useful for future individual or collective action.

[15] Their group must be larger in the local workshops because they form the primary target group of the project at the local level, while at the aggregated level the primary targets are policy makers and service providers.

TABLE 19.1 **Standard table used in the RTIMP evaluation for within-case data collation and triangulation**

Causal link	Secondary data	Primary QUANT data	Primary QUAL data	Strength of evidence[a]	Strength of causality[b]
Impact Claim – Poverty Reduction					
I2→I1	*Insert main findings from the 2010 Ghana Living Standard Survey report, RIMS baseline and the RTIMP M&E*	*Insert main findings from the household survey*			
O3+O2+O1→I2			*Insert main findings from the generic change analysis with community members*		
Contribution Claim 1 – Enhanced Market Linking					
M1c+M1b+O2+O3→C1b	*Insert main findings from the DDA reports, RTIMP Enterprise Record Books, ZOCs progress reports, MoFA and DADU OAs, and the RTIMP M&E*	*Insert main findings from the livelihood analysis and SenseMaker study with intended beneficiaries (farmers and processors)*			
C1a+(M1)→O1 C1b+M1a→C1a			*Insert main findings from the KIIs with DDAs, BACs, SCFs, GPCs, food traders and exporters*		
		Insert main findings from the Constituent Feedback sessions with intended beneficiaries (farmers and processors)			

(continued)

TABLE 19.1 *(continued)*

Causal link	Secondary data	Primary QUANT data	Primary QUAL data	Strength of evidence[a]	Strength of causality[b]
Contribution Claim 2 – Enhanced Production of Roots and Tubers					
C2a+C2b→O2	*Insert main findings from the RTIMP productivity surveys, progress reports from the SRID, GLDB, DDAs and ZOCs, and the RTIMP M&E*	*Insert main findings from the livelihood analysis with intended beneficiaries (farmers and processors)*			
M2a+M2b+(M2c)+M1c→C2a M2c→C2b			*Insert main findings from the KIIs with FFF facilitators, extension agents, DDAs, DADU officers, and CSIR, KNUST & UCC research leaders*		
		Insert main findings from the constituent feedback sessions with (non-)FFF participants (farmers and processors)			
Contribution Claim 3 – Enhanced Processing of Roots and Tubers					
M3b→C3a+C3b→O3	*Insert main findings from the IFAD/FAO 2014 study on matching grant facilities in Ghana, and the RTIMP and REP M&E and supervision reports*	*Insert main findings from the livelihood analysis with intended beneficiaries (farmers and processors)*			
M3b+M3c+C1a→C3c			*Insert main findings from the KIIs with GPCs, BACs and PFI local branches*		
		2–3 constituent feedback sessions with (non-)GPC participants (farmers and processors)			

a. Score 0-6 (cf. rubrics): Justify the score and provide critical notes on remaining data gaps and weaknesses and potential biases.

b. Score 0-6 (cf. rubrics): Justify the score and provide critical notes on the relative influence of RTIMP (through the M-links).

Lessons learned from the Vietnam project helped to improve the sense-making model for Ghana. In Vietnam, discussions took place mostly in mixed-stakeholder groups, and in plenary sessions, which did not give the farmers enough of a chance to collect their thoughts and gain confidence. Learning from this, participants in Ghana first worked in peer groups organized around the part of the TOC that represented their "patch" in the supply chain system—for instance, farmers discussed the production part of the chain. In Vietnam, the reconstruction of the causal flow was done in plenary, which again did not offer sufficient opportunity for farmers to engage in the process. In Ghana, this was done in small mixed groups, organized around geographic areas, with the farmers and processors systematically given the floor first, before all others, to present their views. Plenary discussions took place only on the second day, in a fishbowl set-up, in which beneficiaries constituted the majority of the discussants. This was quite successful and provoked an animated discussion with bankers about the inaccessibility (thus failure) of the microenterprise credit mechanism put into place by the program (Van Hemelrijck 2013b, 2015).

CONCLUSION

The action research around the two IFAD pilots have demonstrated that a participatory and systemic impact evaluation approach such as PIALA can produce rigorous, valid, and credible evidence that is useful for reporting and learning with partners and stakeholders, in contexts where traditional counterfactual analysis is not feasible. Moreover, the pilots have shown that using similar methods engaging beneficiaries in assessing and debating contributions to impact can contribute to enhancing the impact even ex post. Moreover, using similar methods and processes for collecting, cross-checking, and analyzing data in the two impact evaluations made it possible to compare and identify conclusions, and to formulate recommendations that have wider relevance for investments elsewhere, thus are beyond the individual programs in question.

Compared to the usual cost of theory-based, mixed-methods impact evaluations in countries like Vietnam and Ghana, these pilots were done with shoestring budgets. For example, the estimated budget for a one-year randomized controlled trial study in an IFAD-funded program in Ghana similar to the RTIMP was around $200,000. But this study only covered one subcomponent of the program, and eight districts in the northern part of the country. The PIALA evaluation of RTIMP, by contrast, cost $233,000 and covered the entire program, which consisted of three components, each of which had two or three subcomponents. Moreover, it covered 30 districts across the entire country (Van Hemelrijck 2016b).

In every evaluation that aims for greater value with limited resources, trade-offs occur. The PIALA pilots have demonstrated that these trade-offs can be turned into win-wins by carefully considering how rigor and inclusiveness can reinforce each other, and by critically reflecting on the potential loss in value-for-money if one were to be prioritized over the other (Van Hemelrijck and Guijt 2016). Limiting stakeholder engagement in the TOC process to save time and resources, for instance, leads to a substantial loss of rigor in every phase of the evaluation. Conversely, reducing the scale and level of engagement in sense-making limits the cross-validation and thus confidence in the conclusions, while also reducing the inclusiveness and empowerment value of the evaluation (Van Hemelrijck 2016b). Reducing the sample size of the participatory inquiries, as to reduce the cost, not only limits the scale of participation and thus its impact on voice, empowerment, and ownership of the findings (Burns and Worsley 2015), but also thwarts rigorous causal inference. On the other hand, reducing the scope inhibits conclusion validity by confining the systemic analysis and thus the understanding of complex nonlinear impact trajectories (Woolcock 2009).

The most essential conclusion from the action research around the PIALA pilots is that inclusiveness and rigor can reinforce each other, and that this is even more likely to happen when participatory processes and methods are employed at a larger scale (Van Hemelrijck 2016b). People's participation in impact evaluation can contribute to their understanding of the system in which they are functioning (Burns and Worsley 2015), while also adding to the rigor and credibility of findings, if

- It is both inclusive and meaningful, enabling a robust cross-checking of many different *authentic* voices;
- It avoids the dominance of any single truth, power, or particular viewpoint, thus mitigating bias; and
- It creates space for solid debate and *equal* voice, including the voices of those who are the least powerful and least heard (Van Hemelrijck 2016b).

Scale achieved through rigorous sampling and representative inclusion of all stakeholder perspectives makes it possible to generate knowledge that supersedes isolated anecdotes. Moreover, it also makes it possible to build contrasting evidence from "natural counterfactuals" occurring in the sample, thus reducing doubt in causal inference. Rigor emanates from the thoughtful design and facilitation of the participatory processes at scale, in ways that forestall the dominance of a single truth or viewpoint and enable stakeholders to participate equally and meaningfully in the process. Other essential attributes of rigor are methodological complementarity and consistency, and extensive and robust triangulation and cross validation (Van Hemelrijck 2016b).

Critical to the quality of delivery at scale, clearly, is the capacity of the researchers. In the long run, investing in such capacity helps to reduce costs, while enhancing the value of impact evaluation. For the broader development sector, this implies building capacity through providing guidance and support to new experiments with approaches like PIALA (Van Hemelrijck 2016b). For IFAD and its partners, optimizing value-for-money would imply, for instance, using PIALA as a longitudinal approach integrated with program design. This would create more room for building local research partnerships and capacity for impact evaluation, while bringing quality, continuity, and consistency to IFAD's impact learning agenda at the national and global levels.

REFERENCES

Bamberger, Michael, Jim Rugh, and Linda Mabri. 2012. *RealWorld Evaluation*. Sage.

Beach, Derek, and Rasmus Brun Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. University of Michigan Press.

Befani, Barbara, Chris Barnett, and Elliot Stern. 2014. "Introduction—Rethinking Impact Evaluation for Development." *IDS Bulletin* 45 (6): 1–5.

Burns, Danny, and Stuart Worsley. 2015. *Navigating Complexity in International Development: Facilitating Sustainable Change at Scale*. Practical Action Publishing.

Camfield, Laura, Duvendack, Maren, and Richard Palmer-Jones. 2014. "Things You Wanted to Know about Bias in Evaluations but Never Dared to Think." *IDS Bulletin* 45 (6): 49–64.

Chambers, Robert. 2008. *Revolutions in Development Inquiry*. Earthscan.

———. 2015. "Inclusive Rigour for Complexity." *Journal of Development Effectiveness* 7 (3): 327–35.

———. 2017. *Can We Know Better? Reflections for Development*. Practical Action Publishing.

Carugi, Carlo. 2016. "Experiences with Systematic Triangulation at the Global Environment Facility." *Evaluation and Program Planning* 55: 55–66.

Copestake, James. 2014. "Credible Impact Evaluation in Complex Contexts: Confirmatory and Exploratory Approaches." *Evaluation* 20 (4).

Funnell, Sue, and Patricia Rogers. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. John Wiley & Sons.

Gaventa, John. 2004. "Towards Participatory Governance: Assessing the Transformative Possibilities." In Par*ticipation: From Tyranny to Transformation? Exploring New Approaches to Participation in Development*. Zed Books.

Gaventa, John, and Andrea Cornwall. 2006. "Challenging the Boundaries of the Possible: Participation, Knowledge and Power." *IDS Bulletin* 37 (6).

Guijt, Irene, Adinda Van Hemelrijck, Jeremy Holland, and Andre Proctor. 2013b. "PIALA Research Strategy: Improved Learning Initiative." Internal document. IFAD, Rome.

Guijt, Irene, An Nguyen, Lien Bui Ngoc, Huong Dang, and Loi Vu Manh. 2014. "Report on the Participatory Impact Evaluation of the 'Developing Business with the Rural Poor' Project in Ben Tre, Viet Nam (2008–2013): A Pilot Application of a Participatory Impact Assessment and Learning Approach." IFAD, Rome.

Holland, Jeremy. 2013. *Who Counts? The Power of Participatory Statistics*. Practical Action.

IFAD (International Fund for Agricultural Development). 2016. *IFAD Strategic Framework 2016–2025. Enabling Inclusive and Sustainable Rural Transformation*. IFAD, Rome.

Kabeer, Naila. 1999. "Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment." *Development and Change* 30 (3).

Lain, Jonathan. 2016. "Rising to the Challenge—Measuring an Expanding Concept of Resilience in Oxfam's Impact Evaluations." *Policy & Practice Blog*.

Levy, Sarah, and Carlos Barahona. 2002. "How to Generate Statistics and Influence Policy Using Participatory Methods in Research." Working Paper. Statistical Services Centre, University of Reading.

Merrifield, Juliet. 2002. "Learning Citizenship." IDS Working Paper 158. Institute of Development Studies, Brighton.

Mertens, Donna. 2009. *Transformative Research and Evaluation*. Guilford Press.

———. 2010. "Philosophy in Mixed Methods Teaching: The Transformative Paradigm as illustration." *International Journal of Multiple Research Approaches* 4 (1).

Mohan, Giles, and Samuel Hickey. 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, 59–74. Zed Books.

Mosse, David. 2001. "'People's Knowledge,' Participation and Patronage: Operations and Representations in Rural Development." In *Participation: the New Tyranny?* Zed Books.

Newman, Dan. 2015. *From the Front of the Room*. Matter Group.

OECD DAC (Organisation for Economic Co-operation and Development Development Assistance Committee). 2010. "Glossary of Key Terms in Evaluation and Results Based Management." OECD, Paris.

Patton, Michael Quinn. 2011. *Essentials of Utilization-Focused Evaluation*. Sage.

Pawson, Ray. 2013. *The Science of Evaluation: A Realist Manifesto*. Sage.

Picciotto, Robert. 2014. "Have Development Evaluators Been Fighting the Last War… and If So, What Is to Be Done?" *IDS Bulletin* 45 (6): 6–16.

Ravallion, Martin. 2012. "Fighting Poverty One Experiment at a Time: *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*: Review Essay." *Journal of Economic Literature* 50 (1): 103–14.

Stern, Eliot. 2015. *Impact Evaluation: A Guide for Commissioners and Managers*. Bond.

van Es, Marjan, Irene Guijt, and Isabel Vogel. 2015. *Theory of Change Thinking in Practice. Hivos TOC Guidelines*. The Hague: Hivos.

Van Hemelrijck, Adinda. 2013a. "Powerful Beyond Measure? Measuring Complex Systemic Change in Collaborative Settings." In *Sustainable Participation and Culture in Communication: Theory and Praxis*. Intellect Ltd.

———. 2013b. "Improved Learning Initiative for the Design of a Participatory Impact Assessment and Learning Approach (PIALA): Insights and Lessons Learned from the Reflections on the PIALA Piloting in Vietnam." IFAD, Rome.

———. 2015. "Improved Learning Initiative for the Design of a Participatory Impact Assessment and Learning Approach (PIALA): Methodological Reflections Following the Second PIALA Pilot in Ghana." International Fund for Agricultural Development and Bill and Melinda Gates Foundation, Rome.

———. 2016a. "Stretching Boundaries of Evaluation Practice with PIALA in Ghana." *Evaluation for Africa* blog.

———. 2016b. "Understanding Rigour in Participatory Impact Evaluation for Transformational Development: Insights from Piloting a Participatory Impact Assessment and Learning Approach (PIALA)." PhD work in progress paper. Institute of Development Studies, Brighton.

———. 2017a. "Governance in Myanmar. Evaluation of the 'Building Equitable and Resilient Livelihoods in the Dry Zone' Project." Effectiveness Review Series 2015/2016. Oxfam GB, Oxford, UK.

———. 2017b. "Walking the Talk with Participatory Impact Assessment Learning Approach (PIALA) in Myanmar." *Policy & Practice Blog*.

Van Hemelrijck, Adinda, and Irene Guijt. 2016. "Balancing Inclusiveness, Rigour and Feasibility; Insights from Participatory Impact Evaluations in Ghana and Vietnam." CDI Practice Paper 14. Centre for Development Impact, Institute of Development Studies, Brighton.

Van Hemelrijck, Adinda, and Glowen Kyei-Mensah. 2015. "Final Report on the Participatory Impact Evaluation of the Root & Tuber Improvement & Marketing Program (RTIMP): Pilot Application of a Participatory Impact Assessment and Learning Approach (PIALA)." IFAD, Rome.

White, Howard. 2014. "Current Challenges in Impact Evaluation." *European Journal of Development Research* 26.

White, Howard, and Daniel Phillips. 2012. "Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework." 3ie Working Paper 15.

White, Sarah. 2015. "Qualitative Perspectives on the Impact Evaluation of Girls' Empowerment in Bangladesh." *Journal of Development Effectiveness* 7 (2): 127–45.

Woolcock, Michael. 2009. "Toward a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy." *Journal of Development Effectiveness* 1 (1).

———. 2013. "Using Case Studies to Explore the External Validity of Complex Development Interventions." *Evaluation* 19 (3): 229–48.