Managing relationships in qualitative impact evaluation to improve development outcomes: QuIP choreography as a case study.

James Copestake[a], Claire Allan[b], Wilm van Bekkum[c], Moges Belay[h], Tefera Goshu[d], Peter Mvula[e], Fiona Remnant[f], Erin Thomas[g], Zenawi Zerahun[h]

[a]University of Bath, UK; [b]Farm Africa, UK; [c]Self Help Africa, UK; [d]Ambo University, Ethiopia;[e]University of Malawi; [f]Bath Social and Development Research Ltd, UK; [g]Gorta Self Help Africa, Ireland; [h]Mekele University, Ethiopia.

August 2016.[1]

Abstract

Evaluation choreography – or who knows what *when* through the process of impact evaluation - has an important influence on the credibility and usefulness of findings. We explore such choreography from technical, political and ethical perspectives through reflection on a collaborative case study that entailed collaborative design of a qualitative impact evaluation protocol ('the QuIP') and its pilot use in Ethiopia and Malawi. Double blind interviewing was employed to reduce project specific confirmation bias, followed by staged 'unblinding' as a form of triangulation. We argue that these steps can enhance credibility of evidence and that ethical concerns associated with blinding can be addressed by being open with stakeholders about the process. The case study suggests qualitative impact evaluation can contribute to a more deliberative and less rigid style of international development practice.

Key words

Blinding; confirmation bias; impact evaluation; international development practice; qualitative methods.

1. Introduction: rationale and context

Impact evaluation has attracted growing attention in international development practice, both as a contributor to public accountability and as a means to promote learning and organisational effectiveness.[2] Mobilising evidence about what a specific project or intervention is achieving typically entails an often ambiguous mix of collaborative and hierarchical relationships between the commissioners of an evaluation, researchers, project staff, intended beneficiaries and

[1] Please contact j.g.copestake@bath.ac.uk with comments or for permission to cite. Production of this article was supported under research grant ES/J018090/1 from the Department for International Development (DFID) and the Economic and Social Research Council (ESRC). We're grateful for comments on earlier drafts from Graham Room and from participants at a panel session organised by Rachel Hayman on evaluation methods held at the 2016 Conference of the International Society for Third Sector Research in Stockholm on 30 June 2016.

[2] Definitions of impact evaluation vary widely. It can be equated with measured change in a vector of goal indicators (**Y**) arising from a specified set of activities (**X**) compared to a counterfactual of what the change in **Y** would have been in the absence of **X**; but it also refers to a more open-ended process of collecting and interpreting evidence of the impact of a specified activity or project (White, 2010).

others. This paper explores the question of who knows what and *when*. Freely and openly sharing information is generally regarded as a positive attribute of evaluation practice: fostering peer review, building trust, facilitating mutual understanding and strengthening prospects for further collaboration (Fox, 2007). At the same time the credibility of evaluation is also widely perceived to be enhanced by critical detachment: reinventing distance (Camfield, 2014:32) or what Campbell (cited in Pawson, 2013:10) calls "organised distrust". While most advocate a relatively formal separation of interviewer and subject some researchers have sought to enhance credibility through closer immersion into the lives of their subjects.[3]

The aim of this paper is to contribute to the empirical literature on impact evaluation and research as a social process within international development practice (e.g. Bell and Aggleton, 2016; Camfield, 2014; Eyben *et al*., 2015; Hayman *et al.*, 2016; Stevens *et al*. 2013). In particular, we focus on tensions between technical, socio-political and ethical issues arising from managing access to information selectively through the evaluation process. We do so by reflecting on action research carried out between 2012 and 2015 to design and pilot an improved qualitative impact protocol (referred to as the QuIP) for evaluating the impact of livelihood improvement projects in complex rural African contexts (Copestake, 2014). The project included two rounds of pilot testing of the QuIP on four projects (two in Ethiopia and two in Malawi) and entailed collaboration between university-based researchers in both countries and the UK, alongside staff of three international non-government development organisations (INGOs).

This case study is relevant to debate over transparency and the social relations of impact evaluation. First, it was designed explicitly as an alternative to more quantitative impact assessment methods, such as randomized control trials. One of the range of criticisms directed at these has been that the need to identify measurable indicators of treatments and outcomes in advance can influence selection and design of the development interventions to be evaluated: a case of the methodological 'tail' wagging the development 'dog' (Camfield and Duvendack, 2014; Eyben *et al.* 2015). It was therefore important for us to reflect on how the QuIP might also affect power relations between evaluation commissioners, implementing agencies and evaluators. More generally, if impact evaluation is to contribute to less rigid development practice then it needs to be flexible, quick and cost-effective as well as credible (World Bank, 2015:199).

Second, the QuIP pilot studies entailed deliberately blinding researchers and respondents to the full details of the activities being evaluated in order to reduce the risk of project specific confirmation and related response biases. This in turn opened up the issue of how and when to 'unblind' stakeholders. Restricting who knows what and when for technical reasons (such as rendering data and evidence more credible to some users) contravenes the ideal of maximising

[3] A leading example of the latter is the Reality Check Approach (Jupp, 2016), which "... puts intimacy, immersion and consensus at its core" (Camfield, 2014:19; Arvidson, 2014).

transparency to all actors at all times, and also raises ethical questions. For example, Manzano (2016:351) contrasts full and open discussion of programme theory in realist evaluation with traditional advice to researchers to 'amiably' downplay their prior knowledge of the project being evaluated.

The paper is arranged in four sections. The remainder of this section expands on how these issues relate to different theories of international development practice and the role of impact evaluation within it. The second elaborates on the case study, and the third reflects on it from the perspective of field researchers, project staff and intended beneficiaries in turn. The final section draws out more general conclusions about how impact evaluation can be utilised to promote more effective, open, democratic and progressive development practice.

Theories of development and evaluation practice

Debate over impact evaluation forms part of wider debate over international development management practice. Combining Gulrajani (2010) and Stevens *et al.* (2013) this can be introduced by contrasting 'optimistic reformist' and 'pessimistic radical' perspectives. The first seeks to apply universal principles of effective performance management to maximise achievement of measurable goals. Impact evaluation is a means to learn and to improve, with funders also having a legitimate claim to seek credible evidence that they are securing value for money. The second places empowerment and justice at the heart of development, achievable only through political struggle. Finance and associated processes of evaluation are less important as technical means to higher goals than as arenas in themselves for conflict and struggle (Eyben *et al.*, 2015; Hayman *et al.*, 2016). A third 'romantic realist' position views management practice as a process of shared discovery, consensus building and collaboration. It seeks to balance funders' legitimate claim for feedback on outcomes with the rights of intended beneficiaries to know what is being spent in their name, how and to what effect.

These three positions can be further clarified by reflecting on the complexity of development processes. This poses extra challenges for optimistic reformists in the form of a need for ever more elaborate models with which to identify optimal choices. For romantic realists, it opens up the possibility of more holistic understanding and emergent solutions, achievable through complementarity of insights, cross-cultural communication, trust building, joint learning including through use of multiple methods and triangulation of findings. For pessimistic radicals, in contrast, complexity undermines prospects for coordinated action by accommodating divergent ideologies and perceptions of competing interests. Grint (2005) warns that the very decision to characterise a problem as complex (or "wicked") may itself be a device for exercising leadership and power.

As a simple illustration, consider the problem of how to test the knowledge, skill, understanding and wisdom that students have gained from a set programme of study. This can be viewed as a purely technical problem of testing ability to draw up definitive answers that can be assessed using universal marking schemes. Alternatively, assessment can be viewed as a reflection of the interests and authority of powerful examiners, and a means for them to enforce discipline and control. Between these extremes is a realistic romantic view of political deliberation over assessment criteria whose legitimacy rests on building consensus about their reasonableness. Procedural transparency also legitimates assessment schemes, both by contributing to consensus building (by sharing marking schemes, for example) and as precondition for error correction through rights to peer review and to appeal. But we are also familiar with transparency being managed: when markers are required to mark blind or protected by remaining anonymous, for example.

Development context

The more specific context for this research was a programme partnership agreement between the UK Department for International Development (DFID) and selected INGOs, linking core institutional funding to better evidence of recipients' social impact (Coffey, 2012).[4] This was framed in optimistic-reformist language, but the INGOs was delegated responsibility for deciding *how* best to produce additional evidence. A stepping stone towards addressing this problem is clarity over the criteria of what constitutes good impact evaluation. Here we distinguish between four. First, the evidence of causal links between the INGOs' activities X and intended impact Y must be not only *credible*, but offer sufficient *additional* credibility over what they already know.[5] Second, the evidence needs to be *relevant* to decision-making needs: for example, how to prioritize between investments across different fields, or whether to scale-up activities on the basis of initial piloting. In fields that are subject to rapid change the relevance criterion also implies timeliness. Third, the cost of producing the evidence should be proportionate to potential benefits. The challenge to be *cost-effective* is exacerbated by the tendency for costs to be more immediate and certain than potential benefits. Potential benefits of avoiding bad strategic decisions and investments are vast, but they are also inherently hard to predict, as are prospects that credible and relevant additional evidence will affect key choices. Fourth, since it is not self-evident that all aspects of an activity can or should be quantified, valued and aggregated then it is also necessary to locate cost-benefit calculations within a wider *ethical* framework.

In the face of the challenge to supply credible, relevant and cost-effective evidence of impact in

[4] By narrowing the scope of the paper to INGOs we are not suggesting other development agencies have any lesser need for impact evaluation. But operating across national boundaries increases the socio-political and geographical distance between INGO senior staff and intended beneficiaries, and hence the case for more formal feedback loops.

[5] Credibility can be defined as presentation of sufficiently rigorous evidence and argument to enable A to convince B that something is true on the basis of reasonable assumptions, in a way that is also auditable or open to peer review. For the purposes of this paper I define X as having credibly *caused* Y in a particular context if the following conditions are also satisfied: there is strong evidence that X and Y happened; several stakeholders independently (and without prompting) assert or imply that X was an insufficient but necessary part of a causal package that is an unnecessary but sufficient cause of Y; there is no more credible counter-explanation for why they might say this; and explanations of how X caused Y are reasonably congruent with a plausible theory of change. For further discussion of the difference between the terms "reasonable" and "rational" see McGilchrist (2010).

ethically acceptable ways, responsible INGO staff could draw upon a vast body of accumulated experience and literature about a wide range of potentially useful methodological options, particularly across the social sciences. However, adapting this to their needs has proven to be difficult, perhaps most importantly because of contextual complexity. The causal processes leading to social impact are often best characterized as co-evolutionary: in other words they depend on the conjunction of a unique set of necessary conditions that are the produce of dynamics in distinct and weakly linked systems (Room, 2013). While evidence of simpler causal links between variables is often useful, it is rarely possible to infer from it precisely how relevant observed change in one context is likely to be to another. For example, all barley growers may have similar needs for land, labor and seeds. But achieving a breakthrough in yields may also depend on changes in factors affecting supply and demand for alternative crops and other-income earning activities.[6]

The challenge facing INGOs to identify and apply suitable impact evaluation methods is exacerbated by an additional set of social factors. Evaluators with expertise in potentially relevant methods often have a different set of priorities. Commercial interests may bias them towards selection of methods that they find cheaper or easier to use. More academic evaluators may select methods appropriate to publishing work in peer reviewed journals that imply more costly and time consuming standards of rigor. Careful selection, contracting and oversight of the study by the commissioner can minimize such problems. But these supervisory tasks add to costs, as does the effort needed to ensure contracted-out studies are appropriately integrated with INGOs' other data and performance management systems.[7] Meanwhile, by using evaluation studies as means to advance their own preferences and interests, researchers contribute to uncertainty about the credibility, relevance and cost of impact evaluation evidence, weakening incentives to invest in it at all. The market for impact evaluation thereby comes to resemble the market for second-hand cars famously analysed by Akerlof (1970): the cost of identifying poor studies from good ones can be so high that even when forced to commission them they do not invest much time investigating which might actually be useful.

The problems described above act as brakes to the scale and quality of impact evidence generated by INGOs in ways that might improve their practice. They also explain the rationale for this paper, that there is a case for more rigorous and purposeful research (including action-research) into impact evaluation processes and practices themselves. The discussion above illustrates why it is not enough to focus on research and evaluation methodology at the abstract level. Rather, such research needs to combine attention to technical and ethical aspects of

[6] More formally, a working definition of contextual complexity is a setting in which the influence of a vector of factors **X** within the control of the INGO on impact indicators **Y** is confounded by factors **Z** that are impossible fully to identify, hard to measure accurately, interactive and cumulative in their influence on **Y** and impossible fully to control. Additional complexity arises if the nature and values of **X** and/or **Y** are also uncertain – for example, because both are also highly context specific.

[7] The influence of academic values and social norms over impact evaluation methodology also run deeper in the form of inappropriate borrowing of ideas, norms and standards from the so-called 'hard' sciences, for example (Flyvbjerg, 2001; McGilchrist, 2010).

different methods with attention to impact evaluation as a social process in specific contexts. For example, there is scope for researching the possible trade-offs between the credibility that can be derived from appointing independent evaluators versus the benefits from close involvement of potential users to ensure relevance and cost-effectiveness.

This is a large research agenda; more narrowly we explore here the specific issue of the choreography of impact evaluation: not only who needs to know what, but also when. '*Who'* here refers to the hierarchy and networks of both INGO staff (from evaluation commissioners to those directly responsible for implementing actions to be evaluated) and of evaluators (including project managers, field researchers, data analysts, report writers and knowledge brokers). It also includes intended beneficiaries and other stakeholders, raising practical and ethical questions about participation and power along the aid 'value chain'. 'W*hat'* here includes both "confirmatory" evidence about consistency with the theory of change behind the activity being evaluated, and also "exploratory" evidence that goes beyond it (Copestake, 2014). Timing of access is important both in relation to the logical sequence of evaluation activities (from initial scoping and framing through data collection and analysis through to reporting and influencing), allowing also for iteration, duplication and triangulation. Credibility depends here not only on who knows what and under what circumstances, but also on timing: the issue of blinding concerning not only who should and should not be informed about what and why, but also when and for how long.

2. Case study

The purpose of this section is to present a concrete case study through which to explore and elaborate upon the issues raised above. The selected case study is the 'ART Project' (Assessing Rural Transformations) that entailed collaboration between staff at the University of Bath in the UK, Mekele and Ambo Universities in Ethiopia, the University of Malawi, and three INGOs (*Self Help Africa*, *Farm Africa* and *Evidence for Development*).[8] This analysis of the project was first drafted by the lead author, who was also the project's principal investigator. It also draws on unpublished notes and feedback from the other named authors, reflecting on their experience under the project, and on written accounts of stakeholder workshops held in Lilongwe and Addis Ababa in 2015.

The previous section covers much of the ART Project's original rationale. We were interested to explore how to produce evidence of impact to strengthen organisational learning, accountability and adaptability in ways that avoided what Natsios (2010) refers to as "obsessive measurement disorder". More specifically, we set out to address the attribution problem: how

[8] The ART Project ran from 2012 to 2016, and was funded by the UK Department for International Development (DFID) and the Economic and Social Research Council (ESRC) under their joint program of research for poverty alleviation. See go.bath.ac.uk/art. *Self Help Africa* in Ireland has since been renamed *Gorta Self Help Africa*. *Evidence for Development* assisted with complementary quantitative monitoring of the selected projects using an individual household survey method to elicit evidence of changing livelihoods and food security. For reports on this strand of the action research (which is not discussed in this paper) see www.efd.org.

best to produce credible evidence of the causal impact of specified development projects in ways that are also timely, cost-effective and ethical.[9] We aimed to do this in a way that would also be useful for other agencies, and could contribute to wider understanding of the strengths and limitations of qualitative impact evaluation. The vehicle for doing so was to design and pilot a qualitative impact assessment protocol (named the QuIP) appropriate for use to assess project interventions aimed at promoting household level food security in the context of complex rural transformations arising across Africa from rapid climate change and market commercialisation. This also entailed developing a method that would complement other INGO monitoring, evaluation, learning, accountability and performance management activities.

The ART Project started with a collaborative design workshop in May 2013. In the second year, the QuIP was piloted through studies of four rural livelihood promotion projects: two in Ethiopia and two in Malawi (See Table 1). Informants selected from lists of intended beneficiaries were asked to reflect on changes in their lives and livelihoods over the previous year (Copestake and Remnant, 2015). In the third year, a modified version of the QuIP was applied to different samples of intended beneficiaries of the same four projects, with questions extended to encourage respondents to share their perceptions of the main drivers of change they had experienced over the previous two years. Findings from both workshops were written up and reviewed at feedback and dissemination workshops in Addis Ababa and Lilongwe in July 2015.

Insert Table 1 about here.

As the purpose of this paper is methodological, the empirical findings from the two rounds of QuIP impact studies are not reproduced here.[10] Instead Table 2 highlights ten key characteristics of the final version of the QuIP. Here we focus particularly on the blinded interviewing (Step 1), coding (Steps 6&7) and triangulation through staged unblinding (Step 10).

Insert Table 2 about here.

Primary data was collected using semi-structured interviews and focus group discussions (Step 1). These employed a sequence of questions to ask respondents about drivers of change in different domains of their lives over a specified period. Blinding of interviews and focus groups was made possible by the separation of evaluation tasks between field researchers, lead researcher and analyst, as illustrated by Figure 1. The main purpose of this was to reduce the

[9] More formally, the attribution challenge was defined as follows. How do the actions of development agencies (**X**) contribute to improving livelihoods and wellbeing (**Y**) of households allowing for confounding variables (**Z**) associated with diverse, risk and complex contexts? Standard methods seek to do this through statistical inference based on variable exposure across a large samples of number of households to **X** controlling for **Z**. Instead we sought to develop methods based on intended beneficiaries' own self-reported account of multiple configurations of causal mechanisms linking **X** to **Y** and to **Z**.

[10] Copestake and Remnant (2015) summarise findings from the first round of pilot studies. The project web site (go.bath.ac.uk/art) also provides two of the second round pilot QuIP reports, along with full QuIP Guidelines, which run to nearly fifty pages (Remnant and Copestake, 2015).

risks of project related strategic or confirmation bias. This can be defined as explanations based not solely on what respondents and interviewers believe to be the truth, but on what they think may be either in their own interest or consistent with what those carrying out or commissioning the study would like to hear.[11] The nature and extent of such bias is unknown, but its possibility nevertheless seems to be widely viewed as a weakness of self-reported impact attribution, thereby reducing its credibility. Note that even double blind interviewing cannot fully guarantee against this because respondents may choose to share causal explanations on the basis of *assumptions* (whether correct or not) about the purpose of the interview. This might explain, for example, a tendency for respondents to mention the positive impact of government initiatives in Ethiopia.

Figure 1. Institutional relations associated with use of the QuIP.

Blinded data collection also presented researchers with two immediate practical difficulties. First, it precluded them from making use of local project staff to assist in gaining entry into the field and locating respondents. Although this raised the time required for data collection, the extra cost was partly offset by not needing to involve project staff in the task too. Second, as field researchers were not aware of the project being evaluated (or even the name of the agency responsible for it) they could not refer to this to justify the data collection exercise, either to local authorities or to respondents. This problem, and related ethical issues, are discussed further in the next section.

Data coding (Steps 6&7) cannot be similarly blinded because the analyst must have knowledge of the project to be able to code statements in each domain as either attributing impact explicitly to the project, or implicitly to the project (by corroborating the theory of change behind it), or to factors incidental to it. Potential bias here is reduced because the analyst (unlike primary respondents) has no direct personal interest in the project. Their coding work can also be fully and easily audited, challenged and adjusted. The analyst is also directly responsible for production of the draft evaluation report (Step 8) and not having been in the field themselves they are forced to base this analysis solely on the data received from the field research team, including additional written observations and debriefing notes. This again creates a potential audit trail.

A third feature of the QuIP is the opportunity it creates for triangulation through *staged* unblinding of the data (Step 10). This occurred when project staff were given the opportunity to review and discuss the draft report and thereby to offer their own observations and interpretations of the drivers of change identified in it. This served not only as a data check, but also opened up opportunities for more detailed discussion of project implementation,

[11] More precisely the double blinding aims to reduce possible bias in attributing change in an impact domain Y to project related causal factors **X** (relative to other factors **Z**) as a result of the interview being explicitly associated with **X** in the mind of the respondent and/or interviewer. In contrast confirmation bias is generally defined as selectivity in collection and analysis of data in order to support previously held beliefs (World Bank, 2015:182).

particularly explanations supplied by respondents for negative as well as positive explicit and implicit project impact. Incidental drivers were also relevant to reflection on project design and the theory of change underpinning it, particularly the persistence or otherwise of expected risks to project success. These meetings were enriched by also involving the unblinded field researchers, enabling them to enter into dialogue with project staff about the shared evidence in front of them. The presence of more senior staff helped to ensure that the outcome of these discussions contributed directly to learning across the INGO and to follow-up actions.

3. Analysis and discussion

The previous section introduced the QuIP and how the choreography of access to information affects data collection, analysis and use. In this section we draw on experience of testing the QuIP under the ART Project to analyse research relationships from the perspective of appointed field researchers, participating INGOs and intended beneficiaries.

QuIP from the perspective of field researchers

A subsidiary goal of the ART project was to develop a methodology that enhanced credibility and cost-effectiveness by relying on field researchers located as close as possible to the projects being evaluated. Reasons for this were partly instrumental: to benefit from contextual knowledge, field interviewing experience and skills (including fluency in local languages), and to avoid the extra costs of recruiting outsiders from more distant places. Participants in the QuIP design workshop also recognised the potential value of fostering collaborative-horizontal links between researchers and INGOs at national and sub-national levels, as a counter to strong vertical-contractual relations.

Field researchers for the pilot studies in Ethiopia and Malawi were selected by the lead researcher (also the principal investigator) from responses to an open invitation to tender for the work circulated by e-mail through research and NGO networks in the two countries.[12] The four appointees (two of whom responded separately, but agreed to work together) were all affiliated to social science departments of local universities, although they opted to conduct the work as independent consultants, drawing in former students and other collaborators with appropriate language skills and the specified gender balance of one man and one woman per study. Initial briefings with the field researchers covered the rationale for blinding during the field work period, and how to overcome the potential difficulties this might create, alongside discussion of data collection instruments, research ethics and good interviewing practice. All four lead investigators accepted and acquiesced to the blinding approach, recognising the potential *greater good* argument that doing so could enhance the credibility and potential influence of findings. They were also positively motivated by the prospect of participating in a novel methodological experiment.

[12] Selection criteria were cost, relevant experience and evidence of interest in the project. Bidders were invited to read and comment on the draft QuIP guidelines, and to submit an indicative budget. Five bids were received in Ethiopia and four in Malawi.

Actual experience of securing entry into the field was mixed. Two of the three teams proceeded smoothly through gatekeeping conversations with local government officials and headmen. The third encountered significant suspicion, partly inflamed by political protests that were taking place in the region at that time. Appropriate introductory letters had been prepared by the Lead researcher in the UK, but the problem was eventually overcome with the help of personal contacts of the field researcher. This resulted in several days of delay, but recourse to a contingency plan to seek direct support from the commissioning NGO (that would have un-blinded the field researcher) was avoided. Despite this incident, our overall experience was that field researchers' affiliation with a local university was a sufficient source of status, authority and legitimacy to secure the necessary permission for data collection without the need to explain the explicit link to a named development agency or project.

The field research teams' experience of locating farmers without the help of the commissioning INGOs varied considerably according to the extent and reliability of contact details (including sketch maps and cell phone numbers) made available to them.[13] Physical geography and weather were also major determinants of the time required to locate and reach respondents and to arrange focus groups, as vividly illustrated by field diaries and photos. Once located, and after the purpose of the study was explained to them, respondents rarely displayed any reluctance to participate (see further discussion below). Affiliation of members of the research team to a local university, combined with their cultural sensitivity and experience, provided sufficient authority and reassurance. Nor did lack of reference to any specific project or activity impede respondents from articulating their views about the main drivers of change in different domains of their livelihoods and wellbeing.

The lead field researchers all reported remaining unsure of the identity of the INGO and the project they were helping to evaluate throughout the duration of the first round of pilot studies. While able to make a more informed guess a year later, when the second round of studies were conducted, they all remained in the dark about the precise intervention packages and theories of change. However, their reflections on the experience were mixed. They continued to recognise the instrumental value of double blinding in enhancing the credibility of findings, particularly ensuring respondents did not deliberately overstate the importance of the NGOs' activities to their livelihoods and wellbeing. But they also expressed some frustration at the limitation blinding imposed on their ability to probe more deeply into specific aspects of the project being assessed, including why it worked for some respondents and not for others. While the organisation of interview and focus group schedules into domains of impact helped

[13] In most cases a two stage clustered sampling strategy was employed, starting with purposive selection of two to four localities, followed by random sampling of lists of farmers or households located within them and supplied by the INGO, either from lists of project beneficiaries or households covered by the baseline survey. In one case the list was further stratified into individuals who had participated in differing INGO activities (vegetable growing, poultry, goat rearing, beekeeping) and field teams were asked to quota sample from each list within selected localities, but without being told what the different sub-groups signified.

somewhat, the lack of more specific knowledge about project activities as "an explanatory focus" (Pawson, 2013:14) also made it harder to ensure interviews remained focused on specific experiences within the selected time periods.

The blinding of field researchers is also an ethical and political issue. Its usefulness hinges on establishing and maintaining mutual respect, trust, shared commitment with the lead researcher to the ultimate goals of the research and good communication to guard against a slide into a more detached, extractive and ultimately less effective contractual relationship. This applies particularly to the separation (literally across continents) of data collection/tabulation and analysis/reporting. While limiting their role, not having to take responsibility for analysis did have some practical advantages for field researchers, particularly avoiding the contractual uncertainty that can arise with analysis and writing up. But the opportunity both to provide qualitative written feedback on the field work, and to participate in subsequent unblinded discussions of the draft report was symbolically and ethically important, as well as being potentially useful.

QuIP from the perspective of INGO staff

Participation in developing the QuIP was initiated and driven by INGO staff with responsibility for monitoring and evaluation at head office level, and they also oversaw selection of projects to be studied, all implemented through their own country offices. They regarded QuIP studies as a useful "reality check" and "deep dive" into whether selected projects were achieving their intended goals, as an investment in internal learning, and a way of demonstrating to DFID and Irish Aid that such learning was taking place. Growth in the INGOs' scale of operation strengthened these arguments by exposing the limitations of relying solely on internal and more informal monitoring (depicted by the vertical arrows on the left hand side of Figure 1). The demands placed on INGO staff and their collaborators in Ethiopia and Malawi to assist with a wide range of project visits from abroad were considerable, and exacerbated by parallel demands for government oversight, particularly in Ethiopia. For this reason, operational staff were also positive about the limited demand that QuIP studies made on their own time to assist with data collection.[14] But they also recognised the importance of their active participation in three other ways.

First, an initial meeting with the lead researcher was needed to agree on the scope of the study, sample selection and design of interviewing formats. While understanding the reasons for blind interviewing some INGO staff were concerned that if questioning was *too* broad then respondents might simply forget to mention some of the benefits they had obtained from the project. Absence of explicit impact evidence might then be interpreted as absence of impact (a false negative). This concern was partly reduced by adjusting domains and probing questions to increase the likelihood that they would trigger reflection relevant to the project's theory of

[14] In Malawi jokes were also shared about the arrival of 'ghost field researchers' to go with the 'ghost beneficiaries' and even 'ghost villages' that reportedly appeared in response to government input subsidy programmes.

change. Towards the end of interviews respondents were also asked to list and to rank organisations "from outside the village" who had offered them support, and this did indeed prompt more explicit reference to the INGO than other questions, as well as revealing some confusion about the roles of different agencies in the locality.

Second, alongside lists and contact details from which to select interview samples, information from INGO staff about the nature and timing of activities carried out under the project was necessary to enable the lead researchers to identify which causal statements were implicitly consistent with the project's theory of change, and which were incidental to it. Such evidence also helped to verify the activities in which specific respondents participated, and to cross-tabulate this against the various drivers of change they mentioned, thereby also identifying gaps where project activities were not mentioned by those thought to have benefitted from them.

Third, staff participation in discussion of findings provided an important opportunity for two-way learning. Initial briefings emphasised that the studies were intended to promote reflection, learning and improved practice rather than to find fault or to apportion blame. Some defensiveness on the part of staff nevertheless remained, judging by the attention given to interpreting the negative (explicit and implicit) evidence obtained. At the same time, negative evidence did stimulate useful discussion: for example, over the rules by which recipients of goats should pass on the first two kids to neighbours, and over where to locate groundnut shellers to maximise their joint use by people from different localities.

These discussions were enhanced by QuIP reporting formats that enabled staff both to gain a quick overview of the evidence generated and to drill back down to the typed source interview notes. This provided reassurance about the reliability of summary findings, and contributed to the usefulness of triangulation and debriefing sessions. The INGOs also took the opportunity to internalise learning by involving staff from elsewhere in the organisation in data coding and analysis. However, while much of the coding work was relatively straightforward this was not invariably the case, reinforcing the value of it being transparent and auditable in enhancing both internal and external credibility.[15]

QuIP from the perspective of intended beneficiaries

In designing and testing the QuIP the central goal of the ART Project was to contribute to more credible and cost-effective impact evaluation, taking as a starting point the idea of simply asking those who were intended to benefit what had happened to them. Self-reported

[15] A common problem was for a statement to combine both positive and negative elements. For example, a respondent received chickens through a project, some then became diseased and died, but she eventually got help treating them. The analyst's problem is whether to code this as a single explicit impact story (and if so to decide whether it should be positive or negative) or as discrete causal evidence (positive, negative and positive). The choice can also depend on a wider reading of the full interview notes, setting out the respondent's overall view of their participation in the poultry project. For example, one reason for poultry mortality that emerged from discussion was that they were being given to some people simply too vulnerable to be able to look after them adequately.

attribution, we noted, potentially avoids the cost, complications and ethical issues associated with inferring attribution statistically through treatment exposure variation, including reliance on control groups. However, while the QuIP thereby places a high value on what intended beneficiaries of projects have to say, they were not the primary audience for the findings. Thus the QuIP was developed under the ART Project as a "one-way" form of beneficiary feedback (Groves, 2015) to inform those higher up the hierarchies controlling the projects being assessed. QuIP studies aim to benefit intended beneficiaries in the short-term by strengthening their voice, and in the longer-term by strengthening feedback mechanisms to inform future development activities.

The immediate and more certain effect of the QuIP on those project beneficiaries selected as respondents is to make an additional demand on their time. A further ethical complication arises from the double blinding because this means respondents are also not as fully informed about the purpose of the study as they could be. This limits their power to provide feedback more consciously focused on the project, and thereby possibly more directly relevant to the commissioner of the study. The decision to restrict information in this way can be justified by a *greater good* argument that the potential benefits (of thereby broadening the range of findings and enhancing their credibility) outweigh possible extra costs. Thus there are trade-offs between the credibility of findings and their potential relevance, as well as between the rights of respondents and the potential wider benefits of the findings generated. These also involve weighing up the interests of those interviewed, the wider population of intended beneficiaries from which they are drawn and a still wider population of potential beneficiaries of future activities that might be influenced by the evidence generated.

Having outlined some of the issues involved, we now briefly review the experience gained in piloting the QuIP. Interviews were conducted with named individuals selected through clustered random sampling from lists provided by the staff of the projects being assessed. The participation of other household members was neither encouraged nor discouraged and interviews were conducted in the preferred language of the respondent. The interviewers were instructed to open interviews by translating a standard text.[16] Very few respondents refused to participate, or opted to terminate interviews before they were completed. The length of completed interviews ranged from 45 to 90 minutes, with the length of focus groups mostly towards the top end of this range. Respondents were not paid but were offered a small thank

[16] This was as follows: "My name is [...] and I am employed by [...] as a field worker. We are conducting a study into how the income and food security of people living in this area is changing and what can be done to improve this. We are doing this research for [...] and with the approval of the [local authorities]. They have supplied us with a list of households to contact, but we cannot contact all of them, so we have chosen a smaller number at random, including yours. The information we collect will be used for the purposes of this research only, and will not refer to you or to your household by name. You do not have to take part in this study. You can decide if you would like to take part or not. We will not inform anyone else about your decision. If you do decide to take part you can also change your mind and end this interview at any time. And if you do agree to take part, but there are some questions you do not wish to answer this is also fine. You can refuse to answer as many questions as you want." (See page 24 at http://qualitysocialimpact.org/wp-content/uploads/2016/05/QUIP-Full-Guidelines-English-April-2016.pdf).

you gift for participating. Their response to being interviewed in both Malawi and Ethiopia was overwhelmingly positive. Some did ask whether the study was linked to a specific programme or plan (a question the interviewers were unable to answer); but a more common reaction was to appreciate the openness of interviews to learning what respondents' *themselves* thought was important to different aspects of their wellbeing. This may have contrasted with other experiences of being interviewed that were narrower and more rigid, but it probably also reflected at least as much the sensitivity and experience of the field researchers.

In the last year of the ART Project we discussed the option of involving intended beneficiaries in the final workshops in Addis Ababa and Lilongwe, but decided not to do so. One significant factor was cost, but the decision also reflected lack of prior planning of the selection process. With the benefit of hindsight this could have been addressed after the original interviews by asking respondents if they would be interested in attending a final and unblinded focus group meeting to present, discuss and deepen findings. In addition to enabling them to feedback directly and openly on project activities, this would have provided a forum to explore their views on the blinding issue. It wold also have reduced the ethical dilemma alluded to in the previous paragraph, because blinding would then have only been temporary, with the opportunity to provide more directed feedback on project activities delayed rather than denied.

Section 4. Conclusions

In the introduction to this paper we argued for more research into the social relations of impact evaluation. We highlighted two issues: who influences their design, and how the choreography of their implementation affects trade-offs between credibility, cost-effectiveness and relevance. Behind both issues is uncertainty about what evidence impact evaluation can realistically generate, with what levels of credibility for whom, how and at what cost. Driven particularly by public demand for evidence of value for money, evaluation commissioners have generally prioritized confirming how impact goals are being achieved. This search for evidence is expressed in technical language that reflects what we called an optimistic-reformist view of development practice. This is in tension with both a more pessimistic-radical perspective, and a realistic-romantic view that emphasises the role of dialogue and plurality as a response to complexity. INGOs and other development agencies are caught between these views: struggling to reconcile demands for clarity within a hierarchical audit culture with aspirations to be more transformative, adaptable and consensual. Impact evaluation as currently practiced in international development reflects these tensions. Professional evaluators and academics have responded by seeking to develop, elucidate and apply a wide range of approaches that reflect not only epistemological diversity but also cultural diversity in management of inter-organisational relationships from hierarchically extractive (if not coercive) to participatory and egalitarian, via commercial and transactional.

In this wider context, the ART Project case study can be viewed as a realistic-romantic bid to create space for collaboration in developing a form of impact evaluation that addresses and balances these tensions. The QuIP was not intended as a universal solution to the problem of impact evaluation. Rather it aimed to clarify and repackage more generic approaches (including contribution analysis, process tracing and goal free evaluation) to meet specific needs of the participating INGOs. The idea of designing protocols can be criticised for being too prescriptive and rigid. However, they can also offer users and providers of impact evidence a transparent methodological benchmark that adds clarity to their methodological discussion, whether adopted, rejected or adapted. To use a market analogy, our familiarity with leading brands can help us as consumers to decide what to buy and what not to buy within complex and crowded retailing spaces.[17] Of course by introducing another branded product there is also a danger of adding to the confusion and to the alphabet soup of acronyms. This depends on the clarity with which it is presented, can be understood and compared with alternatives, as well as its inherent relative strengths and weaknesses. The more general point is that designing and piloting the QuIP is an example of a consensual and deliberative process in the realist romantic rather than reformist or radical spirit of development practice.

At the same time, our account of the QuIP has highlighted an apparent methodological paradox. On the one hand, we have emphasised that procedural transparency (including the division of labour within the evaluation team) is important to enhancing the credibility of findings by exposing findings (and the methodology behind them) to audit and to peer review. On the other hand, our claims to credibility rest at least in part upon introducing a procedural lack of transparency by temporarily blinding some of these people, as a counter to potential sources of bias. This paradox is not unfamiliar. Blinding and anonymity are transparent and accepted practices in clinical trials and in educational assessment, for example. Adam Smith explored the idea of the *impartial spectator* and this was revived by John Rawls through the device of placing a *veil of ignorance* over the evaluator. One ethical defence of the practice of blinding is to appeal to the greater good: that the end (better evidence) justifies the means (blinding). However, this leaves open the question of the right of those doing the blinding to weigh up the costs and the benefits on behalf of others. It is perhaps reasonable also to expect that blinding should be temporary and reversible (hence better described as *blindfolding*), and does no significant harm to those who are subject to it. One mechanism for guarding against this is to brief respondents and field researchers about the logic behind being blindfolded, and to proceed only if they offer full and ongoing consent to participating on this basis. Going further, commissioners and lead researchers may also agree to offer blindfolded respondents and researchers an option to participate subsequently in blindfold-free debriefing and discussion of the findings, so that they are eventually fully informed about the evaluation, or at least given the option to be so. Experience of such meetings with field researchers under the ART project is that this form of staged *ex post* triangulation can also be very productive in generating further evidence and triggering follow-up action. Scope remains for further action

[17] Of course, the QuIP is far from unique in this respect. For example, a close cousin that also aims to be exploratory as well as confirmatory is PADev (Pouw, et al., 2016).

research into the benefits and costs of extending such activity to include primary respondents also.[18]

To sum up, this paper has sought to broaden debate over impact evaluation by focusing on the importance of the choreography of relationships between those involved. More specifically, we have drawn on the case study of design and piloting the QuIP to explore how blindfolds and their timely removal can enhance the quality and credibility of evidence generated. There is clearly scope for further research into these issues, both with the QuIP and with other methods. Meanwhile, the paper has illustrated how the choreography of impact evaluation can contribute to a romantic realist approach to development practice that emphasises deliberation over more rigid results-oriented managerialism.

References

Akerlof, G. (1970). The market for “lemons”: quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488-500. doi:10.2307/1879431

Arvidson, M. (2014). Ethics, intimacy and distance in longitudinal qualitative research: experiences from reality check Bangladesh. In Camfield, L., editor, *Methodological challenges and new approaches to research in International Development*. London: Palgrave macmillan. Pp.19-37.

Bell, S., & Aggleton, P., editors, (2016) *Monitoring and evaluation in health and social development: interpretive and ethnographic perspectives*. London and New York: Routledge.

Camfield, L., editor. (2014). *Methodological challenges and new approaches to research in International Development.* London: Palgrave macmillan

Camfield, L., & Duvendack, M. (2014). Impact evaluation – are we ‘off the gold standard’? *European Journal of Development Research*, 26(1):1-12.

Coffey International Development. (2012). *Evaluation manager PPA and GPAF: evaluation strategy.* London: Coffey.

Copestake, J. (2014) Credible impact evaluation in complex contexts: confirmatory and exploratory approaches. *Evaluation*, 20(4):412-27.

Copestake, J., O'Riordan, A-M. Telford, M. (2016). Justifying development financing of small NGOs: impact evidence, political expedience & the case of the UK Civil Society Challenge Fund. *Journal of Development Effectiveness*, 8 (2):157-70.

[18] In some contexts there may be scope for using social media to do this more cost-effectively: alerting respondents to where final reports have been lodged and inviting comments on them, and therefore moving closer to full two-way beneficiary feedback.

Copestake, J., & Remnant, F. (2015). Assessing rural transformations: piloting a qualitative impact protocol in Malawi and Ethiopia. In: Camfield, L., & Roelen, K., editors, *Mixed methods in poverty research*. London: Routledge.

Eyben, R., Guijt, I., & Shutt, C. (2015). *The politics of evidence and results in international development*. London: Practical Action Publishing.

Flyvbjerg, B. (2001). *Making social science matter: why social inquiry fails and how it can succeed again.* Cambridge: Cambridge University Press.

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice* 17(4): 663-71.

Grint, K. (2005). Problems, Problems, Problems: The social construction of leadership, *Human Relations* 58(11): 1467-94.

Groves, L. (2015). *Beneficiary feedback in Evaluation*. London: Department for International Development, Evaluation Department. Accessed on 27 July 2016 from: http://r4d.dfid.gov.uk/pdf/outputs/Evaluation/Beneficiary_Feedback_in_Evaluation.pdf

Gulrajani, N. (2010) 'New Vistas for Development Management: Examining radical-reformist possibilities and potential', *Public Administration and Development* 30(2): 136-48.

Hayman, R., King, S., Kontinen, T., Narayanaswamy, L., editors. (2016). *Negotiating knowledge: evidence and experience in development in development NGOs*. Practical Action Publishing: Rugby, with INTRAC, Oxford.

Jupp, D., (2016). Using the reality check approach to shape quantitative findings. Experience from mixed method evaluations in Ghana and Nepal. In Bell, S., & Aggleton, P., editors, *Monitoring and evaluation in health and social development: interpretive and ethnographic perspectives*. London and New York: Routledge. pp.172-184.

Manzano, A., (2016). The craft of interviewing in realist evaluation. *Evaluation*, 22(3):342-360.

McGilchrist, I. (2010). *The master and his emissary: the divided brain and the making of the Western World.* New Haven: Yale University Press.

Natsios, A. (2010). *The clash of counter-bureaucracy and development*. Center for Global Development Essay. Washington DC: Center for Global Development.

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

Pouw, N., Dietz, T., Belemvire, A., de Groot, D., Millar, D., Obeng, F, Rijneveld, W., Ven der Geest, K., Vlaminck, Z. & Zaal, F. (2016). Participatory assessment of development interventions: lessons learned from a new evaluation methodology in Ghana and Burkina Faso. *American Journal of Evaluation*, 1-13.

Room, G. (2013). Evidence for agile policy makers: the contribution of transformative realism. *Evidence and Policy*, 9(2):225-44.

Stevens, D., Hayman, R., Mdee, A. (2013). Cracking collaboration between NGOs and academics in international development research. *Development in Practice*, 23:1071-77.

Taylor, R., Arvidson, M., Macmillan, R., Soteri-Procter, A., & Teasdale, S. (2014). What's in it for us? Consent, access and the meaning of research in a qualitative longitudinal study. In Camfield, L., editor, *Methodological challenges and new approaches to research in International Development*. London: Palgrave macmillan. Pp.38-58.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation 16*(2), 11.

World Bank. (2015). *World Development Report 2015: mind, society and behaviour.* Washington DC: World Bank.

Figure 1. Stakeholder relationships under a QuIP study

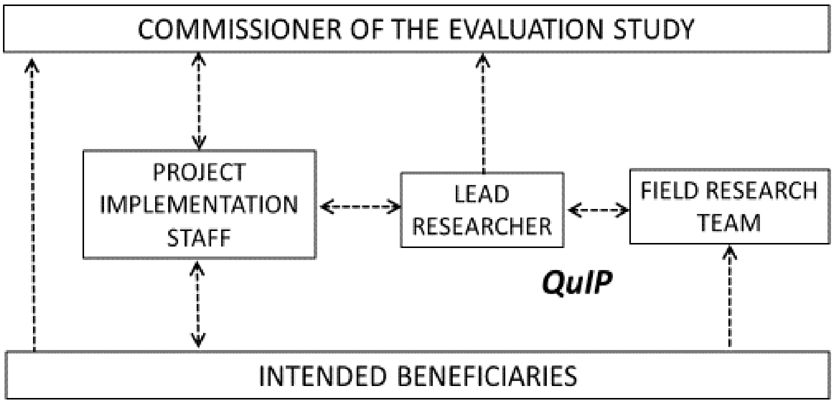
COMMISSIONER OF THE EVALUATION STUDY
PROJECT IMPLEMENTATION STAFF
LEAD RESEARCHER
FIELD RESEARCH TEAM
QuIP
INTENDED BENEFICIARIES

Table 1. ART Project: case study projects

Interventions (X)	Impact indicators (Y)	Confounding factors (Z)
Project 1. Groundnut value chain (Central Malawi). Project 2. Climate change resilient livelihoods (Northern Malawi). Project 3. Malt barley value chain (Southern Ethiopia). Project 4. Climate change resilient livelihoods (Northern Ethiopia).	Food production Cash income Food consumption Cash spending Quality of relationships Net asset accumulation Overall wellbeing	Weather Climate change Crop pests and diseases Livestock mortality Activities of other organisations Market conditions Demographic changes Health shocks

Source: Prepared by authors from ART Project data

Table 2. Ten design features of the QuIP.

	Characteristic	Commentary
1	Blind interviewing Data collection by independent field researchers, without any knowledge of the implementing agency, project or its theory of change.	This entails a division of roles between a lead evaluator and field researchers, with the former acting as an intermediary and a firewall between field researchers and the commissioner of the study.
2	Sampling Stratified random selection of respondents from lists of known beneficiaries of project activities. No need for a control or comparison group.	The lead evaluator again acts as intermediary: agreeing the sampling strategy with the commissioner and passing on beneficiary lists (and contact details for them) to the field researcher.
3	Data collection methods Semi-structured household interviews and focus groups, ideally to complement quantitative monitoring of change using other methods.	Focus groups are stratified to elicit gender and age disaggregated perspectives to complement and triangulate household interview data.

4	Data collection instruments Alternating open and closed question sections for selected impact domains.	Probing questions invite respondents to offer open-ended accounts of the main drivers of change in specified domains. Closed questions allow respondents to sum up whether the overall change was positive or negative for them.
5	Data entry Typed direct from interview records onto pre-formatted Excel sheets to facilitate coding and analysis.	Ability of field researchers to note and type up responses from conversations conducted in local languages avoids additional costs of full transcription and translation.
6	Coding of impact evidence The analyst highlights and codes any text explicitly or implicitly describing project impact (positive or negative), or incidental to project impact.	Explicit evidence refers clearly to the project. Implicit is consistent with the project's theory of change. Incidental is a reality check on other drivers of change, and of confounding factors.
7	Coding of drivers of change Additional coding of positive and negative drivers can be either inductive, based on project theory or both.	Scope for cross-tabulating against data on which project activities the selected households participated in and when.
8	Report generation Excel formulas enable coded data to be sorted and summarised in tabulated form.	Semi-automation speeds the process of doing this. Summary tabulation allows quick assessment of the frequency of different responses as well as an index for checking sources.
9	Data auditing Annexes of sorted source data permit easy auditing of evidence behind identified impacts and other drivers of change.	This opens up the 'black box' evidence behind data analysis, and allows virtual immersion of INGO staff in the perceptions of respondents. It also allows data checking and provides quality assurance.
10	Debriefing Discussion of findings involving researchers and project staff.	Staged unblinding can deepen analysis and provides additional quality assurance.

Source: Prepared by authors from ART Project data