Managing relationships in qualitative impact evaluation of international development: QuIP choreography as a case study.

[17 October 2017]

James Copestake[1]
University of Bath, UK
Claire Allan
Farm Africa, UK
Wilm van Bekkum
Self Help Africa, UK
Moges Belay
University of Addis Ababa
Tefera Goshu
Ambo University, Ethiopia
Peter Mvula
University of Malawi
Fiona Remnant
Bath Social and Development Research Ltd, UK
Erin Thomas
Gorta Self Help Africa, Ireland
Zenawi Zerahun
Mekele University, Ethiopia

Abstract

Who does what and when during an impact evaluation has an important influence on the credibility and usefulness of the evidence generated. We explore such choreography from technical, political and ethical perspectives by reflecting on a 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 'unblindfolding' as a form of triangulation. We argue that these steps can enhance credibility of evidence, and that ethical concerns associated with them can be addressed by being open with stakeholders about the process. The case study illustrates scope for better use of qualitative impact evaluation methods in complex international development contexts.

Key words

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

[1] **Corresponding author**: James Copestake, Centre for Development Studies, Department of Social and Policy Sciences, University of Bath, Bath, UK. Email: j.g.copestake@bath.ac.uk

Introduction

Impact evaluation has attracted growing attention in international development practice as a means to promote learning, improve operational effectiveness and strengthen public accountability. Here we define it broadly to cover the processes of collecting, interpreting and using evidence on the effects of a specified activity or project (cf. White, 2010). This typically involves a mix of hierarchical and collaborative relationships between the commissioners of an evaluation, researchers, operational staff, intended beneficiaries and other stakeholders. The paper contributes to the empirical literature on impact evaluation in development practice as a *social* process (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 technical, cultural and ethical tensions arising from control of *who* gains access to information through the evaluation process and *when*. The word choreography seems appropriate since both dance and impact evaluation aim to reflect on life experiences perceptively, critically and aesthetically through a planned and coordinated sequence of steps (Clammer & Giri, 2017).

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; Picciotto, 2017). 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 evaluators and researchers advocate a relatively formal separation of interviewer and subject, others have sought to enhance credibility through closer immersion in the lives of their subjects. The 'reality check' approach, for example, "… puts intimacy, immersion and consensus at its core" (Camfield, 2014:19; see also Jupp, 2016, and Arvidson, 2014). Manzano (2016:351) further illustrates the range of possibilities by contrasting 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.

These issues are examined further in this paper through reflections on collaborative action research project 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 and climate change adaptations projects in complex rural African contexts (Copestake, 2014). The approach tested was prompted in part by increasing use in development practice of more quantitative impact assessment methods, including randomized control trials. One risk associated with these approaches is that the need to identify measurable indicators of treatments and outcomes in advance risks influencing selection and design of the development interventions evaluated – or the methodological 'tail' wagging the development 'dog' (Camfield and Duvendack, 2014; Eyben *et al.* 2015). To counter this risk there is growing recognition of the case for more open, flexible and adaptive approaches to impact evaluation that can be used in complex and uncertain contexts (Stern *et al.,* 2012; World Bank, 2015:199). This entails finding more flexible alternatives to the attribution problem at the heart of impact

evaluation. One approach is to rely for evidence on self-reported causal claims or statements made by the intended beneficiaries of selected projects, rather than on statistical inference based on observational data about respondents subject to variable degrees of project exposure. Taking a more qualitative approach, however, is open to the criticism that findings are more susceptible to project specific confirmation and related response biases (White and Phillips, 2012). In response to this, the pilot research reported below entailed deliberately concealing details of the project activities being evaluated from researchers and respondents. This in turn poses an ethical dilemma: even temporarily restricting who knows what and when from relevant information about an evaluation may make good technical sense (by rendering data and evidence more credible to some users), but it contravenes the ideal of maximising transparency to all actors at all times.

The paper explores this issue as follows. We first elaborate on the wider context by linking the question of impact evaluation choreography to different theories of international development practice. We next briefly describe the case study research, and reflect on social relations within it from the perspective of field researchers, project staff and intended beneficiaries in turn. The final section draws out more general conclusions for evaluation and international development practice.

Impact evaluation and international development practice

Our discussion of social relationships in impact evaluation is embedded within deeper tensions over the management of international development. Reviewing the literature on development partnerships Stevens *et al.* (2013) distinguish optimistic and pessimistic strands, echoing a distinction made by Gulrajani (2010) between realist and radical perspectives on development management. Combining the two it is useful to contrast 'optimistic-reformist' and 'pessimistic-radical' perspectives. The first applies universal principles of effective performance management to achieve measurable objectives through rational, planned and voluntary collaboration between stakeholders. Impact evaluation is a means to learn and to improve, and for funders to ensure that they are securing value for money. The second places unequal power at the heart of development and emphasises evaluation as a stage for political struggle (Eyben *et al.*, 2015; Hayman *et al.*, 2016). A third 'realist-romantic' position views management practice as a process of shared discovery, consensus building and communicative action (Picciotto, 2017). Evaluation, within this tradition is "developmental" (Patten, 1994) and entails mutual learning that arises from the tensions between professionals' interests, funders' claims for feedback, and the rights of intended beneficiaries to know what is being spent in their name, how and to what effect.

These different positions can be further clarified by reflecting on the challenge posed by the *complexity* of development practice (Patten, 2011; Pawson, 2013; Bamberger, 2016). Optimistic reformists respond with ever more elaborate models to aid identification of 'optimal' choices. For realistic romantics, complexity opens up the possibility of more holistic understanding and emergent solutions, achievable through collaboration, trust building and

reciprocal illumination, including through the use of multiple methods and triangulation of findings. For pessimistic radicals, in contrast, complexity can both accommodates and accentuate divergent and competing perceptions, discursive practices and ideologies. Grint (2005) warns that the very decision to characterise a problem as complex (or "wicked") may itself be a device for legitimising power and exercising leadership.

As a simple illustration, consider the problem of how to evaluate what students have learnt from a programme of study. This can be viewed as a purely technical problem of testing ability to draw up definitive answers to 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 over students. 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 can legitimate assessment schemes both by contributing to consensus building (by sharing marking schemes, for example) and as a precondition for error-correction through rights to peer review and of appeal. But transparency is also hierarchically choreographed: when markers are instructed to mark blind, for example, or required to remain anonymous. The general point is that understanding the social relations of assessment cannot be divorced from wider tensions between rationality and power within a programme of study.

Case study

This section presents a specific case study through which to explore these issues in more depth. It is set within the wider institutional context of "programme partnership agreements" (PPAs) under which the UK Department for International Development (DFID) offered core funding to selected INGOs in return for better evidence of their social impact (Coffey, 2012). These PPAs were framed in the optimistic-reformist language of management-by-results indicative of a strong top-down impetus to the demand for impact evidence. However, DFID did not specify what sort of data INGOs should collect or how, thereby opening up space for discussion about how impact evaluation work could be developmental and exploratory, as well as confirmatory (Copestake, 2014). In reviewing options for doing this INGO staff were able to draw on a vast range of methodological possibilities and sources of expert advice. Indeed the range of options was itself problematic. Specialist evaluators and academics were deemed to have their own methodological interests and biases, potentially adding to costs. Reliance on external experts also risked limiting the scope for linking impact evaluation with INGOs' existing data and performance management systems. These factors widely contributed to uncertainty over the likely credibility, cost and usefulness of impact evaluation, reducing incentives to invest in it beyond the minimum necessary to comply with external funding requirements (Copestake et al., 2016).

These issues prompted two of the INGOs entering into a PPA - *Self Help Africa* and *Farm Africa* - to support an action research approach to the challenge of identifying an appropriate form of impact evaluation. The project (entitled Assessing Rural Transformations, or the ART Project) entailed collaboration with academic staff at Universities in the UK, Ethiopia and

Malawi, as well as a specialist NGO called *Evidence for Development*.[1] Its primary purpose was methodological – to assist the two INGOs by designing and testing a more useful and cost-effective approach to impact evaluation of diverse and complex project activities part-funded through the DFID funded PPA. More specifically, the ART Project set out to explore qualitative approaches to addressing the attribution problem in a way that could be useful for other agencies. The mechanism for doing so was to design and pilot a qualitative impact protocol (named the QuIP) appropriate to assessing project interventions aimed at promoting household level food security in the context of complex rural transformations associated with climate change and rapid market commercialisation.

The research started with a collaborative design workshop in May 2013. In the second year, the draft QuIP was piloted through studies of four 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 respondents encouraged to share their perceptions of the main drivers of change they had experienced over the previous two years. Findings were written up and reviewed at feedback and dissemination workshops in Addis Ababa and Lilongwe in July 2015.

Insert Table 1 about here.

This paper was first drafted by the lead author, who was also the ART Project's principal investigator. It additionally draws on unpublished notes and feedback from the other named authors, and written accounts of the Addis Ababa and Lilongwe workshops. By July 2017 the QuIP has been utilized to conduct fifteen further studies in ten countries, as documented on the website of Bath Social and Development Research (BSDR Ltd), a social enterprise set up to promote its wider use (BSDR, 2017). This is indicative of the feasibility of the approach to a wide range of contexts. However, since the purpose of this paper is methodological, empirical findings are not reproduced here.[2] Instead Table 2 highlights ten key characteristics of the final version of the QuIP. Discussion focuses particularly on its more innovative features: blinded interviewing (Step 1), coding (Steps 6&7) and triangulation through unblindfolding (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 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.[3] The nature and extent of such bias is unknown, but its possibility

is nevertheless 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 could explain, for example, a tendency for respondents in Ethiopia to emphasise the positive impact of government initiatives.

Insert Figure 1 about here.

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* data sharing and sense-making (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, 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 unmasked 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.

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 relied 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 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.[4] 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 a utilitarian or '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.

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. 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.

In most cases a two stage clustered sampling strategy was employed: purposive selection of localities, followed by random sampling of lists of farmers or households located within them drawn either from lists of project beneficiaries or households covered by a

baseline monitoring 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. The separation of sample selection from interviewing was necessary to maintain the double blinding, while allowing the INGO to be fully involved in debate over how best to do it and why. The field research teams' experience of then having to locate respondents without any help from the commissioning INGOs varied considerably according to the extent and reliability of contact details made available to them, including sketch maps and cell phone numbers. Physical geography and weather were also major determinants of the time required to locate and reach respondents and to arrange focus groups.

Once located, and after the purpose of the study was explained to them, respondents rarely displayed any reluctance to participate: affiliation of members of the research team to a local university, combined with their cultural sensitivity and experience, providing sufficient authority and reassurance. Nor did lack of reference to any specific project or activity discourage 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 them 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 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 relevant to 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 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) a useful "reality check" and "deep dive" into whether selected projects were achieving their intended goals, (b) as an investment in internal learning, and (c) as a way of demonstrating to their donors 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. In Malawi, jokes were also shared about the arrival of "ghost researchers" to go with the "ghost beneficiaries" and even "ghost villages" associated with government input subsidy programmes. At the same time, field staff 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 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 relying on specialist analysts and ensuring their work is transparent and auditable. To illustrate, a common issues is for a statement to combine both positive and negative elements: e.g. 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. One reason for poultry mortality that emerged from wider discussion was that they were being given to some people simply too vulnerable to be able to look after them adequately.

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 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.[5] 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 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 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 would 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.

Conclusions

In the introduction to this paper we argued for more research into the social relations of impact evaluation. We explored the difficulties facing INGO staff responsible for impact evaluation design, and focused on the issue of how the choreography of carrying out impact evaluations affects trade-offs between credibility, relevance, cost-effectiveness and ethics. 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 have 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 complex and dynamic contexts. 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 (even coercive) to participatory and egalitarian, via commercial and transactional.

In this wider context, the ART Project case study can be viewed as a 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. The more general point is that its development is an example of a consensual and deliberative process in the realist-romantic rather than reformist or radical spirit of development practice. Drawing on a range of more generic approaches (including contribution analysis, process tracing, goal-free and developmental evaluation) it aimed to develop more detailed guidelines to address the specific needs of the participating INGOs. The idea of designing such 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 – PADev being a parallel example (Pouw et al. 2016). To use a market analogy, 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. While introducing another branded product risks adding to the confusing alphabet soup of acronyms, this can be offset through specifying what it is with sufficient clarity and precision to facilitate detailed and critical comparison with alternatives.

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 both clinical trials and 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 or masking*), 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 research into the benefits and costs of extending such activity to include primary respondents also. 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.

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 more open, flexible and deliberative (or romantic-realist) approach to development practice.

This brief discussion illustrates further why research into impact evaluation needs to combine attention to technical and ethical aspects of different methods with attention to them as social process in specific contexts. This paper focuses particularly on the choreography of impact evaluation: not only who needs to know what, but also when. 'Who' refers here to the hierarchy and networks of 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'. Finally, contextual complexity and procedural uncertainty helps to support the case for investing in collaborative and experimental action research approaches to impact evaluation, framed by a romantic-realist view of development practice.

Funding

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).

Endnotes

[1] The ART Project ran from 2012 to 2016, and was funded by DFID and the UK Economic and Social Research Council (ESRC) - see http://www.bath.ac.uk/cds/projects-activities/assessing-rural-transformations/index.html. It also incorporated quantitative monitoring of changes in food security using the individual household survey method developed by the NGO *Evidence for Development* and described at www.efd.org.

[2] 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.

[3] 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. Confirmation bias is also more generally defined as selectivity in collection and analysis of data in order to support previously held beliefs (World Bank, 2015:182).

[4] 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.

[5] For the actual wording see Page 24 of the QuIP Guidelines at http://qualitysocialimpact.org/wp-content/uploads/2016/05/QUIP-Full-Guidelines-English-April-2016.pdf)

References

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.

Bamberger, M., Voessen, J., Raimondo, E., editors. (2015) *Dealing with complexity in development evaluation: a practical approach*. London: Sage.

BSDR. (2017) *Who we work with*. Bath Social and Development Research Ltd. http://bathsdr.org/about-bsdr/who-we-work-with/ [accessed 17 Oct 2017].

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.

Clammer, J., & Giri, A., editors (2017) *The aesthetics of development: art, culture and social transformation*. London: Palgrave macmillan.

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.

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.

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.

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

Patton, M. (1994) Developmental evaluation. *Evaluation Practice*, 15(3):311-319.

Patton, M. (2010). *Developmental evaluation applying complexity concepts to enhance innovation and use.* New York, NY: Guilford Press.

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

Picciotto, R. (2017). Evaluation: discursive practice or communicative action? *Evaluation*, 23(3):312-322.

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.

Stern, E, Stame, N, Mayne, J, Forss, K, Davies, R, & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations*. London. Department for International Development.

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.

White, H., & Phillips, D. (2012). A*ddressing attribution of cause and effect in 'small n' impact evaluations: towards an integrated framework*. London: International Initiative for Impact Evaluation.

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

Figure 1. Stakeholder relationships under a QuIP study

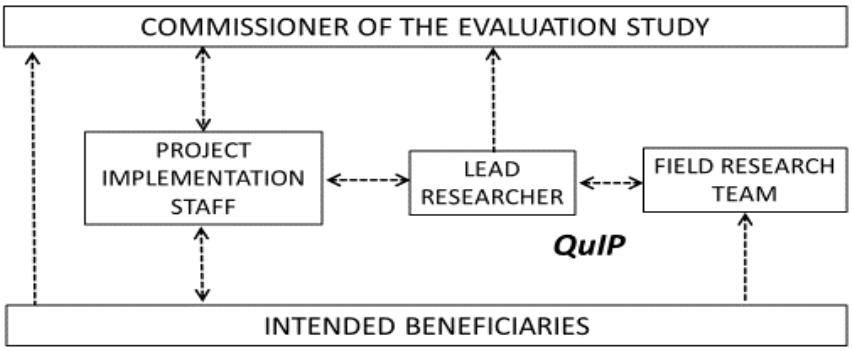

Table 1. ART Project: case study projects

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

Contributors

James Copestake is Professor of International Development at the University of Bath, UK.

Claire Allan is Head of Programme Quality and Impact at Farm Africa, UK.

Wilm van Bekkum is Monitoring and Evaluation Advisory for Self Help Africa, UK.

Moges Belay is a Lecturer in Development Studies at the University of Addis Ababa, Ethiopia.

Tefera Goshu is Lecturer and Head, Dept. of Sociology & Social Work at Ambo University, Ethiopia

Peter Mvula is a Senior Research Fellow in Rural Livelihoods at the Centre for Social Research, Chancellor College, University of Malawi.

Fiona Remnant is Director of Bath Social and Development Research Ltd, UK

Erin Thomas is Deputy Monitoring and Evaluation Advisor for Gorta Self Help Africa, Ireland.

Zenawi Zerahun is Dean of the College of Social Sciences and Languages at Mekelle University, Ethiopia.