Credible impact evaluation in complex contexts: confirmatory and exploratory approaches.

James Copestake
Centre for Development Studies, University of Bath
Draft. 18 October 2013. Comments welcome to j.g.copestake@bath.ac.uk

Abstract

Debate continues over how best international development agencies can evaluate the impact of actions intended to reduce poverty, insecurity and vulnerability in diverse and often complex contexts. There are strong ethical grounds for simply asking those intended to benefit what happened to them, but it is not obvious how to do so in a way that is sufficiently free from bias. This paper explores an approach that combines eliciting such feedback using qualitative methods with quantitative monitoring nested within realist enquiry. In so doing it highlights the tensions between confirmatory and exploratory methodological approaches to the design of a suitable qualitative impact assessment protocol. Confirmatory approaches, based on predetermined theories of change, risk accentuating pro-project bias in a way that can be self-defeating. The paper explores how this can be avoided, as well as possible sources of bias in selection of evaluation methodology.

Key words
Impact evaluation, qualitative methods, development practice, complexity, exploratory and confirmatory analysis, pro-project bias
Introduction

This paper addresses the perennial question of how international development agencies, whose work links people across vast distances and areas, can credibly evaluate the impact of their diverse policies, programmes, strategies, projects and practices relative to declared goals, including the MDGs. The paper draws mainly on collaborative action research with two NGOs – Farm Africa and Self Help Africa (SHA) - into how best they can assess the impact of their programmes promoting sustainable smallholder agriculture and improved food security across twelve countries. However, it is also motivated by my involvement in an on-going evaluation commissioned by the Netherlands Government of its ‘MFS-2’ joint funding commitments of nearly €2 billion Euros to Dutch development NGOs for the period 2011 to 2015: a study that includes baseline and repeat surveys of 158 partner programmes across eight countries with an initial budget of just under €10 million (van der Gaag, W, & Rongen, 2013).

These and other development activities confront political and methodological issues familiar to those interested in impact evaluation of public policies and programmes in many other contexts, including the role of impact evaluation not solely for the purposes of organisational learning but also for building legitimacy through both upward and downward accountability. International development practice is also embroiled in wider debates over the politics of evidence, and results-based management culture (e.g. Eyben, 2013; Green, Roche, Eyben, Dercon, & Witty, 2013; Gulrajani, 2010). In contrast, this paper focuses mainly on narrower methodological issues, particularly the problem of attribution, or how particular outcomes - such as improvements in household level food and nutritional security - can reliably be linked to specific projects, interventions or mechanisms in different
contexts. However, technical and political aspects of impact evaluation can never be fully untangled, and the discussion of methodological issues in this paper are framed by concern with upward accountability on the part of non-government development agencies reliant on funding from government and the public in more affluent countries. This influences, for example, what is meant by ‘credible’ impact evaluation and how much money is realistically available to spend on it.

Before turning to the question of attribution, a brief reference is appropriate to the principle of directly involving intended primary beneficiaries of a development intervention in its evaluation. If we are interested in finding out whether particular men, women or children are less poor as a result of some action it seems common-sense just to ask them. But even putting aside problems of construct validity (over the definition of poverty, for example) it is not obvious how easily they will be able to attribute changes in their experience to specific activities. And there may also be reasons to doubt the reliability of their responses, including strategic bias, confirmation bias (Haidt, 2012:93) and “anchoring” of responses to what is expected (Kahneman, 2011). In this paper I use the term pro-project bias to refer to the possibility that someone will consciously or otherwise conceal or distort what they think they know about a project or activity for whatever reason, including the hope that doing so will reinforce the case for keeping it going. The instrumental value of asking people directly about attribution (benefitting from their close knowledge, but subject to such bias) is ultimately an empirical question, albeit a hard one. But there are also ethical reasons for asking people directly, captured in the rallying cry of the international disability movement ‘nothing about us without us’, and explored in the context of international aid by Anderson et al. (2012). While intended beneficiaries may find ex post consultation a weak substitute for having a role in planning and
sanctioning work done in their name, neither is lack of such involvement a reason for excluding them from evaluation. In sum, the direct involvement of at least some intended beneficiaries in impact evaluation is ethically important, even if it presents methodological challenges. How then, (if at all) can self-reported assessment of impact and its causes assist directly in tackling the attribution problem that lies at the heart of impact evaluation?

**Approaches to impact evaluation**

In relating the issue of self-reported impact assessment to the wider literature I will make two heuristic distinctions: between more quantitative and qualitative approaches to evaluating impact, and between more exploratory and confirmatory approaches. There is of course no shortage of literature on the first of these, including recent reviews of the status of qualitative and mixed approaches in an international development context (Shaffer, 2013; Stern et al., 2012; White & Phillips, 2012). A quantitative approach can be linked to an axiomatic view of impact as the difference in the value of an outcome indicator \( Y_1 \) for a given population after a particular intervention or ‘treatment’ \( X \) compared to what the value would have been for the same population if the treatment had not occurred \( Y_0 \) (White, 2010, p. 154). Putting aside the problem of consistent measurement of \( X \) and \( Y \), a central issue is then how to establish a plausible counterfactual. If the evaluator can make a large number of observations of \( X \) and \( Y \) then there are a range of well-known quantitative approaches to data collection and analysis, including the use of randomised control designs. In contrast, this paper focuses on the scope for more qualitative approaches, on the grounds that there has been less clarity, consistency and consensus about how best to employ these (or ‘small \( n \)’ approaches) in
evaluating the impact of development activities (Stern et al., 2012; White & Phillips, 2012). In passing, it is germane to the argument of this paper to note that even randomisation is not of itself a guarantee against pro-project bias in estimates of impact (White, 2010, p. 156). This is particularly the case if \( Y \) is obtained from respondents (and/or by researchers) who are not blind to whether they belong to the treatment or control sample, and may therefore be prone to different degrees of response bias, including Hawthorne and John Henry effects (Duvendack et al., 2011).

Among the various criticisms of quantitative approaches that rely on experimental or quasi-experimental designs (e.g. Cartwright, 2011; Deaton, 2010; Picciotto, 2012) perhaps the most important concerns are not formal validity but practical applicability.\(^1\) In complex, diverse, fast changing, emergent and recursive social contexts they are often too cumbersome and expensive. It may be possible to measure a large vector of variables \( Y \) for a given population and time period, and to demonstrate how they are affected by exposure to a vector of interventions or treatments \( X \). But each set of results is specific in time and space to a vector of confounding or contextual variables \( Z \) that is too small or too quickly becomes outdated in history (see for example, Pawson and Tilley (1994), and the ensuing exchange with Bennett). Realist evaluation emphasises the need for a cumulative process of broadening understanding of context-mechanism-outcome interactions or knowledge of "...what works for whom in what circumstances, in what respects, over which duration... and why" (Pawson & Manzano-Santaella, 2012, p. 177). Its pursuit of realism can be viewed as being achieved directly at the expense of the quantitative precision gained from artificially restricting variation in treatment and contextual vectors in order to generate statistically significant results (Levins, 1966).
Realist evaluation constitutes only one possible counterpoint to positivist approaches, and can be linked with a broader range of ‘theory driven’ and ‘theory of change’ approaches that emphasise the importance of building evaluation on prior elucidation of programme theory linking mechanism to outcomes (Pawson & Manzano-Santaella, 2012, p. 178; Stern et al., 2012; Ton, 2012). White and Phillips (2012, p. 4 & 34) identify a cluster or group of qualitative approaches, that all “involves the specification of a theory of change together with a number of further alternative causal hypotheses.” They also identify a second group of approaches that “place stakeholder participation at the heart of data collection and analysis”, but conclude that these do not make causal explanation their primary goal, are prone to various biases arising from their reliance on stakeholder perception, and are most usefully employed only as one element within a wider evaluation framework (p. 21).²

This brings me to the second and less widely discussed heuristic distinction explored in this paper: between confirmatory and exploratory approaches to impact evaluation. At first glance this distinction seems close to the one between the two groups drawn by White and Phillips. Their first group can be viewed as confirmatory in the sense that it seeks evidence to either validate or challenge the researcher’s prior theory, and thereby downplays those of other stakeholders. If so, then this is contrasted with participatory approaches which a priori give more weight to the perceptions of other stakeholders. An alternative classification would be to contrast confirmatory small n approaches with those that are more open-ended and exploratory in the sense of explicitly downplaying prior theorisation on the part of the researcher. Combining the two distinctions leaves us with the four possible approaches depicted in Table 1.
White and Philip’s assertion that participatory approaches II and IV are more prone to bias than I and III seems to boil down to an axiomatic statement of faith in the professional judgement of ‘independent’ evaluators, including their capacity to rise above the interests of other stakeholders. This takes us back to questions about the wider politics of impact evaluation that fall beyond the scope of this paper. However, even without challenging the status of the evaluator (thereby ignoring the second column) there remains the possibility that Type III studies could be more appropriate to some tasks than Type I. More specifically, it seems worth considering whether there may be a case for impact evaluation in which the researcher uses more grounded, inductive and exploratory methods, and in which theory emerges through interaction between (and reflection upon) prior theory and elicited stakeholder perceptions, rather than one being axiomatically subordinated to the other. Going further, there does not appear to be any a priori reason why such an exploratory approach should be less credible than a more confirmatory one. Indeed, in contrast to evaluation most qualitative social science research is almost by definition more open-ended, naturalistic and exploratory. A key question is then how far standards of rigour within social science research more generally can be applied to the more tightly scripted context of development impact evaluation.

Table 1: A typology of approaches to ‘small n’ impact evaluation

<table>
<thead>
<tr>
<th></th>
<th>Privilege the professional judgement of the evaluator</th>
<th>Give at least equal weight to the views of other participants or project stakeholders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmatory</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Exploratory</td>
<td>III</td>
<td>IV</td>
</tr>
</tbody>
</table>
To further explore the distinction between confirmatory and exploratory approaches it is useful to be able to review a specific impact evaluation method that falls in category IV in the typology of Table 1. Participatory Assessment of Development (PADev) has been developed over six years by a consortium of Dutch and West African NGOs and researchers coordinated by Ton Dietz at the African Studies Centre at the University of Leiden (T. Dietz, 2013). A core objective of PADev, as set out in published guidelines building on eleven field workshops (T. Dietz et al., 2013), is to address the limitations of impact evaluations that artificially focus on the activities of just one project or agency in areas that have been affected by multiple and overlapping interventions. This problem is particularly acute for those who regard development as the outcome of highly interconnected, hierarchically nested and multi-dimensional systems of activity, and are sceptical of how far it is possible or meaningful to identify (still less quantify) the isolated causal contribution of just one isolated intervention (Bevan, 2013; Patton, 2011; Scheifer, 2008). PADev’s solution to this problem is collective action in the form of joint evaluation of all development interventions in a specified locality over a long period of time. The guidelines offer a suite of interconnected and structured participatory exercises aimed at enabling a reasonably representative cross-section of the population to construct nothing less than a joint history of development in and of the locality over a period of up to thirty years. These activities include constructing a time line of major events, wealth ranking, evaluating long term changes in wellbeing, identifying significant interventions and ranking them according to both immediate and long-term impact. While open-ended and exploratory, the use of preformatted Excel tables for recording outputs aids relatively rapid and structured generation of findings for evaluation purposes.
The PADev method inevitably entails expert facilitation, but also goes to considerable lengths to reflect the views of a cross-section of local people, including the poorest. By moving the spotlight away from any one agency it can claim to go some way towards reducing pro-project bias, even if the possibility of some positive bias towards development itself is hard to completely eliminate, alongside deference to the outsiders mediating the assessment (T Dietz, 2012, pp. 233-234). A third challenge facing this approach is that it entails producing a public good and hence is susceptible to free rider problems. More specifically, individual development agencies may struggle to justify investing in an impact evaluation little of which relates specifically to their own activities: what is gained in terms of contextualisation and balance being offset by reduced detail about their own activities. Overcoming this problem by pooling the evaluation resources of several agencies operating in particular places is possible but difficult, though not insuperable as the case of joint donor funding of the WIDE project in Ethiopia also illustrates (Bevan, 2013). Hence, while inspired by the PADev experience, this paper advocates research into a broader range of exploratory impact evaluation methods. More specifically, while echoing PADev in taking a broader contextual approach in order to mitigate pro-project bias it explores the scope for methods that retain an ultimate focus on the impact of a particular agency or project. But before doing so, it is worth exploring in more depth what is meant by impact evaluation, as well as the criteria for assessing the strengths and weaknesses of different impact evaluation approaches.
Defining credible impact evaluation

White (2010, p. 154) notes that the term impact evaluation is widely used to refer both to any discussion of outcome and impact indicators, and more narrowly to studies that explicitly seek to attribute outcomes to a specified intervention. This paper adopts the second definition, while at the same time taking a broad view about what constitutes a sufficient level of scientific rigour in addressing the attribution problem. Doing so allows for the possibility that specific methods (including those within a positivist tradition) can be nested within broader (including interpretive) approaches. This acknowledges that attributing impact is only one question (if perhaps the most important) that impact evaluations need to answer, another being how an intervention works (Stern et al., 2012, p. 36).

In using the term ‘credibility’ I refer to one party being able to offer a sufficient combination of evidence and explanation to convince another party that something is true, or at least plausible enough to act upon. This recognises that, as a servant of action in a constantly changing historical context, the scientific rigour of impact evaluation has to be weighed alongside cost, timeliness and fit. Anticipating the final section of this paper it also echoes McGilchrist’s distinction between reasonableness and rationality (McGilchrist, 2010). But I do not think this definition entails abandoning the quest for greater consensus about what constitutes quality in qualitative research (e.g. as discussed Hammersley, 2013, p. 83). And while credibility emphasises the importance of context and trust, the rigour with which conclusions about impact are logically derived from stated evidence and assumptions clearly remains important.

A common distinction here is between the validity of an approach, and the reliability of results arising from its application in a particular context (e.g. Lewis &
Ritchie, 2003, p. 270). Given that no impact evaluation can ever be replicated in precisely the same context or setting in time and place, the value of this distinction to qualitative research is questionable, and in the context of international development it is rather analogous to nebulous attempts to distinguish policy from implementation, or theory from practice. The core problem here is again the underlying complexity of the context, where complexity can be defined as a setting in which the influence of $X$ on $Y$ is confounded by factors $Z$ that are (a) impossible fully to enumerate, (b) of uncertain or highly variable value, (c) difficult to separate, and/or (d) impossible fully to control. Additional complexity arises if the nature and value of $X$ and/or $Y$ is also uncertain.

Rather than distinguishing between validity and reliability, an alternative approach to defining credible attribution in impact evaluation is to focus directly on causation. Here I define an evaluator’s claim to establishing impact (i.e. $X$ causing $Y$ in particular contexts) as being credible if: (a) there is strong evidence that $X$ and $Y$ happened in such contexts; (b) $X$ is described by a diverse range of stakeholders as having been a necessary cause of $Y$ in those contexts$^4$; (c) their explanations of the mechanism by which $X$ caused $Y$ in those contexts are independently arrived at and mutually consistent; (d) the counter-hypothesis that they have other reasons for making the statement can reasonably be refuted. Additional, but perhaps not necessary is; (e) a shared understanding of how much $Y$ varies according to exposure to $X$ that is consistent with available evidence.

One further way of exploring the question of credibility is to consider how far any impact evaluation addresses different risks of bias. The above definition addresses this implicitly by proposing structures and processes of evaluation that reduce the plausibility of complicity among different stakeholders. This falls short of
scientific certainty, but in complex contexts it is often as much as we can hope for, particularly given the possibility that efforts to aim higher may be counterproductive in terms of cost, timeliness and usefulness. I am not suggesting that the definition of credible attribution is universal or even widely accepted, rather that it is a realistic one in contexts where overcoming the attribution problem is particularly difficult.

Case study: the ART Project

Origins
This section returns to the core issue of the tension between confirmatory and exploratory approaches to impact evaluation through a discussion of the process of drafting a qualitative impact protocol (QUIP) for credible or ‘good enough’ impact evaluation of NGO activities (hereafter referred to as projects) to promote smallholder agriculture and food security in the context of complex rural livelihood transformations taking place in many parts of Africa. The account covers the initial phase of the ‘ART project’, starting in November 2012. ART stands for ‘assessing rural transformations’ and the project is funded through a three year grant from the UK Economic and Social Research Council (ESRC) and the Department for International Development (DFID) as part of a call for research into ‘measuring development’. It originated from informal discussions between the author and staff of the NGO Self Help Africa (SHA) over growing demand for improved evidence of their impact, arising in part from programme partnerships with DFID and Irish Aid, but also a desire to promote better internal learning.

SHA had previously opted to strengthen their routine monitoring of economic and food security of intended beneficiary households using the individual household method (IHM) developed by the NGO Evidence for Development (EFD). This
approach is based on a combination of participatory rapid rural appraisal, structured household interviewing and real time analysis using standardised software. This generates estimates of how production, exchange and transfer entitlements (in cash and kind) of selected households compare with estimates of their food consumption needs based on standardised nutritional requirements and food conversion ratios. The estimates of adult equivalent entitlement relative to need for a cross-section of households generated by the IHM permit rapid simulation of the heterogeneous impact of price, output, income and other shocks within selected areas, as well as expected impact of project interventions.

While IHM seemed a promising method for monitoring change in key impact indicators and for simulating the effect of planned and unplanned livelihood changes on them, it is insufficient on its own to assess how far observed changes in these indicators can be attributed to SHA’s own work. The ART project was designed (in collaboration with SHA, EFD and Farm Africa) to address this gap by testing the hypothesis that credible impact evidence could be produced at an affordable cost by augmenting the IHM with a standardised qualitative impact protocol (QUIP). In contrast to established quantitative impact methods this aims to generate differentiated evidence (rather than average effects) linked to the individual testimonies of those interviewed, taking into account their variable exposure to project activities, but without the additional requirement to interview a control group. More specifically, we hypothesised room for improvement in use of qualitative methods with respect to sample selection, framing of interviews, structuring of open and closed questions, data interpretation and quality assurance practices.
The ART project has three empirical components. Strand 1 relies on repeat studies using the IHM to measure how cash and food budgets in selected households change in selected households over a three year period. Baselines were conducted for four projects (two in Malawi and two in Ethiopia) between November 2012 and May 2013. Strand 2 comprises piloting of the QUIP by independently contracted researchers, one and then two years after the project baseline. Strand 3 of the research constitutes an evaluation over three years of what evidence of impact different stakeholders elicit from Strand 2 relative to what they knew anyway, including what they learnt from Strand 1.

The next section reports briefly on issues arising (so far) in the design of the QUIP. A design workshop was conducted in Shrewsbury, UK, in June 2013 to produce draft QUIP guidelines for the testing phase under Strand 2 of the project. There were thirteen participants, including representatives from the three collaborating NGOs, Malawi, Ethiopia, Oxfam UK and Irish Aid. An initial draft protocol was produced and circulated by the author prior to the workshop. Each section of this was then subject to detailed discussion, leading to a further round of revisions to the protocol. The QUIP guidelines cover the following: purpose of the study, recruitment of the research team, sample selection, data collection methods (semi-structured interviews and focus groups), briefing and debriefing of the lead researcher, facilitating interviews, data analysis, use of findings and quality assurance. Separate guidelines for the lead researcher go into greater detail about the process of data collection.
Five sets of QUIP design issues

(1) Roles, relationships and the risk of pro-project bias. A central idea behind the QUIP is to collect open-ended narrative about changes in the lives and livelihoods of respondents through which evidence of the impact of a specific project can be inferred without explicit prompting. With this in mind the protocol seeks to minimise contact between researchers and project staff, so that the former are also as far as possible blinded to expected outcomes of the project. Researchers are not briefed about the project and to the extent possible remain ignorant of the theory of change behind it. This emphasis on avoiding pro-project bias appears to be in tension with the argument (discussed in the previous section) for placing project theories of change at the heart of impact evaluation to facilitate formulation of clear and testable impact hypotheses (in contrast, for example, to Ton, 2012). However, the tension is at least partly resolved in the way QUIP (as an impact assessment method) is nested within a larger multi-method evaluation approach, with a strong division of labour within the overall evaluation team.

Within the ART Project this is reflected in the drafting of separate QUIP guidelines for the commissioner and the lead researcher, as well as the contractual separation of responsibility for monitoring outcome indicators using the IHM as part of routing project implementation (Strand 1), independent application of the QUIP (Strand 2), and linking findings of both to each other and to project theory (Strand 3). The resulting channels of upward accountability are illustrated in Figure 1: solid arrows showing formal routes, and dotted arrows showing informal ones. Downward and horizontal communication (e.g. dissemination of findings from the lead evaluator to agency staff and intended beneficiaries) are also important, but omitted for simplicity.
Figure 1. Organisational chart for nesting the QUIP into an impact evaluation.

The diagram illustrates how clear demarcation of roles can contribute to a strategy for reducing pro-project bias and hence achieving greater credibility. However, the lead researcher and field team can never be fully blinded to the identity and character of the project they are evaluating, because it is both natural and morally necessary that they (and indeed respondents) should have some understanding of the reasons behind their questioning. Hence careful structuring of evaluation tasks can never fully substitute for research skills, integrity and professionalism. This adds to the importance of the prior task of selecting the best available lead researcher: methodological protocols can go so far, but ultimately real people have to apply them.

(2) Linking outcome monitoring and assessment. One potential weakness of relying on exploratory narrative accounts of impact (or what is also referred to above as
‘self-reported impact assessment’) is that it is unlikely to generate consistent or reliable estimates of the magnitude of impact on key variables. Quantitative monitoring, such as undertaken here by the IHM, cannot do this on its own, because it is not designed to tackle the attribution problem. However, it does provide an indication of the magnitude of change that might conceivably be attributed to an intervention. For example, if the IHM revealed that an indicator, $Y_1$, of household disposable income on average rose by 2% between baseline and a repeat survey it would still be possible for an intervention, $X$, to have an average impact of more than 2% by offsetting the negative impact of some confounding variable, $Z_1$, such as rainfall. Nevertheless, data on average changes in selected indicators does in practice provide a useful point of comparison for qualitative accounts of the mechanisms underpinning such change, ruling out extreme claims of impact unless these are supported by evidence and credible stories of large confounding shocks or trends.

(3) Timing, seasonality and recall. The QUIP primarily seeks insight into the pattern of impacts over time through respondents' own open-ended accounts of the timing of drivers of change in their lives and livelihoods across the years following a specific project intervention, not precluding views on important events over a longer period too. One way to facilitate this is to begin interviews by constructing a visual timeline of major life cycle related, seasonal events and shocks over the previous three years - including changes in income, expenditure and food availability (cf. T. Dietz, 2013; Locke & Lloyd-Sherlock, 2011; Shaffer, 2013). Reported changes can be compared with annual estimates of key indicators obtained through annual IHM surveys, either generally or at the level of individual households depending on sample overlap. This
is consistent with an inductive and exploratory approach to impact evaluation recognising that impact trajectories are non-linear and differentiated (Woolcock, 2009).  

(4) Selection of respondents within households. While household selection is widely discussed, there tends to be less discussion of who to interview within households. A key issue here is the possibility that projects may raise overall household cash income without necessarily also improving the welfare of all household members: for example, due to substitution of resources from food crops to cash crops, or indirect effects on gender and age specific work allocation. Multiple interviews within each household would provide greater depth and detail of information and gender sensitivity about such possible impacts, but at the cost of additional interviewing time, and with the extra complication of having to find ways to reconcile two or more potentially contradictory sets of data. Second interviews within each household can also be difficult to arrange (due to work related absences from home, for example), and risk creating or accentuating tensions within the household. Primarily for this last reason the QUIP protocol limits interviews to one primary respondent per household, without ruling out involvement of others at the discretion of the interviewer. At the same time interviews are augmented by exploratory gender and age specific focus groups to explore social dynamics. The rationale for this is that respondents are more likely to address sensitive questions about relationships in a setting other than their own home, and without being asked explicitly to refer to their own personal experience.
(5) Interview questions and data analysis. The critical issue here is to clarify the purpose, balance and sequencing of generative, supplementary and closed questions. Generative (more exploratory) questions are designed to stimulate discussion in a way that enables information about drivers of change (intended and unintended, as well as non-project related) to be volunteered without prompting. Supplementary questions sustain the conversation about this topic, but need not be used if the topic is raised anyway. Closed (more confirmatory) questions follow only after the chance for open-ended discussion of a topic has been provided: (a) to maximise the opportunity for respondents to raise unknown and unexpected issues; (b) because information about reasons for change provided voluntarily or without prompting is more credible. The closed questions provide a convenient way to round off discussion of a topic before moving onto the next – hence the case for spreading them through the interview rather than bunching them all at the end.

Linked to the balance between more open and closed questions is the reliability of data recording and analysis. Generally, the greater the exploratory and open-ended component, the more costly and complex the work tends to be. The social science ‘gold standard’ to record, fully transcribe, if necessary translate, and inductively analyse narrative interviews within a qualitative package, such as NVivo, is generally too expensive to replicate within impact evaluation. As an alternative the QUIP is testing an approach that relies mainly on note-taking by the interviewer (structured according to the interview schedule), who then writes it up as a structured narrative report. This allows the interviewer to write-up findings while still remaining relatively blind to prior expectations about project outcomes. A separate analyst can then assess this documented evidence against predetermined hypotheses, as well as responses to closed questions. Within the ART project itself
this second role will be performed as part of Strand 3 of the research. This opens up the possibility of using standardised impact scoring systems and of being able to test the consistency of results. It also raises political and moral questions about “who counts” (Holland, 2013).

**Discussion and conclusions**

This final section draws together the argument concerning the balance between confirmatory and exploratory approaches to impact evaluation. It notes that the preference between different methodological approaches is unavoidably subjective because the relative importance of different sources of potential error and bias is by definition uncertain. While it may be possible to identify particular contextual reasons for favouring one approach over another, this also leaves open the possibility that such choices might themselves be systematically biased.

In this paper I have sought to bring to the foreground a distinction between confirmatory and exploratory approaches to impact assessment on the grounds that it is neglected relative to other relevant dichotomies, including quantitative versus qualitative, and top-down versus participatory approaches. Confirmatory approaches feature more prominently in the impact evaluation literature, where there is a narrower focus on who needs to know what about particular interventions; whereas exploratory approaches have a closer affinity with impact research of a more open-ended and naturalistic kind (Copestake, 2013). More confirmatory processes can aspire to greater rigour by adopting a hypothetico-deductive approach based on predetermined theories of change. In contrast, exploratory approaches give greater weight to the danger of introducing bias towards project theory in the mind of both evaluators and researchers, including being blinkered from considering unintended
consequences. This suggests taking a more open-ended and inductive approach. Cost-effectiveness may appear to be best served by being able to focus impact evaluation narrowly and precisely on specific activities. But in more complex contexts there is the danger that such information will be overwhelmed or rendered redundant by contextual factors that were not even 'on the radar screen'. If so, then seemingly broader and less focused impact research may be very cost-effective indeed. The account of experience to date with designing a qualitative impact protocol within the ART project highlights how these tensions relate both to overall evaluation design and to narrower methodological details. But it also suggests there may be some scope for resolving these through a nested design that encompasses a more confirmatory overall approach with more exploratory components designed to provide additional information.⁹

What this discussion illustrates more generally is that the task of choosing between different evaluation designs entails weighing up potential sources of error and bias that are by definition uncertain. Some sources of bias – e.g. linked to statistical sampling and selection – can be quantified to some extent. This perhaps helps to explain why they get more attention, but if so this suggests cognitive bias towards the known over the unknown. There is also plenty of scope for such choices to be influenced by other factors.

To illustrate, recall the heuristic distinction between quantitative approaches to impact evaluation, relying on statistical inference to address the attribution problem, and more qualitative approaches, relying on qualitative identification of generative mechanisms including self-reported attribution. Distinguish further between sources of bias that arise from the validity of design (e.g. scoping, sampling, attribution system) from those linked to the reliability with which it is implemented
(e.g. due to inadequate training and/or weak incentives on the part of field staff). These simple distinctions can be used to produce the 2x2 matrix shown as Table 2. This also hypothetically suggests that in the face of uncertainty the proponents of each approach tend to adopt a “mental model” that lexicographically orders potential problems of bias in two contrasting ways.\(^{10}\) This suggests there can be quite deep cultural obstacles to achieving interdisciplinary consensus over choice of methods.

<table>
<thead>
<tr>
<th>Methodological approach</th>
<th>Possible source of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Design validity</td>
</tr>
<tr>
<td>Mostly quantitative</td>
<td>First order problems</td>
</tr>
<tr>
<td>Mostly qualitative</td>
<td>Second order problems</td>
</tr>
</tbody>
</table>

Where do these preferences come from? Personal preference for - and familiarity with - particular tools and techniques may be one source, as well as academic background, although the latter is perhaps mitigated by the tendency of disciplines to sub-divide internally to replicate external epistemological differences (Abbott, 2001). More contentious but intriguing are deeper possible connections with the way we think, as set out by McGilchrist (2010), for example. He draws on an extensive survey of the neuro-scientific evidence on hemispheric brain specialisation to suggest humans are all capable of thinking in two distinct if deeply interconnected and complementary ways.\(^{11}\) For most of us, the left hemisphere is associated with a more focused way of thinking that abstracts and simplifies, producing narrower,
more precise and certain models of the world. In contrast, the right hemisphere is
associated with open forms of attention and vigilance, alongside broader,
contextualising, and holistic ways of thinking. Most of the time, we bring both ways
of thinking together in ways that render the distinction between them invisible. This
confers immense potential evolutionary advantages: to be able to think narrowly (as
forensic hunters and gatherers) and broadly (as vigilant and agile evaders of other
hunters) at the same time, for example. But that does not necessarily stop
individuals from having a stronger predisposition towards one way of thinking or the
other. Hence, for example, more quantitative ways of thinking about impact
evaluation might fit more comfortably with left brain dominance (rational, abstract,
precise, generalising, certainty seeking, depersonalised), and qualitative approaches
with right brain dominance (reasonable, concrete, less certain, contextual, person-
oriented rather than idea-oriented, emphasising difference rather than sameness).

McGilchrist suggests one possible influence on the subjective mental models
that mould our methodological preferences. Making the additional jump to a
sociological level of analysis, these may also be reflected in shared mental models of
development, and imprinted in what Eyben (2013) refers to as the “institutional
artefacts” governing how collective responses are structured and managed.
Confirmatory approaches to thinking about impact evaluation are certainly more
congruent with a results-based shared mental model of doing development that is
certainty seeking and inclined to be more controlling. A tension between
confirmatory and exploratory approaches to impact evaluation is also congruent
with the wider tension between “planners” and “seekers” in development thinking
(Easterly, 2006).
While somewhat speculative, this discussion reinforces the more general point that a review of methodological bias in impact evaluation entails reflecting on variation in the understanding, perception and interests of at least four groups: (a) intended beneficiaries of an intervention, or primary stakeholders; (b) those implementing the activity, or secondary stakeholders; (c) commissioners and end users of impact studies; (d) researchers directly engaged in design, collection and analysis of the data. More specifically, this paper has been concerned with impact evaluation in diverse and complex contexts where it is possible that the multiple links between X, Y and Z are at least as well understood by intended beneficiaries as they are by the other stakeholders, but where they may be particularly susceptible to the influence of project implementers.

In making more reasoned methodological choices over impact evaluation approaches at least three contextual factors might favour more emphasis on exploratory over confirmatory methods. First, what level of credibility do commissioners and end-users require? Those more closely connected with an activity are more likely to be able to cross-check evidence of its impact against their own direct experience and other sources. At the same time familiarity may also be a source of bias. The QUIP aims to generate independent evidence that is credible to those without opportunities for cross-checking it against their own direct experience of the activity.

Second, how defined and certain are the project’s impact pathways? The QUIP aims to be better able to reflect complex, uncertain and insufficiently understood impact pathways. This contrasts with more quantitative approaches to impact assessment that work best when project ‘treatments’ and intended
outcomes are clearly understood in advance, thereby making it easier to codify and quantify them.

Third, what is more important: numerical estimates of the average or typical impact on intended beneficiaries; or evidence of the nature of impact and variation between respondents? Exploratory approaches such as the QUIP address the second. While they are likely to be less useful in generating strictly quantitative estimates of typical or average impacts more interviews can be added as necessary to capture different experiences of additional respondents. Sampling and questionnaire design can also be adjusted to focus a QUIP on more specific issues and/or sub-sets of intended beneficiaries, hence offering flexibility and fostering a more adaptable and cumulative approach to learning appropriate to faster moving issues.

Funding and acknowledgements

Production of this paper was supported under research grant ES/J018090/1 from the Department for International Development (DFID) and the Economic and Social Research Council (ESRC). I am also grateful Fiona Remnant for research support, and to participants at the ART project methodology workshop held in Shrewsbury in June 2013, as well as at a joint UEA/IDS workshop on mixed methods research held in London on 1-2 July 2013.

Notes

1 Other issues include the following: (a) the need to achieve a minimum sample size as a precondition for making any inference about impact at all; (b) the emphasis on typical or average effects across samples or sub-samples, hiding what may be important variation within them; (c) ethical concerns about the need for involvement of respondents who receive no direct benefit; (d) the need to pre-commit to testing the outcome of a relatively small and uniform set of treatments (Vaessen, 2011; Picciotto, 2012).
2 Stern et al. (2012:25) also distinguish between ‘theory based’ and ‘participatory’ approaches, but also identify a third ‘case based’ approach, which can to some extent be mapped onto quadrant III in Table 1.

3 Also of interest, but beyond the scope of this paper is the potential for exploratory but primarily quantitative impact evaluation, including decision tree analysis and other forms of data mining (Davies, 2013).

4 In deference to Mackie (1965), ‘necessary’ in this context strictly means 'INUS' or Necessary but Insufficient as a particular causal package that may be Unnecessary but is Sufficient as a cause of Y.

5 This in turn drew upon a QUIP designed during the 1990s to meet the specific needs of microfinance organisations that also linked in-depth impact interviews with routine quantitative monitoring of ‘client level’ indicators (see Imp-Act, 2004).

6 A draft copy of the QUIP is available at http://www.bath.ac.uk/cds/projects-activities/assessing-rural-transformations/index.html

7 A more controversial issue is the reliability of annual estimates of household food security obtained through IHM on the basis of a single annual interview. See Devereux et al. (2012) for discussion of these issues relative to the Household Economic Assessment (HEA) method that was the immediate precursor of the IHM. EFD experience with the IHM is that with careful timing and skilled questioning it is possible to construct surprisingly detailed accounts of household livelihood activities, including seasonal food and cash inflows over a year.

8 More specifically the plan is to pilot four focus group discussions per study (e.g. for younger men, younger women, older men and older women), with a minimum of three people present in each and a maximum of eight.

9 The key issue is then arguably how time and attention is divided between the two. Mayne’s outline of contribution analysis for example, can accommodate exploratory work as part of Step 5 (‘seeking out additional evidence’ within a six step process. This seems to me to overly downplay its potential role in generating information about cause and effect that is less prone to pro-project bias (Mayne, 2012:272).

10 The term mental model is used in the sense explored by North (1990), but also familiar in psychology (e.g. Breakwell, 2007).

11 This left-right brain distinction is quite different from the one explored by Kahneman (2011) between quick and dirty 'System 1’ and slower, more effortful 'System 2’ thinking. McGilchrist suggests that Kahneman ‘cuts the cake horizontally’, the left-right hemisphere distinction cuts it vertically (Rowson and McGilchrist, 2013:30).
References


Stern, E, Stame, N, Mayne, J, Forss, K, Davies, R, & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations (pp. 91): DFID.


