Attributing Development Impact: The qualitative impact protocol case book
https://practicalactionpublishing.com/book/105/attributing-development-impact

Chapter 2. Comparing the QuIP with other approaches to development impact evaluation

James Copestake

1. Introduction

Chapter 1 introduced the QuIP and explained its origins, and the Annex reproduces the QuIP guidelines in full. The purpose of this chapter, in contrast, is to review in more detail how it compares with other approaches to development impact evaluation. This is a potentially huge task that could draw on literature from across the social sciences, philosophy, management, and research methods. There are both a bewildering array of approaches and tools to compare, and a bewildering range of criteria to draw upon in doing so. The guide to impact evaluation produced by BOND, for example, distinguishes between six types of impact evaluation: experimental, statistical, theory-based, case-based, participatory and synthesis (BOND, 2015).[1]

This chapter first narrows the scope of the discussion by defining impact evaluation as an intermediate feedback mechanism falling somewhere between routine performance management and independent research (Section 2). It then classifies the QuIP inductively according to how it compares to a list of other impact evaluation approaches, drawn mostly from the *Better Evaluation* website (Section 3). The chapter then reviews the complex question of what criteria should inform the choice of impact evaluation approach (Section4). Given the complexity of development problems, and the inevitable constraints of time and money on what evidence it is possible collect, we emphasise the importance of (a) balancing depth with breadth of coverage, and (b) choosing an appropriate threshold of credibility or certainty. This affirms the value of approaches to assessing attribution claims (such as the QuIP) that can be exploratory as well as confirmatory, and that can build flexibly and incrementally on what commissioners already know, rather than assuming that they would otherwise know nothing.

2. Defining the field of impact evaluation.

Picking up from Chapter 1, we are primarily concerned in this book with how investors with social or development goals assess whether they are achieving what they intend. Figure 2.1 sets out this problem more precisely. Social investors (top left) employ a project team to carry out specified development activities for a target group of intended beneficiaries, through a project or intervention that is large enough to require at least two layers of

[1] BOND is the leading UK membership body for organisations working in international development. The BOND website also provides a spreadsheet tool for choosing appropriate evaluation methods (www.bond.org.uk/resources/evaluation-methods-tool). This subjects eleven methods to a checklist of 39 questions about the questions that need answering and what requirements must be satisfied for the method to be applicable. The methods are RCTs, difference-in-difference, statistical matching, outcome mapping, most significant change, soft systems modelling, causal loop diagrams, realist evaluation, qualitative comparative analysis, process tracing and contribution analysis.

hierarchy (management and staff). Three feedback loop mechanisms can then be distinguished.[2]

- First, the social investors can rely on what they are told by the project management – both informally and through contractual reporting requirements. We call this the short feedback loop.
- Second, they can compare what they learn from this route with general insights derived from applied research by a relevant knowledge community, much of it in the public domain. We call this the long feedback loop.
- Third, they can commission an evaluator to collect additional evidence about the impact of the project on intended beneficiaries for them. We call this the intermediate feedback loop, and this is the route that the QuIP is designed for. In large organisations this role may be performed wholly or in part by staff who are directly employed, and have specialist expertise in evaluation, but who are not directly involved in management or implementation of the project.

2.1 Short feedback loops.

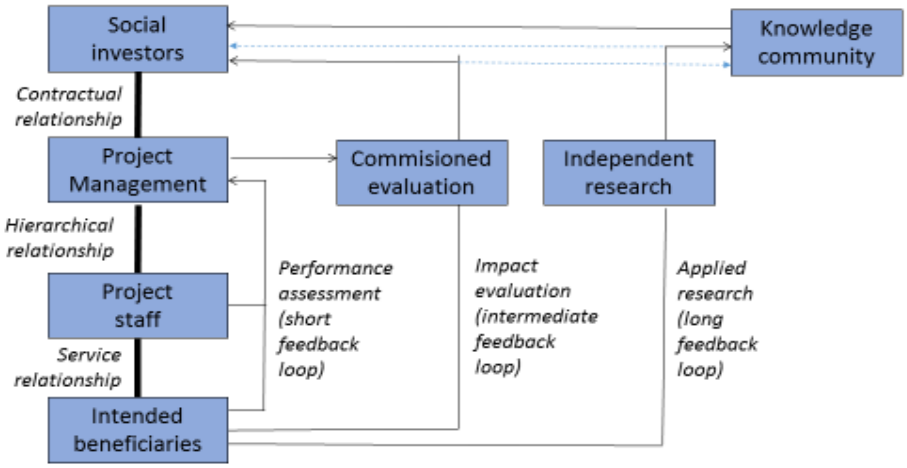
Much of the feedback on development project impact is generated and used through implementing organisations' own routine operational activities. This includes use of data and documents produced through routine planning and performance management activities, as well as evidence mediated verbally through conversations and meetings. Both a strength and a weakness of such feedback is that it will often diverge, leaving project managers and investors with the challenge of deciding who and what to believe. Hence the quality of such feedback critically depends on organisational culture, including levels of transparency, trust and freedom to challenge officially sanctioned views.[3] Another important feature of such evidence is that it is often closely interlinked with detailed and context-specific theories about how the organisation's activities generate impact.[4] Indeed the short feedback loop serves in part to confirm, refine, challenge or contradict such established theory or conventional wisdom.

[2] A fourth channel is for social investors to make contact with intended beneficiaries directly. This is not uncommon, particularly for small projects, and even for very large projects it can be help investors better to understand evidence provided by other channels. While mostly conducted informally, and open to criticism as anecdotal' and prone to 'development tourism' such immersion has also been formalised under the label of the "Reality Check" approach (Jupp 2016).

[3] Academic research into feedback at this level is rich and diverse. Flyvbjerg (2006) uses Aristotle's term *phronesis* to emphasise the importance of contextual-specific and capable judgement or practical wisdom, contrasting it with both abstract scientific knowledge and technical skill. Bordieau's term *habitus*, is broader but similar. Scott (1998) borrows the term *metis*, from Greek mythology, which also suggests the importance of trickery and cunning. Many other writers of development make references to a similar idea, including Eyben (2010) in her discussion of informal practices and "hiding relations" that enhance aid effectiveness in the face of poor policies.

[4] Some of it is also set out more explicitly as a theory of change for the project. There is variation both how specific this is to a particular project or activity, and how far it is internalised within the project's procedures and practices.

Figure 2.1. Development impact feedback loops



Short feedback loops enable organisations to operate most of the time. However, they are fallible. Cognitive traps, herd effects, collective self-delusion are all possible. Larger organisations have to guard against becoming trapped within myths about their performance that nobody within its internal hierarchies has sufficient power and incentive to challenge. Hence a starting point for our discussion of impact evaluation is recognition that short feedback loops need to be supplemented with evidence from other sources.

2.2. Long feedback loops.

It is common sense for social investors to evaluate short feedback loop evidence against evidence available from independent sources. For example, in the context of humanitarian disasters, public media reporting will almost inevitably influence how relief agencies interpret the short feedback loop or operational data provided directly by their own staff on the ground. Long feedback loop data can be very diverse: being defined here to include everything from media reports to academic publications, via official reports and the published outputs of civil society organisations. However, the distinguishing characteristic through which we will differentiate it from both short and intermediate feedback is that it is neither supplied nor commissioned directly by the social investor of the project or activity being assessed. This means that the social investor faces a problem identifying and selecting from it what is most relevant, credible and useful.

A key characteristic of 'long' feedback is that the social investor has limited control over its quality, which is influenced more by the wider peer group or knowledge community, such as an academic discipline or a professional field. An anthropological study may directly address the impact of a development agency and challenge accepted wisdom generated through the short feedback loop. Academic peer review within the knowledge community may also enhance the credibility of its findings. On the other hand, timeliness and cost-effectiveness as well as relevance and sufficiency may all be sacrificed – with much time and effort being devoted to issues that are incidental to the feedback of most value to the project. Of course social investors do nevertheless commission independent research, and doing so generates spill over benefits to other potential users. But investing directly in evidence through the

long feedback loop risks diverts resources away from directly assessing the impact of their investment on intended beneficiaries.[5]

2.3. Commissioned impact evaluation: an intermediate feedback loop.

Impact evaluation generally falls somewhere between the two feedback loops discussed so far. It is distinguished from the short feedback loop by involving staff or hired consultants who are not directly involved in project implementation, and from the long feedback loop because evaluators are directly commissioned, and contractually accountable to the investor (although they may also identify with wider knowledge communities, including the evaluation profession). Securing such feedback is therefore an additional cost to the investor, and hence based on an expectation that this will be offset by benefits derived from the additional evidence obtained, such as improved understanding, better decision-making, strengthened legitimacy, or (more simply) compliance with the demands of higher level funders.

The question of cost-effectiveness of commissioned impact evaluation also depends on what it adds relative to feedback obtained via the other two channels. Two important points arise from this. First, the general market value of the evidence matters less than what it adds to the context-specific knowledge of the investor and commissioner, given their capacity to evaluate its credibility against what they know through internal channels, as well as evidence in the public domain. Second, its value may well depend on how generalizable the evidence is. If the value of (a) short feedback loop performance assessment is partly to review relatively narrow theories of change behind a project, and (b) long feedback loop independent research is to contribute to more general theory, then (c) the case for intermediate commissioned impact evaluation hinges in part on contributing to middle range theory – or evidence that is useful for making decisions over how far a project is likely to be successful in slightly different contexts.[6] These two points are examined in more depth in Section 4.

Comparing impact evaluation with short and long feedback loops also helps us to elaborate on the role of impact evaluation relative to the four challenges of effective action listed in Chapter 1.

- Goal specification and planning is less importance to the evaluator to the extent that the commissioner has already fixed on these.
- The cost-effectiveness of impact evaluation is likely to depend heavily on how it can build upon and complement change monitoring conducted internally by the commissioning agency, as well as in some cases by independent research (e.g. in the form of national household panel survey data).
- Generating additional evidence of causal attribution depends on what the study reveals to corroborate or challenge both the given theory of change of the project and/or more general theories of change associated with independent knowledge communities.
- The role of independent evaluation in enabling organisations to be more agile and adaptive may depend in part on their being able to contribute to useful middle range

[5] Copestake (2013) provides an illustrative discussion of the tension between impact evaluation and applied research (i.e. intermediate and long feedback) for the case of microfinance in India.

[6] For a fuller explanation of the idea of middle range theory see Pawson (2013) and discussion of realist evaluation below.

theory, as already discussed. But it also depends on the social role of the evaluator in relation to the commissioner and other stakeholders: advantaged by gaining additional access and influence compared to fully independent researchers; but also potentially constrained contractually. As we compare the QuIP with other approaches to impact evaluation it will be important to reflect not only on the technicalities of each but on how these influence social and political relationships.

3. Comparing QuIP with other approaches to impact evaluation

An initial classification

Many different approaches to evaluation can be used to generate intermediate feedback evidence. The *Better Evaluation* website (www.betterevaluation) is a useful source of information on a wide range of approaches. It defines an evaluation approach as "an integrated set of options used to do some or all of the tasks involved in evaluation", and then distinguishes between 32 different tasks that this encompasses. These are grouped into seven clusters: how to manage, define, frame, describe, understand causes, synthesise and report/support use. The "understanding causes" task includes checking that results support causal attribution, comparing results to a counterfactual, and investigating possible alternatives. Most of the integrated approaches covered by the website address one or more of these three tasks in some way, and hence can all be defined as a form of impact evaluation.

Having defined what is meant by an evaluation approach, the *Better Evaluation* website lists 24 of them, including the QuIP.[7] The Appendix to this chapter briefly describes each of these approaches in turn, along with six others: 'cost benefit analysis', 'difference-in-difference evaluation', 'goal free evaluation', 'process tracing', 'participatory assessment of development', 'participatory impact assessment for learning and accountability', and 'qualitative comparative analysis'. This provides a useful starting point for readers wishing to compare QuIP to other approaches with which they are already familiar.

Taking the QuIP as a single point of comparison all these approaches have been classified into the four groups distinguished in Table 2.1.[8] Group 1 comprises approaches that are different but have at least one feature that strongly overlaps with the QuIP. In the case of Group 2, the QuIP shares many features, but is generally narrower and more prescriptive in its specification of how different evaluative tasks are completed. In contrast, while the QuIP can be complementary to quantitative approaches to monitoring change its approach to causal attribution differs more fundamentally from most of the approaches in Group 3. Likewise, while the QuIP aims to strengthen feedback from intended beneficiaries to social investors it lacks the strong emphasis on downward accountability and empowerment that

[7] These were listed under the 'approaches' tab, whereas another list on the website omits 'causal link modelling', the 'success case method' and QuIP, but includes 'social return on investment'. The Appendix to this chapter covers them all. For more selective surveys of quantitative approaches see White and Reitzer (2017), and qualitative approaches see Stern et al. 2012) or White & Phillips (2012).

[8] This classification is based on a subjective sorting exercise conducted by one person (the author). This could be done more credibly and formally by combining participatory sorting with network analysis as discussed by Davies (2018).

is a feature of the approaches in Group 4. The following sections selectively explore these similarities and differences in more depth.

Table 2.1. How the QuIP compares with other impact evaluation approaches: summary.

Group 1. Approaches with specific overlapping features with the QuIP.	Appreciative Enquiry; Case Studies; Causal Link Monitoring; Collaborative Outcome Reporting; Critical Systems Heuristics; Goal Free Evaluation; Outcome Mapping; Positive Deviance; Success Case Method; Utilisation Focused Evaluation.
Group 2. Broader approaches, with which the QuIP broadly belongs.	Beneficiary Assessment; Contribution Analysis; Developmental Evaluation; Innovation History; Institutional Histories; Outcome Harvesting; Process Tracing; Realist Evaluation.
Group 3. More quantitative approaches than the QuIP.	Cost Benefit Analysis; Difference-in-Difference Evaluation; Qualitative Comparative Analysis; Randomized Control Trials; Social Return on Investment.
Group 4. Stronger participatory and formative goals than the QuIP.	Democratic Evaluation; Empowerment Evaluation; Horizontal Evaluation; Most Significant Change; Participatory Assessment of Development; Participatory Impact Assessment for Learning and Accountability; Participatory Evaluation and Participatory Rural Appraisal.

3.2. Approaches with features that overlap with QuIP

These approaches differ in emphasis, but generally overlap with the QuIP in specific ways, thereby highlighting both its eclectic character and the scope for improvisation in its use. To give three examples:

- While the QuIP aims to be open to both positive and negative stories of change it could be used more restrictively to focus on the positive, as do both the 'appreciative enquiry' and 'positive deviance' approaches.
- The QuIP first asks respondents what major changes they have experienced in each domain during a specified time period and then encourages them to elaborate on what they think is driving these changes. This feature of working backwards from outcomes connects QuIP strongly with 'outcome harvesting' and 'outcome evidencing' as described respectively by Wilson-Grau & Britt (2013) and Paz-Ybarnegaray & Douthwaite (2016).
- By blindfolding interviewers and respondents to reduce the threat of confirmation and pro-project biases QuIP resembles 'goal-free evaluation', which also avoids being explicit about intervention goals in order to reduce "goal-related tunnel vision" (Youker, 2013)

3.3. QuIP and quantitative approaches to impact evaluation.

QuIP seeks evidence of causation in the form of narrative statements about the impact of selected activities (X) on selected aspects (Y) of the wellbeing of intended beneficiaries of those activities, subject to incidental or confounding drivers of change (Z). Respondent selection can be wider – e.g. to include neighbours of intended beneficiaries, if indirect impact is also anticipated. But attribution claims underpinning the QuIP do not require a control group, nor indeed variation in exposure to the intervention across the sample of respondents interviewed. Rather, causal claims rely on the integrity of statements made by respondents themselves.

Within the wider literature on causal attribution this feature clearly sets QuIP apart from approaches based on RCTs or difference-in-difference evaluation that exploit variation in the exposure of a population to an intervention in order to infer impact statistically.[9] Within this tradition, a change in Y can be attributed to a specified cause, X, only through comparison with a counterfactual of what Y *would* have been in the absence of X, estimated through statistical inference drawing on experimental and/or observational data. See Box 2.1 for some further discussion.

Box 2.1. Impact evaluation based on Randomised Control Trials.

An RCT is widely regarded as the most internally valid way to quantify the impact of a relatively simple intervention across a uniform population in a stable context (Camfield & Duvendack, 2014). Subject to being able to randomly assign the treatment across a big enough sample, then those not treated serve as a counterfactual for those who are treated, of what would have happened to them if they hadn't been. RCTs can then supply an estimate of the average effect of the impact across the sample required for comparing benefits against costs of the intervention. RCTs are relatively simple to interpret because they tackle head-on the risk of selection bias associated with difference-in-difference evaluation and other quasi-experimental approaches. But problems can arise with RCTs too - if sample sizes are too small, perfect randomisation is not possible, the control group is contaminated by treatment effects, responses to interviews are affected by how people feel about being in the treatment or control group, or spillover effects from the treatment group affect the control group (Glennerster & Takavarasha, 2013; White & Raitzer, 2017). RCTs generally also don't reveal much about how impact has arisen, or how it is affected by variation in context and the socio-economic characteristics of respondents (Cartwright & Hardie, 2012; Deacon & Cartwright, 2017). This limits the generalisability (or external validity) of findings, and hence value-for-money of RCTs, given that they are time consuming and expensive. For this reason they are most appropriate to evaluating relatively large investments or testing theory with wide potential relevance, and if a programme or problem is large enough then using them to investigate important implementation issues may also be justified (Duflo, 2017). A positive feature of RCTs is that they require explicit collaboration with the development agency being studied to identify ('prospectively') precisely which activities to evaluate. Nevertheless, there is a risk that design reflects the aspirations and standards of researchers seeking approval of an academic peer group, with correspondingly less weight given to the prior knowledge and credibility thresholds of commissioners, and to the importance they attach to timeliness, sufficiency, relevance, generalisability and cost-effectiveness of evidence. An additional concern is that enthusiasm for RCTS skews investment towards those activities that can be evaluated in this way (Rodrik, 2008) and diverts resources away from other and potentially more flexible approaches to impact evaluation (Stern et al., 2012).

3.4. QuIP and process tracing

The leading alternative to this approach to attribution is theory-based evaluation, also sometimes referred to as the "modus operandi" approach (Scriven, cited in Mohr, 1999). This locates an observed change in Y in a context for which there is a dominant theory that offers only a finite number of possible explanations for it, with X being one of them. Causal claims then hinge on demonstrating that X (or signature characteristics of X) are present, and that this is not so for other possible explanations for Y. The approach can be extended to assessing alternative causal packages, and to situations where both X and other possible causal drivers are present, leaving residual uncertainty as to the relative contribution of each. Following Mayne (2012:273) this can better be referred to as <u>contribution</u> than

[9] This is can also be referred to as a "positivist" and "secessionist" approaches to attribution, relying on "variance" or "regularity" theory (Mohr, 1999; Maxwell, 2004; White, 2010; Gates & Dyson, 2017).

attribution analysis. In other word it asks "…in light of the multiple factors influencing a result, has the intervention made a noticeable difference to an observed result and in what way?" rather than being concerned "…both with finding the cause of an effect and estimating quantitatively how much of the effect is due to the intervention."

One version of this approach is "theory-testing process tracing" (Kay & Baker, 2015). Each additional piece of evidence directly strengthens or weakens a user's confidence in a theory of change linking 'X's and 'Y's, as well as increasing the scope for triangulation. Unprompted positive explicit attribution can be likened to 'smoking gun' evidence of impact, and implicit attribution to 'hoop test' evidence: its presence being less conclusive, but its absence casting doubt on whether the project is working as expected. How strong the evidence is depends in part on the framing of interviews. If respondents are selected because of their participation in the intervention, and interviews take place within the time period for an important expected outcome (Y) to materialise, then not to mention the activity explicitly when asked about change in that specific outcome domain would be surprising. Explicit negative narratives also amount to smoking gun evidence, although isolated instances of this leave open the defence that they are highly context-specific or unusual. Lack of evidence of expected alternative or incidental drivers of a change may also constitute hoop test evidence in support of the intervention.[10]

Table 2.2 suggests that the QuIP conforms reasonably closely to "best practice" in process tracing identified by Bennett and Checkel (2015:261). It also resonates with their argument for greater transparency with respect to the procedures used to collect and analyse evidence, and their call for a "(partial) move away from internally generated practices to logically derived external standards."

Table 2.2. Best practice checklist for process tracing and relevance to the QuIP

Process tracing best practices	Relevance to the QuIP
1. Cast the net widely for alternative explanations.	Multiple interviews and focus groups, combined with masking and use of open-ended questioning to elicit diverse narratives of drivers of change.
2. Be equally tough on the alternative explanations.	Evidence on project related and incidental drivers of change are collected and analysed in the same way.
3. Consider the potential bias of sources of evidence	Masking reduces the threat of project related bias and tunnel vision. Data from intended beneficiaries and project staff are collected separately and systematically compared. Unmasked debriefing meetings provide space for further triangulation.
4. Take into account which explanations are most or least likely to explain a case.	Collection of data for multiple sites, households and focus groups helps to identify more common drivers and mitigate the risk of attaching too much weight to any one source.

[10] This approach can in principle be quantified using subjective scoring or "Bayesian updating" (Befani and Stedman-Bryce, 2017), and be applied to data generated through combinations of quantitative and qualitative methods (Humphreys & Jacobs, 2015). But by aiming to alter a user's confidence in the credibility of a causal claim rather to generate definitive proof it falls short of the highest standards of scientific proof.

5.	Make a justifiable decision when to start.	Interviewing is carefully anchored to a fixed start date – linked to the start of the project being evaluated.
6.	Be relentless in gathering diverse and relevant evidence, but make a justifiable decision when to stop.	Studies are time bound, with sample sizes and selection adjusted to capture diversity. The amount of evidence collected is informed by judgements about marginal returns relative to prior knowledge and ongoing quantitative monitoring.
7.	Combine process tracing with case comparisons when useful for the research goal and when feasible.	Comparisons between households are integral to the approach, and standardization of the interviewing and focus group protocols facilitates this. Informed sampling across different sites is important to address the risk of biased of atypical coverage.
8.	Be open to inductive insights.	Questioning is open to respondents' own unprompted identification of wellbeing changes and their drivers. Coding of these is inductive.
9.	Use deduction to ask "if my explanation is true, what will be the specific process leading to the outcome?"	Interpretation of evidence is aided by triangulating it against the project's theory of change, and staged unmasked triangulation, whereby implementing staff can comment on findings – e.g. offering alternative explanations for negative explicit drivers.
10.	Remember that conclusive process tracing is good, but not all process tracing is conclusive.	The methodology does not rule out being inconclusive about the relative contribution of different causal drivers identified. Evidence of variable impact and lack of overall impact can also be useful.

Source: Compiled by author, using a checklist from Bennett & Checkel, 2015.

3.5. QuIP and realist evaluation

While it is useful to think of QuIP in this, way complex contexts mean it is unlikely that all possible theoretical explanations can be identified and systematically ruled in or out by signature evidence as it suggests. An alternative and more flexible basis for making contribution claims without an explicit counterfactual appeals to our linguistic power to imagine and articulate hypothetical situations. When respondents say 'X caused Y' they often mean more than 'X preceded Y': rather they believe it to be true that if X had not happened then neither would Y. In other words, a tacit counterfactual is implicit in many narrative statements. While confidence in the answer is enhanced if this is made explicit, it is generally impossible to expose and disentangle *all* the possible scenarios respondents may have in mind and be tacitly rejecting.

Taking this additional argument the credibility of causal claims generated using the QuIP in a particular context can be broken down into the following components: (a) there is sufficient evidence that X and the changes in Y happened, (b) several respondents independently - and without explicit prompting – explicitly asserted or implicitly suggested that X was part of a package of factors causing the change in Y, (c) these assertions are congruent with plausible explanations for how this could have happened, and (d) there is no obviously more credible counter-explanation for why respondents might have said what they did. This emphasises the dependence of the methods on respondents' *perceptions*, and reflects the goal of the QuIP to give intended beneficiaries of projects more effective voice through which to challenge development ideas and practices carried out in their name, as argued by

Groves (2015). At the same time, the involvement of field researchers and analysts in interpreting respondents' view reflects a realist position that lies somewhere between the claims to universal truth of positivist science and a constructivist denial of the possibility of establishing any kind of concrete fact independent of the observer (Maxwell, 2004). According to this view truth is 'out there' but hidden; and getting at it entails protracted confrontation of theory with multiple and often inconsistent sources of evidence, kept honest by transparency and peer review, or what Pawson (2013:18) calls "organised distrust". This denial of a strict dichotomy between fact and meaning also supports the view that qualitative methods can usefully employ some strategies associated with variance and regularity theories (Maxwell, 2004:251).

With its rallying cry of "what works for whom in what circumstances" (Pawson, 2013:15) realist evaluation is congruent with the QuIP's granular approach to causation, whereby each case adds independently to understanding multiple causal drivers and outcomes, rather than to confidence levels in one or a few estimates of average treatment effects. An emphasis on the importance of multiple pathways linking X to Y alongside a vector of contextual or confounding factors (Z) is also congruent with Pawson's stress on complexity and on distinguishing between multiple "context, mechanism, outcome configurations". However, the "CMO" terminology does not map perfectly onto the "ZXY" shorthand used in this paper, because from a realist perspective the project actions (X) are part of the context (C) rather than the often more intangible cognitive mechanisms (M) by which X generates outcomes Y.

The underlying conceptualization of complexity is also different, but can be complementary. Pawson (2013:33) defines complexity as variation in project volitions or intentions, implementation, context, time, outcomes, rivalry and emergence ("VICTORE"). A working definition arising from the QuIP research is a setting in which X influences Y in ways that are confounded by incidental factors (Z) that may be impossible to identify, hard to measure accurately, interact with each other in non-linear and/or cumulative ways in their influence on both X and Y, and/or are impossible fully to control. This highlights the point that while correlational data to support binary causal links between variables within one system has its uses, it is rarely possible to infer from such evidence precisely how relevant observed change in one context is to another (Cartwright & Hardie, 2012). Managing this is only possible with the help of "explanatory theory" (ranging from multiple CMO configurations to middle range theory) which realist evaluators appear to treat more fluidly than advocates of a more deductive or "theory-testing" approach to process tracing (Kay & Baker, 2015). Hence while prior theory is important at the initial design stage and in the use of attribution codes, a commitment to open-ended and blinded data collection, and use of inductive coding links the QuIP to "theory-making" as well as "theory-testing" forms of process tracing. It is also at odds with a realist view that interviewers should share their own understanding of project theory as fully and openly as possible with research subjects. (Manzano, 2016).

The above discussion suggests QuIP addresses two out of three common weaknesses in realist evaluation highlighted by Pawson (2013:14): failure to investigate CMOs as configurations, and absence of an explanatory focus. The third weakness he highlights is to work only in one "data medium method" – a point he elaborates by suggesting that "as a

first approximation one can say that mining mechanisms requires qualitative evidence, observing outcomes requires quantitative [data] and canvassing contexts requires comparative and sometimes historical data." (p.19). This suggests the QuIP is primarily a "mechanism miner" best used as part of a mixed evaluation strategy; but also able to contribute to understanding context and outcomes. It also reinforces the argument for using QuIP to complement quantitative monitoring of the frequency and magnitude of change in selected activities, outcomes and contextual factors over time.

Viewed within the broader canvas of realist evaluation, the purpose of a QuIP can be viewed as a more open-ended, exploratory and inductive method than when viewed more narrowly as a form of theory-led process tracing. For example, sampling options are informed not only by the idea of Bayesian updating but also by the criterion of saturation, as reviewed by Guest et al. (2006). The key issue here is how to ensure that additional effort is justified by additional insights - in the form of identification of additional CMO configurations, for example. This logic favours purposive sampling to capture anticipated diversity of experience among intended beneficiaries, including an emphasis on learning from positive and/or negative "deviants" as revealed by prior quantitative monitoring of changes in Y.

3.6. QuIP and participatory approaches to evaluation

The QuIP is a form of beneficiary assessment (Salmen et al. 2002) in the sense that its primary purpose is to document intended beneficiaries' *perceptions* of changes, reasons for these changes, and (at least implicitly) their views on how things could have been different. It thereby gives them voice, although without a firm guarantee that it will have much influence over what other stakeholders do. Voice alone may even have perverse effects: positive feedback from satisfied clients, for example, might even prompt a hard-hearted microcredit agency to tighten the terms of its loans. In this sense, the QuIP is not inherently radical or revolutionary in what it sets out to do: aspiring 'to speak truth to power' but unlikely on its own to challenge that power. Rather, the potential of the QuIP to more transformational development generally depends upon the responsiveness of more privileged actors up the funding chain.

Worse still, while blindfolding may increase the credibility of respondents' voice from the perspective of the QuIP's primary audience this must be offset against the potentially disempowering of not revealing to them everything that could be revealed about the intervention being evaluated. Respondents, for example, might have made more detailed and specific observations about what an agency could have done differently if they had been made fully aware of its identity from the outset. Against this, however, the greater possibility of response bias might have weakened the weight given to their views. In sum, there is a potential trade-off here, and it is a difficult to assess how the arguments each way balance out.

One way to reduce the trade-off is to ensure that blindfolding of both interviewers and respondents is at least only temporary. For example, respondents can be invited to a second meeting at which draft findings from the initial round of interviews are presented and reviewed, ideally in the presence of project staff. Such meetings provide an opportunity to gain deeper insights, strengthen the voice of intended beneficiaries and also provide them with an opportunity for networking and learning.

Informing and empowering intended beneficiaries nevertheless remains a secondary goal of the QuIP relative to 'upward' learning and accountability. This distinguishes it from democratic evaluation, and - to a lesser degree - to other participatory evaluation methods listed in Group 4 of Table 2.4. The extent of this difference depends on how far participatory methods seek a full "reversal" of control over the evaluation process itself (Chambers, 1997); while informing participants is generally more of a priority than informing outsiders, most participatory evaluation approaches continue to be structured and mediated by expert facilitators. This is the case for example, with PADev and PIALA, as described by Pouw et al. (2016), and van Hemelrijck (2016) respectively.

An important feature of participatory approaches is the way a switch in primary purpose towards informing intended beneficiaries and other local stakeholders affects the kind of feedback that is useful, and criteria for evaluating it. Local stakeholders have different prior knowledge against which to triangulate new evidence, including being able to reflect directly on their own experience. To the extent that they are mostly concerned with their own interests then the generalisability of findings will matter less. In these respects indeed Group 4 approaches are perhaps better classified as contributing to performance assessment and a short feedback loop rather than impact evaluation and an intermediate feedback loop.

4. Choosing between approaches to impact evaluation

4.1. How to think about the issue

Section 2 compared and contrasted the QuIP with other approaches to impact evaluation. In doing so, it tried to avoid making value judgements about its relative strengths and weaknesses. This section takes this next step, opening discussion of the conditions under which it could meet potential demand better than alternatives. This entails asking what sorts of questions different approaches can answer and what criteria are appropriate for assessing how well it can answer them.

The purpose of the QuIP should by now be reasonably clear. It has been designed principally to tackle the causal attribution challenge, and to do so for commissioners who need evidence about the impact of specified activities (X) on outcomes in specified domains (Y) that is (a) credible to a wider audience than that generated through routine performance management, but (b) more focused than applied social research for a wider knowledge community. This evidence is based on what intended beneficiaries themselves perceive to be the main drivers of change in their lives in these domains. It is not expected on its own to generate estimates of the magnitude of these effects, although the evidence may assist in modelling and simulating changes in a way that does permit such estimation. Nor is the QuIP design on its own to permit statistically valid estimates of the frequency of different impact mechanisms across a population, although it can assist users in upgrading or downgrading their prior expectations about this. It also aims to cast light not only on X but on other causes (Z) of change in Y, possibly including some that were previously unknown to the commissioner. And it may also generate insight into unintended consequences of X

beyond the initial list of possible outcomes Y. Lastly, it is designed to generate evidence on how these causal patterns vary for different people and contexts.

This specification of the purpose of the QuIP does not come from an axiomatic view of scientific truth or method from which we have derived absolute positions on what constitutes validity, reliability, rigour or even rationality. Rather it has evolved out of a combination of learning by doing and close consultation with actual and potential users about what they consider to 'good enough' to inform their activities, taking into account timeliness, cost and prior understanding. This might appear unduly pragmatic, but it builds on realist philosophical foundations that emphasise complexity. This in turn underpins doubts over the scope for usefully generalising about 'what works' with respect to both development practice and how to assess it. With this comes a preference also for a pluralist and evolutionary view of how to identify and promote good practice.[11] This rejection of a universal solution to the attribution challenge should not be mistaken for the view that anything goes, or that every opinion has equal weight. For a given problem in a given place there will be better and worse way to assess impact, and even a best way – in other words, there is still a role for the technically proficient evaluation specialist. But at the same time this judgement will also depend upon the power, role and interests of the person commissioning the study. Hence professionalism also has a political dimension, including negotiating room to deliver evidence that goes beyond and even challenges what the commissioner is seeking.

How far the QuIP proves a useful addition to the field of impact evaluation will ultimately depend on how well it works, for what and for whom. Impact evaluation is conceived here as a complex and rapidly changing field to navigate through: a contested market for a highly differentiated set of products with distinctive features and combinations of features. Branding and advertising may help to inform users and to signal producers' commitment to different products, but they also reinforce market power and tradition. But at the same time new entrants can emerge, and ultimately we subscribe to the cliché that the proof of the pudding is in the eating. This helps to explain the emphasis in this book on documenting actual use of the QuIP, additionally informed by the view that good development practice (along with good social science) proceeds in part through the accumulation of detailed case studies (Flyvbjerg, 2006:219; Goertz, 2017).

4.2. Balancing breadth and certainty of evidence: a simple model

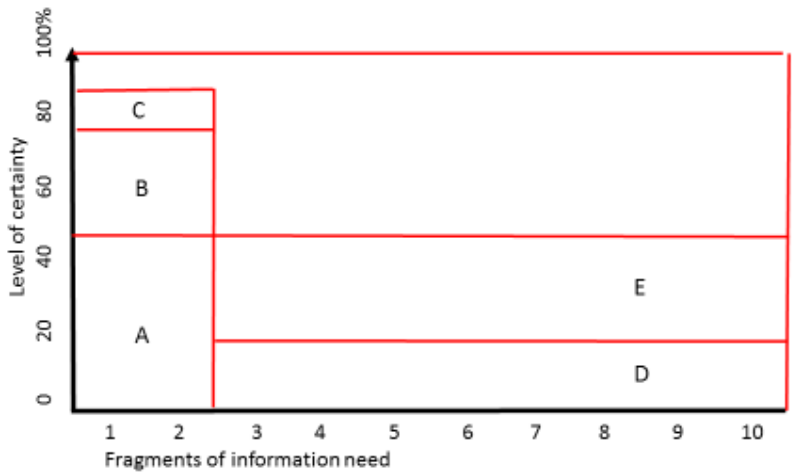
An appreciation of the importance of complexity to choice over method in impact evaluation does enable us to make some tentative generalisations, but based more on analysis of what constitutes an acceptable threshold of evidence for commissioners in different contexts than an absolute view of what constitutes sufficiently rigorous evidence.

[11] This approach to analysing the problem borrows from Andrew Abbott's idea that our collective understanding being made up of "knowledge lineages" that emerge, coalesce, compete, mutate, thrive, evolve and die (Abbott, 2001; Copestake 2015). This evolution takes place simultaneously at multiple levels, with competition between quantitative and qualitative approaches to evaluation, for example, partly reflecting grander controversies over development theory, social science and philosophy. Of course evolutionary practice is moulded by debate (see for example the discussion of David Hume's seminal writing on causation in Chapter 6 of Goertz & Mahoney, 2012).

Decision-makers differ according to their appetite for certainty and uncertainty. They also start out with different levels of prior knowledge. A simple model illustrates the implications of this variation.

Contrast, for example, a potential investor who has recently come across a social enterprise, and who wants to learn more about its social impact prior to investing in it, and its owner and manager who is also interested in finding out more about its social impact. If one or both are also commissioning and paying for a study then their view of the cost-effectiveness will also depend on their certainty appetite and prior knowledge. This is illustrated by Figure 1. For simplicity, the horizontal axis plots ten things the manager and the investor agree it would be useful to know about the social impact of the business, starting with the one they agree is most important (1) and adding less important items of information up to ten. The Y axis plots certainty thresholds for this knowledge, from self-confessed total ignorance (0) to total certainty (100%). Assume the manager already has a view on the two most important items, with 80% certainty, and the other eight with 20% certainty; meanwhile, the investor is ignorant of everything. Two independent impact studies are proposed of equal cost. One will provide 90% certainty about the first two items. The other will deliver 50% certainty about all ten. It would not be unreasonable for the investor to prefer the first study and the manager the second. But the manager may nevertheless agree to contribute to the cost of the first rather than the second if it is necessary to do so in order to achieve a necessary and sufficient level of shared understanding and trust to convince the investor to invest in the business.

Figure 1.1. Criteria for comparing impact assessment studies: an illustrative example.



	Man-ager	Inv-estor
Prior knowledge	A+B+C	0
Incremental learning from first study	C	A+B+C
Incremental learning from second study	E	A+D+E

This model illustrates that choosing how to spend money wisely on impact assessment depends on the commissioners' prior knowledge, (un)certainty preferences and the range of issues they regard as important to cover. There is the choice between studies that set out to confirm known causal pathways or to explore those that are largely unknown. The example, also highlights the importance of trust. As described, the investor did not give any weight to the fact that the manager already knew the two most important facts to be true with 80% certainty. This may have been wise of the investor, given the possibility that the manager might lie about this. Or perhaps the manager was never even asked. If the investor was

aware of what the manager knew and had given it even a low weight then that might still have been sufficient to convince her that the broader study was better value for money, despite delivering less certain evidence on these issues.

Returning to the real world, this example illustrates why impact assessment may be a source of conflict even between like-minded stakeholders along management and financing hierarchies. Potential for disagreement is exacerbated by differences in understanding alternative evaluation methods and in preferences about how precisely impact needs to be measured (Muller, 2018). This helps to explain the widely discussed tendency towards 'overkill' in assessment of development activities (including impact), complete with performance indicators, targets, logical frameworks, reviews and audits. We are sympathetic to the view excessive auditing not only increases implementation costs but can reduce the likelihood of doing anything truly transformative (Natsios, 2010).[12] This helps to explain our quest for an approach that is cheaper, more flexible and supplements what is already known or can be inferred from careful interpretation of a simple monitoring system.

4.3. Scope for generalising about 'what works' in development (and chess)

An additional consideration behind the design of the QuIP is an appreciation of the sheer number of contexts and combinations of drivers of change that is would be useful to understand better. Andrews et al. (2012; 2017) emphasise the same point by depicting the "policy design space" as rugged or non-linear and arguing for an evolutionary approach to development which they call "problem driven iterative adaptation" (see also Room, 2011; Boulton et al. 2015; Bamberger et al., 2016; and the final chapter of World Bank, 2015). In short, as the number of policy design options increases so does the potential advantage of being able to explore alternatives through more agile forms of impact evaluation.

To illustrate the importance of this point consider the game of chess. Evaluating different moves and strategies is obviously relatively simple compared to the reality of doing development: there are only two players and three formal outcomes (win, lose or draw), and play is constrained by only having to think about a maximum of 32 pieces, each with fixed and transparent capabilities. The complexity of the changing context of the board at each move is offset by the simplicity of the ultimate goal, limited resources and the restricted room for manoeuvre of the other player. Yet the number of possible games of chess comprising 35 moves is greater than the number of atoms in the Universe! So how many more possibilities does a development agency have to review when deciding how best to take forward multiple activities with large number of others whose motives, resources, opportunities and understanding are often only weakly understood?

This analogy is useful in thinking through what it is reasonable to learn from formally assessing the causal impact of different moves and strategies. Chess may be complicated, but we nevertheless know a great deal about what enables a player to perform well. Core knowledge comes from simulating simple scenarios – how a knight can use a fork to capture

[12] For wider criticism of overly zealous results-based and measurement culture see Eyben et al. (2015) and Hayman et al. (2016). For a more entertaining and pithy commentary on the downside of the quest for better attribution listen to the "impact blues" by Terry Smutlyo (www.youtube.com/watch?v=5f4rNEsyEYY). Warnings of the danger of going overboard in assessing development effectiveness is of course much older: see, for example, the classic lament about "survey slavery" in Chambers (1983).

a queen, for example. This feeds into case study analysis of complete games, which locate discrete moves in the context of the whole board and a full game. Inductive analysis can also be used to build middle range theory (castle the king early; don't exchange a queen for a knight; avoid doubling up pawns and so on). Likewise, a development agency intervening in a new area can draw upon a broad range of potentially relevant middle-range theory. However, it also needs to guard over-generalisation, or what Scott (1998) calls "thin simplification". No matter how rigorously documented, a policy that worked in one context cannot be relied upon to have the same outcome in a new time or place (Cartwright and Hardie, 2012).

This point about generalisation raises the question of what level of generalisation this book aspires to achieve. We will find it useful in places to generalise about the attribution challenge in logical but simple ways, by exploring what combination of measurable variables X and Z might cause a change in a measurable indicator Y, for example. We also believe it can be useful to explain how to use the QuIP in a broad and generic way. But at the same time we believe it is useful to combine this with real cases studies of how the QuIP has been employed, why and if possible to what effect. These should help reader in thinking about scope for improvisation in adapting the QuIP to new contexts.

To revert to chess. There is much science to learning how to be a better player, both deductively (by building up understanding of how different pieces interact from the basic rules) and inductively (by generalising from passed games). For example, detailed study of possible openings might lead a student to conclude that it is a disadvantage to be black. So might statistical analysis of the outcome of thousands of games. But precisely how disadvantageous it will be for me to be black if I play you tomorrow remains uncertain. Thanks to the relative simplicity of its fundamental elements it has proved possible to build computer programmes that can outperform the best human players. Likewise we strongly advocate employing the full range of logical thinking and computer capabilities to identifying the multiple causal determinants of development outcomes. But ultimately, we think that such analysis will also highlight the limitations of what we know. Scope will remain for performance art, for creative application of good judgement, and for judicious mixing of middle-range generalisations to come up with a good strategy for a particular time and place. And immersion in sufficiently rich contextual case study material will remain an important ingredient for the cultivation of such ability.

5. Conclusions

This chapter has located the QuIP within the wider field of impact evaluation in three steps. First, we looked at the demand side by making a broad distinction between impact evidence produced through routine performance assessment, commissioned impact evaluation and independent research (also referred to as short, intermediate and long feedback loops respectively, with QuIP in the middle category).

Second, it switched to considering the supply of commissioned impact evaluation, classifying different approaches inductively into four groups by taking the QuIP as a benchmark comparator. This clarified how the QuIP selectively incorporates ideas from several approaches, and can be viewed as a more fully specified application of others, including

contribution analysis, process tracing and realist evaluation. By comparing it systematically with alternative approaches this section aimed further to elucidate what the QuIP is.

Third, the chapter opened up the normative question of whether the QuIP adds to the overall portfolio of available ways of tackling the attribution challenge. Is it a useful example of creative synergy, or is it adding to a confusing cacophony of approaches in a crowded space? The important answer to this question, we suggest, will come less through debate and more through case study evidence of its use, including the examples presented in this book. But this section also argued strongly that in a highly complex field and design space it has potential to add value by virtue of having been designed to permit collection of evidence of attribution in a way that is relatively simple, incremental, open-ended and flexible.

Appendix. Comparing QuIP with thirty other approaches to impact evaluation

Approach and brief description.[13]	How the QuIP compares.
Appreciative enquiry A participatory approach that focuses on existing strengths rather than deficiencies - evaluation users identify instances of good practice and ways of increasing their frequency.	The QuIP is more narrowly focused on generating credible impact evidence; it is neutral in eliciting accounts of positive and negative drivers of change.
Beneficiary assessment An approach that assesses the value of an intervention as perceived by the (intended) beneficiaries, aiming to give voice to their priorities and concerns.	The QuIP is a form of beneficiary assessment, but offering more specific and detailed guidelines.
Case study A research design that focuses on understanding a unit (person, site or project) in its context, which can use a combination of qualitative and quantitative data.	The QuIP is based on multiple individual/household case studies, often clustered within purposively selected sites, which may also constitute cases (hence a 'small n' rather than a single case approach).
Causal link modelling This approach integrates design and monitoring to support adaptive management of projects. Managers identify the processes required to achieve desired results and then observe whether they take place along a logic model or results framework.	Elaborating a logic model as part of the theory of change for an intervention is a necessary step for attribution coding and hence using the QuIP to confirm if an intervention is achieving what was intended. The QuIP also focuses on the final causal link from outcomes to impact on intended beneficiaries which is also often the hardest to assess.
Collaborative Outcomes Reporting An approach that builds on contribution analysis, adding expert review and community review of the assembled evidence and conclusions.	The QuIP can be viewed as one way of collecting outcome data for COR. It shares a strong emphasis on multi-stakeholder engagement to validate, interpret and explore potential implications of findings.
Contribution Analysis	

[13] Most of the text in this column is taken from http://www.betterevaluation.org/en/approaches

An approach for assessing the evidence of claims that an intervention has contributed to observed outcomes and impacts.	The QuIP is a form of contribution analysis, but offering more specific and detailed guidelines.
Cost Benefit Analysis A general approach for comparing incremental benefits and costs of an action compared to one or more alternatives. Key steps include: identification of option; scoping of key stakeholders and the impact on them of each option over time; quantification key impacts; valuation and aggregation of costs and benefits.	The QuIP can contribute to identification and scoping of positive and negative causal effects of an intervention on intended beneficiaries and other stakeholders. To go beyond this requires combining it with more precise quantification and valuation of effects based on supplementary data collection, modelling and simulation.
Critical System Heuristics An approach used to surface, elaborate, and critically consider boundary judgments, that is, the ways in which people or groups decide what is relevant to the system of interest.	The QuIP can also expose differences in how implementers and intended beneficiaries perceive a project, including its scope. But it is not so explicitly designed to challenge stakeholders' motivation, power, worldview or legitimacy.
Democratic Evaluation An approach where the aim of the evaluation is to serve the whole community. [The evaluator is accountable to, works with and seeks legitimacy from the members or citizens of this community].	While it enables intended beneficiaries of a project to share their experience with those controlling it the QuIP operates under the authority of the commissioner, rather than insisting on a broader and more democratic mandate.
Developmental Evaluation An approach for evaluations of adaptive and emergent interventions, such as social change initiatives or projects operating in complex and uncertain environments.	The QuIP shares an emphasis on generating timely evidence in a complex and rapidly changing contexts, but is more narrowly specified.
Difference-in-Difference Evaluation Estimates change in specified impact variables for a 'treatment' and 'control' group before and after an intervention, then uses statistical methods (e.g. propensity score matching on observable characteristics) to mitigate selection bias arising from non-random placement of cases into the two groups.	The QuIP attributes causal effects on the basis of self-reported narrative attribution of a 'treatment' group rather than through statistical inference based on comparison to a 'control' group or analysis of variable exposure to an intervention. This limits scope for quantifying the magnitude of impact, but also eliminates the need for a comparison group.
Empowerment Evaluation Provides communities with the tools and knowledge that allows them to monitor and evaluate their own performance.	The core purpose of the QuIP is to provide better evidence to the commissioner, rather than to enable intended beneficiaries to conduct self-evaluation.
Goal free evaluation Open interviews and observation that seeks to understand respondents' lived experience holistically and the meaning they give to it, and to view specific interventions in this light.	Blindfolding is utilised as part of the QuIP to facilitate similarly open ended and exploratory enquiry, within specified domains of respondents' lived experience. QuIP also goes further in then systematically comparing these findings with the theory of change behind a given intervention.
Horizontal Evaluation	The QuIP is not specifically oriented towards locally led activities, and aims to generate evidence that is more

An approach that combines self-assessment by local participants and external review by peers [typically through a three day joint workshop].	credible to a remote audience through a more tightly structured approach to data collection and analysis.
Innovation history A way to jointly develop an agreed narrative of how an innovation was developed, including key contributors and processes, to inform future innovation efforts. Institutional histories An approach for creating a narrative that records key points about how institutional arrangements have evolved over time and have created and contributed to more effective ways to achieve project goals.	The QuIP offers more specific and detailed guidelines for building a narrative account of the impact of a specified intervention, innovation or institutional change. It places more emphasis on intended beneficiaries' own accounts of this, alongside other drivers of change. A potential limitation of the QuIP is that by focusing primarily on the intervening agency and intended beneficiaries the QuIP does not normally engage with network analysis as fully as these approaches.
Most Significant Change Collects and analyses personal accounts of change, includes processes for learning about what changes are most valued by individuals and groups.	The QuIP shares an emphasis on eliciting respondents' own account of causal processes, but without needing to prioritise the most significant. It relies on more formal thematic analysis of causal stories, rather than on a collaborative process of ordering these.
Outcome Harvesting Collects evidence of what has changed and works backwards to determine whether and how an intervention has contributed to these changes. Useful in complex situations when project aims or even specific activities cannot be clearly specified.	The QuIP is a form of outcome harvesting, but offering more specific and detailed guidelines.
Outcome Mapping Unpacks an initiative's theory of change, provides a framework to collect data on intermediate changes that lead to transformative change, and allows for the plausible assessment of the initiative's contribution to results.	Elaborating a detailed theory of change for an intervention is a necessary step for attribution coding and hence for using the QuIP to confirm it an intervention is achieving what was intended and by the expected mechanisms. The use of journals by different stakeholders to monitor changes could be incorporated into the QuIP as an additional source of narrative evidence of drivers of change.
Participatory Assessment of Development Rather than focusing on one intervention or agency PADev simultaneously addresses all interventions in a locality in relation to its overall development. This is done through a structured set of focus group discussions organised through a mediated community workshop [insert reference].	PADev and QuIP are both based on narrative accounts of drivers of change that try to avoid focusing to avoid framing those accounts by reference to a specific activity. PADev does this by taking a community wide perspective, while QuIP does it through blindfolding. Both, but PADev especially thereby produce findings that are potentially relevant to all organisations working in the locality.
Participatory Impact Assessment for Learning and Accountability PIALA is an eclectic approach to gathering data about a development intervention using multiple methods using a range of participatory methods, and also involves intended beneficiaries themselves in analysis and interpretation of data	The two approaches share the goal of generating both formative/exploratory and summative/confirmatory data at the same time, and QuIP could be incorporated into PIALA as a form of data collection. However, it adopts a more transparent and precise approach to deriving and presenting data from primary sources. Representatives of

using the 'Sensemaker' proprietary software developed by the company Cognitive Edge.	intended beneficiaries can be invited to interpret findings, but are not directly involved in generating them.
Participatory Evaluation A range of approaches that engage stakeholders (especially intended beneficiaries) in conducting the evaluation and/or making decisions about the evaluation. (This also incorporates Participatory Rural Appraisal, and Participatory Learning and Action.	QuIP aims to give voice to a sample of intended beneficiaries, and to involve them in interpreting and using findings; but does not to involve them directly in data analysis or management of the evaluation. It primarily responds to demand for upward accountability.
Positive Deviance Involves intended evaluation users in identifying 'outliers' – those with exceptionally good outcomes - and understanding how they have achieved these.	Where changes in key outcome variables is being monitored across a population then QuIP sample selection and data collection can be deliberately biased towards positive deviants. But it can equally be used to illuminate drivers of change more widely across the population, or indeed to focus on gaining a better understanding of reasons for negative deviance.
Process Tracing In its simplest form this is a case study method that starts by identifying a single discrete outcome, such as a murder. It provides guidelines for systematically identifying a package of necessary and sufficient causes to explain the outcome and rejecting alternative packages that could also explain it. [insert reference]	QuIP also seeks evidence to confirm or challenge a theory of change (that an intervention was a necessary condition for impact on an intended beneficiary). QuIP does this for multiple cases and possible impacts, and like process tracing each additional piece of evidence adds to or weakens the commissioners' prior belief in the theory. Though not quantified this can be described as a form of 'Bayesian updating'.
Qualitative Comparative Analysis A statistical approach for identifying packages of necessary and sufficient conditions for achieving a desired outcome across a sample of case studies.	If each QuIP interview is treated as a discrete case, then together they form a 'small n' sample that could possibly be utilised for QCA to analyse multiple factors contributing to specified outcomes, including the contribution of a specified intervention.
Randomised Controlled Trials An approach that produces an estimate of the mean net impact of an intervention by comparing results between a randomly assigned control group and experimental group or groups.	QuIP is based on a fundamentally different approach to impact attribution that avoids the need to compare intended beneficiaries with a control group. However, if sufficient resources are available then there is potential complementarity between the two approaches: e.g. QuIP to elucidate causal mechanisms, unanticipated consequences and reasons for heterogeneity of impact; an RCT to quantify the average impact across a selected population.
Realist Evaluation Realist evaluation is a form of theory-driven evaluation but is distinguished by its philosophical emphasis on the how interventions influence particular decisions (or not). (It also emphasises complexity, heterogeneity and the benefits of combining different methods of data collection and analysis).	The QuIP can be viewed as a narrower and more detailed approach to realist evaluation, or as one method that can be incorporated into realist evaluation. It shares the emphasis on complexity, an appreciation of the benefits from using mixed methods, an interest in 'what works, for whom and in what context', and an appreciation that change occurs through multiple pathways (or what realists call context-mechanism-outcome configurations).

Social Return on Investment Identifies a broad range of social outcomes (not only the direct outcomes for the intended beneficiaries of an intervention) then quantified and values these, and compares them with the investment cost. Hence this is one form of social cost benefit analysis.	The QuIP can help to identify wider outcomes of an investment, and data collection can be extended to possible indirect and unintended beneficiaries (and losers) from an investment. It rarely enables impact to be quantified or valued, so needs be combined with other data (or modelling based on estimated values) to inform a full social cost benefit analysis.
Success Case Method The approach is based on comparing detailed evidence about two case studies: the most successful and least successful subjects of an intervention. It is a useful for understanding what enhances or impedes impact.	The QuIP also relies on comparative case studies, which may be individuals, households, organisations and/or clusters of them. Where data is available for key impact indicators then it is possible to select more and less successful cases (i.e. positive or negative deviants) for analysis.
Utilisation-Focused Evaluation Starts with the intended uses of the evaluation by its primary intended users to guide decisions about how an evaluation should be conducted.	The starting point of a QuIP should also be dialogue with the commissioner over what additional evidence they need and why. This should then influence details of design, including timing, sample size and selection, scope, thematic analysis and data presentation. But a QuIP can also generate useful evidence about an intervention that was not anticipated or solicited for a predetermined purpose.

References

Abbott, A. (2001). Chaos of Disciplines. Chicago, IL: University of Chicago Press.

Andrews, M., Pritchett, L., Woolcock, M. (2012). Escaping capability traps through problem-driven iterative adaptation (PDIA). Paper 299. Washington DC: Centre for Global Development.

Andrews, M., Pritchett, L., Woolcock, M. (2017). Building state capability: evidence, analysis, action. Oxford: Oxford University Press.

Bamberger, M., Vaessen, J., Raimondo, E., editors (2016). Dealing with complexity in development evaluation: a practical approach. Los Angeles: Sage Publications.

Befani, B., Stedman-Bryce, G. (2017) Process tracing and Bayesian updating for impact evaluation. Evaluation, 23(1):42-60.

Behague, D., Tawiah, C., Rosato, M., Some, T., Morrison, J. (2009). Evidence-based policy-making: the implications of globally-applicable research for context-specific problem-solving in developing countries. Social Science & Medicine, 69:1539-1546.

Bennett, A., Checkel, J. (2015). Process tracing: from metaphor to analytic tool. Cambridge: Cambridge University Press.

BOND (2015). Impact evaluation. A guide for commissioners and managers. London: BOND. Prepared by Elliot Stern for the Big Lottery Fund, Bond, Comic Relief and the Department for International Development, May 2015.

Boulton, J., Allen, P., Bowman, C. (2015). Embracing complexity: strategic perspectives for an age of turbulence. Oxford: Oxford University Press.

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

Cartwright, N., Hardie, J. (2012). Evidence-based policy: a practical guide to doing it better. Oxford: Oxford University Press.

Chambers, R. (1983). Rural development: putting the last first. Harlow: Longman.

Chambers, R. (1997). Whose reality counts: putting the first last. London: Intermediate Technology Publications.

Copestake, J. (2013). Research on microfinance in India: combining impact assessment with a broader development perspective. Oxford Development Studies, 41: S17-34.

Copestake, J. (2015). "Whither development studies? Reflections on its relationship with social policy." Journal of International and Comparative Social Policy 31 (2): 100–113.

Davies, R. (2018). Network visualisation of qualitative data. Monitoring and evaluation News. http://mande.co.uk/special-issues/participatory-aggregation-of-qualitative-information-paqi/ [accessed 11 June 2018).

Deacon, A. and Cartwright, N. (2017). Understanding and misunderstanding randomized

Duflo, E. (2017). The economist as plumber. Richard T Ely Lecture. American Economic Review: Papers and Proceedings. 107(5):1-26. Evaluation: A Practical Approach. London: Sage.

Eyben, R. (2010). Hiding relations. The irony of 'effective aid'. European Journal of Development Research, 22(3).

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

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

Flyvbjerg, B. (2016). The fallacy of beneficial ignorance: a test of Hirschman's hiding hand. World Development, 84:176-89.

Gates, E., Dyson, L., (2017). Implications of the changing conversation about causality for evaluators. American Journal of Evaluation, 38(1):29-46.

Glennerster, R., and Takavarasha, K. (2013). Running randomized evaluations: a practical guide. Princeton: Princeton University Press.

Goertz, G. (2017). Multimethod research, causal mechanisms, and case studies: an integrated approach. Princeton: Princeton University Press.

Goertz, G. and Mahoney, J. (2012). A tale of two cultures: qualitative and quantitative research in the social sciences. Princeton: Princeton University Press.

Groves, L. (2015). Beneficiary feedback in Evaluation. London: Department for International Development, Evaluation Department.

Guest, G., Bunce, A., Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. Field Methods, 18: 59-83.

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

Humphreys, M., & Jacobs, A. (2015). Mixing Methods: A Bayesian Approach. American Political Science Review, 109 (4):653-73.

Jupp, D. (2016). Using the reality check approach to shape quantitative findings: Experience from mixed method evaluations in Ghana and Nepal. In: Bell S and Aggleton P (eds) Monitoring and Evaluation in Health and Social Development: Interpretive and Ethnographic Perspectives. London and New York: Routledge, 172–84.

Kay, A., Baker, P. (2015). What can causal process tracing offer to policy studies? A review of the literature. Policy Studies Journal, 43(1):1-21.

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

Maxwell, J (2004). Using qualitative methods for causal explanation. Field Methods, 16:243-264.

Mayne, J. (2012). Contribution analysis: coming of age? Evaluation 18(3):270-280.

Mohr, L. (1999). The qualitative method of impact analysis. American Journal of Evaluation, 20(1):69-84.

Muller, J Z. (2018). The tyranny of metrics. Princeton: Princeton University Press.

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

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

Pouw, N., Dietz, T., Belemvire, A., de Groot, D., Millar, D., Obeng, F, Rijneveld, W., Ven der Geest, K., Vlaminck, Z., Zaal, F. (2016). Participatory assessment of development

interventions: lessons learned from a new evaluation methodology in Ghana and Burkina Faso. American Journal of Evaluation, 1-13.

Paz-Ybarnegaray, R., Douthwaite, B. (2016). Outcome evidencing: a method for enabling and evaluating program intervention in complex systems. American Journal of Evaluation, 38(2): 275-293.

Rodrik, D. (2008). The new development economics: we shall experiment, but how shall we learn? John F Kennedy School of Government: Faculty Research Working Paper 08-055.

Room, G. (2011). Complexity, institutions and public policy: agile decision-making in a turbulent world. Cheltenham: Edward Elgar.

Salmen, L. F. (2002). Beneficiary assessment: an approach described. World Bank, Social Development Paper 10. Washington DC: World Bank.

Scott, J (1998). Seeing like a state: how certain schemes to improve the human condition have failed. New Haven, CT: Yale University Press.

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.

van Hemelrijck, A. (2016). Methodological reflections following the second PIALA pilot in Ghana. Rome: International Fund for Agricultural Development.

van Tulder, R., Seitanidi, M., Crane, A., Brammer, S. (2016). Enhancing the impact of cross-sector partnerships: four impact loops for channelling partnership studies. Journal of Business Ethics, 135:1-17.

White, H. (2010). A contribution to current debates in impact evaluation. Evaluation 16(2), 1-11.

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

White, H., Raitzer, D.A. (2017). Impact evaluation of development interventions: a practical guide. Manila: Asian Development Bank.

Wilson-Grau, R., Britt, H (2013). Outcome harvesting. Cairo: Ford Foundation, Middle East and North Africa.

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

Youker, B (2013). Goal-free evaluation: a potential model for the evaluation of social work programs. Social Work Research, 37(4):432-38.