Case and evidence selection for robust generalisation in impact evaluation

James Copestake, University of Bath
April 2020

Pre-publication version (final version forthcoming in *Development in Practice*)

Abstract

What wider lessons can be drawn from a single impact evaluation study? The paper examines how case study and source selection contribute to useful generalisation. Practical suggestions for making these decisions are drawn from a set of qualitative impact studies. Generalising about impact is a deliberative process of building, testing and refining useful theories about how change happens. To serve this goal, purposive selection can support more credible generalisation than random selection by systematically and transparently drawing upon prior knowledge of variation in actions, contexts, and outcomes to test theory against diverse, deviant and anomalous cases.

Keywords

Case selection, Case studies, Causation, Generalisation, Impact evaluation, Qualitative research

Introduction

It is only human to support a causal claim with a telling example, and in the next breadth to dismiss somebody else's story as merely anecdotal. But this does not mean generalisation from a single case should always be dismissed. The credibility of a causal claim depends on whether it makes logical sense and is consistent with supporting evidence. Indeed if the sole purpose of an evaluation is to explain causation within a single case (e.g. that my action X helped you to achieve Y) then that is enough. More commonly, however, the usefulness of a causal claim also depends on whether it is likely to be true in another context (e.g. that your action X to help somebody else will also achieve Y). Such generalisations inform decisions to close, extend, copy, emulate, scale-up or modify projects because these decisions all entail adapting lessons from one context to another. One way to assess an impact study is to ask how far it strengthens the core ideas we use to make such connections. Yin (2013:327) explains that "…the preferred manner of generalizing from case studies and case study evaluations is likely to take the form of making an analytic or conceptual generalization... The desired generalization should present an explanation for how an evaluated initiative produces its results (or not). The explanation can be regarded as a theory of sorts – certainly more than a set of isolated concepts."

This paper explores how two choices we make in the design of impact evaluation studies affects their power to support robust generalisation: (a) selecting the sources of evidence for a single case study to identify the most important causal processes within it, and (b) choosing which case study to select in the first place. Source selection usually falls to whoever conducts the case study: which farmers to interview when evaluating an agricultural project, for example. Case study selection, in contrast, usually rests with the

organisation that commissions the study – which out of a portfolio of agriculture projects to evaluate, for example. The word 'project' is used here as shorthand for any activity aiming to achieve a development outcome within a defined area and time period.

Any case study is likely to contain causal links between many drivers and outcomes. These can be linked together in a causal map that may include chains with multiple links, branches and loops. But a causal map is also only ever the skeleton of the full story. For example, one showing how an agriculture project affects child nutrition will be a generalisation across the experience of many different farming households, including some affected indirectly. Drawing causal maps involves deciding what level of detail or granularity to include. The researcher must judge whether wider generalisation is facilitated by leaving out some links (and the idiosyncrasies of some individual cases) because they are less relevant to other contexts, and could be a distraction. This process can be aided by comparing new evidence against the commissioner's prior understanding set out in the form of the project's "theory of change" (Vogel, 2012). By confronting prior theory, a study goes beyond demonstrating whether the project worked to addressing *how* it worked, and hence whether it might work in other situations (Cartwright & Hardie, 2012:137; Woolcock, 2013). A key study design issue is how to identify case studies and sources of evidence with the greatest potential to confirm, confound or augment prior theories.

The remainder of this section briefly reviews how this issue is addressed within quantitative, process tracing and realist evaluation traditions.[1] The paper then draws on action research using the Qualitative Impact Protocol (QuIP) to propose some general principles for case study and source selection in qualitative impact evaluation. The concluding section reflects on these, and emphasises the deliberative or judgemental as well as technical nature of these choices. In so doing it demonstrates why random selection is often *not* the best selection strategy.

Quantitative approaches to impact assessment use or exploit variation in the exposure of a population to project activities (X) in order to investigate how this correlates with selected outcomes (Y), controlling statistically – as far as possible - for non-project or confounding factors (Z). Randomised controlled trials (RCTs) are often cited as most able to deliver internally valid impact estimates of this kind. With large enough samples it may also be possible to generalise from an RCT across the wider population from which treatment and control samples were selected. But RCTs are limited in their capacity to explore how and why impact varies within the treatment group as well as across the assessed population (Deaton and Cartwright, 2018). Woolcock comments that "… having expended enormous effort and resources in procuring a clean estimate for a project's impact, the standards for inferring that similar results can be expected elsewhere or when 'scaled up' suddenly drop away markedly." (2013:230)

The generalisability of RCTs can be enhanced by conducting multiple studies, and then conducting a systematic review of them all. However, the more heterogeneous and complex the field the less clear it is how to do this. Woolcock's assessment is that "… development professionals still lack a useable framework by which to engage in the vexing deliberations surrounding whether and when it is at least plausible to infer that a given impact result (positive or negative) 'there' is likely to obtain 'here'" (2013:231). Yin goes further by

suggesting that this requires moving beyond seeing generalisation as a "sample-to-population logic" issue. Instead, he advocates viewing it as a problem of "analytical generalization", or "… the extraction of a more abstract level of ideas from a set of case study findings, ideas that can pertain to newer situations other than the case(s) in the original case study" (2013:325).

At the core of ***process tracing*** is identification of a causal process that is both sufficient to explain a defined outcome within a single case study, and that fits available evidence more plausibly than any alternative explanation. This entails selecting and integrating sources of evidence to build up the best story and reject others, with attention often focused on finding key evidence that permits choosing between them. This sequential approach to source selection based on accumulated prior knowledge departs radically from classical statistical sampling to estimate a population mean. The latter involves no sequential weighting of observations, hence each observation adds equally to confidence in the accuracy of sample-to-population generalisations.

Beyond the single case, process tracing can be used both to build theory and to test whether it applies to a wider population (Beach and Petersen, 2012:3). Gelman and Basboll (2014) suggest that to be more than just illustrative, an exploratory case study should be both *anomalous* (represent aspects of life that are not well explained by existing theory), and *immutable* (documented richly enough to permit critical examination of alternative theories). When it comes to theory-testing then every new case adds something: it only takes one case to refute a prior theory that X is always sufficient to cause Y, for example (Flyvjerg, 2006:228). Dion (1998) uses Bayes' Rule to demonstrate that if a researcher starts out indifferent about whether X is a necessary condition for Y, and just five randomly selected country case studies all reveal X did indeed precede Y, then confidence in the causal claim rises from 50% to 95%. However, causal theories are rarely so simple. Goertz and Haggard (2019:25) conclude a survey of the issue that "…we have barely begun to systematically analyze the crucial decisions and the options in case study-generalization methodologies."

Turning to ***realist evaluation***, generalisation can be equated with the search for "middle range" theory in the form of multiple "Context-Mechanism-Outcome configurations" (Pawson, 2013).[2] These aim to explain "what works for whom in what circumstances" by combining inductive theory building with deductive theory testing in an iterative process referred to as abduction. Truth is hidden, and getting to it entails protracted confrontation of theory with multiple and often inconsistent sources of evidence, kept honest by transparent processes of peer review and "organised distrust" (p.18). Pawson (2013:14) highlights three characteristics of good realist evaluation. First, it should employ multiple "data medium methods", with qualitative approaches often leading the way in identifying cognitive mechanisms within cases, and quantitative data used more to map diverse cross-case contexts and outcomes (p.19). Second, more than one CMO configuration will be needed to do justice to the complexity of most kinds of project. Third, failure to ground a realist evaluation in theory will "…end with explanations that are *ad hoc* and piecemeal" (p.27). As with process tracing, it appears that robust generalisation in the realist tradition of evaluation best builds simultaneously on a triad of within case (small n) causal analysis, cross-case (large n) analysis of variation, and congruent theory (Goertz, 2017).

Lessons from use of the Qualitative Impact Protocol (QuIP)

This paper draws on five years of action research with the QuIP. This is briefly described below, while a comprehensive description can be found in Copestake et al. (2019b Annex). Development of the QuIP is itself an example of a quest for useful generalisation. It was first drafted after a three year period of grant-funded methodological design and testing on four rural development projects in Ethiopia and Malawi. Initial reactions were encouraging, but the awkward question remained: how well would it work in other contexts? To address this question, a non-profit company was established to conduct a larger number of QuIP studies. Over two years (2016-17) it conducted 17 further impact studies in 12 countries across a range of fields. The goal of this action research, encouraged in part by Stern et al. (2012) was not to set the QuIP up as a rival to other approaches to impact evaluation but to employ it as a device for stimulating wider discussion of methodological options. This paper stems from the interest shown by many commissioners of these studies not only in obtaining credible evidence of the impact of specific projects, but doing so in a way that supports more general impact claims.

In brief, design of the QuIP drew on a range of established approaches, including process tracing and realist evaluation, with the goal of offering practical guidelines for checking how a selected project has affected its intended beneficiaries. It does this by asking them - individually and in focus groups - what has changed in the period since it started, across specified domains of their life, and why they think this happened. The guidelines are based on a single visit to the project area by two independent field researchers to collect and document at least 24 semi-structured interviews and four focus groups using established qualitative methods (e.g. Skovdal and Cornish, 2015). Specific studies have been based on multiples or variants of these numbers.[3] A novel feature is that field researchers are given as little knowledge of the project being evaluated as possible, leaving them and respondents unaware of the precise confirmatory focus of the study. This feature is referred to as *double blindfolding*, and aims to reduce confirmation bias (Copestake et al., 2018).

The data collected can be used both to confirm whether reported outcomes are consistent with the project's goals, and to explore incidental drivers of change (Copestake, 2014). Analysis is based on thematic coding focused on identifying and ordering causal claims embedded in reported narrative text.[4] The analyst codes each causal claim in two ways. First, exploratory or inductive coding identifies different drivers and outcomes of change, and whether the respondent perceives them positively or negatively. Second, confirmatory coding classifies causal claims according to whether they *explicitly* link outcomes to the specified project, do so in ways that are *implicitly* consistent with the commissioners' theory of change, or are *incidental* to it. Once the data is coded in this way the software can generate tables, charts and causal maps to visualise what they reveal about causal links. The software also permits instant reference from visualisations of the coded data back to the underlying text. A central part of the analysis is to combine coded causal claims into causal maps that reflect what respondents collectively perceive to be the most important drivers of change in different dimensions of their wellbeing. The larger the number of interviews and focus groups undertaken the more data there is with which to do this, constrained only by the size of the budget and the capacity of a single analyst to integrate the data.[5]

The narrative models generated by the QuIP are built up from statements such as 'Y went down because X went up', and so are quite different from paradigmatic maps of causal relationships within a generalised system. Being based on the lived experience of respondents they may indeed be inconsistent with other sources of evidence, and for this reason QuIP data invariably feeds into a wider process of data integration and interpretation, including fully unblindfolded joint sense-making events.

Sample selection is based on two overlapping criteria, one more exploratory and the other more confirmatory. *Thematic saturation* is concerned with ensuring that enough data is collected to identify the most important causal processes arising from the project, expected or otherwise. *Bayesian updating*, in contrast, seeks evidence to reinforce or undermine prior theory about the causal processes triggered by the project. Both criteria favour samples that capture as much heterogeneity as possible in relation to characteristics of respondents likely to affect impact (see below). Consequently, case selection can benefit from being informed by close knowledge of context, project implementation and characteristics of the intended beneficiaries. But there is also a need for transparency about how the selection is made in order to guard against the risk of 'cherry picking' or other forms of bias.

Source selection within QuIP studies: illustrative examples

Table 1 provides summary information about seven case studies explored more fully in Copestake et al. (2019b). The total number of interviews and focus groups conducted per study ranged from 32 to 96, out of intended beneficiary populations. Selection usually proceeded in two hierarchical steps, cross-cluster and within-cluster (cf. Wilson, 2005). Operational data at cluster level (coop, factory etc) was used for purposive selection, with one cluster often expected to be conducive to project success, and a second less so. Within-cluster source selection then relied on random or opportunistic selection of an agreed quota of interviews from sub-groups defined by gender, age, location or other information available about individual project participants. For example, the Save the Children sample comprised quota samples of six people belonging to each of five different kinds of community group spread across four villages and two districts. In Frome, in contrast, no secondary data was available so respondents were selected opportunistically by approaching them *in situ* in selected parks.

A further issue illustrated by Table 1 is what determines the absolute number of interviews and focus groups per case study, as well as within clusters.[6] In principle, this depends on the prior expectations of the commissioner and their desired "certainty threshold" (Copestake et al., 2019b: 43-45). However, in practice, sample sizes were more influenced by norms internalised by commissioners about what constituted a reasonable budget for the overall study. The evaluator's role is then to design a data source selection strategy to maximise potential to generate useful additional evidence within this budget constraint. Part of the feedback process is then to evaluate whether sufficient evidence had indeed been amassed, or whether collection of additional evidence can be justified: an iterative approach that is also consistent with the reality that impact occurs through time, as does the need for evidence of it. In conducting the C&A evaluation study in Mexico, for example, a mid-course decision was made to add to the number of interviews.

Table 1. Selection of interview and focus groups selected for QuIP studies

Case study	Population frame for source selection	Stratification	Interview sample	Focus group selection
Diageo; malt barley promotion; Ethiopia.	6,000 barley suppliers to Diageo belonging to 39 cooperatives in Oromia Region.	Cooperatives, then farmers classified by value of credit received and quantity of barley delivered.	24 farmers in each of two coops, further stratified by village (total 48).	Younger/older men/women for each selected coop (total 8).
C&A Foundation; garment worker training; Mexico.	23 factories participating in the project over one or two years.	Six factories; operators and supervisors; men and women.	17 supervisors and 16 operators across six factories (total 33).	One in each of four factories (total 4).
Terwilliger Center for Innovation in shelter; housing microfinance; India.	31,629 housing loan recipients from two MFIs, in 12 in Tamilnadu and 140 in Kerala.	Rural & urban branches; first or repeat borrowers.	Quota sample of 9 men & 9 women from rural & urban branches of each MFI (total 72).	Rural/urban first/repeat borrowers for each MFI (total 8).
Tearfund; church and community mobilisation; Uganda.	100,000+ members of two church denominations.	Regions (North and East selected) and congregations from two church denominations.	12 interviews in each of four villages - two in East and two in the North (total 48).	Younger/older men/women in two regions (total 8).
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania.	8,775 project participants in two districts.	Two villages in each district, belonging to one of five different types of project group.	Quota sample of six per project group, spread across the four villages (total 30).	One with men and one with women in each district (total 4).
Global Seed Health Partnership; Malawi, Tanzania & Uganda.	Students taught by 186 volunteers in 27 institutions in five countries.	Medical, nursing and midwifery colleges (four in each of three countries).	Six interviews with students and six with staff per college (total 72).	Four per college (total 24).
Frome Town Council; promoting use of green urban spaces; England.	Park users out of a town population of 27,000.	Day-time users of five parks in a selected week.	18 women & 14 men; 22 with children & 20 with dogs (total 32).	None.

Source: Adapted from Copestake et al. (2019)

Principles of source selection within a case study

This section builds on the foregoing discussion by proposing some general principles for better source selection for qualitative impact evaluation of a project intervention (X_i) subject to variation among intended beneficiaries (i) within a population (N), expected to influence an outcome variable (Y_i), subject to the influence of an index of non-project contextual factors (Z_i). The goal – and it is important that commissioners of studies are clear on this point - is *not* to estimate the average effect of X on Y across N, but to gain as much insight as possible into why the project has a *differential* effect. How best to do this depends on what information is available about X, Y and Z at the study design stage. These are discussed below in turn.

If presented with a list of participants without any other information about them, then source selection is best done randomly, and in a transparent way that minimises the possibility of conscious or unconscious selection bias. In practice, however, random selection is rarely the best option, because additional information is usually available that can be used to select sources likely to add more to prior understanding (Seawright and Gerring, 2008:295). First, there is operational data about how intended beneficiaries were affected by the project. If different people were exposed to two different project components (X^1 and X^2) then these effectively constitute two separate case studies. But if groups overlap, and components interact then it is harder to identify which participants to study to reflect as fully as possible the diversity of exposure across the resulting project "design space" (Pritchett et al., 2013). Stratification of the sample is particularly important if part of the purpose of the study is to inform decisions about which of a range project components and combinations to expand or to stop.

Second, source selection can also utilise information about changes in key outcomes (dY_i), such as might be obtained through comparison of baseline and endline surveys. This data makes it possible to select intended beneficiaries who did better and worse than was typical across the population in order to identify the causal processes explaining this variation. An exploratory study will go further to the extreme in search of thematic saturation, whereas a purely confirmatory study is likely to maximise Bayesian updating by selecting on more frequent cases that are neither extreme nor typical (Seawright, 2016). Unfortunately however, data on dY_i is often simply not available. Not one of the QuIP case studies discussed above, for example, was able to draw on 'before-after' outcome data to inform source selection, although respondents for the Diageo survey were selected on the basis of an analysis of barley deliveries to the company relative to the value of inputs farmers received at the beginning of the season.[7]

The third category of data that can be used to inform source selection concerns non-project influences on project outcomes (Z_i). Such data can be utilised to ensure sufficient evidence is obtained across different intended beneficiary sub-groups - e.g. younger and older women and men. The commissioner may also have a particular interest in feedback on whether the project works better in different contexts – e.g. rural and urban. Where feedback on heterogeneity of this kind is explicitly sought then it is important to ensure a minimum number of participants in each sub-category are included. For example, the threshold suggested in the QuIP guidelines is six.

For confirmatory studies, data about Z_i can also substitute for lack of data about dY_i. If theory suggests that a non-project variable (baseline income, for example) is an important causal driver of dY_i then there is a case for purposefully selecting participants who are positive and negative deviants on this factor. This helps to ensure evidence is collected to test the theory where preconditions for its success are both relatively stronger and weaker. The more elaborate is prior theory about conditions for project success, the more scope there is for selecting a set of respondents to reflect this variation, thereby maximising its potential explanatory power.

Where data is available for Z_i *and* dY_i then an even more nuanced approach to confirmatory assessment would involve first using survey data to model the relationship between the two (e.g. using regression analysis) and then selecting on participants who departed most sharply from the outcome predicted for them. A good strategy for using additional data to test and refine a prior theory is to confront it with anomalous cases or outliers from the model.

These different options are summarised in Table 2. The list of options is not exhaustive, since some case studies may permit selection on X, dY and/or Z, and with different degrees of emphasis on theory-making (exploratory) and theory-testing (confirmatory) goals.

Table 2. Options for source selection depending on availability of data across the specified population

Option	Treatment data (X)?	Outcome data (Y)?	Contextual data (Z)?	Comment
A	No	No	No	Random selection across full population is the only option
B	Yes	No	No	Select randomly from quota samples across categories of treatment or exposure
C	No	Yes	No	Select purposively to include positive and negative deviants
D1	No	No	Yes	Select purposively to reflect important dimensions of variation across the population (e.g. gender, age)
D2	No	No	Yes	Select purposively to include likely positive and negative deviants according to prior theory.
E	No	Yes	Yes	Select purposively to include anomalous cases poorly explained by prior theory linking Z and Y.

QuIP case study selection

Having explored evidence selection within a study, this section considers how to select a single case study in the first place, as well as the grounds for selecting more than one. If two case studies are being chosen across a programme or portfolio of projects with the aim of generalizing across them all then the problem is formally very similar to the cluster selection problem already considered above. Random selection is again unlikely to be the best

strategy because it fails to use available information about inter-project heterogeneity to identify potentially the most insightful pair. This is explained by Seawright and Gerring (2008:295) who list seven possible purposive alternatives to random selection: typical, diverse, extreme, deviant, influential, most similar and most different. Bayesian and saturation criteria point towards selecting for at least some diversity: one positive and one negative deviant, for example. In contrast, if only one case study can be funded, then there are arguments in favour of selecting the one that is most typical, but also most positive (to identify 'best practice'), most negative (to encourage 'falling forward') or indeed potentially the most influential (cf. the two criteria for influential story selection proposed by Gelman and Basboll, 2014). It follows that case study selection can only be discussed meaningfully with reference to how the commissioner envisages it being used.

To illustrate, it is worth returning to the seven QuIP studies already cited. In interviews conducted one to two years after they were completed, commissioning staff mentioned three main potential uses of the studies: to inform learning among project stakeholders, to influence internal operational decisions, and to provide material for dissemination to a wider audience. The third use is particularly relevant to the issue of generalisation. Rather than focusing on within-project decision-making the studies often aimed to inform debate over what the commissioning organisation was claiming to do more generally. In other words, being able to make robust generalisations from QuIP studies beyond the selected project was more than an incidental side-product from project specific learning. Table 3 illustrates this by suggesting generalisations (or middle range theories) addressed by the seven studies.

How explicitly these were articulated by the commissioning organisations varied. The Terwilliger Center, for example, had an established theory of change that it was seeking to test or confirm, whereas Tearfund was interested in finding out more about how communities responded to CCM. The C&A project in Mexico provided an interesting intermediate case. Findings confirmed the implementing agency's aim to empower factory workers by raising their aspirations. But for the commissioning agency the study was more exploratory and revealed to them that the project was badly aligned with its global goals.

In practice, most commissioners combined confirmatory and exploratory objectives, both testing a prior view and open to other possibilities. However, several deliberately opted to evaluate projects that were relatively well established, thereby allowing for identification of causal links and impact pathways over a longer period of time. Where commissioners were also piloting the QuIP as a methodology, it also made sense to select a case study whose impact was already well understood. The first five commissioning agencies listed in Table 3 operate across many countries, and the case studies reviewed constituted only one data point in the flow of evidence they were producing to justify their activities. Indeed four subsequently commissioned further QuIP studies elsewhere. Through a sequence of case studies across an evolving portfolio of projects there is scope to shift the objective from being more exploratory to being more confirmatory, and to select projects with more explicit hypotheses in mind. It also constitutes a radically different approach to case study selection than one aimed at delivering a representative 'sample-to-population' snapshot of impact across a project portfolio at one point in time.

Table 3. Generalisations being tested or explored by selected QuIP studies

Case study	Suggested theory	Potentially wider scope
Diageo; malt barley promotion; Ethiopia.	Purchasing malt barley as a cash crop from small-scale farmers does not have unintended negative social consequences.	Development of sustainable supply chains for commercial brewing operations in many low and middle income countries.
C&A Foundation; garment worker training; Mexico.	Garment factories can offer their employees 'empowerment' training that improves both their relational wellbeing and productivity.	Promotion of decent work in the factories of more than 100 C&A suppliers in Mexico, as well as across its global production network.
Terwilliger Center for Innovation in shelter; housing microfinance; India.	Incremental home improvement funded by commercially self-sustainable housing microcredit benefits borrowers and their households.	A global programme of promoting commercial housing microfinance, including a growing portfolio of support to microfinance institutions (MFIs) in India.
Tearfund; church and community mobilisation; Uganda.	Faith-based community development can have a positive transformative effect, even when not linked to material transfers.	Tearfund has ongoing CCM partnerships in more than 25 countries. It started in Uganda in 2001.
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania.	Important synergies arise from integrating agriculture, nutrition and gender training activities together, rather than intervening in each area separately.	East and Southern Africa account for nearly one third of Save the Children's funding, with nearly half allocated to health, nutrition and livelihood promotion activities
Global Seed Health Partnership; Malawi, Tanzania & Uganda.	American health care volunteer educators can make a positive contribution to the training of doctors, nurses and midwives.	GSHP was placing volunteer educators in 27 health training institutions in five countries across Africa.
Frome Town Council; promoting use of green urban spaces; England.	Council supported amenities and events can have a positive effect on citizens' wellbeing by influencing the way they use parks and other green spaces across the town.	The Council was interested in exploring ways of evaluating the impact of other activities also. Composed entirely of independent (non-party) counsellors, it has built up a wider reputation for innovation.

Source: Adapted from Copestake et al. (2019:223)

This discussion also reveals the judgemental aspect of case study selection. Each QuiP study opened the commissioner up to independent scrutiny, creating space for reflection on what it was trying to achieve and how. This is potentially risky, and hence depends on how open the commissioner is willing to be. Selecting a case study project that is contextually atypical or badly implemented, makes it easier to ignore or to marginalise if findings are disappointing. But from an exploratory perspective, the study of an anomalous project may usefully inform more radical departures from established thinking, activities and areas of operation (Gelman and Basboll, 2014). Selecting case studies of activities that are core to the organisation's identity and field of operation (e.g. CCM in Uganda for Tearfund) is potentially more challenging, although less so if there are strong prior grounds for believing findings are likely to be positive. Finally, perhaps the most interesting case studies are those that focus on the margins of the commissioner's own confidence in its core generalisations,

with respect to context. And while the choice of case study projects usually resides with the commissioner of a study, it is useful for researchers to be aware of their political status, and hence how strongly findings will be contested.

The QuIP as a generalizable evaluation methodology

As a postscript, it is useful to consider how far the seven case studies cited here constitute an adequate basis for generalisation about the QuIP's own usefulness. The case studies chosen were restricted to projects purposively selected by commissioners rather than imposed randomly on a larger population of projects. Selection of the seven (from 17 possible studies) mostly aimed for contextual diversity. More important, however, is the iterative and adaptive nature of QuIP case study selection, with later studies informed by earlier experience, generating sufficient feedback to encourage continued exploration of contexts within which the approach can be useful.

Conclusions

This paper suggests that development organisations rely on core theories about how they can achieve a positive impact. These are documented in strategy statements and theories of change, are part of an organisation's culture, and indeed the moral narrative of staff. They can be equated to the realist idea of middle range theory, linking project activities in different contexts to intended outcomes via causal mechanisms that include changes in the way people think. Experience using the QuIP (as documented in Copestake et al., 2019b) suggests that impact studies can be justified not only by what they reveal about the specific projects evaluated, but also by their role in supporting or challenging the generalisations underpinning the commissioning organisation's wider development practice.

The paper offers insights into both technical and judgemental aspects of how far impact evaluation can contribute to more robust generalisation of this kind. The focus of the more technical discussion has been on source selection within single case studies. An important conclusion is that robust generalisation is not well served by borrowing ideas from sample selection used in quantitative impact assessment. Instead, the paper emphasises the potential superiority of purposive over random selection to support robust generalisation. While, this is not a new finding for specialists in case study research (see particularly Seawright & Gerring, 2008) the paper reinforces the point by assessing selection against both thematic saturation and Bayesian updating criteria, in support of exploratory and confirmatory goals respectively. Table 2 also sets out a structured approach to purposive source selection, based on prior knowledge of the population. It provides a guide to how limited resources can best be deployed to maximise probative value by building on what is already known about variation in project interventions, context suitability and measurable outcomes. This favours strategies for capturing extreme, deviant and anomalous rather than typical cases.

A similar logic also applies to selection of case study projects by the commissioner. By drawing on case study examples, the paper illustrates how in complex settings both case and source selection also depends upon the risk appetite of commissioners. It thereby emphasises the practical limitations of adopting a purely technical approach to case selection. Rather, impact evaluation studies can be viewed as managed spaces for political deliberation and value judgement. Being clear and open about this is particularly important

at the design and contracting stages of a study. Case selection entails exposing cherished generalisations to critical scrutiny, and it is useful for the commissioner to think through how their organisation will respond to negative as well as positive findings. This in turn may influence choice of case studies (e.g. preference for negative deviance and anomalous cases), as well as the number and mix of within-case sources of evidence needed to defend potentially awkward findings.

Accepting a view of impact evaluation as empirically informed political deliberation departs from the norm of referring to it as a purely technical investment in finding out 'what works'. It also highlights the importance of clarifying the logic behind design choices, including why purposive rather than random case selection can contribute to more robust generalisation.

Notes

[1] For fuller reviews of methodological options see BOND (2015) and Stern et al. (2012). Copestake (2019b Chapter 2) also compares the QuIP with 32 other approaches.
[2] This 'CMO' terminology does not map perfectly onto the 'ZXY' shorthand used here, because project actions (X) are defined by realist evaluators as part of the context (C), while the term 'mechanism' refers to the mostly unobservable cognitive causal links through which context (combining X and Z) generate outcomes Y (Blamey and Mackenzie, 2007).
[3] Interviews elicit narrative stories about drivers of change from individuals, while focus groups are organised by age and gender to elicit explanations for change experienced more widely within the location. Individual interviews are usually conducted first mainly to reduce contamination of narrative statements across sources.
[4] This lies somewhere between the mechanical thematic coding of self-evidently real "diamonds in the sand" and the more creative "organic" coding of meanings (Braun and Clarke, 2016).
[5] As the dataset becomes larger it becomes harder for the analyst to select codes inductively in a way that is informed by immersion in all the data (Copestake et al., 2019a). For this reason, larger projects are better evaluated by more than one QuIP study, each coded independently, followed by meta-analysis across them.
[6] Thematic saturation can in principle be measured. For example, Hagaman and Wutich (2016) found that "…16 or fewer interviews were enough to identify common themes from sites with relatively homogeneous groups." However, such findings are likely to be highly context-specific (Braun and Clarke, 2016). Smaller samples may be sufficient for narrowly focused confirmatory studies, as revealed in the earlier discussion of process tracing.
[7] A large literature explores selection of cases within a 2x2 matrix comprising X=1 or X=0, and dY=1 or dY=0, and where X=0 signifies non-participation, and dY=1 signifies a positive outcome (Goertz, 2017). QuIP mostly selects sources where X=1, but selecting cases where X=0 but dY=1 (equifinality) may also be useful.

Acknowledgements

The paper was made possible by all those who contributed to QuIP studies, as acknowledged in Copestake et al. (2019b). Gary Goertz, Steve Powell and Fiona Remnant

also commented helpfully on an earlier draft, and I am grateful for the constructive comments from two anonymous referees.

Disclosure statement

The author is a Director and co-founder of Bath Social and Development Research, a non-profit company set up, under licence to the University of Bath, to promote better evaluation through adaptation and use of the QuIP.

Funding

The original design and development of the QuIP was funded by research grant ES/J018090/1 jointly from the UK Department for International Development (DFID) and the Economic and Social Research Council (ESRC).

Notes on the contributor

James Copestake is Professor of International Development at the University of Bath, UK, and has a particular interest in development finance modalities and their evaluation. He is also a trustee of INTRAC.

References

Beach, D., Pedersen, R. (2012) *Process tracing methods: foundations and guidelines*. Ann Arbor: University of Michegan Press.

Blamey, A., Mackenzie, M. (2007). Theories of change and realistic evaluation. *Evaluation*, 13(4):439-455.

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.

Braun, V., Clarke, V. (2016). (Mis)conceptualizing themes: thematic analysis, and other problems with Fugard and Potts; sample-size tool for thematic analysis. *International Journal for Social Research Methodology*, 19(6):739-43.

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

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

Copestake, J., Remnant, F., Allan, C., van Bekkum, W., Belay, M., Goshu, T., Mvula, P., Thomas, E., Zerahun, Z., (2018). Managing relationships in qualitative impact evaluation of international development practice: QuIP choreography as a case study. *Evaluation.* 24(2):169-84.

Copestake, J., Davies, G., Remnant, R. (2019a). Generating credible evidence of social impact using the Qualitative Impact Protocol (QuIP): the challenge of positionality in data coding and analysis. In *Myths, methods and messiness: insights for qualitative research analysis*

edited by Clift, B., Gore, G., Bekker, S., Batlle, I., Chudzikowskil, K., Hatchard, J. University of Bath, Dept of Health.

Copestake, J., Morsink, M., Remnant, F., editors (2019b). *Attributing Development Impact: the Qualitative Impact Protocol case book*. Rugby: Practical Action Publishing. bit.ly/QuIP-OA

Deaton, A., Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine,* 210:2-21.

Dion, D. (1998). Evidence and inference in the comparative case study. *Comparative Politics*, 30(2):127-45.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2):219-45.

Gelman, A., Basboll, T (2014). When do stories work? Evidence and illustration in the social sciences. *Sociological Methods and Research*, 43(4):547-570.

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

Goertz, G., Haggard, S. (2019). *Generalization, case studies, and within-case causal inference: large-N qualitative analysis (LNQA).* Draft (version 14) prepared for the Oxford handbook on the philosophy of political science workshop, September 2019, Washington State University.

Hagaman, A.K., Wutich, A (2016). How many interviews are enough to identify metathemes in multisited and cross-cultural research? *Field Methods*, 29(1):23-41.

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

Pritchett, L., Samji, S., Hammer, J (2013). It's all about MeE: using structured experiential learning ("e") to crawl the design space. *Center for Global Development, Working Paper 322.*

Seawright, J., Gerring, J. (2008). Case selection techniques in case study research. *Political Research Quarterly*, 61(2):294-308.

Seawright, J. (2016). The case for selecting cases that are deviant or extreme on the independent variable*, Sociological Methods & Research*, 45(3):493-525.

Skovdal, M., Cornish, F. (2015). *Qualitative research for development: a guide for practitioners*. Rugby: Practical Action Publishing.

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.

Vogel, I. (2012). *Review of the use of 'theory of change' in international development*. London: Department for International Development. www.isabelvogel.co.uk.

Wilson, I. (2005). Some practical sampling procedures for development research. In *Methods in Development Research: combining qualitative and quantitative approaches*, edited by Jeremy Holland and John Campbell. Rugby: ITDG Publishing.

Woolcock, M. (2013). Using case studies to explore the external validity of 'complex' development interventions. *Evaluation*, 19(3):229-248.

Yin, R.K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, 19(3):321-32.