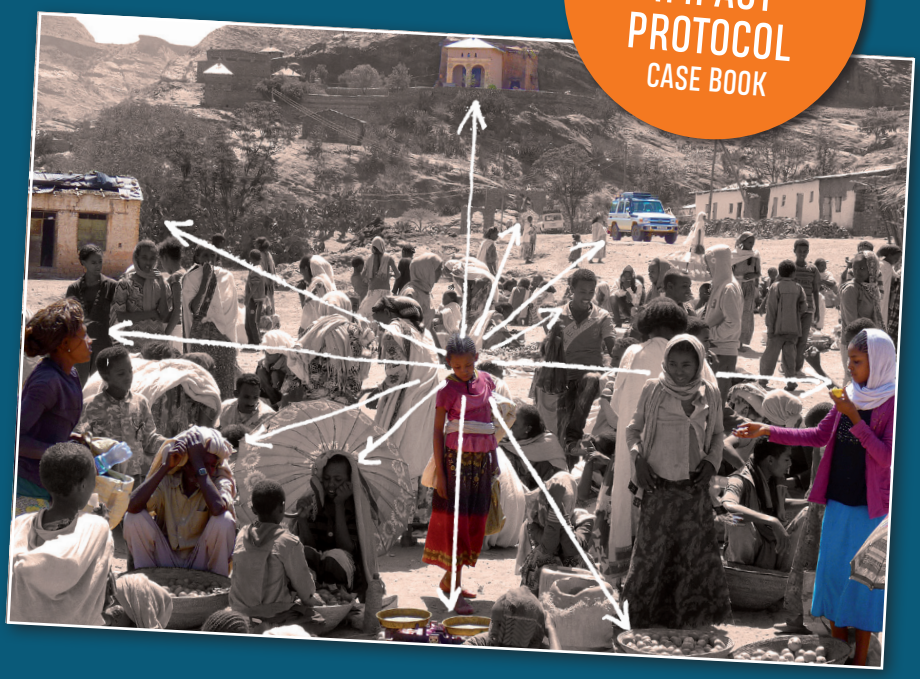
ATTRIBUTING DEVELOPMENT IMPACT

THE QUALITATIVE IMPACT PROTOCOL CASE BOOK



James Copestake, Marlies Morsink & Fiona Remnant

Attributing Development Impact

Praise for this book

'QuIP is well geared to do what it promises: it offers a simple, transparent method to deliver timely, cost-effective and credible causal attributions. And it is well grounded. The theory, history and case studies in this book show why we can trust that it can do what it says. QuIP is a really welcome contribution to methodology for causal inference.'

Nancy Cartwright, University of California San Diego and Durham University, UK

'The assessment of complex interventions is defined by the need to make difficult trade-offs: time, money, talent and support always seem inadequate. But such pressures only intensify the need for good theory, breadth of experience, depth of commitment to professional standards, and giving stakeholders a reasoned basis on which to act. The strategies and cases outlined in this insightful book demonstrate how this can be realized in practice. The Qualitative Impact Protocol enables applied social science to do its job: to faithfully uphold accountability norms while generating sound and usable conclusions.'

Michael Woolcock, World Bank and Harvard University

'An enormously important addition to impact evaluation approaches, with detailed examples and explanation. This book offers practical and theoretically informed guidance on how to bridge the increasing mismatch between the complexity of interventions (and the contexts in which they operate) and the counterfactual impact evaluation methods that are often advocated.'

Patricia Rogers, Director, Better Evaluation

Attributing Development Impact

The Qualitative Impact Protocol (QuIP) Case Book

James Copestake, Marlies Morsink and Fiona Remnant

PRACTICAL ACTION
Publishing

Practical Action Publishing Ltd
27a Albert Street, Rugby, Warwickshire, CV21 2SG, UK
www.practicalactionpublishing.org

A catalogue record for this book is available from the British Library.A catalogue record for this book has been requested from the Library of Congress.

ISBN 978-1-78853-024-8 Paperback
ISBN 978-1-78853-023-1 Hardback
ISBN 978-1-78044-747-6 eBook
ISBN 978-1-78044-746-9 Library Pdf

Citation: Copestake, J., Morsink, M., Remnant, F. (ed.) (2019) *Attributing Development Impact: the Qualitative Impact Protocol case book*, Rugby, UK, Practical Action Publishing, <http://dx.doi.org/10.3362/9781780447469>

Since 1974, Practical Action Publishing has published and disseminated books and information in support of international development work throughout the world. Practical Action Publishing is a trading name of Practical Action Publishing Ltd (Company Reg. No. 1159018), the wholly owned publishing company of Practical Action. Practical Action Publishing trades only in support of its parent charity objectives and any profits are covenanted back to Practical Action (Charity Reg. No. 247257, Group VAT Registration No. 880 9924 76).

Cover photo shows a market scene in Tigray, Ethiopia. Credit: James Copestake
Cover design: RCO.design
Printed in the United Kingdom
Typeset by vPrompt eServices

Contents

http://dx.doi.org/10.3362/9781780447469.000

List of figures, tables, and boxes

Figures

Tables

Boxes

Acknowledgements

Far more people have contributed to this book than the 17 chapter authors. Working backwards from the final outcome, Clare Tawney and colleagues at Practical Action were unfailingly encouraging, as well as flexible about timeframes. Marlene Buchy read and commented on a complete draft, while Ed Howarth, Emma Ford, and Anita Winter assisted with chapter editing. Jo Burman, Luisa Enria, Susan Johnson, and Morten Siersted all played supportive roles within Bath Social and Development Research (BSDR), and participants in the original ART project (Assessing Rural Transformations) are not forgotten as co-producers of earlier versions of the QuIP on which this book builds. Ideas have since been tested with participants in bidding exercises, commissioning discussions, conferences, lectures, seminars, training events, follow-up workshops, and informal conversations. This is an unavoidably incomplete list, but among those who have struck a helpful balance between challenging and supporting the QuIP project have been Catriona Dejean, Claire Hutchings, Clement Sefa-Nyarko, Danielle Dunn, Rachel Hayman, Michael King, Edoardo Masset, Graham Room, John Moffett, Joseph Zulu, Marc Theuss, Martin Walsh, Nazneen Kanji, Peter Mvula, Rick Davies, Tara Bedi, Tom Adams, and Wilm van Bekkum. A particular mention should also go to additional University of Bath students and alumni who have become involved in working on QuIP studies, including Alice Chadwick, Doreen Abalo, Jo Trotter, Kate Pincock, Rebecca Huovinen, and Tumaini Malenga-Duchoslavová. A still larger number of people made it all possible by participating in QuIP studies as managers, researchers, and (above all) as respondents.

We are grateful for financial support received from the Centre for Development Studies at the University of Bath for research and dissemination activities relating to the QuIP under grants ES/J018090/1 and ES/N015649/1 from the Department for International Development (DFID) and the Economic and Social Research Council (ESRC). Writing of the book was also made possible by a one semester sabbatical for James Copestake, while the University of Bath Alumni Fund contributed to the cost of making it freely available as an ebook. Not to be excluded is funding obtained by reinvesting profits earned by BSDR through commissioned studies, including those featured as case studies in the book.

Foreword

It is a pleasure to have been asked to introduce this case-book describing the principles and implementation of the Qualitative Impact Protocol – QuIP – a new and timely approach to impact evaluation in development settings.

In their own words, the authors set out to help 'investors with social or development goals assess whether they are achieving what they intend' and aims to 'explain variation in the wellbeing of intended beneficiaries, rather than quantifying average effects.' QuIP focuses explicitly on causal attribution drawing on beneficiary narrative reports analysed 'in relation to project theory and context (obtained mainly from project staff)'.

After initial development work, the QuIP approach has been refined through repeated application in the 16 planned instances of social development reported in this case-book. Compared with many methodological proposals for improved evaluation practice based on one-off pilots, this should give readers a great deal of confidence! More important in my view is the careful way the authors have analysed and reflected on their now extensive experience with QuIP. This undoubtedly makes it easier to bring together QuIP experience in a user-friendly form. This degree of reflexivity allied with repeated applications of a new approach are unusual in evaluation circles; but we can see with QuIP the payoffs of going down this path.

It is now over 6 years since an international team produced a report for the UK's Department for International Development (DFID) advocating the need for a 'broader range' of approaches to impact evaluation. At the time those of us who authored the DFID report concluded that impact evaluation designs were too often method-led rather than selected for their suitability to the programmes and contexts in which they were set. Furthermore established methods had serious weaknesses with regard to causal attribution that is the main purpose of impact evaluations.

The 2012 report noted how statistical techniques can easily ignore the qualitative realities of development from a beneficiary perspective. Quantitative evaluations and in particular experimental approaches are at risk of obscuring variations in programme effectiveness across intended beneficiaries, classically ending up with net-effects rather than identifying different effects for different subgroups – a preoccupation of QuIP. Qualitative methods are equally open to criticism: of bias or at the very least methods of data collection and analysis that are not transparent and can be challenged as to their quality. In the past before the relatively recent upsurge in interest in comparative case-study approaches, qualitative evaluation approaches have also been more convincing studying the single case. This makes generalisation

and the carry-over of lessons from practice into policy and management innovation difficult.

QuIP is avowedly 'qualitative' although explicitly incorporating concerns for rigour, transparency and replicability that have traditionally been mainly associated with quantitative evaluation approaches. QuIP also avoids over-ambitious claims – concentrating on the demands of causal attribution, rather than claiming to be an all-encompassing evaluation approach. This suggests future opportunities for QuIP to be integrated into evaluation portfolios that need to strengthen their understanding of causal claims.

QuIP confronts questions of bias in qualitative research through an unusual strategy. This consists of 'blindfolding' field researchers i.e. restricting their knowledge of the programmes or interventions that are being evaluated; and separating the roles of those responsible for data analysis and data collection. Combined with semi-structured interviews that focus more on beneficiary reports of *change* rather than on their awareness of projects and programmes, QuIP goes further than most evaluation research to ensure the impartiality and reliability of evaluation findings.

As Nancy Cartwright and her collaborators have argued, assessments of effects in one programme, however precise, are no guarantee that the same results will hold true for other programmes or even for the same programme at another time or in another place. These arguments underpin the need for theory to support generalisation beyond the single case. The growing interest in theory in its various forms, surprisingly prevalent in evaluation thinking these days, is present also in QuIP. Whilst QuIP eschews global theory – universal laws or sweeping generalisations – it *is* concerned with the 'prior theory' of programme commissioners; in Theories of Change that may inform programme design and strategy; and in 'middle level theory' that captures partial generalisation under specifiable conditions.

QuIP is an approach rather than a method - sharing logics in part with other case-based approaches (e.g. realist and QCA) and even Bayesian probability theory. A helpful feature of this volume is the way it makes explicit QuIP's connections – such as shared assumptions and principles as well as obvious differences – with a host of other evaluation methods and indeed whole families of evaluation approaches. In recent years the social sciences have seen a remarkable period of methodological creativity which has now begun to filter through to evaluation practice. QCA, Realist Synthesis, Outcome Harvesting, Process Tracing, Empowerment Evaluation, Agent based Modelling, Congruence Analysis, Systems Mapping, Participatory Action Research (and this is just a starting list) are now commonly deployed in evaluation studies. Things have certainly moved a long way from the time that straightforward surveys and interviews carried the greatest weight in an evaluator's toolkit. Yet in terms of some notion of an evaluation 'product innovation-cycle' many newer evaluation approaches now being field-tested can begin to resemble 'brands' and advocated by their followers to the exclusion of all others. Similarly partisan narratives are common at early stages in the development of

many new products and processes. We know that innovative research methods and techniques have really arrived when they become combined into multi-method packages. QuIP draws on an extensive repertoire of existing methods; delimits its methodological ambition to causal attribution; and focuses on a well-defined application space – investors in social and development goals. All of this makes future multi-approach collaboration more likely.

QuIP betrays its academic origins in Bath University in the way it takes on board so many new and emerging methodology and practice ideas from evaluation and the social sciences more generally. Yet one of the most striking features of QuIP is how after incubation in a University, the approach was refined and systematised in a dedicated market-based company delivering QuIP studies under commercial constraints. These are the kinds of real-world market settings that jobbing evaluators and consultants have to confront routinely: when cash and time are short; and when you only get taken seriously when you address sponsors' real and pressing needs. It is likely to be to QuIP's advantage that the protocol and its use has been honed in these kinds of settings.

One threat to the coherence of the wider evaluation community is a potential schism between academic researchers and practice-based evaluators. The former develop new and sophisticated approaches that practitioners often find it difficult to adopt and replicate. Indeed one of the justifications that commissioners of evaluation use not to adopt innovative yet appropriate evaluation methods is that the suppliers of evaluation, i.e. the consultants who do the work are not well-versed in these new approaches. By laying out a structured approach to applying the protocol, the designers of QuIP have set a good-practice marker in the sand. Methodological rigour, responsiveness to beneficiaries in practical development settings and theoretical self-awareness need not be inaccessible! And well-articulated research-based evaluation approaches can also be made accessible to practitioners.

Elliot Stern

Elliot Stern is Emeritus Professor of Evaluation Research at Lancaster University, and a Visiting Fellow at the University of Bristol. He is editor of the Journal titled 'Evaluation', was the founding President of the UK Evaluation Society and a past President of the European Evaluation Society. He is currently an Academician and Council member of the UK Academy of Social Sciences, and continues high longstanding association with the Tavistock Institute.

CHAPTER 1

Introducing the causal attribution challenge and the QuIP

James Copestake, Fiona Remnant and Marlies Morsink

This chapter provides an overview of the Qualitative Impact Protocol (QuIP) as an approach to impact evaluation. The QuIP was developed under commercial conditions in a wide range of contexts during 2016 and 2017, following a three year period of action research. Studies based on the QuIP help those who commission them to understand the causal links between diverse drivers of change (including their own actions) and wellbeing outcomes within a specified population. They do so by collecting and carefully analysing narrative accounts of change obtained from members of that population. The chapter reviews determinants of the quality and usefulness of such evidence, including: collaborative scoping of studies; 'small n' case selection; reducing bias in interview and focus group data through 'blindfolding' and open-ended questioning; systematic and transparent analysis of drivers and outcomes; and visualization and interactive interpretation of data with stakeholders. Within the field of impact evaluation the QuIP can fill the gap between internal performance assessment and more time-consuming and expensive survey-based or ethnographic research. It recognizes the need to address the challenge of causal attribution in a way that is integrated with commissioners' prior goals and theories of change, investment in routine monitoring, and a commitment to adaptive management. The chapter also provides a short summary of 10 diverse case studies of the QuIP, each of which explains how the QuIP was conducted, illustrates what evidence was generated, and reflects on methodological challenges encountered.

Keywords: impact evaluation, attribution, causality, qualitative research methods, international development, performance management

Introduction

Any organization aiming to bring about positive social change sooner or later confronts the problem of how to confirm whether it is being successful. For many actions, this seems easy enough: we can observe directly the immediate effects of making a gift, for example, being richly endowed with experience and language to help us imagine what would have happened if the gift had not been made. But the full effects of even apparently simple gifts – emotional, social, political, ethical, as well as material – can turn out to be surprisingly complex. For example, in her research into a government cash transfer

http://dx.doi.org/10.3362/9781780447469.001

programme in Ghana, Attah (2017) suggests that the direct material effect of money given to elderly urban beneficiaries mattered less than the opportunities it afforded them to be active givers, and not just takers, within their informal support networks (see also Ferguson, 2015).

If the effects of even simple actions are often more complicated than they appear at first sight, so too is the task of credibly sharing evidence of their effects with others. This book is concerned with the production of useful evidence about whether actions taken in the name of development (variously defined) are contributing to intended improvements in the wellbeing of specified individuals, households, and communities.[1] A first step is to find out how 'intended beneficiaries' themselves think their wellbeing has changed and why. Asking why is complicated, because the influence of any one cause invariably interacts with many other drivers of change. There are many pitfalls to addressing this attribution challenge. In picking one out of a range of possible causal explanations for a change we are open to many biases, including emphasizing the importance of our own actions compared with those of others, giving greater weight to more recent activities than those longer ago, and saying what we think the questioner would most like to hear.[2]

Obtaining 'beneficiary feedback' is particularly important for the many organizations that exist to promote global development goals, including multilateral and bilateral aid agencies, non-governmental organizations (NGOs), and the philanthropic arms and offshoots of international businesses. As they are not paying directly for the activities being carried out, intended beneficiaries cannot simply refuse to 'buy' them, as they could if they were consumers of a commercial product or service. And as the often marginalized citizens of countries remote from those of their 'intended benefactors', they often have limited political opportunities to voice their opinions about what is being done in their name. In other words, it is a characteristic of international development agencies that their activities are subject to a weak or even 'broken' feedback loop (Martens et al., 2002). This means that their moral and political legitimacy depends more heavily on finding other ways to demonstrate the impact they are having, including enabling intended beneficiaries themselves to provide such feedback.

In this book the term 'impact evaluation' is used broadly to refer to this process of collecting, interpreting, and using evidence on the ultimate effects of a specified activity, project or intervention (cf. White, 2010). Taking this broad process as our starting point, we focus on four more specific challenges facing development agencies when they evaluate impact. First, there is *goal specification and planning*, including defining what improvements a project aims to bring about and how. Second, there is *change monitoring*: the empirical task of measuring the direction and magnitude of change in these selected goals over time (or proxy indicators of them). Third, there is the *causal attribution* challenge: assessing the contribution or specific impact of the agency itself on these changes. Fourth, there is the challenge of *adaptive management*: using this evidence to improve what the agency is doing. We will

come back to these four challenges repeatedly in this book. Because they are closely interconnected, there is an obvious limitation to focusing on one of them in isolation. This is evident, for example, in the ongoing struggle to make the Sustainable Development Goals (SDGs) operational. For this reason the case studies in this book locate causal attribution as one part of a response to all four challenges. Nevertheless our central concern is with the attribution challenge; this is because we think it is not just important but also particularly difficult and often neglected.[3]

Why the neglect? Development agencies' incentive to collect and share attribution evidence is mixed. As individuals, our interest in checking up on the impact of charitable donations we make is weakened if the act of making the donation (and being seen to do so) subconsciously matters more to us than what it achieves.[4] In the case of international aid it also makes sense for individual donors and taxpayers to rely (or free-ride) on others to ensure that funds are being well spent. But the integrity of the professionals entrusted with spending funds is in turn tested by an incentive to avoid more rigorous causal attribution of aid – including burying negative findings if such attribution is likely to weaken political support for the aid given (Martens et al., 2002).

Private sector impact investors face similar disincentives, as do social enterprises aiming to promote businesses that can generate positive social as well as financial returns. Molecke and Pinkse (2017) observe that independent change monitoring is regarded by many operators of social enterprises as either too difficult to be credible ('immeasurable') or too expensive to be useful ('imprudent'); whereas causal attribution is regarded as either insufficiently contextualized to be credible ('incomplete'), or insufficiently precise to be useful ('irrelevant'). If selected indicators of social and environmental change (e.g. business growth, job creation, and reduced carbon emissions) are moving in the desired direction then there is limited incentive for impact investors to allocate further resources to clarifying how much of this favourable change can be attributed exclusively to their investment, rather than being fortuitous and something that would have happened anyway.[5]

Countering the incentives organizations have for avoiding the causal attribution challenge are strong reasons for addressing it head on, and for being seen to do so. This book is being written in the UK at a time of increasing demand from politicians, the media, and public, for better evidence of 'what works' and that international aid represents 'value for money'. A recent indication of this demand has been the rise in expenditure on randomized controlled trials (RCTs) to evaluate development interventions by the UK Department for International Development (DFID) and other large donors (Camfield and Duvendack, 2014). This increase suggests that latent demand had previously been constrained at least in part by uncertainty over how best to generate such evidence: a supply side constraint that RCT enthusiasts or *randomistas* claim to overcome. Debate over the merits of RCTs has helped to raise interest in the impact attribution challenge, but we argue it has also unhelpfully narrowed the way this discussion is framed (see Chapter 2).

The case for broadening the range of approaches to impact evaluation was recognized in 2012 by a widely cited DFID commissioned report (Stern et al., 2012). The need for other approaches also motivated the research that led to this book, including a concern that qualitative impact evaluation should be exploratory, rather than too narrowly focused on confirming project theory (Copestake, 2014). This took the form of a collaborative action research project to design and test a more flexible and cost-effective alternative and complement to RCTs in the form of a qualitative impact protocol, or what has become known as the QuIP. Key features of the QuIP and the story of its development are described below.

But first it is important to acknowledge the existence of a wide range of established qualitative approaches to gathering evidence of causal attribution, long before the QuIP came along (see Chapter 2). So why develop another? The main argument for doing so was the perceived widespread confusion and uncertainty about the consistency and credibility of these approaches, reinforced by a tendency for them to be regarded as somehow less scientific than quantitative approaches. While qualitative researchers argued over ontology, epistemology, axiology, etc., the seemingly more pragmatic advocates of quantitative 'solutions' to the attribution problem (including use of RCTs) drew on a clearer, if narrower, mental model of understanding, based on how to overcome problems of sampling and selection bias.[6]

In contrast, qualitative researchers and evaluators seem to have been less successful in establishing clear norms and guidelines to help users discriminate between stronger and weaker studies. A 'black box' between collection of data and generation of findings fuels doubts among practitioners and policymakers about their credibility: doubts that can only partly be overcome by relying on the reputation of the researcher.[7]

Attitudes to different impact evaluation approaches also depend in part on the scale of activities being assessed (Copestake et al., 2016). Many smaller development organizations rely on a pragmatic approach. At its best this relies on a strong and clear internal understanding of what the organization is trying to do and how; this informs a good monitoring system and is supplemented by close personal observation of its activities and effects. Good monitoring data enables staff to track changes in the organization's activities and outputs, as well as correlations with selected indicators of impact. Causal attribution then relies on them being able to interpret this data by triangulating it against both the core theories of change informing their action, and direct observation of what is happening on the ground in real time.

Problems with this approach emerge as organizations grow and their relationship with other organizations (including sources of finance) becomes more complex. Senior managers find it harder to keep in touch with all the organization's activities, and scope for gaps in knowledge and understanding between senior and junior staff grows as organizations become more diversified. The basis for external funding necessarily relies less on personal rapport, and more on compliance with contractual requirements,

including formal reporting on activities, results, and impact. For large organizations and projects, more quantifiable causal claims to impact based on RCTs and other 'large N' statistical methods become both more affordable and easier to justify.

The main focus of this book falls between the extremes of internal performance assessment and large independent quantitative impact evaluation. Qualitative approaches to causal attribution, such as process tracing, can also be time-consuming and expensive. The need for cost-effective intermediate approaches to assessing impact credibly is also greater in the complex and fast-changing contexts characteristic of much international development practice. This is the niche that the QuIP aims to help to fill: as a stand-alone approach; combined with other methods; and also as a means to stimulate further methodological innovation.

How the book is organized and how to use it

The rest of this chapter provides an overview of the QuIP's main features and then a brief account of the backstory of its development and the production of this book. Those readers particularly interested in whether to use the QuIP for a specific study may choose to jump directly from reading this chapter to the book's Annex, which explains in more detail how and why to commission a QuIP study, and how to deliver one. In contrast, Chapter 2 reflects in more depth on how the QuIP relates to the wider field of impact evaluation, compared with other approaches, and critically reviews the many criteria that are used to make such comparisons.

The main body of the book (Chapters 3 to 9) provides a more fleshed-out set of examples of how the QuIP was actually used in a variety of contexts during 2016 and 2017. This illustrates how it has been adapted to different contexts and to serve different purposes for a diverse range of commissioning organizations. Each chapter sets the context, explains how the QuIP was conducted, illustrates what evidence was generated, and reflects on some of the methodological issues faced. These chapters aim both to contribute to the empirical literature on qualitative impact evaluation, and to encourage potential commissioners and implementers of QuIP studies to be creative in adapting it to suit new situations. They also drive home the point that impact evaluation goes beyond mechanical application of a standard methodological formula. Each chapter is largely self-contained, so that readers with limited time can pick out those most relevant to their own interests. For some, it will be useful and important to get a flavour of what the QuIP delivered by way of findings in different contexts, whereas for others this will be less interesting than details of how the QuIP was conducted and why. Chapter 10 provides a synthesis of issues raised and ideas explored in the case study chapters. It also relates them to the wider literature on impact evaluation within development practice, and explores the scope for further methodological research and innovation.

Box 1.1 Ways of using the book

1. Read Chapter 1 to be clear about what the QuIP is and does. Skip to the Annex if you are in a hurry to use the QuIP and need more details about how to go about doing so.
2. Review Chapter 2 quickly or carefully, depending on your prior familiarity with the other methods reviewed and your interest in how QuIP fits into the wider literature on impact evaluation.
3. Sample at least one of the case studies (Chapters 3–9) to get a flavour of the QuIP in context and the sorts of findings it can generate.
4. Read Chapter 10 for an overview of issues addressed by the book and the scope for further methodological research.
5. Dip back into other case study chapters to find out more about different contexts and methodological issues.
6. Refer to the Annex as you read, when you are seeking a more detailed description of different stages of the QuIP.

An overview of the QuIP

The QuIP can be described as doing two things at once.

- First, it comprises clear and practical guidelines for collecting, analysing, and sharing narrative statements from intended beneficiaries about significant drivers of change in their lives, including the impact of specific development actions intended to help them.
- Second, it sets out a flexible approach to generating evidence of whether a *particular action* is having the desired impact, and for whom, including exploring unintended outcomes and identifying unknown drivers of change.

These two activities overlap, but the first works forward from causes to effects, whereas the second works back from effects to causes (cf. Goertz and Mahoney, 2012: Chapter 3). As will become clear, while commissioners tend to be most interested in causes (particularly their own actions), the QuIP enhances credibility by working backwards towards these from outcomes.

Box 1.2 provides a brief description of the QuIP, while the Annex sets out more comprehensive guidelines for how to use it. The rest of this section offers something in between: a fuller description of its key features, as utilized in the studies presented in the remainder of the book. Its potential originality and usefulness, we will argue, depends less on particular characteristics than on the whole package and how it can be adapted to serve different purposes. But first it is necessary to explain what the QuIP is.

A starting point for the QuIP is the premise that those who were intended to benefit from an intervention know a great deal about what has caused and affected changes in their (and their households') lives in the recent past, and what has influenced their active decisions to start or stop doing certain activities. Relying on the narrative testimonies of intended beneficiaries removes the need for an independent counterfactual based on interviews

Box 1.2 A brief description of the QuIP

1. The QuIP is a standardized approach to generating feedback about causes of change in people's lives that relies on the testimony of a sample of the intended beneficiaries of a specified activity or project.
2. The scope of a study is jointly determined by an evaluator and a commissioner, the shared purpose being to provide a useful 'reality check' on the commissioner's prior understanding of the impact of a specified activity or set of activities.
3. A single QuIP is based on the data that two experienced field researchers can collect in around a week. A useful benchmark (that emerged through the design and testing phase) is that a 'single QuIP' comprises 24 semi-structured interviews and four focus groups. Specific studies may be based on multiples or variants of this.
4. Interviewees are selected purposively from a known population of intended beneficiaries, ideally after analysis of what available monitoring data reveals about the changes they are experiencing.
5. Where possible, initial interviews and focus groups are conducted by independent field researchers with restricted knowledge of the activity being evaluated. This means that respondents are also unaware of what intervention is being evaluated, a feature referred to as double *blindfolding* (not blinding, because the blindfolds can be removed at any time).
6. Transcripts of interviews and focus groups are written up in pre-formatted spreadsheets to facilitate coding and thematic analysis.
7. An analyst (*not* one of the field researchers) codes the data in several predetermined ways. Exploratory coding identifies different drivers and outcomes of change (positive and negative). Confirmatory coding classifies causal claims according to whether they *explicitly* link outcomes to specified activities, do so in ways that are *implicitly* consistent with the commissioners' theory of change, or are *incidental* to it.
8. Semi-automated generation of summary tables and visualizations speeds up interpretation of the evidence.
9. It is easy to check back from summary evidence to raw data for purposes of quality assurance, auditing, peer review, and deeper learning.
10. Summary reports of the evidence are a starting point for dialogue and sense-making between researchers, commissioners, and other stakeholders, thereby influencing follow-on activities.

with a control or comparison group. This is because comparisons between what happened and what would have happened otherwise are embedded in causal claims within the narrative – our thinking and language is laden with ways of doing this (e.g. through use of the conditional tense and the many ways we can answer 'why' questions). Collected with care, narrative accounts are full of explicit and latent counterfactuals; the task is to identify and interpret them. Relying on this way of addressing the attribution challenge generally does not yield quantitative estimates of the magnitude of impacts. But it can offer the evaluator a faster, cheaper, and richer route to assessing presence or absence of different causal mechanisms than alternatives that rely on statistical inference across large populations. We refer to reliance on respondents to provide evidence of the causal chain puzzle themselves as *self-reported attribution*, and we distinguish it from *statistically inferred attribution* that generally relies on exposure variation, including comparing treatment and control groups (Hughes, 2012).

Why aren't qualitative impact studies based on this approach more widely used? Addressing four more specific problems faced by evaluators has been central to the design of the QuIP. First, there is *sample selection* – who to interview and why? How does one transcend perceptions of evidence as anecdotal, and how far is it possible to generalize from the testimony of a few respondents, no matter how rich it might be? Second, there are problems of *respondent bias*. For example, trust placed in the stories of intended beneficiaries is often discounted on the grounds that they are likely to say what they think the interviewer wants to hear. How can this source of uncertainty be reduced? Third, there is a problem of *transparent analysis*. Narrative data can be unwieldy, difficult, and time-consuming to analyse – even using qualitative data analysis software. Not being sure how conclusions have been reached (or quotes and case studies selected) in a summary report leaves commissioners having to trust in a 'black box' data analysis process that they can't open. Fourth, is the linked problem of how to make *effective use* of narrative evidence. Long, rich, nuanced, and context-specific findings are harder to interpret and to turn into recommendations for action. This section points to some of the ways the QuIP seeks to address these problems.

Case selection

QuIP is a 'small n' approach that relies mostly on purposive sampling to address questions about how an activity contributes to change, for whom, and in relation to what other complementary or rival causal explanations. This entails departing from the logic of selecting a statistically representative sample to estimate how much an activity has, on average, affected a target impact variable across a known population (Goertz and Mahoney, 2012; Flyvbjerg, 2006). Deciding who to interview, how many, and how best to select them requires clarity about what information is being sought, by whom, and why. Neglecting this leads to misunderstanding about the scope for learning wider lessons from the study.

The QuIP's primary purpose is to explain variation in the wellbeing outcomes experienced by intended beneficiaries, rather than to quantify average effects on them of one particular action. Differences in case selection strategy arise according to what is known in advance about changes (e.g. in a key wellbeing indicator, Y), and whether the priority is to confirm prior expectations about the causal effect of a specific action (e.g. intervention X) or to explore what is happening in a more open-ended way. At one extreme is the situation where there is little information about X, Y or the causal links between them. A QuIP is then largely exploratory, and the best that can be done is to select cases that capture as much variation as possible in whatever is known in advance about the population (e.g. where they live). At the other extreme, case selection can draw both on monitoring data about X and Y across the population and on prior theory about the causal links between them. The QuIP can then also be

designed to confirm or challenge this theory – e.g. by purposively selecting cases where the correlation between X and Y deviates from what is expected.

Deciding on the *number* of interviews and focus groups to conduct depends less on reducing sample bias than on assessing at what point the extra insight gained into the range of causal processes influencing outcomes no longer justifies the cost of collecting more data. This is partly related to the idea of saturation, although determining the cut-off point is never as simple as this term might suggest (Braun and Clarke, 2016). For example, if data is also being used to increase confidence in a prior theory, then new cases that confirm causal links already observed in other cases still usefully add to understanding of how far the claim is transferrable to other situations. If data was being analysed at the same time as it was collected, a research team could perhaps stop collecting data as soon as it became clear that little new information was emerging. However, practical considerations prevent this: budgets and the timing of data collection have to be planned in advance, and there are other reasons for separating it from analysis (explained below). Instead, the QuIP is no different from other approaches to qualitative research and evaluation in relying on prior judgement about how much data to collect and how to select a sample that best tests and augments what the commissioner already knows or believes.

Through the design and testing phase of development of the QuIP, a benchmark was adopted of collecting discrete sets of 24 individual or household level interviews and four focus groups, often split between two locations. This number has the advantage of usually allowing two people to collect data within a full week, as well as being a large enough number to gather enough detailed information within a selected cohort of intended beneficiaries, taking account of the likely diminished marginal returns from a larger sample. It is also feasible for a single analyst to read through and immerse themselves in the volume of data thereby generated. However, multiples or variants of this benchmark for a QuIP study are possible. Indeed such variation is a likely outcome of initial discussions between researchers and commissioners about heterogeneity of the population of intended beneficiaries. Synthesis across independent but parallel QuIP studies is also possible, as illustrated in Chapter 8.

Research in 'small n' sample sizes using other methods is also relevant to these judgements. Morgan et al. (2001) plotted findings from a number of research datasets, finding that no new themes emerged after 20 interviews. Most themes emerged in the first five or six interviews, a pattern that was repeated in QuIP studies. Similarly Guest et al. (2006) showed that of 114 themes identified, 70 per cent came from the first six interviews, and 92 per cent within the first 12 interviews. Namey (2017) drew attention to these and other related empirical studies on saturation sampling in a blog post which concluded that 6–12 good individual interviews, and 3–6 focus groups were often sufficient. QuIP studies tend to split the sample of 24 respondents across two similar locations, leading to groups of 12 respondents in each community.

This fits with these findings, allowing scope for some differences between the two groups.[8]

How should one go about designing a purposive sample? There is no universal best practice method for sample selection for a QUIP study, as it depends upon many contextual factors. The most important of these are (a) the main purpose of the study, (b) availability of relevant data about variation in the characteristics of expected gainers and losers from the project, (c) availability of relevant data about variation in their exposure to project activities, and (d) time and resource constraints.

One good starting point for thinking about sampling for a QuIP study is to look at *contextual variation*. If causal processes are expected to be different for different sub-groups, and there is data to enable identification of those sub-groups prior to sample selection, then there is a case for stratified random sampling. For example, a QuIP study might include a minimum quota of people living in urban and rural areas. Stratification of the sample on these grounds is an art rather than a science, dependent on prior thinking about what contextual factors are most likely to be a source of variation in project outcomes. It also depends on the quality of change monitoring data available.[9]

Another strategy is to look at *exposure variation*. If data is available on variation in who directly received what and when, and it is expected that these differences will have different causal effects, then there is a case for stratifying the sample to ensure it reflects the full range of such exposure. This is particularly the case if one purpose of the study is to aid decisions about which of a range of project activities or components of a package to expand or to stop.

Impact assessment using the QuIP does not require a control group of people completely unaffected by the project, because it addresses attribution by identifying causal claims within each case, rather than by comparing balanced sub-samples of cases. There may nevertheless be an argument for interviewing a sample of non-beneficiaries as a source of extra information about incidental (and potentially confounding) drivers of change. For example, focus groups can be carried out in a 'control' community. Non-direct beneficiaries may also be sampled to ascertain the success of ripple effects on wider communities.

In addition to stratifying according to contextual and exposure variation, a third reason for departing from pure randomization in sample selection is to cluster respondents *geographically*. There is often a strong case for using contextual information (e.g. about agro-ecological zones) to purposefully select or at least stratify area selection. Ultimately, budget constraints may also limit the total number of interviews and focus groups that the QuIP study can cover, and geographical clustering can help to reduce the time and cost of data collection. There may also be a case for staggering studies, with case selection for repeat studies benefitting from what was learned through earlier studies, and the credibility of findings again building incrementally through the addition of each extra piece of evidence.

Reducing bias in interviews and focus groups

The QuIP relies on asking intended beneficiaries of development action what they themselves perceive to be the most important drivers of change in different aspects of their lives. There are strong practical and ethical reasons for taking this direct self-reported approach to addressing causal attribution. However, it does raise questions about possible bias. A common objection to the credibility of self-reported data is that, even if people don't generally lie intentionally for their own gain, we are all socially conditioned to tell others what we think they want to hear.[10] This is known as confirmation or pro-project bias. The QuIP includes three features aimed at mitigating the threat of such bias. First, repeating questions and comparing data from multiple interviews and focus groups (as discussed above) increases the scope for picking up variation in the way participants respond. Second, taking an exploratory or open-ended approach using a semi-structured questionnaire that works back from outcomes with minimal framing and probing, shifts attention away from specified interventions (see 'Open-ended questioning' below). Third, and more radically, where possible the QuIP limits how much knowledge both interviewers and respondents have of the project being evaluated (see 'Double blindfolding' below).

Double blindfolding. Confirmation bias is effectively reduced by creating an appropriate distance between the field researchers and the project being assessed. While fully briefed on the QuIP and trained in use of the questionnaire which has been designed, interviewers are not told who has commissioned the study, or what intervention the interviews aim to evaluate. The training ensures that the team understand why the blindfolded approach is used, to ensure that they feel confident about their role in the process. This helps avoid overly narrow agenda-setting, asking prompting or leading questions, poor listening, and explicitly or implicitly encouraging respondents to emphasize specific causal factors. It also places interviewer and respondent on a more equal footing in relation to prior knowledge.

Respondents are also therefore blindfolded in this process, since they only know as much as the researchers know. Experience has shown that researchers and respondents generally accept the case for being blindfolded in this way, particularly because this means interview questions are more open and reflect a broader interest in respondents' ideas and experiences. However, the degree of blindfolding that is possible does depend on context, including the general level of trust within the field work locality, and on finding appropriate ways of organizing and explaining the study to respondents. For example, it often helps if the researcher is affiliated with an independent institution such as a local university, and can explain that the study is investigating the respondent's experience of how different factors and actors are affecting different spheres or domains of their wellbeing.

Researchers follow standard routines for securing ongoing consent of respondents, and are never asked to withhold any information from them.

But being less than fully transparent about the purpose of the interview is ethically contentious (Copestake et al., 2016). The main defence for nevertheless doing so is that it should generate more credible information, thereby increasing the likelihood that what respondents say will be taken more seriously and hence be more useful. This is a version of the utilitarian 'greater good' argument – that the positives outweigh the negatives – used for blinding in clinical trials. And as with clinical trials it has its limits: the negatives should not include putting respondents at risk of harm or retribution in any way.

Blindfolding does not have to be permanent: indeed it is always preferable to remove blindfolds at the right stage, both on ethical grounds and in order to capture respondents' and researchers' considered views in light of fuller information. To this end, the QuIP includes the option of including 'unblindfolded' feedback workshops once the data has been collected and analysed. These can be extended beyond field researchers to include respondents in the villages where data has been collected, offering an opportunity to share and reflect upon the findings, and to probe further into issues that were unclear or unexpected.

There are of course trade-offs with this approach to blindfolding, not least that without a second round of unblindfolded discussion it limits the scope for interviewers to go into depth about details of particular activities. However, blindfolding is an important riposte to criticisms of self-reported attribution. Mitigating against confirmation and related biases through blindfolding does come at a price, but the decision over precisely how much detail will be hidden and how much revealed will always depend on the context of the project. It is also possible to use a partially blindfolded approach – e.g. by revealing the name of the commissioner, but not details of the project being evaluated.

Open-ended questioning. Taking an open-ended approach to questioning helps to reduce reliance on respondents' ability to recall specific details by allowing them the freedom to recount the stories of change that they perceive to be most relevant and important. Recall periods are generally tied to the implementation period of the project being assessed, but respondents are free to identify the most significant drivers of change *within* that period. This risks bias towards more recent events, but avoids the recall problems associated with seeking precise answers to questions about income, asset ownership and so on during a precise baseline period. The more open approach also allows respondents to draw on those life experiences that they perceive to be most relevant to the questions.

The QuIP employs two data collection instruments: semi-structured household level interviews and facilitated focus group interviews.[11] The questionnaire for both is framed around a series of *outcome domains* based (when available) on the theory of change (ToC) underpinning a project. These domains reflect the broad areas in respondents' lives where some change is expected, including those where unintended negative impacts

may have occurred. Many possible outcomes and causal connections can be covered in one broad domain, so questions do not need to be specific to each aspect of the implementation. Balance entails developing a set of questions that are vague enough not to prompt respondents too explicitly (so as to sustain blindfolding) while being specific enough that it would be surprising if they didn't mention activities they participated in (and that commissioners believe to have been important). Hence co-design of the questionnaire with the commissioner is critically important, as they will later need to accept that it constituted a fair test of the activity being evaluated.

The questionnaires are made up of a series of generative, supplementary, and closed questions. Open-ended generative questions elicit information about changes respondents have experienced within a specified period of time, and are designed to stimulate discussion in an open way. This allows the respondent to talk about change in a given domain in their own terms and to reflect on a range of experiences. Supplementary questions are used to sustain and deepen conversations about changes observed by the respondent and the reasons behind them. The most important of these is 'why did that happen?' given that the goal is to elicit stories of change with implicit causal attribution. Hence supplementary questions should help to expose chains and clusters of causal statements leading back to root sources, which may include but are not restricted to the activities being evaluated. The skill and sensitivity required to conduct and document this process is critical to the whole approach, laying a premium on the ability of the interviewer to develop a respectful rapport with the respondent, building their trust and listening sensitively yet proactively to encourage and unravel causal pathways.

At the end of a set of open questions pertaining to a given domain, the QuIP interviewer asks one or more closed questions to ascertain the overall direction of change in that domain as perceived by the respondent. Given that the open question may have elicited both positive and negative stories, it is important that the respondent is given the opportunity to provide their own summary of whether, on balance, changes are perceived as positive or negative – rather than leaving it to the analyst to make this judgement. These closed questions also provide a useful snapshot of respondents' overall experience of change when it comes to presentation of the data.

Data collection under the QuIP can be likened to an archaeological dig. The first step is to locate very broad sites where artefacts are likely to be found, and cordon off these areas to be explored. With the QuIP this is done through the selection and definition of impact domains. The next step is to start scraping away at the surface in these areas, by asking open questions relating to the given domain. Once a possible find is encountered, the scraping becomes a careful brushing away of the finer earth to expose it, through questions probing for the causes and sources of change. The probability of finding multiple and varied artefacts is vastly expanded by keeping an open mind, i.e. by doing interviews 'blindfolded'. Finds in an interview take the

form of causal claims embedded in the narrative, and although interviews are generally recorded (see Annex), subsequent analysis relies on typed notes of what was said rather than full transcripts. Hence a critical challenge for the field research team is to produce these notes as accurately as they can, relying on a mixture of recall, their handwritten interview notes, and digital recordings. This often also entails translating from the local language in which the conversations took place into another language, adding to the need for close quality assurance.[12]

Analysing and presenting data

The separation of data collection and analysis responsibilities unbundles activities that use two very different skill sets, and makes the process of moving from data to findings more transparent. The analyst's main task is thematic analysis: identifying, analysing, and reporting on patterns within the reams of narrative data (Braun and Clarke, 2006: 6). This entails coding causal claims in the data (a) in an inductive and exploratory way that reflects as accurately and fully as possible what respondents said, *and* (b) in a more deductive and confirmatory way, guided by the theory of change of the activity being implemented.[13] To help analysts manage the tension between these two approaches, the QuIP analysis uses a triple coding approach. This divides each causal pathway as follows:

- *Drivers of change* (causes). What led to change, positive or negative?
- *Outcomes* (effects). What change/s occurred, positive or negative?
- *Attribution*. What is the strength of association between the causal claim and the activity or project being evaluated?

The first two sets of codes are established more inductively, starting with a blank code book and iteratively building up categories according to the stories of change cited by respondents (rather than using predetermined categories). Since outcomes often become drivers themselves, leading to other outcomes, these can be coded as primary, secondary or tertiary outcomes – helping to build up causal chains. This is illustrated by Figure 1.1, which depicts how three linked causal claims are captured for analysis using four codes. In later iterations, additional codes and clusters of codes may also be added more

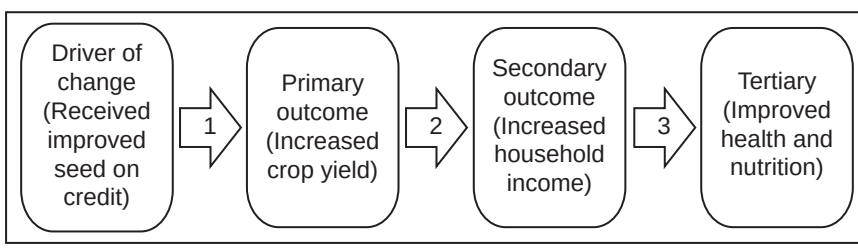


Figure 1.1 Thematic coding of causal claims: an illustration

Table 1.1 Attribution codes for causal claims

Level of attribution	*Positive code*	*Negative code*	*Explanation*
Explicit	1	2	Positive or negative change explicitly attributed to the project or to explicitly named project activities.
Implicit	3	4	Change confirming (positive) or refuting (negative) the specific mechanism (or theory of change) by which the project aims to achieve impact, but with no explicit reference to the project or project activities.
Incidental	5	6	Change attributed to other forces (not related to activities included in the project's theory of change).
Not attributed	7	8	Change not attributed to any specific cause.
Neutral	9		Change that is ambiguous, ambivalent or neutral in its effects: i.e. cannot readily be coded positive or negative.

deductively to reflect the project theory of change. Network diagrams can also be constructed that link multiple drivers and outcomes together in more complex ways.

Attribution coding is carried out purely deductively to explore links in a project's theory of change. Drivers are classified according to whether they *explicitly* refer to project activities, *implicitly* corroborate its theory of change, or are *incidental* to it. For this part of the coding it is obviously necessary for the analyst to be fully unblindfolded and familiar with the theory of change of the project, whether explicitly set out and supplied by the commissioner or implicit in documents supplied by them. Table 1.1 elaborates on the basic set of attribution codes.

Once all change data is coded it is then possible to use frequency counts to tabulate and visualize the data in many ways, as the chapters to follow illustrate. Tables can highlight not only what drivers of change were reported, but also where expected drivers were *not* reported. Thus analysis can reveal how closely respondents' experiences match the project's presumed theory of change, and how different positive and negative drivers interacted. Using qualitative data in this way does inevitably hide much that is meaningful in the coded text, and for this reason reports also review and illustrate important points with quotations. An annex to each report containing all the coded data also enables readers and reviewers to go back to source text, opening up the data to audit and ensuring that respondents' voices are not lost through the quantification of qualitative data.

Analysis along these lines can address many questions, including the following:

- Is the programme having the expected effect on intended beneficiaries?
- What other factors have affected expected outcomes?

- How do these factors relate to each other?
- Has the programme had any unanticipated effects, positive or negative?
- What drivers of change or patterns can be identified that could inform future programme design?
- Are there any 'missing' drivers: interventions not cited or not considered significant by respondents?
- How do the reported causal claims and chains compare with the organization's theory of change, process data on how the project was implemented, and knowledge of impact from other sources?

Even where positive explicit attribution is made to the intervention, the causal chains leading from intervention to that change may not follow the expected pathway. Negative causal chains may indeed help the organization to understand unintended consequences of the organization's interventions, or the mitigating effect of external factors.[14]

One advantage of relying so much on inductive coding is that analysts do not need to be experts in the sector that the project works in; indeed there are advantages to them not having pre-conceived ideas about how a particular theory of change should work. Rather, the job of coding, analysing, and presenting the data is improved by relying solely on the stories being told by the respondents, and representing these stories as accurately as possible. The most important qualities of an analyst relate to their skills in qualitative coding and thematic analysis, not to their prior knowledge of the sector being assessed. However, the lead evaluator putting together a final report or presentation should have relevant expertise, particularly if the QuIP data is being combined with material from other sources. For this reason it is not uncommon to separate out the two tasks.

As set out above, the task of the analyst can appear somewhat mechanical; one objective being to increase the consistency, potential replicability, and reliability of the task. However, in practice the task is more complex than this would suggest. At least four steps can be picked out to highlight why it can be viewed as an art as much as a science:

- First, there is the task of deciding how to group together and distinguish between different causes and outcomes. This includes deciding when to code causal drivers separately or to treat them as an integral package.
- Second, distinguishing between explicit, implicit, and incidental drivers is often difficult because it hinges on just how specific narrative text needs to be about who is driving the identified change. Such coding judgements are best made not in isolation but by viewing them in the context of the whole transcript of an interview.
- Third, difficult choices need to be made between the infinite range of tables and visual outputs that can be derived from any one database. For example, there is much scope for exploring the nature and frequency of observed causal processes for sub-samples, including by gender and age.

- Fourth, and most difficult of all, are judgements about what narrative text to spell out in full (e.g. balancing what is unusual and what appears typical), taking into account the fact that key users of the data will have limited time and attention spans, so cannot be expected to read through long lists of verbatim quotations.

It is these judgements that open up scope for the way thematic coding, analysis, and reporting is affected by the positionality of the analyst.[15] This is illustrated by Figure 1.2.[16]

Placing the analyst at the centre defines their role in converting a range of different sources of data into study findings, with their relationship to other people being mediated by them. The core functional model entails receiving transcripts from the field researchers (supplemented by field reports and photos) and analysing them in relation to project theory and context (obtained mainly from project staff). The role of the lead evaluator typically entails mediating the supply of data and requests from the commissioner and field research team, while ensuring a study meets a minimum set of standards to be considered a QuIP. Neglected in the guidelines – but central to our discussion – is the way the formal analytical task (of turning transcripts and project theory into findings using agreed QuIP guidelines) is mediated in practice by the wealth of other data and relationships that the analyst brings to bear on their work.

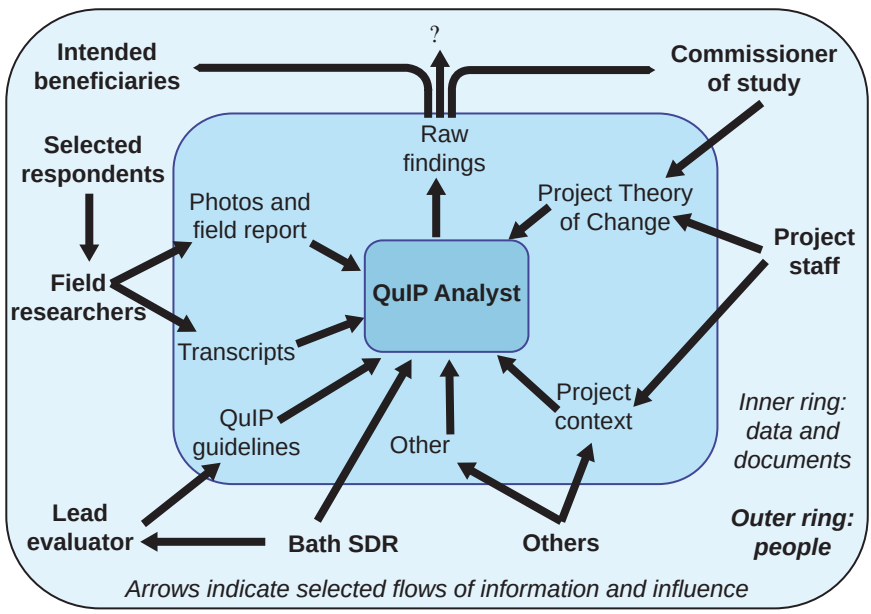


Figure 1.2 The relationship of the QuIP analyst to people and data

From evidence to action

A QuIP study provides its commissioner with an independent check on the diverse drivers of change over (a) an agreed time period, (b) in selected outcome domains, and (c) for a specified group of respondents, including (d) the contribution of a specified project or set of activities. However, this begs the question of what the commissioner should then do with this evidence, whether positive or negative. Most studies only point towards specific actions after further consultation, and/or triangulation against data from additional sources, and it is up to the commissioner to decide how and how far to involve the QuIP team in these discussions. In some cases this involvement may be minimal, and the main benefit to the commissioner may arise from being able to feed evidence of impact upwards to social investors to help sustain flows of funding, and/or downwards to staff and intended beneficiaries to inform debate over how to change implementation processes. In these cases the QuIP study is often only one component of a larger evaluation that directly addresses questions not only about impact but also about relevance, efficiency, cost-effectiveness, and sufficiency of a specified intervention. In practice, even narrowly scripted QuIP studies can throw up useful evidence on the appropriateness of goals, weaknesses in design and implementation processes, and the adequacy of change monitoring systems.

The backstory of the QuIP and this book

The QuIP

An initial experiment with a prototype 'QuIP' was conducted in Peru in 2003 as part of action research into social performance assessment of microfinance organizations funded by the Ford Foundation (Wright and Copestake, 2004; Copestake et al., 2005).[17] The idea was revived in 2011 during discussions with two international NGOs (Self Help Africa and Farm Africa) about how best they could respond to internal and external demand for better evidence of the impact of their projects. They had already adopted a quantitative approach to monitoring changes in household level food security, called the Individual Household Method (IHM), with support from the NGO Evidence for Development. However they were unsure how best to make a more credible case to support claims about their own contribution to changes in farm level food and economic security in areas undergoing rapid and complex livelihood transformations.

These discussions culminated in a collaborative action research proposal called the 'ART Project' (Assessing Rural Transformations) with design and testing of a qualitative impact protocol as its primary goal. This was sponsored by the UK Department for International Development (DFID) and the Economic and Social Research Council (ESRC) under their joint research programme for poverty alleviation. The project entailed collaboration between staff at four

Table 1.2 The ART project: overview of the four pilot projects

Masumbankhunda, Central Malawi (Self Help Africa) Improved groundnut production	*Karonga, Northern Malawi (SHA) Climate adaptation and resilience*	*Assela Southern Ethiopia (SHA) Improved barley production and sales*	*Ahferom, Northern Ethiopia (Farm Africa) Livelihood diversification*
The project promoted access to improved seed and technical advice, but with wide variation between villages and according to gender. Most households experienced modest increases in income, especially poorer households. But they remained susceptible to shocks, particularly in the maize-fertilizer price ratio.	Livelihoods were already diversified but remained highly susceptible to floods and drought. Income did not improve for most households, mostly due to adverse weather. Project effects were delayed and limited, but supportive of the trend out of subsistence staple crops into diverse market-oriented activities.	Selected farmers were generally food secure, but barley yields and prices were initially low. Most households achieved substantial improvements in income, much of it attributable to improved seed and technical advice. Other NGOs entered the area doing similar work.	The project was strongly targeted towards women and youth. Their incomes generally rose, but with high variance. Impact varied sharply between project packages. Respondents remained vulnerable to weather-related shocks.

Source: ART Project data.

universities (of Bath, Ambo, Mekele, and Malawi) and three NGOs (Self Help Africa, Farm Africa, and Evidence for Development). A first draft of the QuIP was agreed at a methodology workshop in Shrewsbury in May 2013, attended by staff from all these organizations plus Irish Aid and Oxfam GB. Two projects were then selected in Ethiopia and two in Malawi for two rounds of pilot testing of the QuIP – one year and two years after carrying out a baseline IHM study, which was also repeated two years later (see Table 1.2). Findings were written up and reviewed at feedback and dissemination workshops in Addis Ababa and Lilongwe in July 2015. Findings from the first round of QuIP pilot studies are summarized in Copestake and Remnant (2015).[18]

The main subject matter of this book is experience with using the QuIP in the two years after the ART project ended, hence after the protocol had passed through this initial period of testing and refinement, and when it was being utilized under more realistic conditions than under an 80 per cent grant-funded action research project. The mechanism for promoting the QuIP after the ART Project ended – and for bringing together experiences thereby obtained – was to set up a social enterprise dedicated to this goal. Bath Social and Development Research Ltd (BSDR) started work in 2016 as an initiative of the Centre for Development Studies (CDS) at the University Bath (see Box 1.3).

Assignments completed during 2016/17 are listed in Table 1.3. This reveals the diversity of commissioners and activities to which BSDR responded

Table 1.3 QuIP studies conducted by BSDR in 2016 and 2017

Commissioner	*Activity evaluated*	*Country*	*Date*	*Type of study*	*Book chapter*
Self Help Africa (INGO)	Community-based cassava production	Kenya	Feb 16 – Apr 16	Single QuIP	–
Oxfam GB (INGO)	Project to empower women through coffee value chain upgrading	Ethiopia	Apr 16 – Jul 16	Double QuIP	–
Diageo Ltd (global company)	Project to strengthen smallholder inclusion in barley production	Ethiopia	Jul 16 – Sep 16	Double QuIP	3
Habitat for Humanity (INGO)	Refinance for housing loans via microfinance institutions	India	Sep 16 – Apr 17	Double QuIP	4
Acumen (impact investor)	Investment in a commercial dairy company	India	Oct 16 – May 17	'Lean' QuIP	–
C&A Foundation (INGO)	Programme to improve health and wellbeing of garment factory workers	Mexico	Oct 16 – Mar 17	Double QuIP	5
Tearfund (INGO)	Church and community mobilization programme	Uganda	Nov 16 – Jan 17	Double QuIP	6
Tree Aid (INGO)	Project to promote non-timber forest products	Ghana	Jan – Feb 17	Single QuIP	–
Save the Children (INGO)	Project to improve agriculture and baby/child nutrition	Tanzania	Mar – May 17	Double QuiP	7
Seed Global Health (INGO)	Placement of Peace Corps volunteers in medical and nursing colleges	Malawi, Tanzania, Uganda	Apr – Jun 17	Three QuIP studies	8
Self Help Africa (INGO)	Integrated area development project	Zambia	Mar – Jun 17	Double QuIP	–
Acumen (impact investor)	Investment in a beauty parlour franchise company	India	May – Nov 17	'Lean' QuIP	–
Voscur/ Bristol City Council (local government)	Technical support for community organizations	UK	Mar – Apr 17	Pilot study	9
Frome Town Council (local government)	Impact of council initiatives to promote use of green space	UK	Jul – Sep 17	Modified QuIP	9
Save the Children (INGO)	Famine early response programme	Ethiopia	Aug – Dec 17	Single QuiP	–
Diageo Ltd (global company)	Support for smallholder cassava and sorghum growers	Uganda	May – Aug 17	Double QuIP	–
Self Help Africa (INGO)	Cereal value chain and nutrition project for smallholder farmers	Burkina Faso	Oct – Dec 17	Single QuIP	–

Source: BSDR Ltd

Box 1.3 Bath Social and Development Research Ltd

BSDR Ltd was set up in 2016 to apply practical ideas arising from the Centre of Development Studies at the University of Bath through training, advisory, and consultancy services in support of policies and practices promoting sustainable local, national, and global development, wellbeing, and social justice. It is a non-profit company with a non-distribution clause, which requires that all revenue earned is reinvested in its activities. To date it has focused on developing and promoting uptake of the QuIP. The University retains ownership of the registered QuIP trademark, and has agreed a non-exclusive licence to BSDR to utilize the QuIP name and to sub-license its use to accredited practitioners. Research underpinning the establishment of BSDR was aided by a follow-up grant under the DFID-ESRC poverty alleviation programme, and from Innovate UK through the ICURe (Innovation to Commercialization of University Research) programme. These financed market research interviews with more than 100 potential users of the QuIP, as well as a two-day workshop to explore its potential for use by impact investors (Niño Zarazúa and Copestake, 2016). No grant money was used for the start-up of BSDR, which sustains itself through commissioned projects and training courses.

during its first two years of operation. Discussions with potential commissioners were based on the benchmark for a 'single' QuIP of 24 interviews and four focus groups conducted by two people in one week of field work. Many studies were based on double or treble QuIPs and/or formed part of a study with wider scope, as explained in the relevant case study chapters. While costs for QuIP studies can vary widely, they are substantially below the norm for many other kinds of evaluation study. Whether impact evaluation studies represent good value for money obviously depends on benefits as well as costs.

The book

The main motivation for this book was to share experiences of using the QuIP in a range of different contexts, and under contractual conditions that were closer to the market for consultancy-led impact evaluation than grant-funded academic research. For this reason the book does not elaborate on the research carried out using the QuIP in Peru in 2003–4, nor under the ART Project in 2012–15. The case studies were selected to maximize diversity of commissioners, fields of activity, and geographical spread within the constraints on time available for writing them up. Box 1.4 briefly describes QuIP studies with three additional commissioning organizations. Time prevented us from writing these up here as case study chapters, but they are referred to in the synthesis discussion in Chapter 10.

Source material for each case study varied and is described in each chapter. They all drew on final QuIP reports, context-setting project documents, and other written material. This was supplemented by key informant interviews with study commissioners and lead evaluators (conducted by Morsink). Most also drew directly on participant observation and on the reflective practice of authors who were also directly involved in the QuIP study. The majority

Box 1.4 Additional QuIP studies (not covered by a case study chapter)

The *Oxfam GB* study was a follow-up to an ex post difference-in-difference evaluation of a fairtrade coffee value chain project. This echoed the Diageo project in that in both cases the commissioner was already confident that increased cash crop sales had boosted the income of farm households, but sought reassurance that this had not been associated with adverse gender, generational or inter-household distributional effects. For a fuller discussion see Mager et al. (2017).

The *Self Help Africa* (SHA) studies in Zambia and Burkina Faso resembled the Save the Children study in assessing an area-based project combining climate-smart agriculture interventions with nutrition education. In contrast, the SHA study in Kenya was similar to the Diageo and Oxfam GB studies in its focus on the social impact of promoting commercial production of a particular commodity (cassava). This was also the case with the study for *Tree Aid* in Ghana, which focused on promoting commercialization of shea beans.

As an impact investor, *Acumen* channels private investment into selected businesses geared to generating a commercial return subject to delivering positive social impact (Dichter et al., 2016). BSDR collaborated with Acumen in developing two 'lean' QuIPs in India: the first studying small farmers supplying a dairy firm, and the second women who joined a beauty parlour business franchise. Within the typology of development finance set out by Reisman and Olazabal (2016) Acumen is classified as an indirect impact investor, whereas Diageo can be viewed as aspiring to be a more socially responsible investor.

of those interviewed agreed to review and comment on drafts, and where this is the case the chapters are also published 'with' these collaborators alongside the names of the main authors. The case study chapters went through varying degrees of vetting and approval by staff within the organizations who commissioned the QuIP studies on which they are based. In a couple of cases this resulted in the removal of discussion about operational decisions made partly as a result of the QuIP studies.[19] However, we have endeavoured to ensure that the voice of the commissioner as an actor in each study remains clear and distinct from that of the other authors. On balance, co-production of these chapters with commissioners resulted in improvements to both quality and accuracy of information and argument – although not without cost in terms of time spent editing and re-editing drafts. The commitment and contribution of commissioners, co-writers, co-researchers, and numerous other stakeholders is acknowledged elsewhere.

Notes

1. On defining development see Clark (2002) and Copestake (2015).
2. For a general discussion of cognitive biases see Kahneman (2011) and Thaler (2015). White and Phillips (2012) also review a range of those that particularly affect impact assessment studies.
3. This also explains how the scope of the book differs from broader introductions to qualitative research methods, such as Skovdal and Cornish (2015), which serve a broader purpose.

4. Economists refer to this as 'warm glow', a term attributed to Andreoni (1989) that has spawned a substantial literature on impure or emotional altruism (e.g. Singer, 2009).
5. In contrast, when indicators of change move backwards, then better causal attribution may enable investors to claim they are nevertheless offsetting or mitigating the adverse trend. But a more opportunistic strategy in this situation is 'mission drift' towards less ambitious goals (Copestake, 2007). Microcredit is a salutary example: when audacious claims that it could simultaneously generate profits and eliminate poverty became harder to sustain, so mainstream microfinance institutions and their backers shifted the goalposts in favour of the more modest goal of promoting financial inclusion (Copestake et al., 2016).
6. Sample bias generally referred narrowly to how to generate statistically significant estimates of the average value of the impact of a measurable 'treatment' X on a measurable outcome variable Y across a known population. Selection bias was concerned with not falsely attributing impact to X that originated in factors determining who gained access to X. For example, graduates of University A might get better jobs than those from University B, not because it taught them better, but because it attracted brighter and/or better connected students in the first place.
7. This can also be described as a 'lemon problem' (Akerlof, 1970). In emphasizing this difference between RCTs and qualitative approaches we are not denying the importance of other positive attributes of RCTs, particularly that they can supply estimates of the *magnitude* of attributable impact.
8. Dion (1998) points to an alternative approach to sample size selection for confirmatory studies using Bayes theorem. If the prior probability of a causal link (from X to Y) and of an alternative hypothesis (from Z to Y) is also 50 per cent then he suggests that only five cases need to be examined to confirm or refute the hypothesis with 95 per cent confidence.
9. For example, stratification might incorporate data on either baseline estimates of income, or estimates of changes in income from baseline to endline, or both. Hence a simple design might quota sample four groups: richer and improving; richer but declining; poorer but improving; poorer and getting worse.
10. See Ozler (2013), for example. Much of the discussion of bias relating to quantitative research methods is implicitly concerned more with estimates of the magnitude of variables, than with the choice and interpretation of causal statements. Nevertheless, it is interesting that 'experimenter demand or social desirability effects' may be less than widely assumed (McKenzie, 2018).
11. Household level interviews seek narrative evidence of changes that have affected members of a specific household, whereas focus groups (mostly more narrowly gender and age-specific) ask about changes experienced by 'people like you', and also aim to reveal social norms. See the Annex for a fuller discussion.
12. All studies reported in this book were analysed and written up in English.
13. Structuring thematic analysis goes against the inductive spirit of qualitative research, but the task is unavoidably influenced to some

extent by the prior thinking and positionality of the analysis. Nowell et al. (2017: 2) also note how '… flexibility can lead to inconsistency and a lack of coherence when developing themes derived from the research data. Consistency and cohesion can be promoted by applying and making explicit an epistemological position that can coherently underpin the study's empirical claims'.

14. Every organization will be pleased to hear that it has had positive impact, and discomfited if it is perceived by intended beneficiaries to have had negative impact. However, a perceived *absence* of any impact can be the most uncomfortable feedback for an organization to hear, and the most difficult to take on board. In this case there is a risk of assuming that the absence of any evidence is not due to the lack of impact, rather due to not having looked hard enough or long enough, or in the right way or place for it. However, if a QuIP has been well designed, one can trust that respondents who experienced a significant change in their lives due to an intervention will mention that intervention.
15. Positionality, thus defined, is not the same as subjectivity. For example, individual characteristics, such as ability to concentrate and attention to detail, also affect the analyst's performance. The term positionality highlights how analysis is affected by the functional relationship we have with others involved in the wider process of generating evidence, including difference in role, socio-economic status, culture, and self-identity.
16. This framing is partly inspired by the institutional ethnography of Dorothy Smith (2005). This takes the specific experience of one subject (e.g. a health care patient receiving an operation) as its starting point and then traces the network of power relationships that define it, with power being mediated both through individual relationships and through the authority vested in documents (such as clinical protocols and medical records).
17. This was also used for a microfinance impact evaluation by Athmer and de Vletter (2006).
18. Two of the second round pilot QuIP reports can be found at the ART Project website http://www.bath.ac.uk/cds/projects-activities/assessing-rural-transformations/index.html. IHM reports for each project are available at http://www.efd.org/expertise/studies-and-reports/. Initial ideas developed at the design workshop are explored in Copestake and Remnant (2015), and the division of responsibilities between participants in the research, including the ethics of blindfolding field teams is explored in Copestake et al. (2018).
19. Evaluating the impact of QuIP studies and/or other sources of evidence on decisions to close or to scale up a project is itself a complex causal attribution challenge, and one that case studies touch on indirectly rather than directly.

References

Akerlof, G. (1970) 'The market for "lemons": quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* 84(3): 488–500.

Andreoni, J. (1989) 'Giving with impure altruism: applications to charity and Ricardian equivalence', *Journal of Political Economy* 97: 1447–89.

Athmer, G. and de Vletter, F. (2006) *Microfinance market in Maputo, Mozambique*. A case study of Novobanco, Socremo and Tchuma. Poverty impact assessment study commissioned by the Netherlands Platform for Microfinance.

Attah, R. (2017) *Significant Others: The Influence of Support Relationships and the Livelihood Against Poverty Cash Transfer Programme on the Wellbeing of Vulnerable Urban People in Ghana*, doctoral thesis, Bath: University of Bath.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology* 3: 77–101 <http://dx.doi.org/10.1191/1478088706qp063oa>.

Braun, V. and Clarke, V. (2016) '(Mis)conceptualising themes, thematic analysis, and other problems with Fugard and Potts' sample-size tool for thematic analysis', *International Journal of Social Research Methodology* 19(6): 739–43 <http://dx.doi.org/10.1080/13645579.2016.1195588>.

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

Clark, D. (2002) *Visions of Development: A study of Human Values*, Cheltenham: Edward Elgar.

Copestake, J. (2007) 'Social performance management in microfinance: from mission drift to mainstream', *World Development* 35(10): 1721–38.

Copestake, J. (2014) 'Credible impact evaluation in complex contexts: confirmatory and exploratory approaches', *Evaluation* 20(4): 412–27 <https://doi.org/10.1177%2F1356389014550559>.

Copestake, J. (2015) 'Whither development studies? Reflections on its relationship with social policy', *Journal of International and Comparative Social Policy* 31(2): 100–113 <http://dx.doi.org/10.1080/21699763.2015.1047396>.

Copestake, J. and Remnant, F. (2015) 'Assessing rural transformations: piloting a qualitative impact protocol in Malawi and Ethiopia', in L. Camfield and K. Roelen (eds), *Mixed Methods in Poverty Research*, London: Routledge.

Copestake, J., Dawson, P., Fanning, J.P., McKay, A. and Wright-Revolledo, K. (2005) 'Monitoring the diversity of the poverty outreach and impact of microfinance: a comparison of methods using data from Peru', *Development Policy Review* 23(6): 703–23 <http://dx.doi.org/10.1111/j.1467-7679.2005.00309.x>.

Copestake, J., O'Riordan, A-M. and Telford, M. (2016) 'Justifying development financing of small NGOs: impact evidence, political expedience & the case of the UK Civil Society Challenge Fund', *Journal of Development Effectiveness* 8(2): 157–70 <https://doi.org/10.1080/19439342.2016.1150317>.

Copestake, J., Remnant, F., Allan, C., van Bekkum, W., Belay, M., Goshu, T., Mvula, P., Thomas, E. and Zerahun, Z. (2018) 'Managing relationships in qualitative impact evaluation of international development practice: QuIP choreography as a case study', *Evaluation* 24(2): 169–84 <https://doi.org/10.1177/1356389018763243>.

Dichter, S., Adams, T., and Ebrahim, A. (2016) 'The power of lean data', *Stanford Social Innovation Review* Winter: 36–41.

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

Ferguson, J. (2015) *Give a Man a Fish: Reflections on the New Politics of Distribution*, Durham, NC: Duke University Press.

Flyvbjerg, B. (2006) 'Five misunderstandings about case-study research', *Qualitative Inquiry* 12(2): 219–45 <https://doi.org/10.1177%2F1077800405284363>.

Goertz, G. and Mahoney, J. (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*, Princeton, NJ: Princeton University Press.

Guest, G., Bunce, A. and Johnson, L. (2006) 'How many interviews are enough? An experiment with data saturation and variability', *Field Methods* 18(1): 59–82 <https://doi.org/10.1177%2F1525822X05279903>.

Hughes, K. (2012) *Where There is No External Control: Exploring the Use of Treatment Exposure Variation and Programme Theory to Manage for Outcome Success*, doctoral thesis, London: London School of Hygiene and Tropical Medicine.

Kahneman, D. (2011) *Thinking, Fast and Slow*, London: Penguin.

Mager, F., Remnant, F. and Walsh, M. (2017) *Cash Cropping and Care: How Cash Crop Development is Changing Gender Relations and Unpaid Care Work in Oromia, Ethiopia*, Policy and Practice note, Oxford: Oxfam GB.

Martens, B., with Mummert, U., Murrell, P. and Seabright, P. (2002) *The Institutional Economics of Foreign Aid*, Cambridge, UK: Cambridge University Press.

McKenzie, D. (2018) 'More on experimenter demand effects' [blog], *Impact Evaluation Blog*, World Bank, 25 July <https://blogs.worldbank.org/impactevaluations/more-experimenter-demand-effects> [accessed 14 October 2018].

Molecke, G. and Pinkse, J. (2017) 'Accountability for social impact: a bricolage perspective on impact measurement in social enterprises', *Journal of Business Venturing* 32: 550–68.

Morgan, M., Fischhoff, B., Bostrom, A. and Afman, C.J. (2001) *Risk Communication: A Mental Models Approach*, Cambridge, UK: Cambridge University Press.

Namey, E. (2017) 'Riddle me this: How many interviews (or focus groups) are enough?' [blog], Research for Evidence <https://researchforevidence.fhi360.org/riddle-me-this-how-many-interviews-or-focus-groups-are-enough> [accessed 3 October 2018].

Niño Zarazua, M. and Copestake, J. (2016) *Social Impact Investment and the Attribution Challenge* [pdf], CDS Briefing Paper, Bath, UK: Centre for Development Studies, University of Bath <http://bathsdr.org/wp-content/uploads/2016/05/QUIP-SII-Briefing-Paper-May-2016.pdf> [accessed 3 October 2018].

Nowell, L., Norris, J., White, D., and Moules, N. (2017) 'Thematic analysis: striving to meet the trustworthiness criteria', *International Journal of Qualitative Methods* 16: 1–13 <https://doi.org/10.1177%2F1609406917733847>.

Ozler, B. (2013) 'Economists have experiments figured out. What's next?' [blog], *Impact Evaluation Blog*, World Bank, 14 January <https://blogs.worldbank.org/impactevaluations/economists-have-experiments-figured-out-what-s-next-hint-it-s-measurement> [accessed 14 October 2018].

Reisman, J. and Olazabal, V. (2016) *Situating the Next Generation of Impact Measurement and Evaluation for Impact Investing*, New York: Rockefeller Foundation.

Singer, P. (2009) *The Life You Can Save: Acting Now to End World Poverty*, New York: Penguin Random House.

Skovdal, M. and Cornish, F. (2015) *Qualitative Research for Development: A Guide for Practitioners*, Rugby, UK. Practical Action Publishing.
Smith, D. (2005) *Institutional Ethnography: A Sociology for People*, Lanham, MD: AltaMira Press.
Stern, E., Stame, N., Mayne, N., Forss, K., Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, Working Paper No. 38, London: Department for International Development.
Thaler, R. (2015) *Misbehaving: The Making of Behavioural Economics*, New York: W.W. Norton & Co.
White, H. (2010) 'A contribution to current debates in impact evaluation', *Evaluation* 16(2): 1–11 <https://doi.org/10.1177%2F1356389010361562>.
White, H. and Phillips, D. (2012) *Addressing Attribution of Cause and Effect in 'Small n' Impact Evaluations: Towards an Integrated Framework*, New Delhi: International Initiative for Impact Evaluation.
Wright, K. and Copestake, J. (2004) 'Impact assessment of microfinance using qualitative data: communicating between social scientists and practitioners using the QUIP', *Journal of International Development* 15: 355–67 <https://doi.org/10.1002/jid.1082>.

About the authors

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Fiona Remnant, MSc International Policy Analysis, is Managing Director of Bath Social and Development Research (BSDR), and has worked in development for over a decade, specializing in the application and communication of academic research to practitioners and policymakers. She has worked for the Centre for Poverty Analysis in Sri Lanka, Oxfam in the UK, and the Centre for Development Studies at the University of Bath. She collaborated with James Copestake on the Assessing Rural Transformations action research project at the University of Bath between 2012 and 2016 which culminated in the development of the QuIP and the creation of BSDR.

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of QuIP across a range of contexts and countries.

CHAPTER 2

Comparing the QuIP with other approaches to development impact evaluation

James Copestake

This chapter positions the Qualitative Impact Protocol (QuIP) as an approach to producing intermediate feedback that provides an independent reality check on internal performance assessment, but is nevertheless focused on the commissioner's specific evidence needs. It then compares and contrasts the QuIP with 30 other approaches to impact evaluation. These are classified inductively into four groups by using the QuIP as a benchmark: similar but narrower in scope, similar but more general, more heavily quantitative, and more thoroughly participatory. The chapter warns against over-generalizations about the relative merits of different impact evaluation approaches without reference to context, timeliness, cost, prior understanding of impact, and what the commissioner deems to be 'good enough' evidence.

Keywords: impact evaluation, causality, attribution, qualitative research methods, international development, performance management

Introduction

Chapter 1 introduced the QuIP and explained its origins, and the Annex reproduces the QuIP guidelines in full. The main purpose of this chapter is to compare and contrast the QuIP with other approaches to development impact evaluation. This chapter also locates the QuIP more precisely in relation to different kinds of feedback about impact, and explores some of the criteria that inform selection of different approaches.

There are a huge number of approaches, methods, and tools that could be compared, and many criteria can be employed in doing so. For example, BOND (the UK's leading umbrella body for development NGOs) distinguishes between six approaches to impact evaluation: experimental, statistical, theory-based, case-based, participatory, and synthesis (Stern, 2015). It also provides a spreadsheet tool for choosing between 11 specific methods, using a checklist of 39 questions that need to be answered and requirements that must be satisfied for the method to be applicable.[1] In contrast, this chapter starts with the QuIP, and limits itself to exploring how it fits within this pantheon of impact evaluation approaches and methods.

http://dx.doi.org/10.3362/9781780447469.002

The chapter first defines impact evaluation as an intermediate feedback mechanism falling somewhere between routine performance management and independent research. It then classifies the QuIP inductively according to how it compares to a more comprehensive list of other impact evaluation approaches than that covered by BOND, drawn mostly from the *Better Evaluation* website. The chapter then tackles the question of what criteria should inform the choice of impact evaluation approach. Given the complexity of development problems, and the inevitable constraints of time and money on what evidence it is possible to collect, we emphasize the importance of choosing an approach that (a) combines testing and exploring theories of change, and (b) selects a threshold of credibility or certainty to suit the main user. The chapter also affirms the value of the QuIP as an approach to assessing attribution that builds flexibly and incrementally on what users already know, rather than assuming that they would otherwise know nothing.

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 an implementing agency to carry out specified development activities for a target group of intended beneficiaries. Three feedback loop mechanisms can then be distinguished.[2]

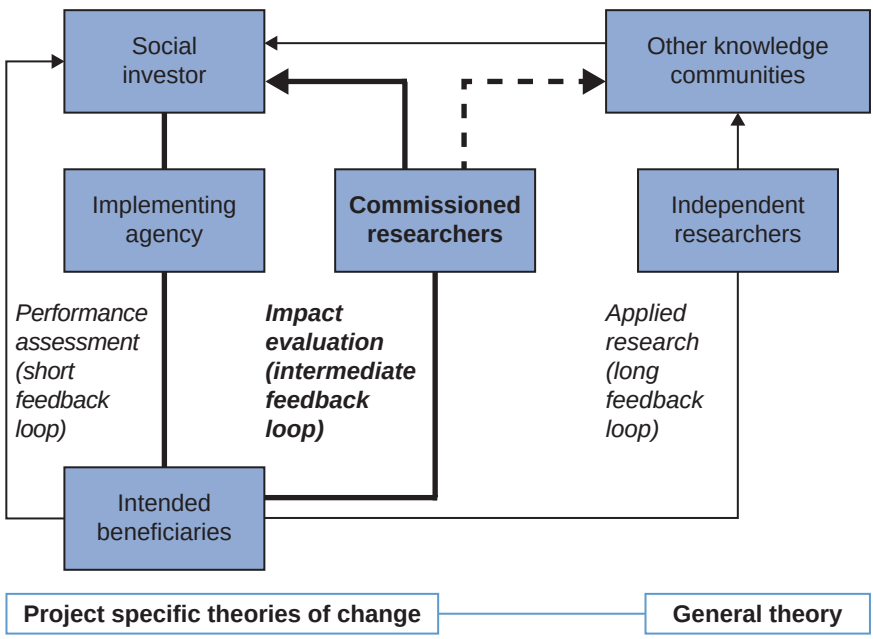


Figure 2.1 Development impact feedback loops

- First, the social investors can rely on what they are told by the implementing agency – 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 produced by relevant knowledge communities, 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 for them. We call this the intermediate feedback loop, and this is the route for which the QuIP is designed. In large organizations this role may be performed wholly or in part by staff who are employed and have specialist expertise in evaluation, but who are not directly involved in management or implementation of the project.

The bottom of the diagram sets up a spectrum of the types of theory underpinning these feedback loops: from theory that is highly specific to a particular project, to theory that is much more general. All three feedback loops are informed by prior theories or ideas about what the intervention is, or should be, achieving and how, although this may not be formally expressed or agreed. There has been a trend in international development over the last decade towards a more explicit statement of 'theories of change' or 'ToCs' (Vogel, 2012). These include a statement of the programme logic through which an investor expects to have an impact on intended beneficiaries via the actions of the implementing agency; but also incorporate a wider understanding of how these actions interact with other dynamics within a wider specified context or system (Prinsen and Nijhof, 2015). Project specific theories of change, in turn, draw upon more general theories, both from specific areas of professional practice and from broader knowledge communities. For example, a causal claim that adopting a new crop variety will raise grain yields may be built into project specific theory, but also depends on theories (and indeed scientific laws) drawn from agronomy and biology.

Short feedback loops

Much of the feedback on development project impact is generated and used through the implementing organizations' 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 be divergent, leaving project managers and investors with the challenge of deciding who and what to believe. Hence the quality of such feedback critically depends on organizational 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 organization's activities generate impact. The short feedback loop serves in part to confirm, refine, challenge or contradict such theories of change.

Short feedback loops enable organizations to operate most of the time. However, they are fallible. Cognitive traps, herd effects, and collective self-delusion are all possible. Larger organizations 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.

Long feedback loops

It is common sense for social investors to evaluate short feedback loop evidence against evidence available from independent sources. Long feedback loop data can be very diverse, including everything from media reports to academic publications, via official reports and the published outputs of civil society organizations. It can be distinguished from both short and intermediate feedback on the basis that it is neither supplied by the implementing agency, nor commissioned directly by the social investor. This means that the social investor faces a problem identifying and selecting from it what is most relevant, credible, and useful.

A defining characteristic of long feedback is that the social investor has limited control over its scope and quality, even if they invest in it directly. This is because independent researchers are first and foremost beholden to a wider peer group or knowledge community, such as an academic discipline or a professional field. For example, 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 and ideas that are irrelevant or incidental to the project.[4]

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 or 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 of (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 intermediate or *middle-range theory*, which falls somewhere along this spectrum. This is useful for making decisions over how far a project is likely to be successful in slightly different contexts.[5]

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 important 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 how far a study corroborates or challenges 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 facilitating *adaptive management* depends in no small part on its contribution to useful middle range theory – i.e. judgements about how and to what extent impact achieved under the specific project is likely to be achievable in other contexts. But it also depends on the social role of the evaluator in relation to the commissioner and other stakeholders; they have the advantage of gaining additional access and influence compared with fully independent researchers, but are 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 also on how these influence social and political relationships.

Comparing the QuIP with other approaches to impact evaluation

An initial classification

Many different approaches to evaluation can be used to generate intermediate feedback. The *Better Evaluation* website (www.betterevaluation.org) is a useful source of information on a wide range of evaluation approaches.

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

Group 1. Approaches with some *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; utilization focused evaluation.
Group 2. More *quantitative* approaches than the QuIP	Cost benefit analysis; difference-in-difference evaluation; qualitative comparative analysis; randomized control trials; social return on investment.
Group 3. *Broader* approaches, with which the QuIP is congruent	Beneficiary assessment; contribution analysis; developmental evaluation; innovation history; institutional histories; outcome harvesting; process tracing; realist evaluation.
Group 4. Approaches with 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.

It defines these 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. These are grouped into seven clusters: how to manage, define, frame, describe, understand causes, synthesize, and report/support use. The 'understanding causes' task includes checking that results support causal attribution, comparing results with 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.[6]

The Appendix to this chapter briefly describes each of these approaches in turn, along with seven others.[7]

Taking the QuIP as a single point of comparison, we have classified these approaches 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. Turning to Group 2, the QuIP departs more radically from quantitative approaches to causal attribution, but with some scope for complementary use. QuIP shares more than one important feature with approaches in Group 3, but is generally narrower and more prescriptive in its specification of how different evaluative tasks are completed. Likewise, while the QuIP aims to strengthen feedback from intended beneficiaries to social investors, it lacks the primary emphasis on downward accountability and empowerment that is a feature of the approaches in Group 4. The following sections selectively explore these similarities and differences in more depth.

Approaches with features that overlap with the QuIP (Group 1)

These approaches differ in emphasis, but overlap with the QuIP in at least one important way. This highlights both the QuIP's eclectic character and the

scope for improvisation in its use. To give three examples: first, the QuIP aims to be open to both positive and negative stories of change. However, it can easily be adapted to be used more restrictively to focus on the positive, as do both the 'appreciative enquiry' and 'positive deviance' approaches. Second, 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 and Britt (2013) and Paz-Ybarnegaray and Douthwaite (2016). Third, 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).

QuIP and quantitative approaches to impact evaluation (Group 2)

The QuIP seeks evidence of causation in the form of narrative statements about the impact of selected activities (X) on selected dimensions (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 'within-case' statements made by respondents themselves.[9]

Within the wider literature on causal attribution this feature clearly sets the QuIP apart from 'secessionist' quantitative approaches to evaluation that exploit variation in the exposure of a population to an intervention in order to infer impact statistically by relying on observed regularities between selected variables (Mohr, 1999; Maxwell, 2004; White, 2010; Gates and Dyson, 2017). 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 from experimental and/or observational data. The most widely espoused quantitative approach to impact evaluation is to rely on randomized controlled trials (RCTs). This is discussed in Box 2.1.

Qualitative and quantitative approaches to impact evaluation and research have distinct purposes: the first is more concerned with identifying, explaining, and interpreting causal processes, the latter with more narrowly codifying data in order to facilitate measurement and mathematical analysis (Moris and Copestake, 1993). And while the argument that they constitute 'incommensurate paradigms' has been widely rejected (Morgan, 2007), few would disagree that they embody 'distinct cultures' (Goertz and Mahoney, 2012). For this reason there are grounds for not directly comparing the QuIP

Box 2.1 Impact evaluation based on randomized controlled trials (RCTs)

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 and Duvendack, 2014). Subject to being able to randomly assign the treatment across a large enough sample, then those not treated serve as a counterfactual for those in the treatment group, of what would have happened to them if they hadn't been treated. RCTs can then supply an estimate of the average impact of being in the treatment group, across the sample. This can, in turn, be given a monetary value and incorporated into cost benefit analysis. 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: perfect randomization is not possible if sample sizes are too small; the control group may be contaminated by treatment effects; responses to interviews may be affected by how people feel about being in the treatment or control group; or spillover effects from the treatment group may affect the control group (Glennerster and Takavarasha, 2013; White and 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 within the assessed sample or beyond it (Cartwright and Hardie, 2012; Deaton and Cartwright, 2018). This limits the generalizability (or external validity) of findings, and hence the value-for-money of RCTs, given that they are time consuming and can be hugely expensive. For this reason they are most appropriate when evaluating relatively large investments or testing theory with wide potential relevance; indeed, if a programme or problem is large enough then using them to investigate important implementation issues may also be justified (Duflo, 2017). An important 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 their use reflects the aspirations and standards of researchers seeking approval of their academic knowledge community, with correspondingly less weight given to the prior knowledge and credibility thresholds of commissioners, and to the importance they attach to timeliness, sufficiency, relevance, generalizability, 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, potentially more flexible approaches to impact evaluation (Stern et al., 2012).

with quantitative methods at all, and certainly for doing so only cautiously and carefully.

However, the distinction between them can also be usefully deconstructed in order to open up avenues for using them in complementary ways, and for integrating aspects of both in the same study (Tashakkori and Creswell, 2007; Fetters and Molina-Azorin, 2017). This entails first distinguishing between the different characteristics and attributes associated (or conflated) with each: parsimonious/complex; deductive/inductive; numbers/words; facts/meaning; generalized/contextualized; open/closed; narrower/broader in scope, and so on. Doing so then opens up possibilities for transcending the broad distinction between them through a more nuanced analysis of the specific attributes of different tasks within any research process. For example, QuIP coding is both inductive and deductive.[10] This more open view of the relationship provides

more scope for reflecting on how the QuIP can contribute to mixed method evaluation. Chapter 10 returns to this issue.

Approaches with which the QuIP broadly belongs (Group 3)

The leading alternative to 'large N' impact evaluation based on statistical inference is theory-based evaluation, also referred to as the 'modus operandi' approach (Scriven, cited in Mohr, 1999).[11] 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 the change, the presence of 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 true 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 at the same time, leaving the relative contribution of each uncertain.

QuIP and process tracing. Process tracing as a form of theory testing entails assessing the extent to which discrete pieces of evidence cumulatively strengthen or weaken a user's confidence in a theory of change linking X and Y (Kay and Baker, 2015). Process tracing is particularly powerful in assessing the causes of important singular events, like a change of policy or the outbreak of conflict, but can also be applied more generally. It can be linked mathematically both to logic and set theory (e.g. Goertz and Mahoney, 2012) and to Bayesian statistics (e.g. Fairfield and Charman, 2017; Befani and Stedman-Bryce, 2017).[12]

The link between the QuIP and process tracing becomes clear if unprompted positive explicit attribution in the QuIP is likened to 'smoking gun' evidence of impact in process tracing; and implicit attribution to 'hoop test' evidence – where its presence is not conclusive, but its absence casts 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 materialize, 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.

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	*How incorporated into the QuIP*
1. Cast the net widely for alternative explanations.	Multiple interviews and focus groups, combined with blindfolding 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.	Blindfolding reduces the threat of project-related bias and tunnel vision. Data from intended beneficiaries and project staff are collected separately and systematically compared. Unblindfolded 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.
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 or 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 and Checkel (2015)

QuIP and realist evaluation. In complex contexts it is unlikely that all possible theoretical explanations for selected outcomes Y can be identified and systematically ruled in or out by signature evidence as process tracing aspires to do. An alternative and more flexible philosophical basis for making contribution claims relies less on ruling out alternative explanations and more on weighing up the positive causal claims of trusted observers. This appeals to our linguistic power to imagine and articulate hypothetical situations.[13]

The credibility of causal claims generated using the QuIP in a particular context can be broken down into the following components:

- there is sufficient evidence that X and changes in Y happened;
- several respondents independently – and without prompting – explicitly asserted or implicitly suggested that X was part of a package of factors causing the change in Y;
- these assertions are congruent with plausible theory explaining how this could have happened; and
- there is no obviously more credible counter-explanation for why respondents might have said what they did.

This formulation emphasizes the dependence of the QuIP on respondents' perceptions, and reflects its aim to give intended beneficiaries 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' views reflects a realist position that lies somewhere between the claims to universal truth of positivist science, and denial of the possibility of establishing any kind of concrete fact independent of the observer (Maxwell, 2004; Wynn and Williams, 2012). 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' (CMO). However, the 'CMO' terminology does not map perfectly onto the 'ZXY' shorthand used here, 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 of complexity 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 context (Cartwright and Hardie, 2012).

Managing all this complexity is only possible with the help of explanatory theory. Thus for Pawson (2013: 27), '… theory-driven means what it says … designs that attempt to utilize the realist explanatory apparatus without prior grounding in programme theory will end with explanations that are ad hoc and piecemeal'. At the same time, realism is flexible in combining deductive hypothesis formulation with inductive theorizing about causal processes in an iterative way, by positing generative mechanisms that can plausibly explain different configurations of contexts and outcomes. This seems to parallel the emphasis in the QuIP on both confirmatory analysis, based on prior theories of change, and exploratory analysis of causal explanations offered by respondents. In both cases, prior understanding informs questioning but is also refined by it. Identifying multiple CMO configurations informs 'middle-range' theory that is both more general than programme theories of change and more contextualized than the general theories associated with different strands of social science.[14]

The above discussion suggests the QuIP shares two out of three principles of realist evaluation highlighted by Pawson (2013: 14): to have a strong explanatory focus and to acknowledge the complexity of CMO configurations. His third principle is to employ more than 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' (Pawson, 2013: 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 the 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 the QuIP can be viewed as more open-ended, exploratory, and inductive 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.

QuIP and participatory approaches to evaluation (Group 4)

The QuIP is a form of beneficiary assessment (Salmen, 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 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 produce 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' voices from the perspective of the QuIP's primary audience, this must be offset against the potentially disempowering effect of not immediately revealing to respondents 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. One way to reduce this 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 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 – from other participatory evaluation methods listed in Group 4 of Table 2.1. 'Most significant change' and the QuIP, for example, both share a reliance on causal claims elicited from respondents. However, the former gives more weight to doing so in a way from which the respondents can themselves more immediately benefit, rather than analysing data for use by the commissioner and other stakeholders. 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 Participatory Development (PADev) and Participatory Impact Assessment, Learning and Accountability (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 the criteria for evaluating it. Local stakeholders have different prior knowledge against which to weigh up the value of new evidence, including being able to reflect directly on their own experience. To the extent that they are mostly concerned with their own interests, the generalizability of findings will matter less. In these respects, 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.

Choosing between approaches to impact evaluation

How to think about the issue

The previous section aimed to compare and contrast the QuIP with other approaches to impact evaluation in a way that avoided making strong 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 they can answer them.

The QuIP 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 social research produced for a wider knowledge community. QuIP evidence is based on what intended beneficiaries themselves perceive to be the main drivers of change in their lives. It is not expected on its own to generate estimates of the magnitude of these effects, although the evidence may contribute to model-based simulation of impact magnitudes. Nor is the QuIP designed 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 prior expectations about this. It also aims to cast light not only on X but on other causes 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 variation in these causal patterns between people and contexts.

The QuIP has been designed and has evolved through a combination of learning-by-doing and close consultation with actual and potential users about what they consider to be 'good enough' to inform their activities, taking into account timeliness, cost, and prior understanding. This approach has been pragmatic and eclectic, but builds on realist philosophical foundations that emphasize complexity. This in turn underpins a fear of the danger of

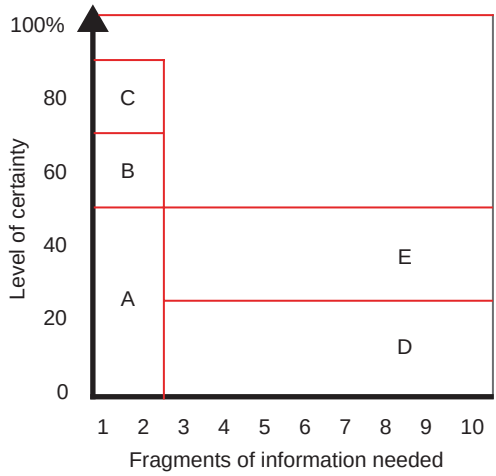
over-generalizing about 'what works' with respect to both development practice and how to assess it. With this comes a preference also for a pluralist and evolutionary view of how to identify and promote good practice.[15] This rejection of a universal solution to the attribution challenge should not be mistaken for the view that every opinion has equal weight. For a given problem in a given place, there will be a better and a worse way to assess impact, and even a best way: hence also 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 purposes and for whom. Impact evaluation is conceived here both as a complex and rapidly changing field, and as 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 commitment to different products, but they also reinforce market power and tradition. But 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).

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 generalizations, but based more on analysis of what constitutes an acceptable threshold of evidence for commissioners in different contexts than an absolute view. Decision-makers differ according to their appetite for certainty and uncertainty. They also start out with different prior levels of knowledge.

A simple model illustrates the implications of this variation (Figure 2.2). This contrasts a potential investor in a social enterprise, who wants to learn more about its social impact, and its owner/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 views of cost-effectiveness will also depend on their *certainty appetite* and prior knowledge. In Figure 2.2, 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 enterprise, starting with the one they agree is most important (1) and adding less important items of information up to 10. The Y axis plots certainty levels for this knowledge, from self-confessed total ignorance (0 per cent)



	Manager	Investor
Prior knowledge	A+B+C+D	0
Incremental learning from first study	C	A+B+C
Incremental learning from second study	E	A+D+E

Figure 2.2 Criteria for comparing impact assessment studies

to total certainty (100 per cent). Assume the manager already has a view on the two most important items, with 80 per cent certainty, and the other eight with 20 per cent certainty; while, in contrast, the investor is ignorant of everything. Two independent impact studies are proposed of equal cost. One will provide 90 per cent certainty about the first two items. The other will deliver 50 per cent certainty about all 10. 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 sufficient level of shared understanding and trust to convince the investor to invest in the business.

This model illustrates that choosing how to spend money wisely on impact assessment depends on the commissioner's prior knowledge, (un) certainty preferences, and the range of issues they regard as important to cover. For example, there is the choice between studies that set out to confirm known causal pathways or to explore those that are largely unknown. The figure also highlights the importance of trust. In the example, 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 per cent 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 small 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 two key issues.

Returning to the real world, this example illustrates why impact assessment may be a source of conflict even between like-minded stakeholders. Potential for disagreement may also be exacerbated by differences in understanding of

alternative evaluation methods, and disagreement over how precisely impact needs to be measured (Muller, 2018). It also speaks to the widely held view of development practitioners that their paymasters are prone to 'overkill' in assessing development activities through endless requests for information, missions, performance reviews, and audits – a tendency that is not only costly but can reduce the likelihood of their being able to do anything truly transformative (Natsios, 2010).[16] This explains an emphasis in designing the QuIP on finding ways to assess impact that is cheaper, more flexible, and supplements what is already known.

Scope for generalizing about 'what works' in development (and chess)

An additional consideration behind the design of the QuIP is an appreciation of the huge range of contexts and combinations of drivers of change that it would be useful to understand better. Andrews et al. (2012, 2017) emphasize 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 more agile approaches to exploring alternatives.

To illustrate the importance of this point consider the game of chess. Evaluating different moves and strategies is obviously relatively simple compared with 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 capabilities. The complexity of the changing context of the board at each move is offset by the simplicity of the ultimate goal, and by the restricted room for manoeuvre of each 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 numbers of people whose motives, resources, opportunities, and understanding are often only weakly understood?

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 a queen, for example. This feeds into case study analysis which locates discrete moves in the context of the whole board and a complete game. Inductive analysis can also be used to build 'middle-range' theory (castle the king early; do not 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 generalizations about what to do and what not to do. However, it also needs to guard against over-generalization, 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 raises the question of what level of generalization it is possible to achieve about the relevance of different approaches to impact evaluation. We will find it useful in places to generalize 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. It can also 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 case studies of how the QuIP has been employed, why, what findings were generated, and (what is more difficult) how they were used. These should help inform thinking about the scope for adapting the QuIP for use in other 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 generalizing from past 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, I 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 generalization in coming 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.

Conclusions

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

Second, it considered 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 version 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 approaches to 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. This section also argued strongly that in a highly complex design space there is a particular need for 'agile' approaches to impact evaluation, like the QuIP, that are relatively inexpensive, simple, incremental, open-ended, and flexible.

Appendix: comparing the QuIP with 30 other approaches to impact evaluation

Approach and brief description[17]	*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 intended beneficiaries, aiming to give voice to their priorities and concerns.	The QuIP is a form of beneficiary assessment, but offers 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 it is 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 focuses on the causal links from activities to outcomes and impacts on intended beneficiaries.
Collaborative outcomes reporting (COR)	
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.

(Continued)

Table Continued

Approach and brief description[17]	*How the QuIP compares*
Contribution analysis	
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 offers more specific and detailed guidelines.
Cost benefit analysis	
A general approach for comparing incremental benefits and costs of an action compared with one or more alternatives. Key steps include: identification of options; scoping of key stakeholders and the impact on them of each option over time; quantification of 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 judgements, 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' motivations, power, worldviews 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 evaluation 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 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 to estimate average treatment effects while mitigating 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 with a control group. 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.

Approach and brief description[17]	*How the QuIP compares*
Goal-free evaluation	
Open interviews and observation that seeks to understand respondents' lived experiences holistically and the meaning they give to them, and to view specific interventions in this light.	Blindfolding is utilized as part of the QuIP to facilitate similarly open-ended and exploratory enquiry within specified domains of respondents' lived experiences. The QuIP goes further by then systematically comparing findings with the theory of change behind a given intervention.
Horizontal evaluation	
An approach that combines self-assessment by local participants and external review by peers, typically through a three-day joint workshop.	The QuIP is not specifically oriented towards locally led activities, and aims to generate evidence that is more 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.	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 stakeholder network analysis as fully as these approaches.
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.	
Most significant change	
Collects and analyses personal accounts of change, and includes processes for learning about what changes are most valued by individuals and groups.	The QuIP shares an emphasis on eliciting respondents' own accounts of causal processes, but without prioritizing the most significant. It relies on more formal thematic analysis of causal stories rather than on participatory processes for ordering and interpreting 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 offers 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	Elaborating a detailed theory of change for an intervention is a necessary step for attribution coding and hence for using the

(Continued)

Table Continued

Approach and brief description[17]	*How the QuIP compares*
transformative change, and allows for the plausible assessment of the initiative's contribution to results.	QuIP to confirm that an intervention is achieving what was intended. The use of journals by different stakeholders to monitor changes can be incorporated into the QuIP as an additional source of narrative evidence.
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 organized through a mediated community workshop.	PADev and the QuIP are both based on narrative accounts of drivers of change that try to avoid framing those accounts by reference to a specific activity. PADev does this by taking a community-wide perspective, while the QuIP does it through blindfolding. Both produce findings that are potentially relevant to all organizations working in the locality, but the QuIP is more strongly tailored to the information needs of the commissioning organization.
Participatory impact assessment for learning and accountability PIALA is an eclectic approach to gathering data about a development intervention using a range of participatory methods, and also involves intended beneficiaries themselves in analysis and interpretation of data using the 'Sensemaker' proprietary software developed by the company Cognitive Edge.	The two approaches share the goal of generating both formative/exploratory and summative/confirmatory data at the same time, and the QuIP could be incorporated into PIALA as a form of data collection. But they employ different approaches to deriving and presenting data from primary sources. The QuIP does not involve intended beneficiaries directly in the initial analysis of data, but they can be consulted on how to interpret data.
Participatory evaluation A range of approaches that engage stakeholders (especially intended beneficiaries) in conducting the evaluation and/or in making decisions about the evaluation. (This also incorporates *participatory rural appraisal*, and *participatory learning and action*).	The QuIP aims to give voice to a sample of intended beneficiaries and can involve them in interpreting and using findings, but it does not aim to involve them directly in data analysis or management of the evaluation. It primarily responds to demand from a commissioning organization, and hence the primary focus is on upward rather than downward accountability.
Positive deviance Involves intended evaluation of users in identifying outliers – or cases with exceptionally good outcomes – and then understanding how they have achieved these.	Where change in key outcome variables is being monitored across a population, sample selection and data collection for the QuIP can be deliberately biased towards positive deviants. But the QuIP can also be used to illuminate drivers of change more widely across the population, and/or to focus on gaining a better understanding of reasons for negative deviance.

Approach and brief description[17]	*How the QuIP compares*
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.	The QuIP also seeks evidence to confirm or challenge a theory of change (e.g. that a project was a necessary condition for a specified impact on an intended beneficiary). The QuIP does this for multiple cases and possible impacts, and, like process tracing, each additional piece of evidence adds to or weakens the commissioner's prior belief in the theory.
Qualitative comparative analysis (QCA)	
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 be utilized for QCA to analyse multiple factors contributing to specified outcomes, including the contribution of a specified intervention. But the QuIP avoids *ex ante* specification of the drivers of change to be covered, so may leave data gaps that limit the scope for using the data for QCA.
Randomized 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 treatment group or groups (see Box 2.1).	The 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. using the QuIP to elucidate causal mechanisms, unanticipated consequences and reasons for heterogeneity of impact; and the RCT to quantify the average impact across a selected population.
Realist evaluation	
Realist evaluation is a form of theory-driven evaluation, distinguished by its philosophical emphasis on how interventions influence particular decisions or not. It emphasizes complexity, heterogeneity, and the benefits of combining different methods of data collection and analysis.	The QuIP is a more precisely specified approach, but shares many features with realist evaluation, and 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.

(Continued)

Table Continued

Approach and brief description[17]	*How the QuIP compares*
Social return on investment	
Identifies a broad range of social outcomes (not only the direct outcomes for the intended beneficiaries of an intervention) then quantifies 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 to 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 useful for understanding what enhances or impedes impact.	The QuIP also relies on comparative case studies, which may be individuals, households, organizations, and/or clusters of them. Where data is available for key impact indicators then it is possible to select more and less successful cases.
Utilization-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 the QuIP is 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 the QuIP can also generate useful evidence about an intervention that was not anticipated or solicited for a predetermined purpose.

Notes

1. See www.bond.org.uk/resources/evaluation-methods-tool. BOND is the leading UK membership body for organizations working in international development. The 11 methods reviewed are randomized controlled trials (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.
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 help investors to better 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 formalized under the label of the 'Reality Check' approach (Jupp, 2016).

3. Academic research into feedback at this level is rich and diverse. Flyvbjerg (2001) uses Aristotle's term *phronesis* to emphasize the importance of context-specific and capable judgement or practical wisdom, contrasting it with both abstract scientific knowledge and technical skill. Bourdieu'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. Copestake (2013) explores the tension between impact evaluation and applied research (i.e. intermediate and long feedback loops) for the case of microfinance in India.
5. For a fuller explanation of the idea of middle-range theory see Blamey and Mackenzie (2007), Pawson (2013), and discussion of realist evaluation in the section headed 'Approaches with which the QuIP broadly belongs'.
6. 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 Raitzer (2017), and for qualitative approaches see Stern et al. (2012), or White and Phillips (2012).
7. The additional seven are: 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. These have been added because they all entered explicitly into discussion during the process of designing and testing the QuIP.
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).
9. If the purpose of a study is to test a causal theory that the presence of an intervention X is necessary to an outcome Y (as a necessary part of a sufficient package of causes) then investigating cases where X is absent is irrelevant. However, doing so may help to identify alternative or 'equifinal' packages that also lead to Y. For systematic discussion of this see Goertz (2017).
10. Pushing the point one step further, deductive specification of domains for a QuIP study, based on a prior theory of the different dimensions of wellbeing, can contribute to data collection that can then be analysed inductively to suggest revisions to this theory: the sequential use of deduction and induction being an example of 'abduction' (Pawson, 2013).
11. Following Mayne (2012: 273), theory-led evaluation can be extended also to include contribution analysis, which 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?'
12. Bayesian updating of prior expectations can also be used in principle to integrate evidence obtained using quantitative and qualitative impact assessment methods (Humphreys and Jacobs, 2015).

13. 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. While confidence in the answer is enhanced if the tacit counterfactual is made explicit, it is generally impossible to expose and disentangle all the possible scenarios respondents may have in mind and be rejecting.
14. Some realist enquiry appears to differ from the QuIP by advocating full and open sharing of researchers' and their research subjects' understanding of project theory (Manzano, 2016). However, the QuIP ideal is to achieve the same, only in two stages: blindfolded then unblindfolded encounters. This enables similarities and differences in understanding to be exposed more clearly to third parties (chiefly the commissioner).
15. This approach to analysing the problem borrows from the idea that our collective understanding is 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. For example, competition between quantitative and qualitative approaches to evaluation partly reflect grander controversies over development theory, social science, and philosophy (e.g. see the discussion of David Hume's seminal writing on causation in Chapter 6 of Goertz and Mahoney (2012)).
16. 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 are of course much older: see, for example, the classic lament about 'survey slavery' in Chambers (1983).
17. Most of the text in this column is taken from http://www.betterevaluation.org/en/approaches

References

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

Andrews, M., Pritchett, L. and Woolcock, M. (2012) *Escaping Capability Traps through Problem-driven Iterative Adaptation (PDIA)*, Paper 299, Washington, DC: Centre for Global Development.

Andrews, M., Pritchett, L. and Woolcock, M. (2017) *Building State Capability: Evidence, Analysis, Action*, Oxford: Oxford University Press.

Bamberger, M., Vaessen, J. and Raimondo, E. (eds) (2016) *Dealing with Complexity in Development Evaluation: A Practical Approach*, Los Angeles: Sage Publications.

Befani, B. and Stedman-Bryce, G. (2017) 'Process tracing and Bayesian updating for impact evaluation', *Evaluation* 23(1): 42–60 <https://doi.org/10.1177%2F1356389016654584>.

Bennett, A. and Checkel, J. (2015) *Process Tracing: From Metaphor to Analytic Tool*, Cambridge, UK: Cambridge University Press.

Blamey, A. and Mackenzie, M. (2007) 'Theories of change and realistic evaluation', *Evaluation* 13(4): 439–55 <https://doi.org/10.1177%2F1356389007082129>.
Boulton, J., Allen, P. and Bowman, C. (2015) *Embracing Complexity: Strategic Perspectives for an Age of Turbulence*, Oxford, UK: Oxford University Press.
Camfield, L. and Duvendack, M. (2014) 'Impact evaluation – are we "off the gold standard"?' *European Journal of Development Research* 26(1): 1–12 <https://doi.org/10.1057/ejdr.2013.42>.
Cartwright, N. and Hardie, J. (2012) *Evidence-based Policy: A Practical Guide to Doing it Better*, Oxford, UK: Oxford University Press.
Chambers, R. (1983) *Rural Development: Putting the Last First*, Harlow, UK: 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 <http://dx.doi.org/10.1080/13600818.2012.689818>.
Copestake, J. (2015) 'Whither development studies? Reflections on its relationship with social policy', *Journal of International and Comparative Social Policy* 31(2): 100–13 <https://doi.org/10.1080/21699763.2015.1047396>.
Davies, R. (2018) 'Network visualisation of qualitative data' [online], *Monitoring and Evaluation News* <http://mande.co.uk/special-issues/participatory-aggregation-of-qualitative-information-paqi/> [accessed 11 June 2018].
Deaton, A. and Cartwright, N. (2018) 'Understanding and misunderstanding randomized control trials', *Social Science and Medicine* 210: 2–21 <https://doi.org/10.1016/j.socscimed.2017.12.005>.
Duflo, E. (2017) 'Richard T. Ely Lecture: The economist as plumber', *American Economic Review: Papers and Proceedings* 107(5): 1–26.
Eyben, R. (2010) 'Hiding relations: The irony of "effective aid"', *European Journal of Development Research* 22(3): 382–97.
Eyben, R., Guijt, I. and Shutt, C. (2015) *The Politics of Evidence and Results in International Development*, Rugby, UK: Practical Action Publishing.
Fairfield, T. and Charman, A. (2017) 'Explicit Bayesian analysis for process tracing: guidelines, opportunities and caveats', *LSE Research Online* <eprints.lse.ac.u/69203/>.
Fetters, M. and Molina-Azorin, J. (2017) 'The *Journal of Mixed Methods Research* starts a new decade: principles for bringing in the new and divesting of the old language of the field', *Journal of Mixed Methods Research* 11(1): 3–10 <https://doi.org/10.1177%2F1558689817696365>.
Flyvbjerg, B. (2001) *Making Social Science Matter: Why Social Inquiry Fails and How it Can Succeed Again*, Cambridge, UK: Cambridge University Press.
Flyvbjerg, B. (2006) 'Five misunderstandings about case-study research', *Qualitative Enquiry* 12(2): 219–45 <https://doi.org/10.1177%2F1077800405284363>.
Gates, E. and Dyson, L. (2017) 'Implications of the changing conversation about causality for evaluators', *American Journal of Evaluation* 38(1): 29–46 <https://doi.org/10.1177%2F1098214016644068>.
Glennerster, R. and Takavarasha, K. (2013) *Running Randomized Evaluations: A Practical Guide*, Princeton, NJ: Princeton University Press.
Goertz, G. (2017) *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*, Princeton, NJ: Princeton University Press.

Goertz, G. and Mahoney, J. (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*, Princeton, NJ: Princeton University Press.

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

Guest, G., Bunce, A. and Johnson, L. (2006) 'How many interviews are enough? An experiment with data saturation and variability', *Field Methods* 18: 59–83 <https://doi.org/10.1177%2F1525822X05279903>.

Hayman, R., King, S., Kontinen, T. and Narayanaswamy, L. (eds) (2016) *Negotiating Knowledge: Evidence and Experience in Development NGOs*, Rugby, UK: Practical Action Publishing with INTRAC.

Humphreys, M. and Jacobs, A. (2015) 'Mixing methods: a Bayesian approach', *American Political Science Review* 109(4): 653–73 <https://doi.org/10.1017/S0003055415000453>.

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

Kay, A. and Baker, P. (2015) 'What can causal process tracing offer to policy studies? A review of the literature', *Policy Studies Journal* 43(1): 1–21 <https://doi.org/10.1111/psj.12092>.

Manzano, A. (2016) 'The craft of interviewing in realist evaluation', *Evaluation* 22(3): 342–60 <https://doi.org/10.1177%2F1356389016638615>.

Maxwell, J. (2004) 'Using qualitative methods for causal explanation', *Field Methods* 16: 243–64 <https://doi.org/10.1177%2F1525822X04266831>.

Mayne, J. (2012) 'Contribution analysis: coming of age?' *Evaluation* 18(3): 270–80 <https://doi.org/10.1177%2F1356389012451663>.

Mohr, L. (1999) 'The qualitative method of impact analysis', *American Journal of Evaluation* 20(1): 69–84 <https://doi.org/10.1177%2F109821409902000106>.

Morgan, D. (2007) 'Paradigms lost and pragmatism regained: methodological implications of combining qualitative and quantitative methods', *Journal of Mixed Methods Research* 1(1): 48–76 <https://doi.org/10.1177%2F2345678906292462>.

Moris, J. and Copestake, J. (1993) *Qualitative Enquiry for Rural Development*, Rugby, UK: ITDG Publications.

Muller, J.Z. (2018) *The Tyranny of Metrics*, Princeton, NJ: 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.

Paz-Ybarnegaray, R. and Douthwaite, B. (2016) 'Outcome evidencing: a method for enabling and evaluating program intervention in complex systems', *American Journal of Evaluation* 38(2): 275–93 <https://doi.org/10.1177%2F1098214016676573>.

Pouw, N., Dietz, T., Belemvire, A., de Groot, D., Millar, D., Obeng, F, Rijneveld, W., Ven der Geest, K., Vlaminck, Z. and Zaal, F. (2016) 'Participatory assessment of development interventions: lessons learned from a new evaluation methodology in Ghana and Burkina Faso', *American Journal of Evaluation* 38(1): 1–13 <https://doi.org/10.1177%2F1098214016641210>.

Prinsen, G. and Nijhof, S. (2015) 'Between logframes and theory of change: reviewing debates and a practical experience', *Development in Practice* 25(2): 234–46 <https://doi.org/10.1080/09614524.2015.1003532>.
Rodrik, D. (2008) *The New Development Economics: We Shall Experiment, But How Shall We Learn?* Faculty Research Working Paper 08-055, Cambridge, MA: John F. Kennedy School of Government.
Room, G. (2011) *Complexity, Institutions and Public Policy: Agile Decision-making in a Turbulent World*, Cheltenham, UK: Edward Elgar.
Salmen, L.F. (2002) *Beneficiary Assessment: An Approach Described*, 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. (2015) *Impact Evaluation: A Guide for Commissioners and Managers*, London: BOND.
Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, London: Department for International Development.
Tashakkori, A. and Creswell, J. (2007) 'Editorial: the new era of mixed methods', *Journal of Mixed Methods Research* 1(1): 4–7.
van Hemelrijck, A. (2016) *Methodological Reflections Following the Second PIALA Pilot in Ghana*, Rome: International Fund for Agricultural Development.
Vogel, I. (2012) *Review of the Use of 'Theory of Change' in International Development*, London: Department for International Development.
White, H. (2010) 'A contribution to current debates in impact evaluation', *Evaluation* 16(2): 1–11 <https://doi.org/10.1177%2F1356389010361562>.
White, H. and 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. and Raitzer, D.A. (2017) *Impact Evaluation of Development Interventions: A Practical Guide*, Manila: Asian Development Bank.
Wilson-Grau, R. and 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.
Wynn, D. and Williams, C. (2012) 'Principles for conducting critical realist case study research in information systems', *MIS Quarterly* 36(3): 787–810.
Youker, B. (2013) 'Goal-free evaluation: a potential model for the evaluation of social work programs', *Social Work Research* 37(4): 432–8.

About the author

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

CHAPTER 3

A deep dive into Diageo's malt barley supply chain in Ethiopia

James Copestake, Gabby Davies and Tefera Goshu

This chapter presents a case study of the Qualitative Impact Protocol (QuIP). It was used as a 'deep dive' into the social impact of global beverage company Diageo's investment in supply chain development in Ethiopia. The Sourcing for Growth (S4G) programme promoted barley production by smallholder farmers for use in commercial beer production. Two farming localities were selected for comparison to capture variation in offtake of grain per hectare relative to supply of inputs. This highlighted wide variation in social and economic impact. In addition to illustrating the value of informed and purposeful 'small n' sample selection, the chapter emphasizes the importance of locating impact evaluation in a wider political economy context. It is one of seven case studies exploring how the QuIP was used in different contexts during 2016 and 2017.

Keywords: impact evaluation, causal attribution, Ethiopia, cash crops, barley, beer

Introduction

Travelling south-east from Addis Ababa the road climbs out of the Great Rift Valley onto the Didda Plateau, spiritual home of Ethiopia's world class athletes and a landscape of undulating fields, winding lanes, eucalyptus, and majestic juniper trees. It was also the setting for one of the four projects on which the QuIP was first tested – a Self Help Africa project designed to increase smallholder production and sale of barley, Ethiopia's third most important crop. That study had confirmed the potential to raise rural incomes by building on long-established government farm extension services and farmer cooperatives to improve barley yields and marketing channels (Copestake et al., 2015).

This chapter draws on a follow-up QuIP study commissioned by Diageo, the global beverage company. Diageo is the largest spirits company in the world: in 2017 it generated £3.6 bn profit on global sales of £12 bn (net of excise duties) from 200 brand alcoholic drinks, including Johnnie Walker, Smirnoff, and Guinness. Growing barley for the brewing industry may seem an incongruous model of rural development for those who hold on to the

http://dx.doi.org/10.3362/9781780447469.003

image of an Ethiopia mired by conflict and famine (Gill, 2010), and the chapter concludes with a discussion of this. But first, it tells the story of how and why the QuIP study was commissioned by Diageo in 2016, and what the findings were. It then elaborates on the issue of sample selection, on the grounds that this is of wider relevance. Lastly, we return to political economy and public policy. The chapter was drafted by Copestake, mostly drawing on the QuIP report (BSDR, 2016). It was then revised in the light of comments and suggestions from Goshu and Davies (who were responsible for data collection and analysis, respectively) and David Croft, who commissioned the study on behalf of Diageo.

One feature of rapid economic growth in Ethiopia, particularly in Addis Ababa and other major cities, is the increase in disposable income that the middle and upper classes can spend on consumption goods. Prominent among these are alcoholic beverages, including beer. This is true of many African countries, and a small number of multinational companies are competing fiercely for shares of this burgeoning market by importing established international brands, producing them locally, and by buying up often previously state-owned local breweries and brands (Jernigan, 2015). Sourcing raw materials as cheaply and reliably as possible is an important driver of profitability. Despite being land-locked and having an economy dominated by agriculture, much of the demand for malted barley for commercial beer production in Ethiopia has until recently been met by imports.[1] Hence, while it has encouraged foreign investment in brewing, the Ethiopian Government has at the same time been keen to increase the value of local content in domestic beer production.

Diageo consolidated its niche in Ethiopia by purchasing both the local Meta brand and the Meta Abo brewery near Addis Ababa in 2012, for US$225 m.[2] Whether or not increasing local procurement of barley was a condition for this, or indeed is profitable in the short term, the company clearly has a long-term strategic interest in doing so. This also partly aligns its interests with those of NGOs such as Self Help Africa and major donors such as DFID in seeking ways to strengthen the livelihoods of Ethiopia's many smallholder farmers. Diageo collaborates with several international non-governmental organizations (INGOs) alongside private companies and with the Ethiopian Government itself, to promote increased smallholder barley production, including through a programme called 'Sourcing for Growth' (S4G) (Diageo, 2017). Smallholder barley yields in Ethiopia are rarely more than 2 tonnes per hectare, and can readily be doubled through adoption of a package of improved seeds, chemical inputs, and technical support. Among the INGOs promoting smallholder barley production are Farm Africa, Self Help Africa, and Techno Serve. Much of their work has been carried out with government-sponsored agricultural extension staff and farmers' cooperatives, and supported by official aid donors. But the goal of raising farm incomes also partly aligns their activities and interests with those of big brewing firms like Diageo.

The study

Commissioning

It was through the pilot study of Self Help Africa's malt barley project that a senior corporate social responsibility (CSR) manager for Diageo became aware of the QuIP and decided to commission a study. During an initial conversation he described it as a potentially useful way of conducting what he called a 'deep dive' down the company's supply chain. He was interested to explore possible wider (and possibly negative) social and economic consequences of smallholder malt barley farming, as well as finding ways to do this that would not unduly divert Diageo's staff in Ethiopia away from their routine commercial responsibilities and which would provide valuable insight into the impact of the programmes.

Data collection took place in the middle of 2016, shortly after the fourth season of Diageo's involvement in promoting enhanced production and local barley procurement. Being a commercial operation the evaluation was not tied to a particular project period or reporting deadline, timing of the study being less important than its role as an independent audit through which the head of CSR could assess what was happening at the field level. In addition, there was potential for it to assist in reviewing Diageo's ongoing relationship with the INGO Techno Serve, with whom they were in partnership to implement the S4G programme. Although not specifically adapted to do so, interviews with selected farmers did indeed generate significant insights into that relationship, as described below.

This account illustrates two important points about the commissioning of a QuIP study. First, the commissioner identified a range of potential benefits, including public relations and reputational risk management, as well as evidence-based assessment of whether collaborative investment in supply chain development was working as expected. This highlights the benefits of an approach that can be both confirmatory and exploratory at the same time. Second, potential benefits depended on prior knowledge and expectations of the commissioner and other stakeholders inside and outside the company. The important issue was whether the study could provide insights over and above what was already known, and how this additional information could be used to manage stakeholder relationships. We return to these aspects of the study in the section entitled 'Political economy and public policy context', below.

Data collection

The commissioner was happy to delegate most details of data collection to the QuIP evaluation team, apart from a request to include a question designed to throw light on whether children were more or less involved in farming activities during the period being assessed. To avoid asking farmers outright about child labour, the questions on this topic asked about changes in both adults' and children's use of time for work and other purposes.

Having previously conducted a malt barley study in the region for Self Help Africa, questionnaire design was otherwise relatively straightforward, and also meant that the evaluation team was able to draw upon an experienced team of field researchers to conduct a 'double QuIP' of 48 interviews and eight focus group discussions. All but one of the named primary respondents were men (reflecting male dominance of membership of the farmers' cooperatives), but the focus groups were segmented to include younger and older women, as well as younger and older men in each of two locations. Questions addressed changes in respondents' lives over the past three years in the following domains: food production, cash income, cash spending, food consumption, time working, children's activities, intra- and inter-household relationships, assets, and overall wellbeing.

The population frame for this study comprised more than 6,000 farmers from whom Diageo had purchased malt barley during the 2014/15 season, all of them belonging to one of 39 primary farmers' cooperatives affiliated to five cooperative unions. This lent itself to a two-stage sampling strategy, comprising purposive selection of two primary cooperatives followed by selection of 24 farmers within each. In addition to names and locations, the list of farmers provided for sampling also specified the area they intended to cultivate (as a basis for calculating in-kind credit supply of seed, fertilizer, and other inputs) and how much malt barley they subsequently delivered to Diageo through their cooperative. Overall they delivered 5,498.5 tonnes of barley from 2,802 hectares financed, giving a notional average 'yield' of 1.96 tonnes per hectare, or 0.91 tonnes per farmer. However, the average area financed per farmer and the tonnes procured per hectare financed diverged widely between cooperatives. This led to the decision to purposefully select one primary cooperative (Gese Bilbilo) to represent a more 'extensive' buying pattern (larger area financed, but lower procurement per hectare) and another (Oddo Leka) to represent a contrastingly more intensive pattern of procurement. Table 3.1 draws on the secondary data supplied by Diageo to highlight the difference between the two clusters.

Findings

This section elaborates more fully on the substantive findings generated by the study. Data analysis focused particularly on highlighting and explaining divergent responses between farmers in the two selected localities (as discussed in

Table 3.1 Characteristics of the two selected cooperatives and the household sample

Cooperative	*Gese Bilbilo*	*Oddo Leka*
Pattern of procurement	Extensive	Intensive
Total membership	215	137
Mean estimated area cultivated of all members (hectares of barley)	0.6	0.2
Mean delivery per hectare of all members (tonnes of barley/ha)	1.67	3.66
QuIP household level interview sample	24	24
Mean estimated area cultivated of sampled households (ha of barley)	0.8	0.2
Mean crop delivery of sampled households (tn/ha of barley)	2.42	6.17

the section on sample selection, below). Abundant positive feedback, particularly from Gese Bilbilo, strongly corroborated the implicit theory of change behind the S4G programme, and no evidence arose of negative social impact on education or household nutrition.[3] But in sharp contrast, many respondents in Oddo Leka provided strong and explicit negative feedback about the S4G programme.

Particularly striking was the contrasting evidence of change between the two clusters revealed by closed questions. When asked to sum up whether their experience of change across eight impact domains had been positive, negative or neither, the respondents from Gese Bilbilo answered positively 158 times out of a possible 192 (82 per cent of responses) and negatively only three times. This contrasts sharply with the much more mixed responses for Oddo Leka, shown in Table 3.2. Four respondents were positive or neutral about change across all domains (OL1, OL2, OL3, and OL15), whereas 11 were negative or neutral, with the remaining nine reporting both positive and negative change in different domains.

Table 3.2 Household responses to closed questions in Oddo Leka

HH	*Age*	*FP*	*CY*	*CS*	*FC*	*TA*	*CT*	*A*	*WB*
OL1	1	+	+	+	+	+	=	+	+
OL2	3	+	=	+	=	+	=	+	+
OL3	1	+	+	+	+	+	=	+	+
OL4	1	–	=	–	=	=	=	–	–
OL5	3	+	–	–	=	=	–	+	+
OL6	2	+	+	–	+	+	=	+	+
OL7	4	=	–	=	+	+	=	+	=
OL8	4	=	=	=	=	+	–	+	=
OL9	2	=	=	–	=	=	=	=	–
OL10	4	=	=	–	=	–	=	–	–
OL11	3	=	=	=	=	=	–	=	–
OL12	1	–	–	–	–	=	=	+	+
OL13	2	=	=	–	=	=	=	–	=
OL14	2	=	=	–	=	=	=	=	=
OL15	1	+	+	+	+	+	=	+	+
OL16	1	–	=	–	=	–	=	=	=
OL17	4	+	+	+	+	+	=	–	+
OL18	3	+	–	–	=	+	=	–	=
OL19	2	+	+	–	+	+	=	+	=
OL20	3	–	–	–	–	=	–	–	–
OL21	3	=	=	–	=	=	=	=	=
OL22	4	–	–	–	=	=	–	=	=
OL23	4	=	=	–	=	=	=	+	=
OL24	2	=	–	–	=	–	–	–	=

Notes: HH = household, FP = food production, CY = cash income, CS = cash spending, FC = food consumption, TA = time on agriculture, CT = children's study, A = assets, WB = wellbeing. Age is shown in quartiles: 1 = 24–35; 2 = 37–46; 3 = 47–56 and 4 = 59–75. All but one respondent were men.

Table 3.3 Frequency counts of causal statements by cluster, domain, and attribution tag

Domain	*Positive*			*Negative*		
	Explicit	*Implicit*	*Other*	*Other*	*Implicit*	*Explicit*
Gese Bilbilo (extensive procurement pattern)						
C1. Food production	11 (1)	23 (4)	2(2)			
C2. Cash income	13 (4)	24 (4)				
D1. Purchasing power	3 (1)	21 (4)	6 (1)		7 (1)	
D2. Food consumption	1	16 (3)	1 (3)			
E1. Work time (adults)		19 (4)	1			
E2. Children's work/study balance		1 (1)			2	
F1. Intra-HH relations	2	8 (1)	2 (3)	1		
F2. Inter-HH relations		1 (2)	1 (1)	1 (1)		
G1. Assets	3	20 (4)	(2)	1		
H1. Wellbeing	3	22 (4)		1 (1)		
Oddo Leka (intensive procurement pattern)						
C1. Food production	8 (1)	7(2)	6	8 (3)	(4)	7
C2. Cash income	16 (3)	4	6	5 (1)	2	16 (3)
D1. Purchasing power	1		4	18 (4)	8 (2)	
D2. Food consumption		3	2	1 (1)	1 (1)	
E1. Work time (adults)	2	5	8	5	3 (2)	(1)
E2. Children's work/study balance			0	2	2 (2)	
F1. Intra-HH relations		2	4 (2)	2	2 (2)	
F2. Inter-HH relations			5 (1)	1	1 (2)	
G1. Assets		5 (1)	7	1	5 (2)	
H1. Wellbeing		4	7 (1)	4 (1)	4 (2)	

Note: The first number refers to semi-structured interviews (24 per cluster) and the number in parenthesis to focus group discussions (four per cluster)

The two selected clusters also delivered sharply contrasting sets of stories about the drivers of these changes, including the impact of S4G. This is summarized by Table 3.3, which indicates how many households provided positive and negative causal statements to explain change in each domain, coded according to whether they *explicitly* referred to the S4G programme, were *implicitly* consistent with its theory of change or were *incidental* to it. In Gese Bilbilo, where malt barley production for sale was more established, farmers' feedback was generally very positive. Thirteen out of 24 respondents and all four focus groups explicitly linked S4G to rising cash income, and all of them did so implicitly. Most respondents also identified drivers of increased food production, purchasing power, food consumption, asset ownership, and overall wellbeing that were implicitly consistent with the project's theory of change. In contrast, they reported very few negative drivers of change, the most widely cited (by seven respondents) being that purchasing power had been eroded by increased land rents that they attributed in turn to rising production of malt barley and other cash crops.

In Oddo Leka, two-thirds of interviews and three of the four focus groups also generated evidence of a positive and explicit causal connection between S4G and an increase in cash income. However, exactly the same number of respondents attributed a recent *fall* in cash income explicitly to the programme. Smaller numbers of respondents linked loss of cash income in turn to reduced purchasing power, increased work demands, lower food consumption, poorer relationships within and between households, loss of assets, and a reduction in overall wellbeing.

These findings are based on analysis of predetermined outcome domains using attribution codes (explicit, implicit, other) that are standard for all QuIP studies. Presenting the data in this way provides a quick overview of the evidence generated, but conceals much of the detail of what was actually said. The following excerpt illustrates the point, being just one of eight statements from farmers in Oddo Leka that were coded *negative explicit* under domain C1.

> Over the last three years, income has fluctuated in my household. For example, in 2015 [E.C. 2007 by the Ethiopian Calendar] harvest period, my income was very good. This was due to the fact that the organization called Techno Serve gave us improved Meta beer barley seed and its productivity was very good. In relation to 2006 E.C harvest, the harvest of [2015] was double. This is because the improved seed provided was adaptable to our area's soil type and the provision of sufficient fertilizer and pesticides improved the productivity of the harvest in [2015]. Due to this I was awarded solar panels from the organization. In the next year, I took a huge amount of this improved seed and ploughed the lion's share of my land for barley. Then, the harvest of [2016] totally failed and this has put me into great crisis. Almost all of my land was ploughed for barley, but I didn't harvest one sack of crop, rather it was eaten by livestock as a grass. Even the organization did not ask me for the money for the improved seed and fertilizer that I took. The reason behind the fluctuation in income over the last three years is that, in the first round when the Techno Serve organization gave us the improved Meta beer barley, they provided us with seed which was treated with chemicals. But, in the coming year they provided us non-treated seed. Hence, it failed to be productive and put me into crisis.

Comparing the quotation and Table 3.3 highlights the dilemma that all qualitative analysts face between generalization and simplification of rich primary data. Users of a study cannot avoid their reliance on the professionalism and skill of the analysts in deciding what to aggregate and what to highlight. But the dilemma can partly be mitigated through appropriate juxtaposition of data in different forms, and also by presenting it in ways that make it possible to switch from one to the other quickly and flexibly – an issue to which we will return in later chapters. In addition, the selection process can be formalized by augmenting analysis based on predetermined codes (such as those used in Table 3.3) with more inductive analysis. Table 3.4 takes one step in this direction

Table 3.4 Inductive analysis of positive drivers of changes in income

	Interviews		*Focus groups*	
	Gese Bilbilo	*Oddo Leka*	*Gese Bilbilo*	*Oddo Leka*
Provision of improved seed, fertilizer, insecticide, and herbicide to raise productivity	23	16	OM, YM, YW	OM, YM, OW
Improved farming techniques and productivity due to advice from Development Agents	4	2	OM, YM	
Diversification of crops	14	3	YM	
Renting additional land to produce more crops	2	0		
Good price paid for malt barley crop	16	3	OM, YM, OW, YW	
Diversification into non-farm livelihood activities	1	7	OW, YW	
Specializing in malt barley production	6	1	OW, YW	
Increased commitment to farming (time and methods)	4	3	YM	
Increased livestock trading and/or rearing	11	1	OM, OW, YW	

Note: OM, YM, OW, YW refer to older men, younger men, older women, and younger women. Numbers represent the number of respondents (out of 24) who cited the given driver.

by listing distinct drivers of increased cash income identified by the analyst. It confirms that access to improved inputs was perceived to be the major driver of rising income in both clusters. But it also reveals differences: not surprisingly there was less discussion of drivers of rising income in Oddo Leka and more emphasis on diversification into non-farm livelihoods. In contrast in Gese Bilbilo there was more discussion of crop diversification, livestock activities, and the importance of malt barley prices.

Table 3.5 extends the analysis to negative drivers of change in Oddo Leka across all domains. It reveals that the effect of the malt barley crop failure on cash income was compounded by the simultaneous failure of the *enset* (false banana) crop due to bunchy top disease. The effect of lower income was in turn exacerbated by price inflation, particularly the effect of increased fertilizer prices. These shocks in turn increased indebtedness, forcing some households to sell assets, and others to take children out of school in order to work in the fields. Only a couple of respondents mentioned rising inequality as a problem, whereas in Gese Bilbilo this came up as an issue in three of the four focus group discussions (for younger women, younger men, and older women).

Recommendations from the report were developed by Diageo in response to concern over these negative findings. For example, seed quality assurance was

Table 3.5 Frequency count of negative drivers of change in Oddo Leka

	Food production	*Cash income*	*Cash spending*	*Food consumption*	*Time spent on agriculture*	*Time children spent studying*	*Intra-household relationships*	*Village relationships*	*Assets*	*Wellbeing*
Inflation		1	18 (4)							1
Malt barley crop failed	5 (3)	14 (2)								
Different malt barley seed supplied	(1)	7 (3)								
Crop disease decimated *enset* (false banana) harvest	8 (2)	(1)		1 (1)				1		1
Increased debt and inability to cope with shocks	4 (2)	8 (3)					1 (2)		3 (2)	2
Increased time-wasting and drinking due to despondency					3 (3)			1 (2)		
Children spending more time on agriculture		1			3	4 (2)				
Higher expenditure on fertilizer	(2)		8 (2)							
Reduction in productivity negatively affects livelihood	1 (2)	1		1 (2)	(1)		1 (1)			2 (1)
Increased differences between HHs		1								1

Note: First number refers to interviews, number in parenthesis to focus groups

strengthened. Some staff argued that farmers shared part of the responsibility for lower than expected yields, but also recognized that other factors were out of their hands. Diageo partly offset the farmers' losses by being more generous with payments of quality premiums. It also piloted a crop insurance scheme, although it was not clear how much of a difference this would have made to the farmers in Oddo Leka, or indeed how feasible it would have been for them to obtain independent loss estimates against which to make claims.

Negative as well as positive findings were fully and openly presented in the public version of the evaluation report, made available for downloading on Diageo's website (Diageo, 2017). Under the section headed 'Improving Our Programme' this discusses how to improve seed supply, including 'working with government to improve the regulation and quality of seed production and supply throughout the country' (p. 9). It also asserts that they shared the problem with farmers by agreeing to buy crops even though these didn't meet their quality standards. Second, the report highlighted the importance of risk management by asserting that 'as the cost of not having it can be calamitous, crop insurance is now a non-negotiable part of the S4G support package' (p. 9).

Sample selection

A problem for all qualitative evaluation methods, the QuIP included, is how to generalize credibly from a relatively small number of cases. How best to do this (as discussed in Chapter 1 section 'Case selection') depends in part on the primary purpose of the study. If this is confirmatory, then the challenge is to use available resources as effectively as possible to augment users' prior beliefs – as formalized by the Bayes theorem. If it is exploratory, then the challenge is to get as close as possible to saturation, i.e. to be as confident as possible that there are no major gaps in what is revealed about the most important drivers of change across a population. Both purposes were in evidence for this study, with the commissioner seeking as much reassurance as possible that they were not left unaware of major problems or risks down the supply chain, and hence could not be accused of negligence in assessing them.

Given these goals, a fixed budgetary constraint, and a fixed time frame for the study, the sampling problem boiled down to how to capture within the sample as much diversity of experience as possible among the 6,000 + farmers known to have sold to Diageo in the previous season. One theoretical option would have been to select 48 respondents purely at random, possibly also using probability proportional to population stratification of the sample across the five cooperative unions to ensure representative geographical coverage. This was rejected mostly on the logistical grounds that finding and visiting everyone in such a sample would have used up far more time (and fuel). In addition, it would have failed to make use of all the data supplied. This revealed both significant and weakly correlated variation on two scales. First, the mean value of finance provided per farmer varied across the 39 primary

cooperatives by a factor of more than four (from 0.18 to 0.79 hectares) and in a region dominated by barley production this could be taken as a weak possible proxy for overall farm size. Second, barley procurement per hectare funded also varied widely (from a mean per cooperative of 0.09 to 2.42 tonnes per hectare). This could be explained by variation in actual yields – hence an indicator of farmers' technical performance. However, another explanation could have been variation in the extent of side-selling of malt barley to other traders, this being a reflection of farmers' market power. Either way it seemed sensible to select a sample that purposefully captured variation along both scales. This explains the decision to cluster interviews to one cooperative where procurement was 'extensive' and one where it was 'intensive' (refer back to Table 3.1), with an additional criterion being to select them from different cooperative unions.

With hindsight, it can be argued that clustering the sample (with 24 interviews in just two cooperatives) limited the potential to capture variation in impact across the full sample. For example, covering four cooperatives (with 12 interviews in each) could have been more effective in this respect. One reason for not doing so was that it would have entailed reducing the number of focus groups in each cluster from four to two, and hence the scope for differentiating them by age and gender. In addition, analysis of variation between farmers in both input supply and procurement per hectare revealed a high level of variability both *within* the cooperatives as well as *between* them. To address this, the list of farmers within the selected primary cooperatives was also ranked according to both variables, and random sampling was stratified to ensure that the final sample more fully captured variation in both.

It was methodologically positive that the strategy adopted of purposefully selecting contrasting cooperatives with respect to prior data did throw up two sharply contrasting stories. It is interesting to note, however, that the more negative account came from Oddo Leka, despite the fact that average procurement of malt barley per hectare was more than double that in Gese Bilbilo. Without repeating the study and comparing results there is no way of knowing if a different sampling strategy would have yielded different findings.

Uncertainty over the effect of sampling on evaluation findings does not end here. First, there is the issue of how frequently such 'deep dives' should be repeated given annual variations in agricultural outcomes due to weather and market conditions, as well as longer-term trends in the supply chain. Second, Ethiopia is only one country in which Diageo operates, and Diageo is only one brewing company. A year after this study was completed Diageo commissioned a second QuIP study, this time of sorghum and cassava procurement for its brewery near Kampala in Uganda. In this instance they were sourcing through traders or aggregators, and so visibility of the farmers was limited. Data available was primarily at this trader or aggregator level, which is not uncommon in commodity crops. This created a less reliable sampling frame as the farmers involved were usually unaware of the final destination of their harvest.

Political economy and public policy context

There are few debates in development studies that go back as far as that over the impact of agricultural commercialization. Advocates of green revolution strategies assert the possibility of achieving almost 'magical' transformation in productivity, generating the surplus of affordable food needed to sustain urbanization. Critics of such modernization strategies emphasize the 'tragic' and polarizing effects of such change: profitable commercialization for a few, depriving many of access to land and other resources, and forcing them to scrape together insecure livelihoods in the informal economy and as wage labourers. A third view, labelled 'romantic' or 'populist' by its critics, suggests that when combined with appropriate support services, commercialization and sustainable intensification of land can deliver greater economic security to large numbers of small-scale farmers, shifting them from uncertain subsistence farming into more profitable commodity value chains, supplying not only export markets (for tea, coffee, tobacco, etc.) but also rapidly growing domestic and urban demand for foodstuffs.

Of course, the agrarian world is large and diverse enough to accommodate magic, tragic, *and* romantic histories; and it would require far more space to provide a balanced review, even for contemporary Ethiopia.[4] But we can note elements of each in the Diageo study. The main story in Gese Bilbilo (also strong in Oddo Leka) was of rising farm income driven by investment in improved seed, fertilizer, and other inputs to produce more barley as a cash crop. These benefits were not being grabbed by large-scale farmers only, and were taking place alongside diversification into other crops and livelihood activities. New malt barley varieties were being grown alongside traditional varieties, with both being consumed as food as well as sold. Rising incomes were having a positive influence on family nutrition and school attendance.

However, even within the restricted time frame of the study some longer-term adverse effects could be discerned. In Gese Bilbilo several farmers did hint at how tightening land rental markets were accentuating wealth and income disparities between farmers. And from Oddo Leka we were reminded that investing heavily in a cash crop because it was successful one year can have devastating effects when the crop fails in the next.[5] Farmers have been managing such risks for generations, but the Oddo Leka case does raise the political question of how farmers and buyers share such downside risk, and how it could and should be shared. The priority for Diageo staff on the ground is to meet imminent supply targets for the brewery, even if the programme itself is intended to support long-term rural development goals too. This sets up a tension for more senior staff between managing short-term goals and building longer-term partnerships through risk sharing to enhance the security of local supply chains.

These issues resonate with many other food vs. cash crop debates (e.g. see Mager et al., 2017), but our concern here is not just with barley but with alcohol, which Babor et al. (2010) remind us is 'no ordinary commodity'.

For advocates of economic transformation in what Collier (2013) refers to as 'frontier markets', commercial brewing can play an important pioneering role in industrial development, as it did during the first Industrial Revolution. And this kind of thinking does indeed resonate strongly with the Government of Ethiopia's strategy of 'agriculture led industrialization', crafted during the presidency of Meles Zenawi. Large-scale industrial brewing can establish linkages that 'crowd-in' private investment; it can be an important source of public revenue, and drive improvement in product quality. Although Diageo points out its commitment to promoting responsible consumption, other commentators note that international brewing companies are also powerful catalysts for promoting alcohol as a lifestyle choice in the face of strong evidence linking 'harmful use' of alcohol with the risk of dying from non-communicable diseases (WHO, 2018). Without entering into debate over what constitutes advertising, drinking and influencing public health regulation 'responsibly' (and what constitutes a harmful level of use), it is important to locate whatever public relations benefit Diageo and other beverage companies derive from promoting rural livelihoods within this wider public policy context. The net welfare benefits to local sourcing along alcohol value chains ultimately constitute only one element of a wider food and public policy issue.

Conclusions

Our self-reflections on this first example of a QuIP study were positive. The study confirmed positive income effects of Diageo's malt barley procurement in one area, but highlighted unanticipated problems in another. It provided reassurance to the commissioner that induced uptake of child labour was not an issue. It delivered findings on time and on budget, and prompted a repeat study in another country. Methodologically, the study piggy-backed effectively on monitoring data to capture significant diversity in respondents' experiences, although the possibility that farmers' experiences in other localities may have been different in other ways cannot be eliminated. And the commissioner did publish the findings on its website for all to see, even if the scope of the study was narrowly framed relative to the wider public policy issues presented by the commercialization of alcoholic beverages in Ethiopia and indeed elsewhere.

Notes

1. In 2011/12 nearly half of the 67,500 tonnes required by the brewery industry in Ethiopia was imported (Copestake, 2013).
2. Diageo was not alone. The market leader, BGI Ethiopia, is owned by French drinks company Group Castel. Heineken purchased two breweries from the government in 2011 (Bedele and Harar, for $85 m and $78 m, respectively), and in 2013 the local Habesha brewery constructed a new plant

with support from the Lehui Group, China's leading beer manufacturer (Copestake, 2013).
3. Diageo (2017: 6) acknowledges other negative outcomes '… such as more expensive fertilizer and higher land rents due to increased demand' but attributes these directly to the success of the barley crop.
4. For a comprehensive review of rural transformations taking place in Ethiopia see Pankhurst (2017).
5. The QuIP study for Diageo in Uganda produced more evidence of this kind, but linked to boom-bust or 'cobweb cycles' in grain markets.

References

Copestake, J., Goshu, T. and Remnant, F. (2015) *QuIP Report for Self Help Africa Malt Barley Project in Oromia, Ethiopia* [pdf] <http://bathsdr.org/wp-content/uploads/2016/04/SHA-Assela-QuIP-Report-2015.pdf> [accessed 2 January 2018].

Babor, T., Caetano, R., Casswell, S., Edwards, G., Giesbrecht, N., Graham, K., Grube, J., Grunewald, P., Hill, L., Holder, H., Homeal, R., Osterberg, E., Rehm, J., Room, R. and Rossow, I. (2010) *Alcohol: No Ordinary Commodity: Research and Public Policy*, 2nd edn, Society for the Study of Addiction and Pan American Health Organisation, Oxford: Oxford University Press.

Bath Social and Development Research Ltd (2016) *QuIP Report on Diageo's Sourcing for Growth (S4G) Programme: Oromia District, Ethiopia*, Bath, UK: BSDR Ltd (unpublished report, August 2016).

Collier, P. (2013) *Aid as a Catalyst for Pioneer Investment*, WIDER working paper no. 2013/004, Helsinki: World Institute for Development Economics Research.

Copestake, J. (2013) *Behind the Aid Brand: Distinguishing between Development Finance and Assistance*, Working Paper 24, Bath, UK: Centre for Development Studies, University of Bath.

Diageo (2017) *The Impact of Sourcing for Growth (S4G) in Ethiopia* [pdf], London: Diageo plc <http://bathsdr.org/wp-content/uploads/2017/07/Diageo_Sourcing_for_Growth.pdf> [accessed 18 October 2018].

Gill, P. (2010) *Famine and Foreigners: Ethiopia since Live Aid*, Oxford, UK: Oxford University Press.

Jernigan, D. (2015) 'The concentration of the global alcohol industry and its penetration in the African region', *Addiction* 110(4): 551–60 <http://dx.doi.org/10.1111/add.12468>.

Mager, F., Walsh, M. and Remnant, F. (2017) *Cash Cropping and Care: How Cash Crop Development is Changing Gender Relations and Unpaid Care Work in Oromia, Ethiopia*, Oxford, UK: Oxfam GB <http://dx.doi.org/10.21201/2017.1329>.

Pankhurst, A. (ed.) (2017) *Change and Transformation in Twenty Rural Communities in Ethiopia: Selected Aspects and Implications for Policy*, Ethiopia Wide team Addis Ababa: Pankhurst Development Research and Consulting Plc.

World Health Organization (2018) *Noncommunicable Diseases: Key Facts* [online], Geneva: WHO <http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> [accessed 31 May 2018].

About the authors

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Gabby Davis, PhD, is a research fellow in the digital skills observatory of the Institute of Coding at the University of Bath. Previously, she was a Senior Project Manager at Bath Social and Development Research, where she led on training and worked on ten QuIP studies in seven countries. Her main expertise is in qualitative research, wellbeing research, socio-ecological relations and post-conflict settings.

Tefera Goshu, MA Social Anthropology, is on the teaching staff at Ambo University in the Department of Sociology, and is a PhD candidate at Addis Ababa University in the Department of Social Anthropology. His main expertise and research interests include rural development, food security, gender, social inequality, adolescents and youth, and socio-cultural change. He has worked on five QuIP studies in the capacity of lead field researcher.

CHAPTER 4

Improving working conditions in the Mexican garment industry

Marlies Morsink, James Copestake and Max Niño-Zarazúa, with Savi Mull

The Qualitative Impact Protocol (QuIP) contributed to C&A Foundation's final evaluation of a two-year project designed to promote wellbeing and improve the productivity of garment factory workers in four provinces of central Mexico. This was the Yo Quiero Yo Puedo cuidarme y mejorar mi productividad *(YQYP) project implemented by the Mexican Institute for Family and Population Research (IMIFAP). The QuIP study provided C&A Foundation with rich insights into how intended beneficiaries perceived their lives to have changed, and reported positive impacts of the project on job satisfaction and productivity, gender equality, and interpersonal relationships at work and in the home. Methodologically, the chapter reflects particularly on the challenge of conducting interviews blindfolded, and on how the specification of theories of change affects interpretation of findings. This is one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.*

Keywords: impact evaluation, causal attribution, Mexico, garment industry, factory workers

Introduction: commissioner and project background

C&A is a global fashion clothing chain and family owned company, named after Clemens and August Brenninkmeijer who founded it in 1841. It has more than 2 million customers, 60,000 employees, and a supply chain that encompasses 788 suppliers employing more than a million people, as of 2018.[1] Sustainability and social responsibility feature prominently in its mission. 'Our customers shouldn't have to choose between looking good, feeling good and doing good. They deserve great fashion that's also good for the people who make their clothes, and good for the environment. No decision or trade-off should be necessary and at no extra cost to the customer.'

Public pressure to improve the transparency of big brand clothing supply chains has increased sharply since the Rana Plaza building collapse in Bangladesh in April 2013, and C&A has positioned itself as a leader in the sector by agreeing to publish details of all its Tier 1 and Tier 2 suppliers.[2]

http://dx.doi.org/10.3362/9781780447469.004

Transparency, of course, is only one step along the path towards improving working conditions throughout the supply chain. Complementing private charitable operations of the family, the C&A Foundation provides another channel for taking proactive philanthropic steps. The Foundation was established in 2014 with a remit to rationalize C&A's philanthropic activities around a unified global vision to transform the fashion industry.[3] Its website boldly acknowledges the challenge:

> Right now, our industry isn't working for the good of the 150 million people who make our clothes. Our preference for fast, trendy and affordable fashion leads to severe forms of just-in-time production at the lowest cost possible … We believe this can change. We believe fashion has the power to improve the lives of the men and women behind our clothes. We believe fashion can be a force for good. Our mission is to transform the industry to make that happen.

With the overarching goal of 'making fashion a force for good', C&A Foundation supports activities aimed at five main areas: improving working conditions in garment factories, promoting sustainable cotton production, eliminating forced and child labour, fostering a transition to circular fashion, and strengthening communities through employee engagement, humanitarian aid, and disaster risk reduction.

In focusing on the first of these challenges – improved working conditions – C&A Foundation funded a project to promote the wellbeing of factory workers in four provinces of central Mexico. Called *Yo Quiero Yo Puedo cuidarme y mejorar mi productividad* ('I want to and I can take care of myself and improve my productivity', and referred to for short as YQYP), it was implemented by the Mexican Institute for Family and Population Research (IMIFAP) between March 2014 and August 2016, and funded by Fundación C&A.[4] Its overall objective was 'to promote the integral wellbeing of the workers in the Mexican textile industry in order to improve their productivity and support the guidelines of codes of conduct of international companies attached to the principles of the 2020 Global Pact' (BSDR, 2017: 47).

IMIFAP specializes in applying a psycho-social approach to human development that seeks enhancement of life skills (e.g. decision-making, control of stress, empathy), acquisition of relevant knowledge (e.g. health, self-care, safety, wellbeing, positive working environment, and labour rights), and reduction of psychological barriers (e.g. fear, shame, guilt). Activities are designed to trigger changes in attitude and behaviour that enable participants to gain more control over their lives, empowering them to improve individual, family, and community health and self-care, job satisfaction and productivity, gender equality, and interpersonal relationships. 'If workers are enabled with the knowledge to develop their emotional and cognitive social skills — whether through a focus on health, education, citizenship or work — they not only experience life benefits, but they also become active contributors to their work environments' (Pick, 2015, see also Pick and Sirkin, 2010). The idea behind developing a YQYP project for the

C&A Foundation was to 'apply the methodology to achieve sustainable changes within factories, training middle managers and employees to promote the overall well-being of staff and increase productivity'.[5]

Table 4.1 elaborates on the YQYP project activities intended to trigger these changes. These started with sensitization workshops aiming to secure the willing collaboration of senior management and owners of selected *maquilas* (apparel factories). Those who agreed to participate then nominated *maquila* supervisors (middle level managers) to participate in formative workshops. These were delivered by IMIFAP trainers through weekly sessions over 10 weeks (lasting a total of 40 hours) based on seven units about health and productivity. In the second year, continuing factories followed four new units about gender equality and violence, industry hygiene and safety, and skills promotion, lasting for a further 56 hours. Once trained, supervisors were tasked with

Table 4.1 Components of the C&A Foundation funded YQYP programme

Activity	*Target audience*	*Objective*	*Led by*	*First year activities*[6]	*Second year activities*
Sensitization conferences					
Two hour interactive presentation	Sector leaders, *maquila* owners and senior managers	To gain interest and to build high level support for the programme	IMIFAP project team	Two conferences involving 25 *maquilas*. 14 joined the project	Three conferences involving 22 *maquilas*. 6 follow-on and 9 new joiners
Formative workshops					
Experiential workshops using *ludic* and participatory methods	*Maquila* supervisors	To develop life skills, personal agency and to empower them as change agents	IMIFAP project team	10 × 4-hour weekly sessions on health and productivity	Same for new joiners, or 14 new sessions for follow-on *maquilas*
Replica workshops					
Fifteen minute sessions using experiential and *ludic* activities	Operators	To change attitudes and behaviour, promote personal agency and self-empowerment	Super-visors	Daily sessions over 18 weeks (6 learning units)	Daily sessions over 18 weeks (6 learning units)
Accompaniment visits					
Participant observation and feedback on replica sessions	Supervisors and operators	To enhance the quality of replica workshops	IMIFAP project team	79 sessions	245 sessions

Source: BSDR (2017: Table 4.1).

running daily 15-minute 'replica' workshops with their operators (i.e. factory floor workers) over 18 weeks; during this stage, supervisors received 'accompaniment' visits and supervision from IMIFAP to monitor and improve the quality of these workshops. The training at both levels employed participatory and *ludic* (game based) methods, based on handbooks prepared by IMIFAP.

This chapter reports on an external, final evaluation in 2016 of the two year project. The next section describes the QuIP component of this evaluation, followed by some of its findings. We then reflect on interpretation of the findings, principally from the perspective of the international commissioner of the study. This chapter was first drafted by Morsink and Copestake, drawing on the final evaluation study (authored by Niño-Zarazúa) and key informant interviews with Niño-Zarazúa and Mull conducted by Morsink. It was then revised in the light of their comments and suggestions.

The 2016 external evaluation

Design and implementation of the YQYP project was led by IMIFAP, with C&A Foundation facilitating contacts with industry leaders, but otherwise maintaining an oversight role. By the middle of 2016, the initial two year project grant was coming to an end, and the evaluation was commissioned to assess outcomes and learn from the experience. The Request for Proposal (RFP) covered the five criteria laid out in the OECD-DAC framework for evaluation: relevance, effectiveness, efficiency, impact, and sustainability.[7] In line with the principle of tailoring evaluation to fit the characteristics and context of the project being assessed, C&A Foundation also consulted closely with IMIFAP and relevant foundation staff about the appropriate methodology for the evaluation. More specifically, consultations between C&A Foundation and Susan Pick, founder and president of IMIFAP, resulted in agreement that the evaluation should have quantitative and qualitative components. The process evaluation component of the evaluation would also be able to utilize existing monitoring data previously collected by the project team, including records of supervisors' attendance at formative workshops, and 344 written reports of accompaniment visits.

The project implementation team had laid the foundation for a quantitative psychometric assessment of the project by collecting closed questionnaire data from a sample of supervisors and operators designed to assess their knowledge, skills, and attitudes across 12 domains. The plan was to collect this data before and after training in all participating *maquilas* in both years of the project, as well as for a comparison group of employees in five non-participating *maquilas*.[8] Analysis of the data, IMIFAP hoped, would yield estimates of psychological changes associated with project participation. At the same time, they recognized the potential for utilizing qualitative methods to gain complementary insight into causal mechanisms behind the observed changes. Pick had previously come across the QuIP in 2016 and was hopeful it might be an appropriate way to explore the drivers behind the intangible or difficult-to-measure planned outcomes of YQYP.

The bid submitted for the evaluation by BSDR (and selected from a pool of more than a dozen bids) adopted a two-pronged approach to the attribution question: it employed the QuIP to assess causal drivers of impact, and sub-contracted a separate team to conduct quantitative analysis of the psychometric data. This division of labour and acute time pressure to initiate the study meant that integration of the two approaches was very limited. Partly for this reason, this chapter focuses on the QuIP component of the evaluation in isolation.

Data collection for the QuIP sample started in September 2016, and a draft report was produced by February 2017. The population frame for the QuIP study comprised data from all *maquilas* which participated in the second year of the project: six of these participated in YQYP activities over two years (Y1&2), while nine joined only in the second year (Y2 only). One objective of the study was to explore differences between the two groups. In addition to focusing only on Y2 participants, sample selection was also deliberately skewed towards those *maquilas* that met a minimum threshold level of project implementation as measured by participation in formative workshops, recorded accompaniment visits and completed evaluation questionnaires.[9] During the course of the QuIP study the sample size was increased, in order to ensure a larger number of women was interviewed (BSDR, 2017: 19). Table 4.2 indicates that the sample of 33 interviews across six *maquilas* included 16 women, and was divided nearly equally between supervisors and operators,

Table 4.2 YQYP 2017 QuIP sample size and composition

	Y1&2				*Y2 only*				
	IC	*PS*	*TA*	*Sum*	*FM*	*LP*	*IT*	*Sum*	*Total*
Supervisors									
Total registered	30	12	8		7	3	22		
Total interviewed	3	4	1	*8*	2	3	4	*9*	*17*
(of which women)	*(3)*	*(0)*	*(1)*	*(4)*	*(1)*	*(0)*	*(3)*	*(4)*	*(8)*
Participants in focus groups		5	2	7			4	*4*	*11*
(of which women)		*(1)*	*(1)*	*(2)*			*(0)*	*(0)*	*(2)*
Operators									
Total registered	145	22	11		26	45	Not known		
Total interviews	3	4	1	8	3	0	5	8	16
(of which women)	*(3)*	*(0)*	*(1)*	*(4)*	*(0)*	*(0)*	*(4)*	*(4)*	*(8)*
Participants in focus groups		7					6		*13*
(of which women)		*(2)*					*(4)*		*(6)*
Total participants									33
(of which women)									*(16)*

Source: BSDR (2017: Tables 2.4 and 2.5).
Notes: IC = Industrias COS (Mexico City); PS = Poin S.A. de C.V. (Puebla); TA = Telas Asturcón (Puebla); FM = Fábrica María (Puebla); LP = La Poblana (Puebla); IT = Inova Textiles (Mexico State).

as well as between those from *maquilas* that had participated in the project for one and for two years. The table also provides details of the four focus group discussions, involving another 11 supervisors and 13 operators.

During sample selection for the QuIP study (which doubled as a component of the quantitative process evaluation), gaps and anomalies in the records of participants, particularly operators, came to light. Partly for reasons of record-keeping, the task of contacting participants for QuIP interviews was far from straightforward. This problem was exacerbated by high levels of staff turnover at the *maquilas*, and initial suspicion over the purpose of the study. The QuIP research team addressed these problems as far as they could by working through the *maquilas'* human resources departments, but some compromises proved necessary, including undertaking individual interviews at the workplace and not blindfolding focus groups. The section entitled 'Interpreting the findings' explores this further.

The QuIP had not previously been used in a factory setting, nor with a focus on employees as intended beneficiaries. This meant it was necessary to formulate a new set of outcome domains and questions to guide interviews and focus groups. To do this, the lead evaluator first helped the IMIFAP team to reconstruct the project's theory of change more formally (see Figure 4.1). This clarified the two purposes of the qualitative component of the evaluation: first, to learn what respondents identified by way of changes in their physical

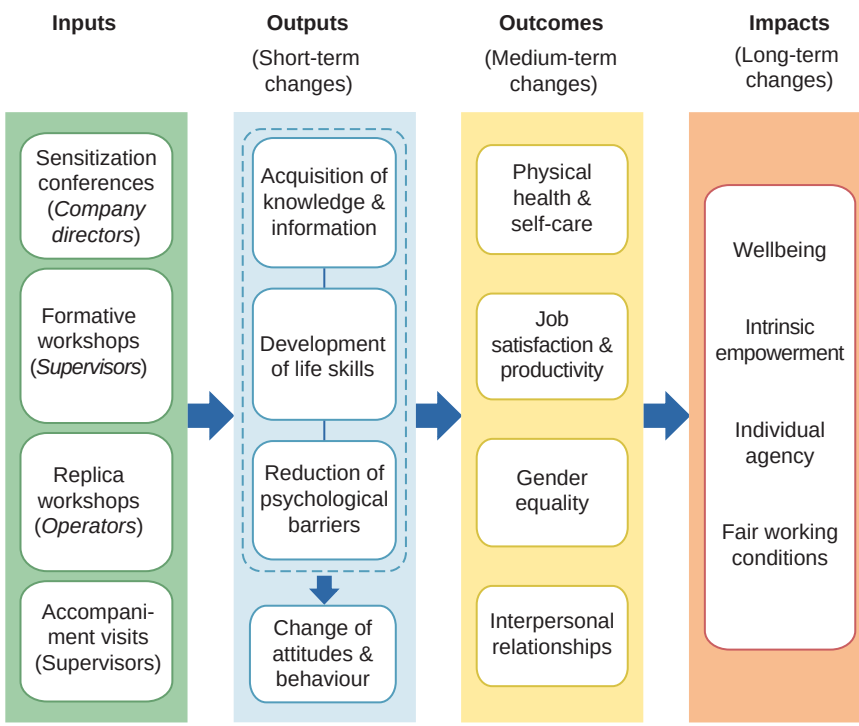


Figure 4.1 Theory of change for the YQYP project

health and self-care, job satisfaction and productivity, gender equality, and interpersonal relationships; and second, to find out how far they linked these to knowledge and information obtained through the YQYP life skills workshops over one or two years, on topics such as personal development, self-care, working environment, human and labour rights, labour obligations, safety in the workplace, equality, prevention of violence at work, and personal economy and finance. The study thereby aimed not only to confirm the theory of change set out in Figure 4.1, but to flesh it out with a more detailed understanding of the project's causal mechanisms and time lags.

Selected findings

Many of the 33 blindfolded interviewees, as well as participants in all four unblindfolded focus groups, did link YQYP explicitly to positive changes in their lives, as well as making statements that implicitly supported the project's theory of change (see Table 4.3).

The number of negative changes reported is much smaller overall, and most were attributed to drivers not related to the project.[10] The most widely perceived positive project impacts were spread across three domains: job satisfaction and productivity, gender equality, and relationships. In contrast, the project was not identified as a source of improved economic security: improvements in the economic realm were linked to other external factors including diversification of income outside work, contributions from other household members, pay rises, and doing overtime work.

Delving deeper into the data, Table 4.4 highlights the five drivers of change reported most frequently in the individual interviews and the focus group discussions.

This highlights the relative prominence of positive statements about the effect of the YQYP programme on working relationships – a finding that was

Table 4.3 Attribution of positive and negative change to the YQYP project

Outcome domain	*Positive changes*			*Negative changes*		
	Project explicit	*Project implicit*	*Other*	*Project explicit*	*Project implicit*	*Other*
Health and self-care	9 (4)	8 (3)	22 (1)	0 (0)	6 (2)	18 (2)
Job satisfaction and productivity	21 (4)	15 (2)	25 (1)	0 (1)	2 (1)	2 (2)
Gender equality	18 (4)	15 (1)	12 (0)	0 (1)	3 (1)	0 (1)
Economic security	1 (0)	1 (0)	28 (2)	0 (1)	1 (2)	3 (1)
Relationships	22 (4)	15 (2)	10 (0)	0 (1)	3 (1)	4 (0)
Overall wellbeing	3 (3)	7 (1)	13 (1)	0 (0)	0 (0)	0 (1)

Source: BSDR (2017).
Note: The first number indicates the number of respondents out of 33 interviewed (and four focus groups in parenthesis) who made at least one causal statement in the impact domain indicated.

Table 4.4 Top five coded drivers of positive change

	Y1&2		*Y2 Only*		*Total*
	OP	*SP*	*OP*	*SP*	
Individual interviews					
YQYP training in effective communication and working relationships	13	16	10	14	53
YQYP training in tolerance, values, equality, and working responsibilities	1	8	10	9	28
YQYP training in productivity and motivation to achieve better results	4	9	4	6	23
YQYP training in teamwork	9	5	2	5	21
Pay rise/better income/remuneration	3	8	2	8	21
(Sample size)	*(8)*	*(8)*	*(8)*	*(9)*	*(33)*
Focus groups					
YQYP training in effective communication and working relationships	5	5	3	4	17
Increased sense of how to work/live better and happier	4	3	1	3	11
YQYP training in balance between work and personal life	4	1	4	1	10
YQYP training in tolerance, values, equality, and working responsibilities	1	3	1	4	9
YQYP training in teamwork	2	2	3	1	8
(Sample size)	*(1)*	*(1)*	*(1)*	*(1)*	*(4)*

Source: BSDR (2017: Table 4.4)
Notes: 'OP' refers to operators, and 'SP' to supervisors. Numbers indicate the frequency with which the code was ascribed to text across all respondents, including multiple use across the six outcome domains by the same respondent.

robust across interviews and focus groups, and across operators and their supervisors. In the corresponding analysis for negative drivers of change (not shown) the category coded 'pressure/workloads/stress and conflicts' was most widely cited across all these categories. This evidence strongly suggests that at least in some contexts the project was successfully addressing a problem of poor working relationships, if not doing so fully.

To illustrate the narratives of change, from drivers to outcomes, that back up frequency counts in Tables 4.3 and 4.4, Box 4.1 contains a set of quotations which demonstrate 'project explicit attribution' to YQYP in the domain of 'relationships' (refer to the frequency count of 22 in Table 4.3).

Interpreting the findings

The commissioner's perspective

Savi Mull is an evaluation specialist in C&A Foundation's Effective Philanthropy Team and regards her role as 'ensuring that the Foundation is able to measure results in a robust way across all the thematic areas globally, and to learn what

Box 4.1 Selected causal claims linking YQYP to improved relationships

'There has been positive change in how we support each other, in how we resolve conflicts, in helping as a team, in greater willingness to work. The workshops helped us work as a team, we did some activities in team dynamics which were very useful to understand this …' [female supervisor].

'… we've changed, we respect the five-minute space that people need when they are upset or angry, if we see them like that we return later. I tried to have a loving, respectful and trustful relationship with my daughters. With my husband I tried to have respect, trust and tolerance with one another, although sometimes it's difficult … it was the workshop, it opened my eyes, we all are like that, we help each other and we keep reminding each other what we learnt in the workshops. We know each other better, we learnt how to control our emotions, respect each other, be empathetic, and have better communication. So, all this works well, you just need to put it in practice at work and with your family …' [female supervisor].

'… it's changed for the good. I've improved my relationship and ability to talk with my partner and my daughters about what's going on or what we're going to do, whether everyone goes or just one, we talk about everything now. This change is because I reflected on what to do after my mum died, and because of the workshops we received. The workshops helped me to relate with my family better. I understood well that we need to leave our work problems at work, they shouldn't reach our homes, and also our family problems shouldn't be brought to work …' [male operator].

'… we learnt not to be violent or use bad manners and expressions with our work colleagues such as shouting or swearing. We have to be respectful when we talk; now our work colleagues have a better relationship. We now resolve our conflicts talking to each other in order to avoid fights …' [male supervisor].

'When there are conflicts, that we normally have, we always start talking, we always have a dialogue. We saw this in the workshops: to talk always and in the right moment … because if we try to talk when we are stressed, we won't progress at all, on the contrary, we will make things more complicated. So, this is what we learnt in the workshops …' [male supervisor].

'When there is a conflict of interest between two people we try to mediate based on who is wrong and we demonstrate them why. We don't say 'you're wrong' and that's it, we say them why they're wrong. Before we didn't do that, now it's different' [male supervisor].

'Yes I have changed, with the new ideas from the engineers [Directors] as well as the talks and training sessions we have had, we now work better as a team. I feel good with all these changes because the working environment has changed, our productivity has increased and working as a team has improved' [male operator].

'I feel very well, very happy. I enjoy my job, I feel that working in teams is good, the more communication we have as a team the easier we achieve our objectives, right? Before we didn't have much communication, so we didn't work as a team, we weren't well organized and there was a bit of a mess. We had returns and we were delayed in our deliveries, so we had to work extra time and all this was because of a lack of communication and teamwork. Since we have the training courses we have improved and we're still improving' [female operator].

works and what doesn't.' Having joined C&A Foundation in April 2016 she was able to provide a first-hand account of how the Foundation interpreted and responded to the findings of the BSDR-led evaluation. She recapped the primary reason for commissioning an external evaluation of YQYP in the first place. 'This was to see what results were coming out of the project after two years, on the basis of which funding decisions are also made. Even though C&A Foundation's investment in Mexico is much smaller than in other parts

of the world, it is still wanting to make sure that learning and results are emerging out of the investment being made.'

Turning to the findings, Mull reported that at one level C&A Foundation responded very positively, judging that the QuIP study provided a wealth of insight into ways intended beneficiaries perceived their lives to have changed, and the degree to which they attributed these changes both explicitly and implicitly to the YQYP project.

> What we did learn is that there is clear evidence that the initiative contributed to different aspects in the lives of workers across both years. And for some domains it was larger for supervisors, and in other domains for workers. One of the positive nuances we saw was on the gender equality piece. That came out very well. We picked out a quote from a female supervisor on how they resolved conflict and supported each other positively, there was a sense of team-building. And there were female workers who had felt they were being harassed, and were able to speak up because of the training. The initiative gave them 'voice'. So the benefit that came out, of increased agency or empowerment, is what we got out of the report; the QuIP gave us that.

Two months after the completion of the evaluation, C&A Foundation published the report on its website along with a brief 'Lessons Note' which was also disseminated using Twitter and LinkedIn (see BSDR, 2017). Distilled from the evaluation, it highlighted the positive outcomes of the YQYP project that were revealed through the QuIP study, emphasizing reported improvements in how factory employees felt about various aspects of their lives, including on-the-job satisfaction and productivity, relationships at work and in the household, gender equality, and health and self-care. It also mentioned a greater sense of teamwork among workers and supervisors in both years, and reported that more than half of the supervisors and workers viewed YQYP as having given them a better sense of personal development and job satisfaction.

External evaluations of larger grant-funded projects are presented to the C&A Foundation Board, which guides its grant-making, supervises its processes, and awards the funds that support work by its partners. It is beyond the remit of this chapter to document how this particular evaluation influenced subsequent operational decisions, this being both an internal matter for the C&A Foundation, and also one that is methodologically complex in itself given the many different sources of information that contribute to particular decisions. Instead, the remainder of this section explores two methodological problems relevant to interpreting the QuIP findings.

Sampling problems and prospects for scaling up

According to CANAIVE (the Mexican National Chamber of the Apparel Industry) there were 8,613 registered apparel companies in Mexico, employing 450,000 people (BSDR, 2017: 10). By comparison the YQYP Project aimed to

collaborate with 23 companies and to benefit 2,500 workers, and was limited geographically to central Mexico. Its relatively small scale does not diminish its absolute achievements, but it does beg the question of how typical the *maquilas* in which it conducted trainings are, relative to the whole sector, and what the implications would be for scaling it up. Even within the smaller set of *maquilas* targeted, the project experienced sharp variation in the level of collaboration it received, with most activity being concentrated in half a dozen more cooperative firms.

Some factory owners were convinced that the YQYP project would contribute positively to worker wellbeing and productivity, and two of these even requested that the project be extended to other sites, including one in Honduras. However, IMIFAP encountered others who expressed scepticism; even when owners were supportive, individual factory managers could be unwilling, complaining that the training was a distraction that diverted time away from production. In an attempt to address such concerns, YQYP adapted the structure of the training, offering supervisors a condensed version of the programme in fewer but longer sessions over a shorter total time span. It is likely that any attempt to scale up would have required further modifications of this kind.[11]

In the process of conducting the evaluation it also became clear that the statistic of 2,500 intended beneficiaries was a rough estimate based on the number of workers assigned to each supervisor trained by YQYP staff. And it turned out to be impossible to establish a more accurate estimate because of incomplete records of supervisor-led sessions for workers. High turnover of employees raised questions about attendance rates, and may also have skewed the sample of those interviewed to those who were more established in their jobs, as well as more regular and enthusiastic about the training. Overall, this large 'attrition funnel' – with 57 QuIP respondents at its narrow end, to anything up to 450,000 workers at its wide end – leaves considerable room for doubt about how generalizable the positive study findings might be to implementation on a larger scale (White, 2014). This in turn opens up questions about the effectiveness of the YQYP approach to improving working conditions in the sector compared with other possible strategies.

Blindfolding and the scope of data collected

Blindfolding emerged as a source of contention during this QuIP study. Members of the IMIFAP implementation team feared that if the evaluation team approached workers directly, this would jeopardize the relationships they had struggled to establish with factory management. Consequently, IMIFAP handled communication with factory management, explained the nature and purpose of the study to them and sought their assistance in identifying supervisors and workers, under instruction not to divulge the ultimate purpose of the research to them. Meanwhile Mull, representing C&A Foundation, was wary of blindfolding the researchers on the grounds that it would limit the

scope for probing in the interviews, and hence the interviewer's ability to obtain relevant data. C&A Foundation also has the practice of sending a representative to sit in on data collection for evaluation studies, and in this context Mull asked for a QuIP interview to be observed by a local staff member from the Mexico office. The aim was to ensure data collection quality and to hear findings first-hand, so as to better understand and interpret the final report. The QuIP field team reported that the presence of an additional person in the room, though not identified as a representative from the C&A Foundation, inhibited the conversation with the respondent. This paradoxically served to reinforce the view of the C&A Foundation representative that blindfolded interviews would not reveal much about the project. This trial observed interview resulted in a C&A Foundation decision *not* to blindfold focus group discussions, and to increase the total number of interviews. Another outcome was that, in order to strengthen triangulation, respondents in the individual interviews would not be included in the focus group discussions.

It might appear that this particular methodological discussion between the QuIP team and C&A Foundation was ultimately resolved when the final QuIP report demonstrated that the blindfolded interviews did indeed elicit a wealth of data explicitly attributing positive outcomes to the YQYP project. In reflecting on the experience, Mull acknowledged both sides of the argument, and made a case for sequentially combining blindfolded interviewing with un-blindfolded debriefing, feedback, and/or sense-making meetings, so as to potentially get the best of both worlds:

> Not asking specific questions about the project means respondents are less likely to be biased in their feedback; however if the questions regarding change are too broad and open-ended there is a risk that the intervention won't get mentioned. It's really an art to do the initial deep dive in such a way that doesn't make the interviewee think that you're driving at something specific. One thing QuIP does fantastically is manage bias. But what could be good, is if the QuIP researcher first takes a deep dive into everything under the sun for agency creation. Then the researcher does a debrief with the principal investigator who knows about the specific project. And then the researcher goes back and does another deep dive to understand what happened with the project. Then you're doing real-time, adaptive M&E – even with the caveat of blindfolding.

In subsequent debriefing discussions about the study it became clear that C&A Foundation had a second and more fundamental concern about the YQYP evaluation, one not directly related to evaluation methodology. As already explained, design of the QuIP study was a collaborative process, involving both C&A Foundation staff and IMIFAP, and drawing on the theory of change for the YQYP project as set out in Figure 4.1. The focus of the evaluation, as a result, was on whether work-based group training had had a transformative effect on knowledge, social norms, attitudes, agency, and

ultimately the behaviour of participants.[12] As the QuIP emphasizes elicitation of the perceptions of intended beneficiaries, it seemed a suitable choice of approach to cast light on whether and how this mechanism was working in practice. The quantitative component of the evaluation (not discussed here) meanwhile focused on psychometric assessment of cognitive changes among employees.

However, the QuIP study did not go far enough in confirming or exploring the mechanisms of the theory of change of C&A Foundation's *new* global strategy, which had been articulated after the commencement of the YQYP project. Figure 4.2 reproduces from the C&A Foundation website the theory of change for improving working conditions (the strategic goal most closely aligned to this project).

There is some overlap between the two theories of change: for example, YQYP can be said to have been helping 'to amplify worker voice and participation in improving working conditions, especially for women'. But it is clear that YQYP was not designed explicitly with 'Working Conditions' theory of change in mind, and this undermined its potential fit with this particular signature programme of the Foundation. Mull commented:

> At the Foundation we make it very clear in our theory of change and results measurement that the outcomes do not stop at 'awareness and knowledge'. We always push the programme managers and our partners to ask, 'You've been trained… so what? What does your training translate into?' As you move along the 'KAP' model your Knowledge is built, and then your Attitude changes – you become more confident because you know something – but then what Practice is that translating into? Actions are what we're using as evidence that the knowledge is being built. For example, to measure 'leadership skills of women' you should count the number of women workers who have been trained in leadership, and who are now part of Worker Welfare Committees negotiating for collective bargaining agreements, etc.

It is of course possible that the YQYP project might have contributed to concrete and measurable actions to improve working conditions. However, such outcomes would be a bonus, going beyond the scope of what the project set out explicitly to do and to demonstrate. In subsequent discussions, Mull emphasized this point in three different ways:

> When we looked at the YQYP project, we realized it was creating worker agency, but wasn't necessarily improving working conditions in the factories. In the Foundation's ToC [theory of change] we are looking at improving working conditions – the number of women who are becoming leaders, for example; but the YQYP programme was not looking at that.
>
> There are different initiatives that we run, that have clear results towards action. It doesn't stop at training. Take for example another initiative where we said that at the end of the three year grant, there

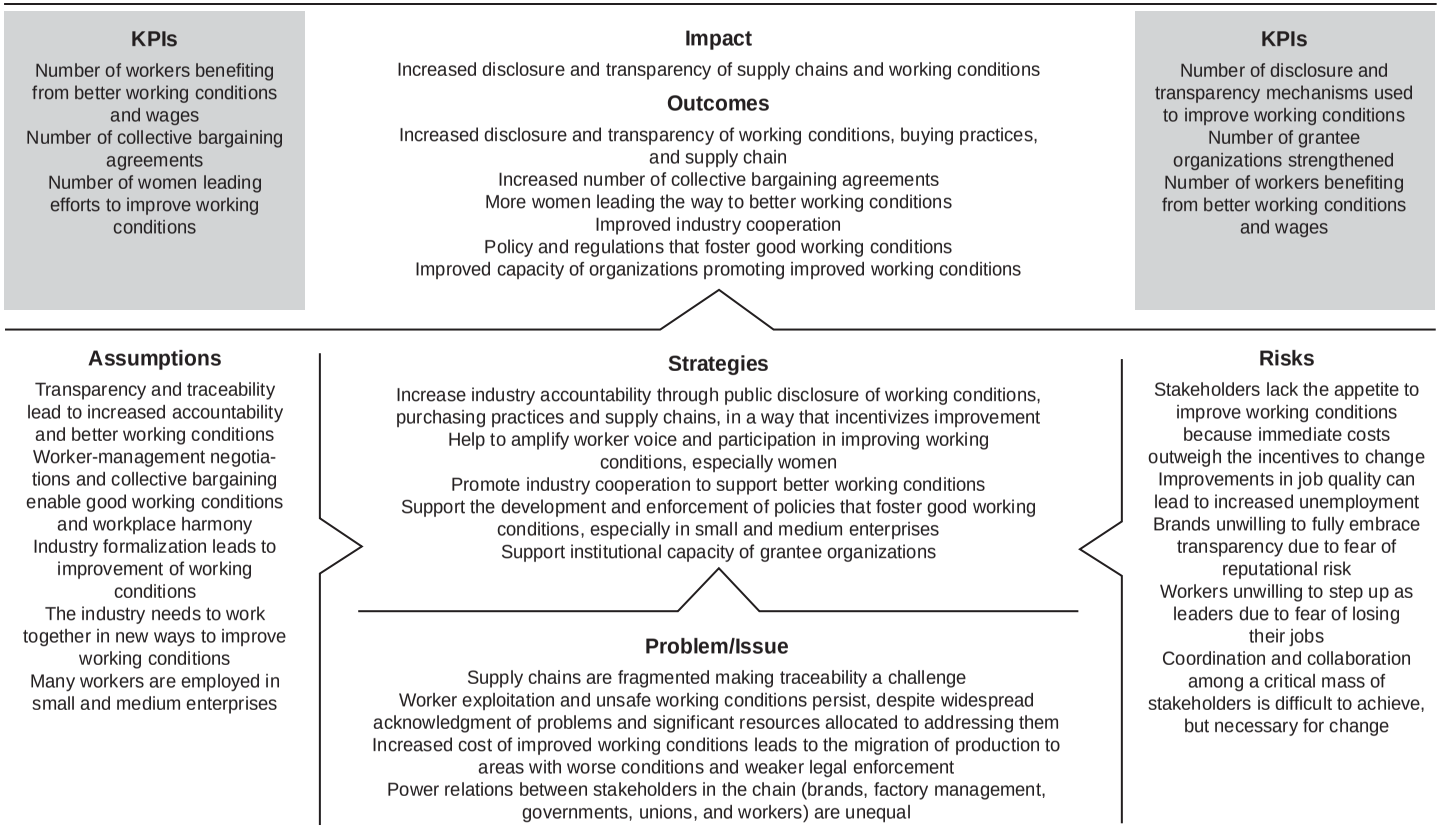


Figure 4.2 Theory of change for C&A Foundation Signature Programme 'Working Conditions'
Source: C&A Foundation website (http://www.candafoundation.org/)

> has to be change within the factories, and you have to measure the change. With a three year programme, you have to pin that down right from the start.
>
> In the QuIP evaluation of YQYP, even if concrete actions were reported, they would be reported by the QuIP as an *unintended* outcome. They would have happened by chance or by luck, not *as a result* of the programme. But we clearly expect such actions as a result of the programme.

Mull however did make clear that the QuIP approach – which aimed to identify changes in employees' working conditions as well as changes in what they thought and felt – helped the Foundation to better understand the aims of the YQYP project.

> This has been a *huge* learning experience for us. The report turned out to be very helpful. The QuIP gave us what we needed once we adapted the methodology. The nuances came out very clearly in the report, there was no doubt of how the workers *felt* once they were trained. And I emphasize the word 'felt' because that's about agency, right? Capability, confidence, empowerment.

The primary usefulness of the QuIP study to C&A Foundation was to illuminate the ways in which the YQYP project stopped short of bringing about change in working conditions. To achieve this would have entailed going beyond creating a sense of agency or empowerment among a very mobile population of factory workers, to generating evidence of observable actions taken by YQYP-trained workers at the factory where the training had taken place, on the basis of a new-found sense of agency. The YQYP project did not incorporate this additional step, and this would have remained the case even if the QuIP study had uncovered through its exploratory methodology a wealth of actions taken by workers.

Conclusions

It is beyond the scope of this chapter to assess C&A Foundation's evolving global strategy for promoting positive change in the garment industry, or to deliberate about the relative merits of the YQYF project compared with other ways of improving working conditions in Mexico's *maquilas*. This section reflects instead on two more general lessons that can be drawn from the chapter: one concerning the QuIP and impact evaluation methodology, and the second on the role of independent evaluation in policy.

On the technical subject of the evaluation methodology, it is important to emphasize the innovative nature of C&A Foundation's inclusion of a QuIP component in the full evaluation study. Not surprisingly, it was the issue of blindfolding that attracted most discussion. On the positive side, the study demonstrated that with appropriate introductions, training, and piloting, experienced social researchers can conduct blindfolded interviews

with factory employees that generate relevant evidence. Consultation also led to the decision to triangulate across blindfolded and un-blindfolded data collection. An alternative might have been to carry out un-blindfolded follow-up meetings to blindfolded interviews, but this could not have been fitted within the study time-frame.

Of course, it is always possible to argue with the benefit of hindsight that an evaluation could have been strengthened through greater prior consultation. In practice, scope for doing so is always constrained, and methodological adaptation through ongoing consultation during a study (as happened here) can be as or more effective.[13] What the evaluation did do was give IMIFAP the opportunity to demonstrate what the YQYP project was doing, based on a combination of quantitative and qualitative methods. In addition, and as summed up by Mull, 'the adapted QuIP methodology delivered an evaluation that helped C&A Foundation to acknowledge the results, learn from their investment, and understand the extent to which the results aligned with its mission'.

Notes

1. http://www.candafoundation.org/about
2. Tier 1 suppliers cover 'cut and sew' production units, Tier 2 cover printing, laundries, and embroidery firms, and Tier 3 cover fabric mills, spinning mills, and dye houses. C&A lists 119 suppliers in Mexico (see http://sustainability.c-and-a.com/supplier-map/). For details of the transparency pledge that C&A has signed up to see Clean Clothes Campaign (2017). C&A also made headlines by commissioning historian Mark Spoerer to throw light onto its deplorable wartime collaboration with the German Nazi Party (*The Economist*, 2016).
3. Its board is made up of C&A directors Edward Brenninkmeijer (Chairman), Albert Brenninkmeijer, Martin Rudolf Brenninkmeijer, Bart Brenninkmeijer, and Jeffrey Hogue. See https://www.c-and-a.com/uk/en/corporate/company/sustainability/
4. Fundación C&A was established by C&A in Mexico in 1999, operating after 2011 as part of the global philanthropic network coordinated by the C&A Foundation. In this chapter we use the name C&A Foundation to refer to both.
5. https://yoquieroyopuedo.org.mx/en/our-journey. The project was also informed by a research study of the Mexican textile and footwear industry conducted by the consulting firm INSITUM from March to May 2014.
6. The 14 participating *maquilas* in the first year included four who did not send representatives to a sensitization conference. The *maquilas* who participated in the second year comprised six continuing from the first year, four who participated in new sensitization conferences and five who did not attend.
7. On the OECD-DAC framework see Austrian Development Cooperation (2009).
8. For supervisors, the pre- and post-treatment sample sizes were 179 and 153, and the corresponding comparison group samples were

40 and 32. For operators the pre- and post-treatment sample sizes were 938 and 431, and corresponding comparison group samples were 206 and 254. Comparison groups were drawn from five non-participating *maquilas* in the first year and six in the second (BSDR, 2017: Tables 2.6 and 2.1).

9. The final selection was further complicated when one of the Y2-only participants selected (Trajes Mexicanos) declined to participate after the original champion of the YQYP project within the company left. This was a major setback because it was by far the largest participant, accounting for 66 out of 118 supervisors and 1,208 out of 1,536 operators in the list used for sample selection. These statistics indicate it was an outlier not only with respect to scale but also in having a supervisor-operator ratio of 18.3 compared with the weighted mean for all other factories of just 2.2.
10. The negative explicit comments about the project referred, in one case, to the failure of the training to do anything to raise pay; and in a second to how one participant's response to the training led to conflict with his partner, who attributed his changed behaviour to being with another woman.
11. The extent to which the YQYP training could be conducted at a given factory also depended on how organized factory management was, and their level of cooperation. Trainings had to be scheduled – requiring administrative responsiveness; and supervisors and workers had to attend them – requiring managerial follow-through. The IMIFAP implementation team experienced non-compliance and delay at both of these interfaces. Sometimes they would travel to a factory to conduct a training, only to find the appointment had been forgotten or unilaterally cancelled at the last minute by a factory manager or the supervisor. This was especially troublesome given the geographical distances between the factories, and the factories and the IMIFAP office.
12. To use the 'context, mechanism, outcome' (CMO) language of realist evaluation this assumption can be termed a causal 'mechanism' generating improved personal agency as an 'outcome' from the carefully constructed 'context' of factory-based training (Pawson, 2013).
13. This resonates with the debate triggered by Hirschman about 'beneficial ignorance' and the extent to which planning, in a complex world, should precede action or follow it. See Flyvbjerg (2018) and the papers that preceded it.

References

Austrian Development Cooperation (2009) *Guidelines for Project and Programme Evaluations* [pdf], Vienna: Austrian Development Agency <https://www.oecd.org/development/evaluation/dcdndep/47069197.pdf> [accessed 14 October 2018].

BSDR (2017) *External Evaluation of the Fundación C&A Initiative 'Yo quiero, yo puedo … cuidarme y mejorar mi productividad'* [pdf], submitted to the C&A Foundation and Fundación C&A México, Bath, UK: Bath Social and

Development Research Ltd <https://www.candafoundation.org/global/our-work/results-learning/finalreportimpactevaluationyqypprogrammemexico120617-vfwebversion.pdf> [accessed 14 October 2018].
Clean Clothes Campaign (2017) *Follow the Thread: The Need for Supply Chain Transparency in the Garment and Footwear Industry* [online], Amsterdam: Clean Clothes Campaign <https://cleanclothes.org/resources/publications/follow-the-thread-the-need-for-supply-chain-transparency-in-the-garment-and-footwear-industry/view> [accessed 8 October 2018].
Flyvbjerg, B. (2018) 'Planning fallacy or hiding hand: which is the better explanation?' *World Development* 103(March): 383–6 <http://dx.doi.org/10.1016/j.worlddev.2017.10.002>.
Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*, London: Sage.
Pick, S. (2015) 'Workers' wellbeing can create a more sustainable apparel industry — and boost profits' [online], *Triple Pundit*, 20 October 2015. <http://www.triplepundit.com/2015/10/investing-workers-wellbeing-can-catalyze-sustainable-apparel-industry-boost-profits/> [accessed 19 January 2017].
Pick, S. and Sirkin, J. (2010) *Breaking the Poverty Cycle: The Human Basis for Development*, Oxford: Oxford University Press.
The Economist (2016) 'How to confront a dark corporate past: a Dutch case suggests firms with horrible stains on their history are better off facing up to them', *The Economist*, 29 October 2016.
White, H. (2014) 'Current challenges in impact evaluation', *The European Journal of Development Research* 26: 18–30.

About the authors

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of QuIP across a range of contexts in seven countries.

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Max Niño-Zarazúa, PhD, is an independent consultant specializing in financial inclusion and international development. His main expertise is in strategy, product development, institutional assessment, and impact evaluation of microfinance products and services for low-income people in Latin America, Africa, and Asia. He was lead evaluator and project manager on Habitat for Humanity's QuIP study of housing microfinance in India;

and on C&A Foundation's QuIP study of working conditions and wellbeing of textile factory workers in Mexico.

Savi Mull, MA Sociology, is Evaluation Specialist at C&A Foundation where her main role is ensuring measurement of results in a robust way to generate evidence of what works and what does not work. She has close to 20 years of work experience in evaluations and grant management, and her main expertise lies in mixed methods evaluation. She was responsible for commissioning and overseeing the external evaluation using the QuIP approach of the YQYP initiative implemented in Mexican apparel factories.

CHAPTER 5

Exploring the social impact of housing microfinance in South India

Jitendra Balani, James Copestake, Marlies Morsink, Max Niño-Zarazúa, Sandra Prieto and Greg Skowronski

This chapter reports on a study using the Qualitative Impact Protocol (QuIP) commissioned by the Terwilliger Center for Innovation in Shelter (TCIS) of Habitat for Humanity International (HFHI). TCIS promotes a market-led approach to meeting the huge global demand for better housing. It does so by providing wholesale loan financing and advisory services to microfinance institutions (MFIs), and assistance in designing housing finance products for low-income households. This study evaluated the impact of housing improvement loans (HILs) provided by two MFIs in India: ESAF Microfinance and Investments (EMFIL) in Kerala, and Growing Opportunity Finance India (GOF) in Tamil Nadu. The study found that HILs led to improved housing conditions and social relations, and increased feelings of security and privacy, but also reported important context-specific variation in impact, particularly between rural and urban areas. This is one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.

Keywords: impact evaluation, causal attribution, India, microfinance, housing improvement loans, shelter

Introduction[1]

Housing is often proclaimed to be one of the 'big three' priorities for low-income households around the world, along with food and primary education. A comprehensive report on the global housing crisis from McKinsey estimated that 330 million urban households around the world live in substandard housing and more than 200 million households in the developing world live in slums (Woetzel et al., 2014).

This deficit exists predominantly in the developing world among low-income populations where, for many reasons, new and formally financed units are unattainable for all but a minority of the population. As a result, there is great demand for improvements and repairs to existing shelters, which are often deemed inadequate. Indeed, in much of the world, the predominant pattern for building and upgrading homes is progressive — by means of small, incremental stages in accordance with a household's priorities and financing abilities.

http://dx.doi.org/10.3362/9781780447469.005

Many households work on the improvement and extension of their homes first to obtain the minimum standards in size and quality, and later to accommodate changes in household size or to obtain income from their investment in the house (Green and Rojas, 2008). Incremental housing can be described as an inverted version of the traditional, formal process of building and financing a house. For example, in the traditional process, the complete features of the house are available to the owners from the first day of occupancy. In the incremental construction process, households begin residing in a home with the most basic features and build at the pace their financing capacities allow. However, many of these households do not have access to conventional mortgages, and government-financed housing programmes are usually constrained by limited resources.

Meanwhile, microfinance has become an increasingly important means to access capital for low-income populations, often the same populations lacking adequate shelter. While originally intended mainly to promote microenterprise, the attention of the microfinance industry has broadened to address other client needs and preferences by offering a wider range of financial products. For those institutions serving the 'base of the pyramid' and committed to positive financial and social results, housing microfinance emerged organically as a market-based approach to addressing substandard housing by helping the millions of people living in the world's slums to incrementally improve their living conditions (Mayank et al., 2012). This understanding is strengthened by evidence that between 20 and 30 per cent of microenterprise loans are diverted into housing, despite not always being tailored appropriately to this purpose.

Although housing microfinance shows much promise, the supply of such services still lags far below client demand. Launched by Habitat for Humanity International (HFHI)[2] in October 2016, the Terwilliger Center for Innovation in Shelter aims to make affordable housing possible for 8 million families by 2020, by facilitating more efficient and inclusive housing market systems. The Terwilliger Center works with numerous market actors within housing market systems by supporting local firms and expanding innovative and client-responsive services, products, and financing, so that low-income households can improve their shelter more effectively and efficiently. A leading example is the Center's MicroBuild Fund: the first impact investing fund dedicated exclusively to enabling housing microfinance.[3]

> We started as an NGO in 1976 through philanthropic efforts, but the housing deficit has outgrown those efforts. There is a huge gap between demand and supply, and with that model, we would never have been able to meet the demand for affordable and decent shelter. So in 2012 we established a unit within HFHI (known back then as Center for Innovation in Shelter and Finance) to work with private sector actors to help them in the design of housing microfinance products, since we realized that access to affordable financing was one of the main barriers preventing low income people from building and/or improving their housing conditions. The Terwilliger Center is carrying on this work [quote from key informant interview with Jitendra Balani].

The Terwilliger Center aims to help financial institutions design products that fit the affordability levels, preferences, and needs of low-income households: not just a financial product, but one that will support the incremental building practices that prevail around the world, to improve both the physical condition of the house and the quality of life of those living in it. To help achieve these outcomes, it offers microfinance institutions (MFIs) a comprehensive set of guidelines and advisory services for housing microfinance product development.

Alongside its advisory and financing roles, the Terwilliger Center also focuses on advancing knowledge about housing markets by conducting research studies, creating publications, developing tool kits, and scheduling public events to promote the sector. Assessing the impact of access to housing microfinance on low-income households is extremely relevant to the Terwilliger Center to inform its strategy and to help validate its theory of change. In 2016, it commissioned Bath Social and Development Research Ltd (BSDR) to conduct an evaluation study of housing microfinance loans provided by two partner MFIs in southern India.[4] The present chapter is based on this study. The following section outlines the context of microfinance in India, introduces the Terwilliger Center's activity in the country, and provides background information on the two MFIs. Next, we outline the QuIP component of the study and provide an illustrative overview of findings. The final section draws out some general lessons about the QuIP, and about how the study relates to the Terwilliger Center's market-based approach to improving housing.[5]

The India context and a profile of the selected MFIs

In India housing is not only a basic need but also an important vehicle for social, cultural, and economic development (Arya, 2013; NHB, 2013). In the past decade, the Government of India has introduced various financial reforms and long-term programmes to improve the housing finance market. For example, the National Housing Bank has been instrumental in giving licences to affordable housing finance companies that cater to low and middle income households. But while significant progress has been made towards the goal of 'housing for all' by 2022, a shortage of housing has remained a major problem across the country (Khan, 2012). One estimate is that the housing shortage in rural areas is 43.67 million units, and in urban areas 18.78 million units (NHB, 2013). India's large microfinance sector has started to recognize the opportunity to address this need through design and delivery of loans to help borrowers improve their shelter units incrementally.

HFHI and the Terwilliger Center in India

Given the magnitude of the housing problem and a well-established microfinance market, India is an important country for Habitat for Humanity's Terwilliger Center and its market-led approach. In 2010 HFHI incorporated MicroBuild India (MBIND) with the objective of increasing access

to housing microfinance for low-income households across India. MBIND offers wholesale loan financing for housing microfinance products and works closely with Indian financial intermediaries serving low-income populations. By the end of 2017, it had disbursed US$9 m to more than 15 MFIs across the country.

In addition to providing wholesale lending, MBIND provides advisory services to financial institutions to help them develop and refine housing microfinance products, following the Terwilliger Center guidelines. Members of the Terwilliger Center also serve on the board of MBIND and advise the CEO on strategy and operations. Given the amount of finance available to MFIs from other sources, the Terwilliger Center recognizes the importance of its technical advisory role, as Jitendra Balani highlights: 'Because getting into housing microfinance is an altogether new thing for MFIs in India, we also provide advisory services along with the capital; advisory services help MFIs launch a differentiated housing microfinance product'.

The Terwilliger Center's business model is based on its theory of change. This states that providing debt capital and technical assistance to MFIs will increase their provision of housing microfinance loans and housing support services (such as basic literacy training and construction advice) to poor and low-income households who will in turn improve their housing conditions. These short-term outcomes are expected to lead to the following medium-term outcomes: improved safety from hazards, better health and security, increased educational security, improved economic security, and greater social inclusion. Figure 5.1 illustrates this.

The Terwilliger Center commissioned an impact assessment study of two microfinance institutions, ESAF Microfinance and Investments Pvt Ltd (EMFIL),

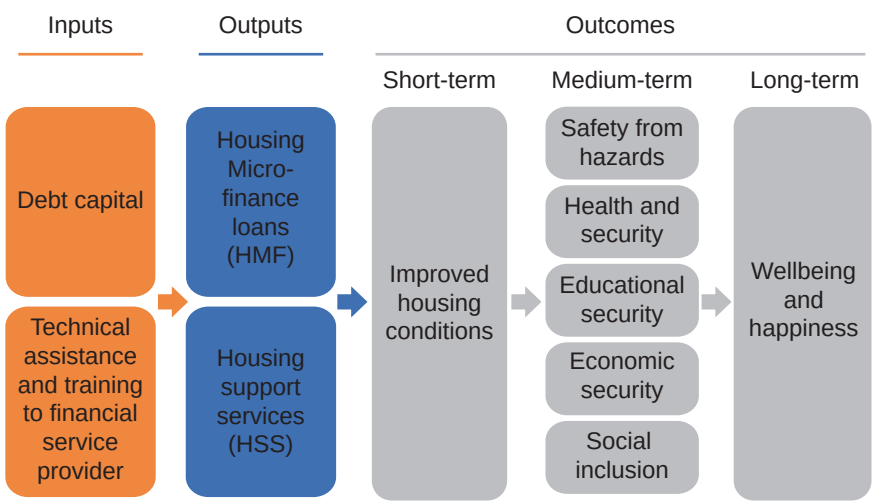


Figure 5.1 Theory of change for the Terwilliger Center
Source: TCIS, August 2016.

and Growing Opportunity Finance (India) Private Ltd (GOF), together with their housing microfinance clients in 2016. Center staff took responsibility for selecting two of MBIND's MFI partners, then eight in number. The primary criterion was that the MFIs should have issued housing loans at least three years previously, to allow sufficient time for the medium-term outcomes outlined in the theory of change to be realized. Second, the number of loans of this age needed to be sufficiently large to allow both some purposive stratification of respondents (see below) and some randomization. Third, it was important that the MFIs responded positively to the opportunity to collaborate in the study, given that doing so had not already been established as a precondition for partnering with the Terwilliger Center. These criteria led to the selection of EMFIL and GOF.

The study's overall objective was to evaluate the impact of housing microfinance products and services offered by EMFIL and GOF (two recipients of support from MBIND) to fund the development of housing microfinance product lines. Bath Social and Development Research Ltd (BSDR) conducted the research, with the assistance of Micro-Credit Ratings International Ltd (M-CRIL), on behalf of the Terwilliger Center. The study was conducted from September 2016 to March 2017 (see section headed 'The QuIP evaluation study', below).

A profile of the two selected MFIs[6]

EMFIL was established in 2007 as an offshoot of the Evangelical Social Action Forum (ESAF).[7] Kerala is its main area of operation, but it also operates in eight other states and a union territory. It offers a mixture of 15 credit and non-credit products. It introduced a pilot home improvement loan programme in December 2013, based on a market study that showed high demand among EMFIL's existing borrowers.

GOF is an implementing partner of the Opportunity International Network, dedicated to women's empowerment and microfinance around the world. GOF was formed through capital contributions from four Mutual Benefit Trusts promoted by Inter-Mission Industrial Development Association (IIDA) which currently holds a 49.40 per cent share in GOF. In 1996, IIDA started a microenterprise development programme with the help of Opportunity International Network to promote employment and income-generating opportunities to the poor through microfinance lending. GOF was incorporated in February 2006 and obtained its non-bank finance company (NBFC) registration in November 2006. It was one of the first MFIs to pilot test a housing microfinance product, receiving funding support from Habitat for Humanity and Opportunity International USA (2.5 m Indian rupees or $38,461) in 2011. Before the pilot, a market study revealed the need for housing microfinance loans among GOF's existing clients.

While both EMFIL and GOF are partners of MBIND and offer housing microfinance products to their clients, there is a vast difference in the

Table 5.1 EMFIL and GOF compared (2016)

	EMFIL	*GOF*
Year of establishment as an NBFC	2007	2006
Housing microfinance pilot	2013	2011
No. of loan products	15	3
Primary/major loan product	Income generation loan (64.1% of gross portfolio)	Income generation loan (97% of gross portfolio)
Operational area	Nine states and one union territory	Two states and one union territory
Number of branches	264	21
Total active borrowers	1,772,628	60,078
Compound annual growth rate (CAGR) of borrowers since 2013	49.3%	17.3%
Total portfolio (Rs million)	23,784.01 ($366 m)	1,119.15 ($17.2 m)
CAGR of portfolio since 2013	63.1%	50.7%

Note: EMFIL data as of 30 September 2016 and GOF data as of December 2016

size and scale of operations between the two institutions. EMFIL operates in nine states with a gross portfolio of around 23.8 bn Indian rupees ($366 m) whereas GOF is active in two states and has a much smaller gross portfolio of 1.1 bn Indian rupees ($17.2 m). Table 5.1 charts the evolution and size of the two MFIs.

In terms of outreach, EMFIL operates on a much larger scale. However, both EMFIL and GOF have concentrated in their home states (Kerala and Tamil Nadu, respectively) for housing microfinance loans due to the larger number of mature clients and stronger client relationships. EMFIL's housing portfolio is around 1.5 bn Indian rupees ($23 m) whereas GOF's housing portfolio amounts to 33.3 m Indian rupees ($0.5 m). Growth of GOF's housing portfolio has been patchy as it was severely impacted during the Andhra Pradesh crisis (see CGAP, 2010) and had to stop offering housing loans from 2014 to 2015, only restarting in February 2016. The outreach and growth pattern of the housing microfinance portfolio of the two institutions is captured in Table 5.2.

Both EMFIL and GOF offer single housing microfinance products to their clients, which are similar in features. The main difference is that GOF treats the housing microfinance loan as a non-qualifying asset and has hence priced it comparatively higher than EMFIL's product, as shown in Table 5.3.

In terms of operational practices, the processes used for delivering housing microfinance products to the clients are part of the general microfinance processes for both MFIs. While there are dedicated managers to monitor the housing programme at the field level, the conventional staff who are in charge of group lending are mainly responsible for the promotion and

Table 5.2 EMFIL and GOF housing improvement loan portfolios compared (2016)

	EMFIL	*GOF*
Number of branches with home improvement loans	140 (mostly Kerala)	12 (only Tamil Nadu)
Active borrowers of home improvement loans	31,073	556
Home improvement loan portfolio (Rs million)	1,525.48 ($23 m)	33.33 ($0.5 m)
Percentage of housing to total borrowers	1.8	0.9
Percentage of housing to gross portfolio	6.4	3.0
CAGR of borrowers (%)	570 (base year 31 March 2013)	70 (base year 31 March 2012)
CAGR of portfolio (%)	901 (base year 31 March 2013)	50 (base year 31 March 2012)

Table 5.3 Characteristics of housing improvement loans offered by EMFIL and GOF (2016)

	EMFIL	*GOF*
Who is eligible?	At least two years of association, with a good credit history	Successful completion of first cycle of loans with good credit history
Lending method	Individual	Individual
Purpose of loans	Plastering, tiling, kitchen maintenance, parapet maintenance, roof, fencing, room/house extension, and toilet construction	Room extension, roofing, flooring, plastering, tiling, septic tank, toilet, bore-well, and compound wall construction
Location	Urban/Rural	Urban/Rural
Loan size (Rs)	25,000 ($384) to 75,000 ($1,153)	50,000 ($769) to 75,000 ($1,153)
Interest rate (%)	23.0	28.0
Repayment	Monthly	Monthly
Loan term	24 months for loans up to Rs 50,000 ($769), and 36 months for larger loans	24 months
Collateral	Group recommendation and guarantor	External guarantor
Moratorium	One instalment	One instalment

delivery aspects. Table 5.4 compares the performance of the housing microfinance portfolio of the two MFIs.

The performance of EMFIL's housing microfinance portfolio is better than GOF's, with lower priced loans, lower operating cost, lower financial cost, excellent portfolio quality and higher profitability ratios. This is mainly because EMFIL's housing microfinance portfolio has steadily grown over the years and has better economies of scale while GOF had a setback during 2014 and 2015 when it had to discontinue the housing microfinance products and rebuild the product starting in February 2016. However, GOF has progressed well since then and has already attained operational self-sufficiency for housing loans.

Table 5.4 EMFIL and GOF performance of housing improvement loan portfolio (2016/17)

	EMFIL (%)	*GOF (%)*
Earnings/expenses and portfolio quality		
Yield (ratio of income from loans to loan portfolio)	21.4	25.7
APR (annual percentage rate)	24.9	29.1
FCR (financial cost ratio)	12.0	13.5
OER (operating expense ratio)	6.3	12.7
PAR (portfolio at risk) >60 days	0.0	0.0
Profitability		
Spread (net operating margin)	5.2	0.3
OSS (operational self-sufficiency)	128.3	101.2
ROA (return on assets)	4.8	3.0

Note: Ratios annualized for the fiscal year 2016–2017

The QuIP evaluation study

The QuIP study of the two selected MFIs was commissioned in August 2016. Lack of client-level baseline data ruled out a quasi-experimental approach, but the Terwilliger Center was also deliberately seeking something different. '… We wanted to learn from different approaches, and thought a more qualitative approach – where people have a greater say about the impact that they are experiencing in their own lives – would be better. We wanted insights from the field so we could re-visit our theory of change' (JB).

BSDR secured the contract in partnership with M-CRIL, a consultancy firm based in Gurgaon with a long record of working in India's microfinance sector. In addition to the impact assessment of the households accessing housing microfinance loans, the study included an additional component, to make an institutional assessment of the performance and sustainability of each MFI's housing loan portfolio.[8] This component, based on a desk review of relevant documents and key informant interviews, was mostly conducted by M-CRIL. It is the source of the financial data provided in the previous section but is not discussed further in this chapter. The aim of the other study component, which utilized the QuIP, was 'to understand the social impact of housing microfinance and how such loans are changing the social, economic and housing conditions of low income households' (HFHI, 2017: 6). The Terwilliger Center staff also hoped this part of the study would provide information to review their theory of change in the light of actual experience with disbursing housing microfinance loans.

> There were two features of QuIP that made us choose the approach: it's definitely a self-reported approach, where people are telling their stories from their own perspective; and another thing is that it's a blinded approach. When we do product development and refinement, we also adopt a qualitative approach … at the Terwilliger Center we use 'Human

> Centered Design Principles' to interview a very small set of respondents, and adopt different techniques where the users are actually co-creating the product. So we found the QuIP approach very familiar in terms of the way questions were asked. Like 'what, why, how'. I think there was plenty of opportunity for respondents to think and then raise questions instead of just answering 'yes' or 'no' (JB).

The interview guidelines for the QuIP were developed to align with the theory of change (ToC) set out in Figure 5.1, and included six outcome domains: housing conditions; health, safety, and security; housing finance and services; economic security; relationships; and wellbeing. These made the focus on housing clear to both interviewers and respondents, but the narrower confirmatory purpose of the study (namely to assess the contribution of the MFI's housing microfinance loans) was not made explicit. The ToC also influenced the selection of MFIs and borrowers, by highlighting expected lags in realizing intended outcomes (see below).

Each MFI was first asked to prepare a list of past housing microfinance borrowers from selected branches, from which a sample of 32 borrowers would be drawn for interview. EMFIL readily provided a sampling frame of 162 clients who received their first house improvement loans (backed by MBIND) between April and November 2014 – no clients had yet taken a second loan.[9] Obtaining a similar list from GOF proved harder (for reasons explained in Box 5.1) but they eventually supplied a list of 314 housing

Box 5.1. Delays in implementing the study

Planning of evaluation studies is often made against tight deadlines set by commissioners and funders. This was not the case here, but the study does usefully illustrate how factors beyond the control of those directly involved mean that such studies in practice often take longer than anticipated. The original plan here was to conduct the evaluation between September 2016 and January 2017. In fact, data collection was conducted in December/January (for EMFIL) and February/March (for GOF), with the final report agreed only in May. Delays arose for various reasons. First, selection of EMFIL was not finalized until late October, due to delays in hearing back from other possible MFIs in the context of the many other operational demands being made on them around that time. This made it necessary to train a new team of field researchers with the necessary language skills. Second, delays occurred in securing lists of borrowers for sampling, partly due to limitations within the management information systems of the MFIs. An initial list supplied by GOF included an insufficient number of longstanding clients. Before it could provide a fresh list, in November, the Indian Government announced its demonetization initiative – withdrawing all Rs 500 and Rs 1,000 notes as legal tender. This created an immediate loan recovery crisis for GOF, because most of its clients only repay in cash. All other work was put on hold until the situation stabilized. When a second list was provided, only 36 clients were in their second loan cycle. Only in January was it possible to secure the final list. The postponement also resulted in having to recruit and train additional field researchers. A third source of delay occurred between finalization of a draft report in March and agreement on its final version in May, in part based on securing further data and clarifications from the two MFIs. Once accepted, the Terwilliger Center was very quick in producing a published version of the findings (HFHI, 2017), drawing on support from the IKEA Foundation.

improvement loan clients who had received loans during 2012–13, of which 31 clients were in their second loan cycle. However, some of the addresses on the client lists were not up-to-date, while others were incomplete. Consequently, a loan officer had to personally accompany the QuIP interviewer to seven interviews, thereby revealing to respondents that the interviewer was somehow related to GOF.

The QuIP study was not intended to be statistically representative of all housing loan recipients of the selected MFIs, but to generate evidence to compare against the theory of change. Nevertheless, the Terwilliger Center was interested in exploring sources of variation in impact for each MFI according to rural–urban location, land tenure status of the borrower's dwelling, their education, marital status, gender, and age. However, no data was available on land tenure or education, whereas that for marital status was limited to three categories (single, married, widowed) with no reference to co-habiting or being divorced. Moreover, GOF lent only to women and EMFIL mostly so; hence there wasn't a sufficiently large pool of men available to study impact by gender of the borrower. Consequently, the GOF sample was divided equally between rural and urban respondents cross-tabulated against first or second housing loan recipients, with a quota sample of nine in each sub-category (making 36 in total). One focus group discussion was arranged for each of the four sub-categories, adding 34 respondents, of whom four had participated in the individual interviews. The EMFIL sample of 36 was also divided into four, but using respondents' ages (up to 44, or over 44) instead of number of loans received to cross-tabulate against rural and urban locations. Similarly, one focus group discussion was also arranged for each of the four sub-categories, involving an additional 29 borrowers, of whom only three were also interviewed individually.[10]

Illustrative findings[11]

EMFIL

Most significant outcomes. The most frequently reported outcome of taking out a home improvement loan was improved living conditions, with 24 respondents reporting this positive impact, particularly the rural cohort. Over half the clients (22) interviewed in urban and rural localities, particularly older clients, reported that having access to housing loans through their groups also contributed to improved social relations with their peer members. An important outcome of home improvements was the increased feeling of security, cited by 22 mostly rural respondents.

Home improvement loans also contributed to increased access to credit, which over half of interviewed clients (20) cited as a positive outcome, particularly those in urban localities. Respondents chose EMFIL loans for several reasons, including an easy loan application process, more money that could be borrowed, no security required to access loans (i.e. the house was not a guarantee), savings group offered an opportunity for women to socialize,

and that EMFIL has been working in the area for a long time and has a good reputation. The main reasons for taking out a housing loan were to construct a new concrete house, repair an existing house, repaint the house, extend the house or to purchase household goods.

Generally, respondents took out more than one loan from different financial organizations to undertake home improvements and/or to cover other expenses. EMFIL was favoured for home improvements and also Kudambasree – a Kerala Government microcredit programme based on women's self-help groups. Cooperative bank loans were also used to build or repair houses. Gold loans were, on the other hand, primarily utilized for emergencies and everyday expenses.[12] A few respondents stopped using different financial providers and consolidated multiple loans or paid off expensive loans by taking out one large loan from EMFIL due to the lower rate of interest.

Housing quality standards. During household interviews, field researchers observed the nature and quality of the housing improvements undertaken by the clients and recorded whether the changes and characteristics of the dwelling met the criteria of Habitat's housing quality standards. All 36 of EMFIL's clients who were interviewed used locally sourced materials and labour to maintain and upgrade their houses. Similarly, all respondents met the sanitation criteria; that is, they all had access at all times to properly constructed, safe, and hygienic toilets sufficiently close to their dwellings with proper drainage systems. However, five households had problems accessing sufficient water, and three had water that did not meet the water quality standard. Four households did not meet the minimum standard for usable space (or covered area) in their dwellings, and three older respondents' houses were not safely located to protect their families against natural hazards.[13] One lesson from the study was that particular attention should be paid to the older urban respondents among whom the lowest housing standards were found.

Factors related to negative outcomes. The results did not show explicit evidence of negative outcomes as a result of the house improvement loans, but some factors were reported that appeared to have negatively affected some of the outcomes the programme had aimed to improve. Although these factors were not all caused by the programme and some were outliers, references to indebtedness across urban and rural clients and housing quality (particularly among the older urban clients) may merit attention in future programme design. The majority of respondents — primarily from the older rural cohort — reported reduced income mainly due to reducing or stopping work because of ill health and, to a lesser extent, family members moving away or business failures. Ill health was also a negative driver of change, generally related to personal or family health conditions and/or water contamination caused by a local gold factory in Cherpu.

Some interviewed clients (8), particularly the younger ones, stated that they were less economically secure, often because of increased debts from

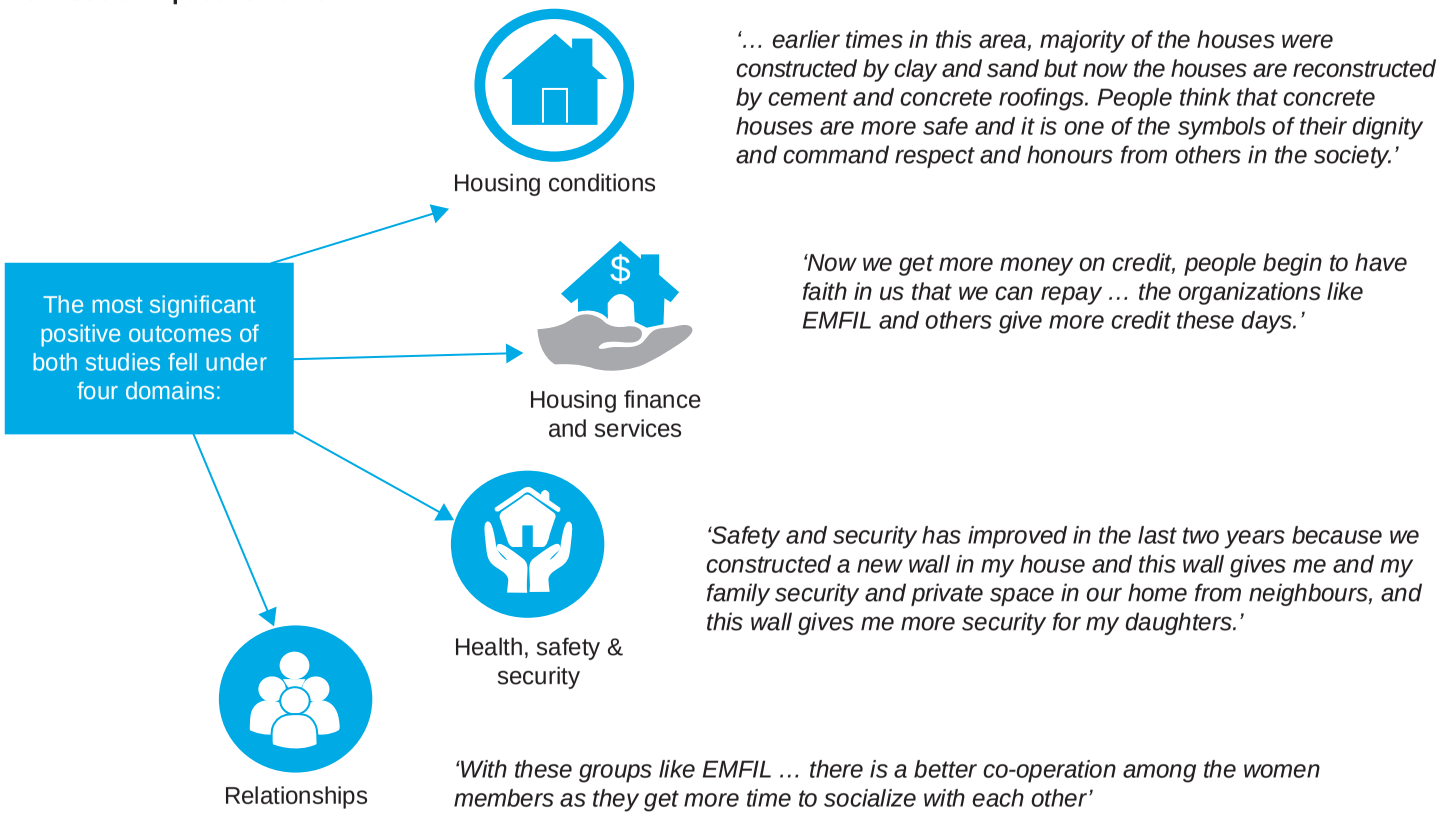


Figure 5.2 Illustrative quotations of significant positive outcomes among housing improvement loan clients

different borrowing sources. Several respondents (7) across the urban and rural sub-categories reported higher levels of debt and consequently increased stress levels as they were worried about making loan repayments, particularly those with multiple loans. The water contamination in Cherpu, and jealousy between neighbours in rural areas over having improved their dwellings, were the main factors affecting community relations.

Growing Opportunity Finance (GOF)

Most significant outcomes. The QuIP study showed that GOF was the most dominant MFI in both areas sampled, and that GOF's housing improvement loans have a positive impact on the lives of first- and second-cycle clients across urban and rural localities in Tamil Nadu in a variety of ways. The most significant positive outcomes relating to GOF's home improvement loan programme were reported in four areas.

- The most cited positive outcome was improved housing conditions, with 32 of 36 respondents in both urban and rural areas explicitly stating that they had taken home improvement loans from GOF for making home improvements. The majority of respondents (30), particularly urban second-cycle clients, reported having extended their house or built more rooms. For several urban clients, building extensions helped to house small businesses; for rural clients, the extension helped to better accommodate the whole family.
- A large proportion of interviewed clients (26), particularly rural second-cycle clients, no longer borrowed from local moneylenders, demonstrating a significant change in their borrowing habits. These changes were due to the increased presence of GOF and other MFIs in the area, greater availability of more reliable loans with lower interest rates, and the easy and simple procedure for accessing loans.
- A further positive contribution cited by 24 respondents, particularly those in their second cycle, was that having a house with enough space for all family members meant that children, in particular, had their own space to sleep and study. The majority of clients (23), particularly second-cycle clients (urban and rural), said that housing improvements, particularly building compound walls, had provided greater protection for their children. In addition, respondents reported that they felt more secure, especially from natural hazards such as floods and monsoon. These housing improvements also contributed to an increased sense of privacy, social status, pride, dignity, and respect from others.
- Expanding or starting a new business due to home improvement loans was a further positive outcome for 18 respondents, particularly second-cycle clients. As people started or expanded a business and diversified their livelihood activities, they also experienced increased income in their household that made them feel more financially secure.

Housing quality standards. The housing quality standards of GOF's clients who were interviewed in the Tamil Nadu areas had scope for further improvement because most of the clients lived near disaster-prone areas. A relatively larger number of households who were interviewed met the quality standards in water and sanitation. Good quality water was accessible to 30 of the 36 households interviewed. Similarly, 28 households across urban and rural localities had access to toilet facilities properly designed and constructed with drainage systems. Nevertheless, only a third of these households were built with durable materials to protect them in case of a natural disaster. Only two clients interviewed lived in houses built with appropriate construction and material specifications to mitigate the risks associated with living in disaster-prone areas. Over half of respondents met the covered area quality standard, that each person in the household should have a usable covered floor area of no less than 3.5 square metres.

Factors related to negative outcomes. The negative outcomes cited by five people from the rural cohort were related to reduced income and savings as well as ill health caused by illness (diabetes) and accidents. These outcomes were interrelated as clients reported that being ill affected their ability to work or prevented them from working, and that led to other problems such as worrying about loan repayments, reduced ability to buy enough food, using their savings to pay loans, and increased levels of debt. The researchers did not find explicit evidence of negative outcomes because of GOF's housing improvement loans. However, the issues of increased levels of debt, using savings to repay loans and worrying about loan repayments, together with the lower housing quality standards found among GOF clients, suggested that special attention should be paid to future design and provision of home improvement loans.

Comparison of the social impact of the two MFIs' housing loans

Perceptions of overall change. At the end of each section of the interview, respondents were asked closed questions intended to summarize the changes they had experienced over the previous two years, in the case of EMFIL; and five years, in the case of GOF. The range of answers were limited in that respondents were only given three choices (better, worse, the same), but they provided a useful snapshot of the overall direction of changes experienced. Generally speaking, GOF's clients reported more positive changes compared with EMFIL's clients, particularly in housing conditions, economic security, community relationships, and overall wellbeing domains. This is an interesting finding, particularly in the housing condition domain, given that housing quality standards of GOF's interviewed clients were markedly lower than those of EMFIL's clients.

Interestingly, the urban clients from both MFIs perceived more negative changes to their safety and security compared with the rural clients. While

access to housing finance was generally perceived as leading to positive changes among rural and urban GOF clients in Tamil Nadu, three EMFIL clients, particularly from rural Kerala, reported that access to housing finance services had worsened. Finally, there was a marked similarity between the two samples in the perception of health: overall, respondents from both MFIs across all cohorts felt that their health had worsened, with more than half of respondents from EMFIL and a quarter from GOF, particularly the urban clients, expressing concern. As discussed above, EMFIL clients' ill health was related to personal or family health conditions and/or the water contamination caused by a local gold factory in Cherpu, whereas GOF clients generally attributed ill health to diabetes and accidents.

Outcomes and drivers of positive change. Respondents from both EMFIL and GOF revealed important patterns and trends concerning the most commonly cited drivers of change that led to positive outcomes. Respondents said that taking out a repair loan from an MFI, constructing a new concrete house, extending the existing dwelling or improving the housing conditions were the most important factors that led to positive change in their lives. For respondents from both MFIs, the driving force behind decisions to take a loan and improve their homes were: insufficient space in the house as the family size had increased; children growing up and needing their own space to sleep and study; the forthcoming marriage of a son or daughter; and an aspiration to live in a concrete house in order to increase their social status.

Achieving improved living standards increased the sense of security of both EMFIL and GOF respondents, particularly in rural areas, as they were better shielded during the monsoon season and could live in their homes without fear of being flooded. In addition, while EMFIL respondents, particularly from the younger rural cohort, reported that extending their houses resulted in improved family relations and increased socializing, GOF respondents across urban and rural localities felt that they increased their privacy and provided their children with their own space for sleeping and studying.

Taking a loan from EMFIL and GOF also meant that women were able to be part of a savings group, which yielded other benefits to the respondents such as having increased confidence by being a member of a self-help group, and opportunities to discuss and share problems with others. This is important, because women across rural and urban areas not only had a better social life but also developed a greater sense of solidarity, which is crucial for lending methodologies based on joint liability.

The increased presence of MFIs in the area and the greater availability of loans had an effect on borrowing patterns. The majority of GOF's respondents, particularly those in their second cycle of home improvement loans, cited that having better access to loans had resulted in changes in their borrowing habits, since they had stopped borrowing money from local moneylenders and instead preferred to borrow from GOF. This in turn led to increased feelings of financial security and the knowledge that they would

be better able to repay the loans as the interest rates were lower and they could pay the principal amount and interest rate at the same time. Similarly, the majority of EMFIL's respondents, in particular from the urban cohort, felt that having better access to funds and/or credit increased a feeling of financial security, particularly because the process of taking out loans was simple, with no security being needed.

While a degree of improved financial security was realized because of better access to credit, it would be overstating the case to say that economic security had been achieved as a result of EMFIL programmes. On the other hand, the economic security of a number of rural and urban women in the second cycle of home improvement loans had been positively affected by GOF project activities. Four women stated that they had been able to start or expand their own home-based businesses after they had extended their houses, thereby increasing their income. The increased access to credit had also improved their financial security, adding to overall economic confidence.

Outcomes and drivers of negative change. Although clients from both MFIs reported far fewer drivers of change that led to negative outcomes, there are important issues that need to be addressed. Increased levels of debt and subsequent worries and stress over repaying loans were mentioned by urban and rural respondents (particularly the older cohorts) in relation to both MFIs and informal moneylenders. This contrasts with the observation made earlier that the majority of respondents from both MFIs felt that their financial security increased as a result of having more available and accessible loans. However, a few people also felt that having debts, and having a fear of not being able to repay them and losing an asset, were increasingly affecting their health and sense of economic security. One older EMFIL rural respondent pointed out that having debts was the main problem in her family. Several respondents, particularly the younger urban cohort, also reported being very worried about repaying multiple loans and becoming increasingly stressed about their level of debt.

These contrasting views address the important issue in the microfinance sector of increased availability of credit facilities leading to multiple loans and over-indebtedness. The difference between access to and use of financial services needs to be addressed for future programme improvements in the sense that any effort to expand the access to more credit products should be accompanied by other support services such as consumer education programmes, including financial management and financial education. Through these, MFIs would help people make more informed decisions about how to use their financial service options more wisely.

Similarly, several of GOF's clients reported that they experienced a reduction in income and savings as a result of increases in their level of debt or in the amount of loan repayments. Although these negative drivers were not solely attributed to GOF loans, particularly home improvement loans, it is important to take into account that multiple borrowing was the major reason for negative impacts in the Indian context.

The main negative impact associated with GOF project activities was that business loans were being used for other purposes. These included paying for home improvements, paying off older debts, and covering everyday living expenses. While this is not directly attributed to GOF's housing improvement loans, it illustrates the *fungibility* of credit. This brings an opportunity to adopt or improve verification practices in the lending process carried out by the MFI. Finally, ill health and the reduced ability to work and subsequent loss of income were other negative impacts reported by EMFIL and GOF's clients that were attributed to other factors outside of the MFIs' project activities. Nevertheless, these negative impacts could affect MFIs' operations by decreasing repayment rates, with clients using loans for other purposes, such as health care, and becoming over-indebted. Ill health challenges MFIs to adapt the services they offer to help clients deal with health issues and the financial difficulties that arise from them.

Discussion

The Terwilliger Center published summary findings from the study in June 2017 (HFHI, 2017). At the invitation of the Terwilliger Center, the lead author of the QuIP report also presented these findings at the Sixth Asia-Pacific Housing Forum in Hong Kong in September 2017. While positive findings outweighed the negative, he was also able to draw attention to operational implications of the study to mitigate potential risks of housing microfinance, including the case for investing in product development, of complementing credit with financial education and counselling for clients, and of providing housing support services to improve the quality and impact of the home improvements. However, the author was not involved in follow-up meetings on issues raised by the report with EMFIL or GOF – nor was this planned.

> Follow up meetings can raise awareness, but not operational influence. Some of the findings in the report – especially on multiple lending/over indebtedness – compelled the MFIs to review their loan appraisal mechanisms. Those types of issues we generally address through our advisory services to MFIs. Our experience suggests that it takes around 12–15 months of dedicated and extensive support to implement such change management initiatives (JB).

As intended, the QuIP component supported the Terwilliger Center in assessing its theory of change by providing an independent evaluation of the complexity and diversity of impact pathways from housing microfinance to wellbeing and happiness. While the overall findings of the study were broadly consistent with the Center's prior theory of change, there were interesting differences too. These included the lack of evidence of positive spillovers from improved housing to greater economic security, and limited reference to technical advice on construction to complement the loans. The study also highlighted that the theory of change understated the extent

to which achievement of medium- and long-term outcomes at the client level are contingent on their capacity to avoid ill health, maintain income streams, and respond to incidental shocks. While the QuIP field team did not have the technical skills to assess the vulnerability of respondents' homes to shocks, the high proportion of clients who were assessed as being vulnerable to natural disasters was particularly striking. This highlights the importance of the Terwilliger Center's role in providing advice to MFIs and other housing market actors on the quality of housing materials and building standards alongside its role as a wholesale provider of housing microfinance.

There are also potential methodological lessons to be learned from the study for the QuIP. First, it is important to allocate time for commissioners and implementing agencies (in this case the two MFIs) to discuss impact evaluation studies prior to carrying them out, to ensure these are as closely aligned as possible to their mutual interests and needs. For example, a clearer alignment between the Terwilliger Center's theory of change and the domain structure of the data collection instrument could have facilitated tighter and more incisive confirmatory analysis. This could even have been formalized by subjecting the data to a Bayesian analysis in which incremental change in the commissioners' confidence in key causal links within the theory of change are subjectively estimated and compared across different contexts.

Second, even with a relatively small sample size, there is scope for follow-up analysis of differential social impacts based on an *ex post* classification of outcomes. In other words, rather than ask how outcomes varied between rural and urban clients, it is also possible to explore how a larger set of exogenous characteristics of the sample (age, location, family composition, baseline poverty and housing status, health records, experience of shocks) explain variation in reported outcomes. Even without having a representative sample of the wider borrowing population this can be used to identify statistically significant sources of variation in impact among the sample of interviewees. This and the previous point also illustrate the limitations of labelling studies as purely quantitative or qualitative, thereby underplaying the scope for quantitative analysis of qualitative data and vice versa – a point raised in Chapter 1 and explored further in the final chapter of this book.

Third, there is the issue of timing. It is useful for organizations such as the Terwilliger Center periodically to subject their theory of change (and associated practices) to empirical reality checks, as happened here. But this feedback loop is relatively slow, and faster feedback is also needed to ensure wholesale funders of microcredit institutions do not exacerbate poor lending practices and over-exuberant growth strategies, or contribute to pockets of over-indebtedness and debt bubbles. Externally commissioned and independent studies such as the one reported in this chapter are no substitute for the internal social performance assessment and management of MFIs themselves (Copestake, 2007). Hence, it is as important for external sponsors of microfinance to contribute to social auditing of these systems (alongside financial auditing of internal financial management systems) as it is to invest in their own independent impact evaluation studies.

Notes

1. This introduction draws heavily on HFHI (2015).
2. Habitat for Humanity International is an international NGO which has helped more than 9.8 million people meet their affordable housing needs across the globe.
3. See https://www.habitat.org/impact/our-work/terwilliger-center-innovation-in-shelter. The Microbuild Fund is dedicated to helping low-income families by 'lending to microfinance institutions, which in turn provide small loans to families to build safe, decent and durable homes as their finances allow'. The website states that the MicroBuild Fund has 'already provided access to better housing for more than 415,000 people', and that as of 30 June 2017, it had approved $90 m across 49 institutions in 28 countries.
4. The call was originally put out by Habitat for Humanity's International Center for Innovation in Shelter and Finance, which was renamed the Terwilliger Center in 2016. For simplicity we refer to the commissioner of the study here as the Terwilliger Center.
5. An initial draft (by Copestake) drew directly on the BSDR study led by Niño-Zarazúa, as well as key informant interviews conducted by Morsink with Niño-Zarazúa (the lead evaluator) and Jitendra Balani (JB) of the Terwilliger Center. This draft was then amended by Balani, to directly incorporate findings set out in HFHI (2017), and comments and suggestions from Skowronski and Prieto.
6. This section is based on HFHI (2017).
7. ESAF was established as an NGO in 1992. It initially focused on promotion of livelihood activities among the marginalized sections of the society and gradually diversified to microfinance, microenterprise development, natural resource management, education, health, and relief and rehabilitation. HFHI first partnered with it in 2004 as a response to the great tsunami of that year.
8. Another issue for the Terwilliger Center (not explored here) is the 'additionality' of its support for the MFI: in other words, the extent to which it had contributed to expansion in the scale and quality of its work in the housing sector relative to what it would otherwise have been.
9. This EMFIL list was drawn from three communities – Amballur, Cherpu, and Wadakanchery. The final list supplied by GOF covered nine communities served by branches in north and south Chennai.
10. The purposive sample selection procedure used means that it is not possible to claim that findings were statistically representative of all longstanding HIL clients of the two MFIs. However, there are grounds for believing they reflect important aspects of the diversity of that population within the communities covered. Some findings are presented as frequency counts in order to indicate how widespread different responses were. However, it is not possible to extrapolate from them to the wider population with a known degree of statistical significance.
11. This summary of findings is taken from HFHI (2017).
12. Gold loans are financial transactions using gold as the guarantee or deposit against an amount of money lent to the customer. One of the key characteristics of gold loans is that they are disbursed quickly and without the hassle of loan appraisal and checks. Gold loans in Kerala are mainly provided by

Muthoot Finance Ltd, which is an Indian financial corporation that claims to be the largest gold financing company in the world.

13. Breaking this down further: young and old rural respondents had access to good quality water; at least one household in each cohort did not have durable structural materials to allow for safe refuge and exit in case of a natural disaster; and the houses of young rural and urban respondents were safely located.

References

Arya, V. (2013) *Housing Microfinance in India: Benchmarking the Status*, New Delhi: ACCESS-ASSIST.

CGAP (2010) *Andhra Pradesh 2010: Global Implications of the Crisis in Indian Microfinance*, Focus Note 67, November, Washington, DC: CGAP.

Copestake, J. (2007) 'Mainstreaming microfinance: social performance management or mission drift?' *World Development* 35(10): 1721–38 <http://dx.doi.org/10.1016/j.worlddev.2007.06.004>.

Green, M. and Rojas, E. (2008) 'Incremental construction: a strategy to facilitate access to housing', *Environment and Urbanization* 20(1): 89–108 <https://doi.org/10.1177%2F0956247808089150>.

Habitat for Humanity International (HFHI) (2015) *Housing Microfinance Product Development: A Handbook*, 3rd edn, Atlanta: Habitat for Humanity International's Terwilliger Center for Innovation in Shelter.

HFHI (2017) *The Impact of Housing Microfinance: An Independent Institutional and Social Impact Evaluation of Two Housing Microfinance Products in South India*, Atlanta: Habitat for Humanity International's Terwilliger Center for Innovation in Shelter, supported by the IKEA Foundation.

Khan, H.R. (2012) *Enabling Affordable Housing for All: Issues and Challenges*, inaugural address delivered at *International Conference on Growth with Stability in Affordable Housing Markets*, New Delhi: The National Housing Bank and the Asia Pacific Union for Housing Finance.

Mayank, H., Nanavaty, M., Chakraborty, S., Mitra, S. and Limaye, A. (2012) *Affordable Housing in India: An Inclusive Approach to Sheltering the Bottom of the Pyramid*, On.point, Jones Lang LaSalle <http://www.jll.co.in/india/en-gb/Research/Affordable_Housing_in_India_2012.pdf> [accessed 14 October 2018].

National Housing Bank (NHB) (2013) *Scaling up Housing Micro-Finance*, Institute for Financial and Management Research Capital in partnership with the National Housing Bank and DFID, New Delhi: NHB.

Woetzel, J., Ram, S., Mischke, J., Garemo, N. and Sankhe, S. (2014) *A Blueprint for Addressing the Global Affordable Housing Challenge* [online], Executive Summary, McKinsey Global Institute, October, McKinsey and Company <https://www.mckinsey.com/featured-insights/urbanization/tackling-the-worlds-affordable-housing-challenge> [accessed 14 October 2018].

About the authors

Jitendra Balani is Manager for Financial Inclusion and Capital Markets, Asia Pacific at Habitat for Humanity's Terwilliger Center for Innovation in Shelter. He has more than nine years of experience in providing advisory,

research, and training solutions to financial institutions that serve low-income households across Asia and Africa. Prior to Habitat, he was associated with a consulting-cum-research firm specializing in the financial inclusion sector.

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of the QuIP across a range of contexts and countries.

Max Niño-Zarazúa, PhD, is an independent consultant specializing in financial inclusion and international development. His main expertise is in strategy, product development, institutional assessment, and impact evaluation of microfinance products and services for low-income people in Latin America, Africa, and Asia. He was lead evaluator and project manager on Habitat for Humanity's QuIP study of housing microfinance in India; and on C&A Foundation's QuIP study of working conditions and wellbeing of textile factory workers in Mexico.

Sandra Prieto, MA Non-Profit Management and Social Economics, is Global Director of Financial Inclusion at Habitat for Humanity's Terwilliger Center for Innovation in Shelter, where she oversees the global strategy of financial products and services for affordable housing for low-income households. She had general oversight of the QuIP study on the impact of housing microfinance products in India.

Greg Skowronski is Director, Asia-Pacific at Habitat for Humanity's Terwilliger Center for Innovation in Shelter. He leads the regional growth strategy for market-based approaches that increase access to products, services, and financing for affordable housing. He was responsible for commissioning the QuIP study on the impact of housing microfinance products in India.

CHAPTER 6

Faith-based rural poverty reduction in Uganda

James Copestake, Michelle James, Marlies Morsink and Charlotte Flowers

The Qualitative Impact Protocol (QuIP) was commissioned by the faith-based charity Tearfund to gain deeper insight into its Church and Community Mobilisation (CCM) programme in Uganda. CCM is based on a theory of development which is centred on self-empowerment and community-based social improvement, fostered through theological resources and religious spaces. The QuIP was conducted in four villages in the east and north of the country, where Tearfund had partnered with Pentecostal Assemblies of God (PAG) and Church of Uganda (CoU), respectively. The case study illustrates the scope for combining faith-based and evidence-informed approaches to rural poverty reduction. A priority of Tearfund's was to share what it learned through the QuIP not only within the organization, but with its partners and community participants. To do so, it organized feedback and 'unblindfolding workshops'. This chapter presents one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.

Keywords: impact evaluation, causal attribution, Tearfund, Uganda, faith-based development, community development

Introduction

SOME TIME AGO, THE LEAD AUTHOR of this chapter took a taxi ride across Kampala, intending to have dinner with a friend. It didn't go well. The traffic was gridlocked, and in nearly three hours he advanced less than three miles; eventually the dinner was abandoned, and he returned to the hotel where he had started. During the journey the driver maintained a quite extraordinary serenity; but more remarkable still, he held on to an unshakeable faith that the traffic was about to clear: '… just round the next corner'; 'after this roundabout'; 'once we get through those traffic lights'; 'past this junction …'

Where does such faith come from? How about belief in the emancipatory power of faith? Does it have to be blind? And how well informed is doubt in the power of faith? These are interesting questions to address in a book about impact attribution and the scope for a more evidence-informed approach to development. Scepticism runs deep in social science, particularly towards positive evidence generated by those who have a vested interest in demonstrating success, whether to justify their salary, or to sustain the 'warm glow'

http://dx.doi.org/10.3362/9781780447469.006

they derive from what they do.[1] If social scientists ever felt the need for a patron saint then Thomas the Apostle – latterly dubbed 'Doubting Thomas' – would be a good candidate: 'Thomas […] was not with the disciples when Jesus came. So the other disciples told him, "We have seen the Lord!" But he said to them, "Unless I see the nail marks in his hands and put my finger where the nails were, and put my hand into his side, I will not believe"'.[2]

While strongly influenced by the principle of separating religion and state, the field of international development is nonetheless replete with 'faith-based organizations' (FBOs), and the issue of how faith affects their performance has attracted considerable scholarly attention. Clarke (2006), for example, concludes his review by suggesting that

> FBOs … have a number of characteristics that distinguish them from their secular peers. They draw on elaborate spiritual and moral values that represent an important and distinct adjunct to secular development discourse. As a result, they have a significant ability to mobilise adherents otherwise estranged by secular development discourse. They are highly networked both nationally and internationally and are highly embedded in political contexts and in processes of governance in both horizontal and vertical terms. They are less dependent on donor funding and they have well-developed capacity and expertise in the key areas of development practice (Clarke, 2006: 845).[3]

Tearfund is a UK-registered Christian charity, established in 1968 and currently working in over 50 countries to eradicate poverty.[4] In 2016 it had a total budget of over £70 m, allocated between disaster response (35 per cent), community development (29 per cent), church mobilization (9 per cent), and advocacy (7 per cent).[5] A large component of the community development budget is allocated to the Church and Community Mobilisation (CCM) programme, a partnership-based development process that Tearfund has promoted and supported through local churches and rural congregations for over 15 years, and in 41 countries. Its aim is 'to envision local churches to mobilise communities and individuals to achieve "holistic transformation" in which people flourish materially, physically, economically, psychologically and spiritually' (Tearfund, 2018: 2). Unencumbered by targets and timeframes, CCM is mostly funded through private donations, and can also be viewed as a leading example of an explicitly faith-based approach to development practice.[6] Its emphasis on 'social transformation' rather than on 'managerial' institutional logic also makes it an interesting case study of impact evaluation methodology (Elbers et al., 2014).

This chapter reports on an evaluation of CCM in Uganda using the QuIP. The next section elaborates on the project. This is followed by an overview of how the QuIP study was designed, implemented, and utilized by Tearfund. The chapter continues with a review of the empirical findings from the study and concludes with further reflections on the relationship between evaluation methodology, evidence, faith, truth, learning, accountability, and legitimacy. The chapter was drafted by James Copestake and Marlies Morsink, incorporating material from the QuIP study report (BSDR, 2017)

produced by lead evaluator Michelle James (2016). It also draws on a key informant interview with Charlotte Flowers in November 2017 (cited as CF), who played a leading role for Tearfund in commissioning, overseeing, and disseminating findings from the QuIP study. James and Flowers also reviewed and commented on the initial draft. The lead QuIP field researcher for the study was Moses Mukuru.

The theory and practice of Church and Community Mobilisation (CCM)

As with the YQYP programme in Mexico (see Chapter 4), CCM is based on a theory of development centred on empowering people to help themselves. It proposes that to reduce material poverty, attitudes of helplessness and dependency need to be replaced by self-belief and agency. It draws partly on the theory and practice of participatory development going back to Paulo Freire (1970); but it also draws heavily on Christian theology. 'It's about building self-esteem, and trying to break that emotional poverty where people see themselves as too poor to do anything; it's about saying, in Christian terms, "You are made in the image of God, you are of value"; and encouraging people to think about what they can do' (CF).

The more specific theory of change (see Figure 6.1) underpinning CCM is to foster a dynamic interaction between theological resources, religious spaces, and their context, to promote social mobilization based on the rationale that 'when the church is envisioned to provide a space for people to be empowered, to understand their self-worth, to build relationships with others and work together for change, initiatives and projects will bring about a change in holistic wellbeing' (Tearfund, 2017).

CCM is not a programme with clearly defined physical deliverables or time frames. Rather, through the utilization of bible studies, discussion tools, and group activities, it seeks to 'awaken' local church leaders, congregations, and poor rural communities and encourage them to collaborate in realizing their own development.

Tearfund's involvement in CCM is mediated by partnerships with local churches, to whom they look for close understanding of rural communities, commitment to sustained relationships, and capacity to provide leadership and training. Its own role is primarily to support partner churches in training facilitators. To this end, Tearfund publishes relevant material, including a CCM manual that its partner churches can adapt to suit their own denominational traditions.

CCM facilitators are equipped with a set of questions, techniques, and stories (many drawing on or illustrated from the Bible) to help community members think about what they need and what their community would look like if it were the best community it could be. They then coach community members to take up roles as information-gatherers, review the resources already at their own disposal, reflect on how to use them, and decide on priorities for collective action. This process is left in the hands of the community, but Tearfund and its partner churches remain open to requests for help.

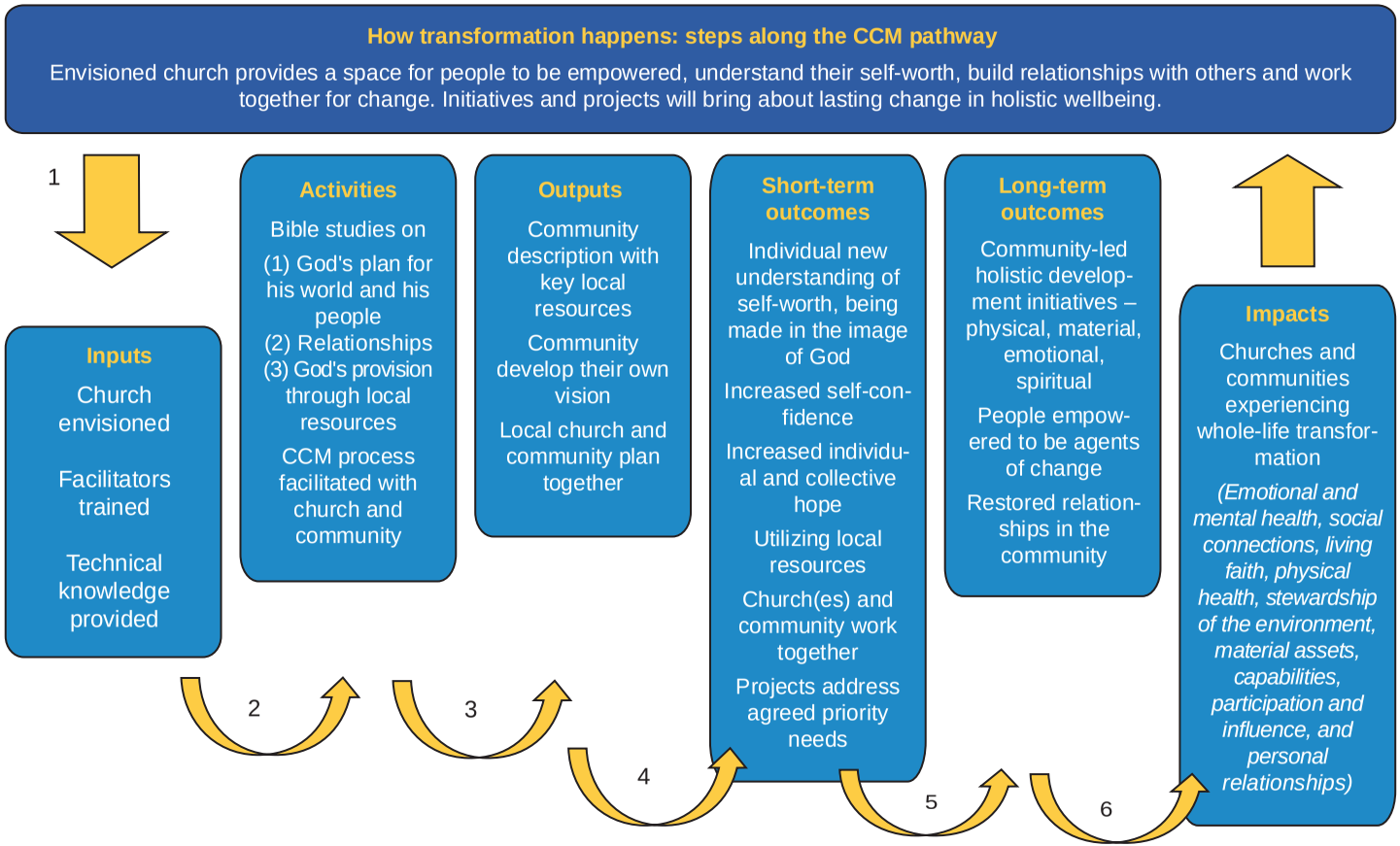


Figure 6.1 Theory of change for Tearfund's CCM process
Source: Tearfund (2017)

> The idea is that the help Tearfund provides is demand-driven rather than supply-motivated; Tearfund is keen to support the community development process, but without trying to lead it in any particular direction. […] For example, if we hear from a partner that a church is really interested in having a well, then that church or community needs to come to us to say what it wants, and we might provide training on how to dig a well, or provide specialist technical support if that is what's needed. But we'll never say, 'We think you need a well, and we're coming to dig it for you' (CF).

In a new locality CCM starts with bible studies, touching on such topics as justice, self-worth, and community-spiritedness. This serves as the basis for broader discussion of how the church can serve its members as well as the larger community, and how to work together to address issues of common concern. The first project that an 'awakening' congregation decides on has often been to build its own church, with members of the congregation making the bricks and doing all the construction work. From here, what direction CCM takes depends on the priorities and decisions of the specific congregation and its wider community. 'This is where things can go in lots of different directions. Even though Tearfund has developed a theoretical CCM process that is standardized, it aims simply to facilitate how communities can recognize and prioritize their own needs. They are the ones doing the development and deciding what is needed in their context' (CF).

In some cases, Tearfund supports livelihood training – in how to fix mobile phones, make and lay bricks, or adopt different agricultural techniques, for example. 'It's alright to get people coming up with ideas, but sometimes they just don't have the competencies to implement them, and we can help them build competencies' (CF). In Uganda, Tearfund has also developed a specific CCM programme of advocacy training to foster local-level social accountability and governance (Tearfund, 2016).[7]

The QuIP study in Uganda

Commissioning the study and country selection

Tearfund's interest in conducting an impact evaluation using the QuIP was to gain deeper insights into CCM for the organization, its partners, and intended beneficiaries; the request arose from within the organization, rather than being prompted by an external funder. 'We get lots of nice impact statements from our visits and internal evaluations, but wanted to dig deeper; we were keen to see what robust research would reveal' (CF). Tearfund had already conducted an external evaluation of CCM in Tanzania in 2015 using a difference-in-difference approach (Scott et al., 2014; see also Chadburn et al., 2013). This generated information about what was happening on the ground, but left unanswered the questions about how observed changes were taking place.

> We have our theory of change, and our idea of what we think happens in CCM – that is: Tearfund does facilitation training, the church gets inspired, they work with the community, and then lots of nice things happen. We wanted to test those steps and the links between them: Is the training in fact leading to these other steps? How does that work? That's what really appealed to us about the QuIP: it would help us really learn, because it would help us understand more about the drivers of change (CF).

Congruence in the values behind the CCM and QuIP was also important.[8]

> QuIP methodology ticked a lot of boxes for us, because CCM is led by the people themselves. The ethos is about empowering people to take charge of the process, and not have us trying to control the process. The fact that the QuIP allows those beneficiary voices to be at the forefront of the research, we thought was really special (CF).

Tearfund's decision to invest in a pilot QuIP study entailed extensive internal consultation and discussion, extending beyond those specialized in monitoring, evaluation, and learning.

After an initial meeting, Tearfund created a working group, including representatives from several countries doing CCM, and people from Tearfund's Technical Team. There was some apprehension about trying a new and qualitative approach (one that would not generate statistically robust evidence of CCM's impact on poverty, for example) and it was important to ensure clarity about what could realistically be expected, and to secure wide support based on this understanding.

Another issue that prompted internal discussion was country selection. CCM varies substantially from country to country; hence there are good grounds for conducting several QuIP studies. Tearfund decided to start with a pilot study, leaving open the decision to then repeat it elsewhere. Uganda was chosen, as one of the countries in which CCM was oldest and most established (along with Tanzania and Kenya), and because there was strong support from the country representative.

> We know the situation and the context in Uganda better than some other countries. Uganda is a bit of a flagship for CCM because it's been going there so long. We wanted to learn what's worked well there, and then move on to how things can be adapted or done differently in places where it's a bit more of a challenge (CF).

Approximately 84 per cent of Uganda's population self-report as Christian.[9] The incidence of absolute poverty (defined as living on less than $1.90 a day) is high but has been falling quite fast – from 62.2 per cent to 33.2 per cent between 2002/03 and 2012/13, for example (World Bank, 2016). These figures suggest that Uganda is likely to have provided relatively favourable conditions for CCM to flourish in recent years. However, in Soroti and Kitgum, the areas of the study, there was evidence that poverty had worsened – against the trend in the country as a whole.[10]

Consultation with local partners and sample selection

Tearfund first started working in Uganda in 1973 and currently partners with 11 local Christian agencies in 30 districts. CCM was introduced in 2001 and by 2017 Tearfund estimated it had reached 300 churches and 105,000 individuals. Tearfund's main CCM partner is Pentecostal Assemblies of God (PAG), followed by the Church of Uganda (CoU). Having no country office in Uganda, Tearfund relied on virtual communication to invite its partners to participate in the study. While this led to contrasting initial involvement (see Box 6.1), both were actively involved in subsequent unblindfolded meetings to discuss the findings.

In line with the hands-off philosophy of CCM, Tearfund had very little monitoring data to offer the QuIP research team to aid sample selection. What they did have was a list of villages where there were known CCM facilitators in two eastern districts where PAG had been operating CCM since 2012, and for three northern districts where CoU had been operating since 2011. They also had household survey data with CCM beneficiary names, although this

Box 6.1 Involvement of local partners in the study

Tearfund relied mostly on virtual communication to brief their partners on the study and invite them to participate. A fortuitous face-to-face meeting with relevant PAG staff helped. 'I asked them whether they thought it would work, and for their ideas about what sorts of questions it made sense to ask. I wanted to make sure what we were asking covered the areas where they wanted to see change.'

In contrast this wasn't possible with CoU.

> Whereas I had previously worked with PAG and had met face-to-face on another research project, I was unable to meet with the Diocese of Kitgum staff, and only managed a few very bad reception phone calls and emails. They were on-board for the research to happen, and they gave us sampling information for the churches and people involved. But they didn't really understand the process: for example they couldn't understand why they didn't have to meet the researchers, or introduce them to the participants. As a result they didn't invest in the research process the way they could have. We sent them samples of interview questions to get their input, but they didn't really engage as much as PAG did.
>
> It wasn't until after the field research had been completed that the Diocese of Kitgum central management really came on board. They really 'got it' once we met face-to-face at the unblindfolding meetings and I could explain more. When they heard all the good feedback from the local churches about the experiences they'd had with the QuIP interviews and the participatory events we organized in the villages to explain the project afterwards, they could see how the research provided so much learning. CoU as well as PAG got very involved in building recommendations during the workshop. At the start of the process they were a bit unsure, but they really bought in by the end, and contributed a lot during the workshop. I think CoU were really pleased with how the study went, and really understood afterwards why we'd done it the way we did. We really want this kind of buy-in from our partners, because we don't want the learning to stay with us, we want it to be with them. It's about our partners thinking about what they can learn from this research, and what they are going to do differently.

Source: Charlotte Flowers

was 5 years old and didn't have any addresses. 'Kitgum District in Northern Uganda is where the LRA (Lord's Resistance Army) had been very active and a lot of people had been forced from their homes. It's a poorer area than Soroti District in Eastern Uganda for example. We wanted to have that as a comparison' (CF). The QuIP field research team – recruited through academic contacts and trained over two days in Kampala – was provided with the name of an independent gatekeeper at sub-county level to assist with identifying selected villages. But the team was not provided with the names of CCM facilitators and remained unaware throughout of the identity of the programme being evaluated, of the involvement of Tearfund, and of the names of the two partner churches. They carried with them an introductory letter from BSDR and Makerere University explaining the background to the study, but not naming Tearfund, CCM or the partner churches. The household survey names proved difficult to use: because of the civil conflict and the time elapsed, many people had moved, and therefore the researchers had to use snowball sampling. Box 6.2 provides further information about sample selection.

Box 6.2 Sample selection

The two villages selected in each area were where the number of known CCM participants was greatest. This may have biased selection towards villages that had been more active, although this turned out not to be the case for one of them (Kweyo). In the east the villages were Angopet in Soroti district and Omagara in Serere district. In the north they were Lubene and Kwewyo in Kitgum district.

Two teams of field researchers (one man and one woman) were trained to collect data in each region: one fluent in Atkeso for the eastern villages, and the other in Acholi for the northern villages. Once in each village, they relied on snowball sample selection to identify 12 people for interview, and additional participants for the focus groups. The final sample for each region comprised 24 interviewees per region, plus four focus groups – one each for older and younger men, and for older and younger women.

Overall, the sample size and selection procedure were not sufficient to permit generalization across the more than 100,000 people believed to have participated in CCM in some way over the years. On the other hand, the analyst reported a lot of repetition in statements from respondents drawn from the same village. This may partly reflect a tendency for snowball sampling to include similar people and/or extended family members. The best way to improve on the scope for credible generalization would be to cast more light on the characteristics of the four selected villages relative to the 300 estimated to have participated in CCM.

Domain selection and data analysis

Given the broad and deliberately under-specified goals of CCM, the structure of interviews and focus groups was necessarily broad. It was also influenced by an initiative within Tearfund to develop a standard normative framework for assessing 'whole-life transformation' across its entire programme of activities,

Table 6.1 CCM activities in Uganda

CCM initiatives across Uganda	*PAG in Soroti and Serere districts*	*CoU in Kitgum district*
Building permanent churches	Apprenticeship skills training (construction, electrical repair, citrus trees management)	Child care programmes
Building permanent brick houses		HIV education and care
Infrastructure: clearing roads and digging shallow wells	(Re)training nursery and primary school teachers, and chaplains for PEP schools	'Ot me Gen' (faithful house) training for married couples
School building		
Adult education, including teaching gender equality	Adoption of energy-saving stoves	Formation of savings and loan groups for parents of children with nodding syndrome*
Savings and loan groups		
Environmental protection	Planting trees to reduce flooding	
Improved sanitation		
Support for vulnerable people (orphans, widows, people living with disability, people living with AIDs)	Advocacy and disaster risk reduction training	Energy-saving stoves
New livelihoods (fruit growing, livestock, crops, fish farming, brick-making, motor bike taxi, carpentry, radio/phone repair, shops/kiosks)		

Source: PAG and CoU via Tearfund
Note * Nodding syndrome is a neurological condition with unknown aetiology. In addition to northern Uganda, it occurs in Tanzania and South Sudan.

called the Light Wheel (Tearfund, 2017).[11] This was largely compatible with the domain structure of previous QuIP studies conducted in rural areas, except that 'living faith' was added.

The QuIP analyst referred to secondary data provided by the two partner churches about community initiatives conducted by CCM groups in the two regions, as shown in Table 6.1.

In the absence of specific data on CCM-inspired activities in each village, the QuIP analyst was asked to code causal statements as implicitly consistent with Tearfund's theory of change if it was clear from reading the whole interview that specific actions were triggered by the respondent's participation in CCM activities, even if this was not repeated explicitly in each and every statement. This reading was supplemented by secondary data provided by the two partner churches about support activities conducted in the two regions, also shown in Table 6.1. 'CCM is like a cascade effect, which is part of what makes it so hard to monitor' (CF). Both PAG and CoU appointed their own trainers, who received coaching from Tearfund, and in turn trained facilitators in local churches. While difficult to assess all the individual activities, there was strong commitment to CCM. 'PAG wants all its pastors to be trained in CCM now, it's part of the 2020 Vision that all PAG churches will be facilitating CCM' (CF).

Blindfolding, unblindfolding, and feedback

Blindfolded data collection took place in October 2016. When subsequently informed about the activity being evaluated, the QuIP field researchers reported that they had not guessed the commissioner was Tearfund, assuming instead that it was another NGO (World Vision, to be specific) that had been mentioned in the interviews.

> What was interesting was that the field researchers didn't know Tearfund at all before or during the research, despite CCM being explicitly mentioned frequently. Tearfund wasn't mentioned that often in the interviews, it was more the church or CCM itself that was mentioned – but the programme isn't advertised as Tearfund's, it is run by the partner (CF).

The QuIP field researchers met with PAG and CoU project staff to discuss and verify the initial QuIP findings. This allowed the partners to challenge any initial coding they disagreed with and also began the process of engaging with the findings to build recommendations.

Given the participatory ethos of CCM, it was important to Tearfund that participants in each village should also be able to engage with the evaluation findings, and thanked for their participation.

> In December 2017, I visited each community, firstly to thank them for taking part in the research. Then the main thing was to share the findings and celebrate their success, reinforcing the message that 'you have done this, not us.' I told them we'd been a bit reticent about doing it in a way where we weren't telling them who the research was for. We were concerned people might feel we were deceiving them, but at the same time we wanted people to feel completely free to talk about their whole wellbeing. It turned out that people were very understanding. They said 'yes, that makes sense, because this way we could be more honest with you.' They really understood why we'd done the interviews blindfolded, so that was good (CF).

Participants were furthermore encouraged to give feedback about preliminary findings from the study.

> I facilitated mini workshops where we went through a five-year timeline of CCM and created a pictorial diagram of what had happened and how the community had grown. Then we talked about the findings from the QuIP and dug a little deeper. For example, sometimes participants had mentioned things Tearfund didn't know about, like a small local NGO; and we wanted to verify those kinds of things. So we got some really good stories, which were helpful in understanding some of the results. People shed a bit more light on things that had come up in the interviews, and it was nice to go deeper where we were unsure of some of the results (CF).

The visits also provided the commissioner with a chance to obtain feedback on the research process. This was reassuring: 'The field team really related to the participants; they built up a good rapport, which I think is vital for the study to work well. One lady said to me, "Oh, he was my son – he can come anytime" which confused me at first before I realized she was just saying they really got on, which is brilliant. You really need people who are not only adept at the interview process, but know how to build that rapport' (CF). Holding feedback workshops in the villages fulfilled a double purpose: not only was this commensurate with Tearfund's participatory ethos, it deepened and enriched the study findings. 'Going back to the communities and doing the unblindfolding was great. We'll definitely be doing that again' (CF).

These were not the only follow-up dissemination events that Tearfund sponsored and organized following completion of the report. Findings were also presented at a conference of the Joint Learning Initiative (JLI) on Faith and Local Communities in Dublin in December 2016, as well as being shared internally within Tearfund and disseminated to a wider audience through a summary report (Tearfund, 2017). In November 2017, findings were shared at a workshop in Kampala attended by staff from PAG, CoU, and other CCM partner organizations, with time devoted to thinking through recommendations for doing the programme differently.

> It's about the partners "owning" some of those recommendations. Tearfund was there to play a supporting role. It was really good for our partners to get a secular or non-Christian point of view, an "outsider" view, via the researchers. It's good for the secular and the Christian worlds to meet – as well as the academic and NGO worlds (CF).

In February 2018, the annual 'Facilitator Conference' brought together people trained in CCM, and five days were devoted to presenting the Uganda QuIP report (Tearfund, 2018), talking about what could happen differently and what could be improved, and creating action plans.

Illustrative findings

Reported change across different domains of wellbeing

Individual interviews with 48 people included 10 closed questions about the direction of change in different domains of wellbeing over the last five years. Responses from the 48 respondents (25 women and 23 men) were strikingly mixed. Those living in the two villages in the east were fairly evenly balanced between positive and negative. In contrast in Kitgum district in the north, one village (Kweyo) reported strongly negative change overall, while respondents in the other were on balance positive.[12] This serves as a reminder of how sharply the fortunes of even nearby villages can diverge during the same period. The question most widely answered positively across the whole sample, was 'overall, how do you feel that community relations and decision

making changed over the past five years?' (44 better, 2 worse, 2 same/don't know). This was true even in Kweyo, where it contrasted sharply with mostly negative responses across all other domains. At the other extreme, responses to the question 'overall, how much are you eating as a household compared to this time five years ago?' were mostly negative (10 better, 30 worse, 8 same/don't know). This illustrates the limitation of relying on a single indicator to capture overall changes in wellbeing.

Explicit attribution of change to CCM-related activities

Many narrative statements about the drivers of these changes in people's lives explicitly mentioned PEP (the local name for CCM), Tearfund's two partner churches, and/or associated village level facilitators and faith leaders. Not surprisingly, these causal connections were made most often during discussion of people's 'living faith' and of links with organizations from outside the village. Respondents frequently also made explicit causal connections from personal faith to household and wider community relationships, with some connections made also to livelihood activities and material outcomes. This is shown by Table 6.2 and the illustrative quotations in Box 6.3.

There are only a small number of negative changes explicitly linked to PAG and CoU and none related to CCM/PEP specifically. In one instance an interviewee was asked to stand down from her church position as a result of her husband abandoning her and the church. The other references were to tensions or lack of collaboration between different religious groups. For example, a 32-year-old woman in Lubene commented: 'The different faith groups do not work together. Each one has its own programme and works for its followers. The only time we see them teaching together is when someone dies in the community and the different groups come to pray for the dead. Beyond that, each one works on its own.'

Table 6.2 Frequency counts of explicitly attributed causal statements[13]

	Positive explicit		*Negative explicit*	
	Interviews	*FGDs*	*Interviews*	*FGDs*
Household composition	12	–	–	–
Ability to produce food	9	2	–	–
Ability to earn money	5	–	–	–
How you spend money	6	1	–	–
Household and village relationships	29	3	1	–
Overall wellbeing	19	3	–	–
'Living' faith	40	7	6	4
Links with external organizations	41	4	–	–

Source: BSDR (2017)
Note: Totals from 48 interviews and eight focus group discussions (FGDs)

Box 6.3 Illustrative positive explicit statements

Omagara, 36-year-old woman

'In the past, people didn't care about faith, but it is now a fountain of comfort, peace and hope. Faith leaders also counsel us to be strong and to help us overcome our difficult situations. Yes, in the past, I was a drunkard. From 2010, when I got saved, I became a much more focused person.'

Kweyo, 45-year-old man

'When you belong to a faith group you can have peace of mind because you get consolation in the word of God. Sickness has reduced because we pray, conflicts have also reduced because we have hearts of forgiveness. When you respect the word of God, you don't waste money on alcohol.'

Angopet, 59-year-old woman

'Five years back we were in absolute poverty. Now we are much better in all these respects. Our relations are also good, and we have learnt a lot on health, human relations and our rights from the different programmes from government, CCM, World Vision and even our VSLA (Village Savings and Loan Association) meetings.'

Lubene, 47-year-old man

'The excess food that I produce, I also sell to earn more income for the household. One of the reasons for these changes has been the support that we received from AVSI [Italian NGO], LWF [Lutheran World Federation NGO] and Church of Uganda. As a group member I got training, which increased my knowledge in financial planning and management.'

Angopet, 59-year-old woman

'There is an improvement in our relationship with other people in the village because only a few still drink but the majority are now saved. In addition, when PEP came here, they didn't target only members of PAG. Everyone was targeted, and the message was, "everyone is of value and useful". Out of this message, community relations have improved. We also now speak well. We share problems and we visit each other. In the past it was not the case. There was also theft. If I came out, I would also be beaten. There were many bad people. Further, previously some differences in the village were religious. But now, even when we are building our church, members from other churches, especially the Catholics and Anglicans, invited us 'come to our homes, we will contribute to the building of the church of God'.

Angopet, 53-year-old man

'PEP gave us comprehensive mind-transforming functional education that touches every aspect of life from bible studies to self-help. After PEP came here, there is a lot of behavioural change towards self-help and development.'

'… with the PEP training we got we have started a boda boda [motorbike taxi] business and we now sell firewood as an income generating activity right from October 2014.'

Omagara, 50-year-old man

'What we are doing now is to make manure and put it in the gardens, but the challenge is that there are many trees and one person cannot make all that manure, it needs some support where manure can be made on a large scale. To reduce the impact of drought, I have continued to plant trees, but it cannot be of help if I do it as one person. It needs everyone in the community to do it. So I thank the PAG church that has helped support communities to carry out their activities of planting trees. They support by facilitating transport and providing teaching materials that are used in the community.'

(Continued)

Box 6.3 Continued

Lubene, 43-year-old woman

'Church of Uganda trained me and other community members in making local energy-saver stoves, the church has also supported group savings by training its members but also by providing small startup kits.'

Source: BSDR (2017)

Other drivers of change

The important context for this generally positive evidence of project impact was an abundance of accounts of livelihoods being adversely affected by weather and climate change, with adverse knock-on effects on food consumption and asset ownership. The second most widely raised problem area focused on rising costs, particularly of schooling but also of health care. Hence, what the study documented were often grim stories of people, families, and communities struggling in adversity, in which religion and the support that can be derived from it can be viewed as a coping strategy. This fuller picture is captured more comprehensively by the inductive drivers of change analysis set out in Tables 6.3 and 6.4, with the negative driver data deliberately shown first in order to place the more complex data on positive drivers in this context. Investing in children's education could be viewed as a way for many respondents to offer them a more secure long-term future, but one that entailed high risks and suffering in the short-term. A 47-year-old man in Lubene illustrates this: 'The older children have dropped out of school and they are now helping me with farm work. The reasons for the significant change have been because I spend all my earnings to send my other children to very expensive schools in Kampala. I sold all the assets that I had to pay to put my children in the good schools … I even sold a motorcycle, 20 cows and two oxen.'

The contribution of different external agencies to change

At the end of the interviews, respondents were asked to name key external organizations operating in their area and to rank them according to how much they valued them. The results are reproduced in Table 6.5, and are consistent with the frequency of coded citations in the narrative data.

The number of organizations referred to came as something of a surprise to Tearfund: 'Over 60 different organizations were mentioned. CCM can work in a bit of a vacuum sometimes, not relating to other NGOs and other things that are happening out there, including what the government is doing. It could definitely be better at understanding the context and the different stakeholders' (CF). When combined, Tearfund's two partner churches ranked as the most important positive influence over the households in the sample group by a significant margin (322 references), followed by World Vision (129) and Village Savings and Loan Associations (VSLAs) (83).[14]

Table 6.3 Most commonly cited negative changes and associated drivers of change

Drivers	*Outcomes*						
	Livelihood vulnerability or inability to grow income sufficient for needs	*Decreased material assets and resources and productivity*	*Reduced food consumption and variety*	*Worse community relationships*	*Worse physical wellbeing*	*Worse family relationships*	*Lack of peace/ inner turmoil*
Lack of skills or equipment to improve livelihoods	8	2	1	–	–	–	–
Reliance on market for food	–	6	–	–	–	–	–
Lack of capital/being in debt	7	4	–	–	–	1	–
Lack of employment	2	3	–	–	–	–	–
Climate change/irregular weather patterns	39	48	17	5	4	4	7
Poor soil health	22	21	2	–	–	–	1
Ill health	4	11	2	–	12	1	1
Selling personal assets	3	4	–	–	–	1	1
Higher percentage of income spent on food	–	8	1	–	–	–	–
Destroyed/failed crops or livestock dying	6	12	–	–	–	1	–
Profit margin on business reduced	2	6	–	–	–	–	–
High health care costs	2	17	1	–	1	–	–
Increased cost of schooling and materials	3	30	3	1	2	2	4
Private school fees	1	13	2	–	1	1	–
Anti–social behaviour	–	2	1	2	–	4	–
Robbery/corruption	–	8	1	2	2	–	2
Conflict over land	1	1	–	3	–	1	1
Family breakdown/tensions	2	2	1	–	1	4	3
Individualism in the community	1	2	–	12	–	–	–
Inter–faith tension/fear	–	–	–	3	–	–	–

Source: BSDR (2017)

Note: Totals refer to number of times selected change was cited by respondents across all domains (can be cited in up to six domains across 56 interviews).

Table 6.4 Most commonly cited positive changes and associated drivers of change

Drivers	*Outcomes*											
	Livelihood resilience	*Skills acquisition/ education*	*Increased material assets and productivity*	*Hope in the future*	*Improved wellbeing*	*Improved self–worth and confidence*	*Improved family relationships*	*Improved inter–faith relations*	*Improved community relationships*	*Increased empowerment*	*Changed perceptions*	*Reduced anti–social behaviour*
Moving to cash crops and new markets	15	2	29	2	–	–	–	–	–	–	–	–
Increased livestock rearing/trading	11	1	13	–	–	–	–	–	–	–	–	–
Farming a larger area	4	–	5	1	–	–	–	–	–	–	–	–
New farming methods/training	31	3	31	6	–	1	–	–	–	2	–	–
Livelihood diversification, incl. paid employment	14	5	24	1	–	1	–	–	–	1	–	–
Training in business skills/leadership	–	6	6	1	1	4	–	–	1	6	–	–
Training in advocacy and human rights	–	–	–	–	–	15	–	–	4	8	15	1
Ot me gen training in family relations	–	–	3	1	2	3	14	–	1	1	6	2
PEP (CCM)	9	11	11	9	8	14	10	12	19	24	11	–
HIV training	–	–	–	2	1	2	2	–	1	–	2	–
House improvements and new assets	–	–	10	2	4	–	–	–	–	–	–	–
Membership of savings groups	4	4	17	2	4	4	–	3	18	14	1	1
Faith groups and holistic ministry	–	2	3	1	5	–	2	1	2	2	3	1

Table 6.4 continued

Drivers	*Outcomes*											
	Livelihood resilience	*Skills acquisition/ education*	*Increased material assets and productivity*	*Hope in the future*	*Improved wellbeing*	*Improved self–worth and confidence*	*Improved family relationships*	*Improved inter–faith relations*	*Improved community relationships*	*Increased empowerment*	*Changed perceptions*	*Reduced anti–social behaviour*
Increased church involvement	1	4	1	14	15	13	10	1	19	2	14	21
Becoming a Christian/actively pursuing a Christian faith	–	–	2	12	10	5	8	–	8	4	6	16
Spiritual wellbeing improved	–	–	–	1	3	2	–	–	–	–	–	–
Improved family relations	–	–	–	1	1	–	2	–	–	–	–	–
Better relations with govt, police and law	–	–	–	–	–	1	1	–	4	4	–	–
Interfaith collaboration	–	–	1	–	3	–	–	17	7	–	3	1
External support for development projects	2	3	12	2	5	–	–	–	2	1	–	–
Improved community cohesion	–	1	9	3	3	1	1	1	18	3	3	–
Taking on community responsibility	–	–	–	1	–	10	1	–	1	2	–	–
Increased education/ attainment	1	12	–	11	4	3	1	–	–	–	2	–

Source: Adapted from BSDR (2017: Table 4).
Notes: Totals refer to number of times selected change was cited by respondents across all domains (can be cited in up to six domains across 56 interviews)

Table 6.5 Ranking of external organizations by importance

Organization	*Ranking*							
	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*Total*
World Vision	11	7	5	–	3	–	–	26
Village Savings and Loan Associations	3	2	6	5	4	3	–	23
PEP: local name for CCM	6	5	1	3	1	2	1	19
Church of Uganda: Tearfund partner	13	1	1	–	–	1	–	16
Lutheran World Federation	2	1	6	2	1	–	–	12
National Agricultural Advisory Services	1	4	2	1	2	–	–	10
AVSI (Italian International NGO)	4	3	–	1	1	1	–	10
Pentecostal Assembly of God: Tearfund partner	3	2	1	1	1	1	–	9
Soroti Rural Development Agency (SORUDA)	2	1	1	–	1	1	1	7
Send a Cow	2	3	1	–	–	1	–	7

Source: BSDR (2017: Table 6.1)
Note: The frequencies give equal weight to the rankings of 48 individual respondents and eight focus groups.

Box 6.4 Illustrative quotations on the role of savings groups

Omagara, 60-year-old man

'Being a member of the SACCO [Savings and Credit Cooperative Organization] has also improved my physical and financial wellbeing. Relationships with people have changed because for example in the SACCO where I work we treat people fairly. The knowledge and skills that I have now have greatly improved my wellbeing and this goes together with my faith.'

Lubene, 43-year-old woman

'There is change after the Diocese training on saving. We learnt how to save money … I now see that alone, I could not address the challenges that I was facing in my household.'

Lubene, 47-year-old man

'Most of the trainings came through the church, AVSI and World Vision. All these three NGOs have helped in income generation. Putting people in groups has increased productivity of the group members.'

Source: BSDR (2017)

Village Savings and Loan Associations

VSLAs and other savings groups were widely reported to be a positive driver. CCM was often not the sole or even main instigator of these, but respondents did often link them, for example as a means by which they were able to

respond to church-based encouragement to save. That several respondents reported falling income but increasing assets can also be attributed to the role of savings groups in enabling them to save and buy assets such as goats as security against future shocks. Box 6.4 provides some illustrative examples of how different organizations contribute to promoting savings groups, and how the groups in turn contribute to diverse outcomes.

Discussion and conclusions

Tearfund's pilot use of the QuIP in Uganda provided a rich body of evidence about their faith-based approach to rural poverty reduction. In its own publication (Tearfund, 2017) based on the findings, it picked out five positive drivers of change (faith, self-esteem, relationships, new knowledge, and local savings groups) and two negative (environmental change and school fees). It also highlighted four general insights:

- Changing hearts and minds is vital to impact all aspects of people's lives.
- The local church encourages faith in action.
- Changing weather patterns are restricting progress.
- The QuIP provided excellent evidence of positive impact and its causes.

Not surprisingly, the publication presented a generally positive message about the direction of development in the four villages, as well as Tearfund's own contribution, for example, by highlighting that '52 per cent of households cited CCM as a positive driver of change in the last five years' (19 per cent through PAG and 33 per cent through Church of Uganda) – without framing this statistic in the context of the citations of other organizations, or emphasizing the non-representative nature of the sample. Nevertheless, it did provide a relatively holistic account of the complex combination of drivers of change in four villages. Tearfund was also particularly innovative in finding a range of different ways in which to use the findings, not only to inform external audiences but also for its own internal learning and to feed back to respondents in the four villages studied. In reflecting on the methodological benefits of using the QuIP, the publication picked out six features in particular: alleviating bias through blindfolding, understanding attribution, rigorous coding, accountability to beneficiaries, use of local research expertise, and scope to inform internal learning associated with the supply of evidence on causal drivers of change rather than just their magnitude. This chapter has also documented how it was possible to adapt the QuIP to evaluate a programme with a deliberately open and fluid (faith-based) theory of change, particularly to throw light on the mechanisms by which intended outcomes were being achieved. Furthermore, it illustrated the scope for utilizing the QuIP, through unblindfolded follow-up meetings, as a participatory evaluation approach to support community-based development action.

Of course, it is impossible to be entirely objective in the interpretation of rich qualitative data sets, and subjectivity inevitably also introduces some selectivity

on the basis of the data user's own interests and values. Chapter 10 returns to this issue. Nevertheless, this case study did illustrate that faith-based and evidence-informed approaches to development practice are not antithetical – in other words, there is scope for combining them. The study was not designed to be a piece of research into the efficacy of faith-based organizations in development, but it did provide evidence to illustrate how shared religious values and discourse can contribute to positive outcomes, particularly in a context such as Uganda, where this cultural resource is shared not only within and between development organizations but more widely within society. To take such an analysis further it would be necessary to reflect also on the counter-factual question of how a secular (i.e. not religiously 'faith-based') NGO might have performed in reducing poverty in the same area with similar resources – although of course the resources deployed by Tearfund were also a product of shared values with many of their supporters.

This case study also illustrated how the role of impact evaluation extends beyond the supply of better empirical evidence on what is working and how – important though this is. The introduction to this chapter made clear that Tearfund's motivation in commissioning the study was primarily to promote internal learning and improvement rather than external accountability. However, the chapter also illustrated the way in which the study was able to serve a legitimating purpose by affirming Tearfund's broader theory of change, including its partnership model (James, 2016). Using the typology of approaches to producing and maintaining NGO legitimacy proposed by Thrandardottir (2015) it can be argued that the QuIP study demonstrated the potential to conduct impact evaluation in a way that is compatible with the more democratic and political culture of the 'social change model' of legitimacy, rather than the more functional and technocratic 'market model'. This is also consistent with Tearfund being able to maintain what Gulrajani (2010) describes – and not in a pejorative way – as a more romantic view of development management, as an alternative both to a colder managerial culture or one that is more radical in its critique of global and national power structures. In short, the sustainability and efficacy of faith-based approaches to development is of interest not only in itself, but also as a leading example of the potential to do development differently.

Notes

1. See Copestake et al. (2016: 6) for a discussion of the concept of 'warm glow' in this context.
2. New International Version of the New Testament, Gospel according to John, Chapter 20, Verse 24.
3. Tomalin (2012: 609) is more cautious, concluding that 'further assessments of the characteristics, roles, and activities of all types of non-governmental organisations (NGOs) are needed to assist in the choice of development partners and to test claims of distinctiveness and comparative advantage.'

4. 'A short history of Tearfund', https://www.tearfund.org/en/about_us/history/
5. 'Where your money goes', https://www.tearfund.org/en/about_us/#changing-policies-section. The remaining 20 per cent goes to fundraising (13 per cent) and support and running costs (7 per cent). 'Envisioning' is widely used by Tearfund to refer to 'awakening local church leaders and subsequently parishioners to their God-given mandate for integral mission' (Tearfund, 2018).
6. CCM has been evolving within Tearfund since 1973. It has been funded from a wide range of sources, including some Christian grant-making institutions in the USA and 'integral mission partners' in the Netherlands, Belgium, and Australia. Tearfund has undertaken regular evaluations internally which are required by donors as part of their standard programme cycle, but had not commissioned an external impact study since 2014.
7. Field work under the QuIP study reported here, however, took place in areas not yet covered by this programme.
8. Staff at Tearfund were also aware of the tradition of research at the University of Bath into wellbeing in developing country contexts (Gough and McGregor, 2007; Copestake, 2008; White and Blackmore, 2016). This resonated with its own attempts to develop a normative framework for assessing its work (see below).
9. 84 per cent of the population is Christian (according to the 2014 census) and 14 per cent Muslim. Roman Catholicism was the largest denomination (40 per cent), followed by Church of Uganda (32 per cent), with 11 per cent belonging to Pentecostal congregations (Government of Uganda, 2016).
10. The proportion of poor living in the eastern or northern region has risen from 68 per cent in 2013 to 84 per cent in 2016 (World Bank, 2016).
11. Arranged as spokes in a wheel this comprises nine domains: personal relationships, social connections, participation and influence, emotional and mental health, physical health, material assets, capabilities, stewardship of the environment, and living faith.
12. Subsequent to the research it was found that in Kweyo there has not been as much engagement in CCM as in other places. The CCM process began in the central church while people were still living in the displacement camp (during the Lord's Resistance Army conflict). Once people went home the main CCM members were dispersed and the programme lost momentum. A change in church leadership also meant less backing from the church.
13. Given the transformative aspirations of CCM (encompassing individuals' attitudes and beliefs, social relationships, and material circumstances), a high proportion of the remaining narrative data was coded as 'implicitly' consistent with CCM's theory of change, both positively and negatively. But, being consistent with so many other possible causes, this is hard to interpret, and for this reason frequency counts for implicit coding are not shown.
14. Although mentioned in open interviews, Tearfund itself was not named in this section, but this was as expected, given its approach of supporting local church partners to be the active agents in the community.

References

BSDR (2017) *QuIP Report on Tearfund's Church and Community Mobilisation (CCM): Kitgum and Soroti Region, Uganda*, Bath, UK: Bath Social and Development Research Ltd.

Chadburn, O., Anderson, C., Venton, C. and Selby, S. (2013) *Applying Cost-Benefit Analysis at a Community Level: A Review of its Use for Community-based Climate and Disaster Risk Management*, Oxford: Oxfam/Tearfund.

Clarke, G. (2006) 'Faith matters: faith-based organisations, civil society and international development', *Journal of International Development* 18: 835–48 <http://dx.doi.org/10.1002/jid.1317>.

Copestake, J. (2008) 'Wellbeing in international development: what's new?' *Journal of International Development* 20(5): 577–97 <https://doi.org/10.1002/jid.1431>.

Copestake, J., O'Riordan, A-M. and Telford, M. (2016) 'Justifying development financing of small NGOs: impact evidence, political expedience and the case of the UK Civil Society Challenge Fund', *Journal of Development Effectiveness* 8(2): 157–70 <https://doi.org/10.1080/19439342.2016.1150317>.

Elbers, W., Knippenberg, L. and Schulpen, L. (2014) 'Trust or control? Private development cooperation at the cross-roads', *Public Administration and Development* 77: 713–32 <http://dx.doi.org/10.1002/pad.1667>.

Freire, P. (1970) *Pedagogy of the Oppressed*, London: Penguin.

Gough, I. and McGregor, J. (eds)(2007) *Wellbeing in Developing Countries: From Theory to Research*, Cambridge, UK: Cambridge University Press.

Government of Uganda (2016) *The National Population and Housing Census 2014*, Kampala: Bureau of Statistics.

Gulrajani, N. (2010) 'New vistas for development management: examining radical-reformist possibilities and potential', *Public Administration and Development* 30: 136–40 <https://doi.org/10.1002/pad.569>.

James, M. (2016) *Affirming and Legitimising Partnership Models of Development: A Case Study of Tearfund's Church and Community Mobilisation Programme*, MRes (Global Political Economy) dissertation, Bath, UK: University of Bath, Department of Social and Policy Sciences.

Scott, N., Foley, A., Dejean, C., Brooks, A. and Batchelor, S. (2014) *An Evidence-based Study of the Impact of Church and Community Mobilisation in Tanzania*, London: Tearfund and Gamos.

Tearfund (2016) *Bridging the Gap: The Role of Local Churches in Fostering Local-level Social Accountability and Governance*, London: Tearfund.

Tearfund (2017) *An Introductory Guide to the LIGHT Wheel Toolkit: A Tool for Measuring Holistic Change*, London: Tearfund.

Tearfund (2018) *Flourishing Churches, Flourishing Communities: Church and Community Mobilisation in Uganda*, London: Tearfund <https://learn.tearfund.org/~/media/files/tilz/churches/ccm/2018-tearfund-flourishing-churches-flourishing-communities-ccm-in-uganda-en.pdf?la=en> [accessed 9 October 2018].

Thrandardottir, E. (2015) 'NGO legitimacy – four models', *Representation* 51: 107–23 <https://doi.org/10.1080/00344893.2015.1023102>.

Tomalin, E. (2012) 'Thinking about faith-based organisations in development: where have we got to and what next?' *Development in Practice* 22: 603–703 <https://doi.org/10.1080/09614524.2012.686600>.

White, S. and Blackmore, C. (2016) *Cultures of Wellbeing: Method, Place, Policy*, Basingstoke, UK: Palgrave Macmillan.

World Bank (2016) *Uganda Poverty Assessment 2016: Fact Sheet* [online], Washington, DC: World Bank <http://www.worldbank.org/en/country/uganda/brief/uganda-poverty-assessment-2016-fact-sheet> [accessed 9 October 2018].

About the authors

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Michelle James, MRes Global Political Economy, is an independent consultant specializing in research, strategy, and evaluation in the development sector. Her main interests include partnership models of development, community empowerment and mobilization, and holistic wellbeing measurement. She was lead evaluator on two Tearfund QuIP studies of the Church and Community Mobilisation programme, in Uganda and Sierra Leone.

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of the QuIP across a range of contexts and countries.

Charlotte Flowers, MSc Social Development Practice, is a Design, Monitoring, and Evaluation Officer at Tearfund, with a focus on Tearfund's Church and Community Transformation (CCT) programmes. She supports and advises country staff and local partners in developing and implementing monitoring and evaluation plans, including local participation in collecting and reviewing programme information. She was responsible for commissioning and overseeing the Tearfund QuIP study of CCT in Uganda, as well as subsequent studies of CCT in Sierra Leone and Bolivia.

CHAPTER 7

Harnessing agriculture for better nutritional outcomes in southern Tanzania

James Copestake, Gabby Davies, Marlies Morsink and Martin Whiteside, with Amy Schmidt

This chapter reports on the use of the Qualitative Impact Protocol (QuIP) to evaluate the Harnessing Agriculture for Nutrition Outcomes (HANO) programme. This was implemented by Save the Children (with funding from Irish Aid) in the Lindi Region of southern Tanzania over five years from 2012, to help reduce chronic malnourishment among infants and young children. The QuIP study explored HANO's peer-education approach to promoting behaviour change in key infant and young children feeding practices, and the programme's efforts to strengthen the capacity of local government and civil society organizations (CSOs). The chapter focuses particularly on two innovations to the QuIP: combining household-level data collection with blindfolded key informant interviews of CSO and government staff; and collaborative interpretation of findings after the study was completed. It also highlights how achieving potentially transformational attitudinal change hinged on delivering a package of complementary activities spanning agriculture, nutrition, and gender-aware social development. This is one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.

Keywords: impact evaluation, causal attribution, Save the Children, Tanzania, agriculture, nutrition

Introduction

THE BUNDLING OF INTERVENTIONS is a recurring issue in development practice. Add fertilizer to maize and yields will increase, but only if the crop has enough water. And of course that is only the start. The complementarity of inputs can explain both the magic and the tragedy of so-called green revolution strategies in agriculture: combine all the ingredients in a timely fashion and the combined impact can be transformative; but fail to supply just one and net returns to the whole package can quickly turn negative.[1] This logic also applies to less closely connected inputs or activities: how far do social returns to better education hinge on simultaneously improving child health, for example? And how far does improving child nutrition benefit from promoting new food production and infant feeding practices alongside each other?

http://dx.doi.org/10.3362/9781780447469.007

This chapter focuses on use of the Qualitative Impact Protocol (QuIP) in early 2017 to evaluate such an integrated project. The Harnessing Agriculture for Nutrition Outcomes project (HANO) was implemented by Save the Children in Lindi Region of southern Tanzania over five years from 2012. This section sets out the background to the project, and explains the decision to evaluate it using the QuIP. The next section describes the QuIP study. This provides a further example of the bundling of activities with the goal of achieving synergies, because the study combined standard household-level QuIP data collection with key informant interviews of staff belonging to the project's local implementing partners. We then report on findings, focusing particularly on triangulation across these two levels of data collection. The chapter ends by sharing reflections on both the findings and on the methodological issues raised by the study.

The chapter draws heavily on the final QuIP report (BSDR, 2017), authored by Davies (lead analyst on the study) and Whiteside (lead evaluator). This was supplemented by two key informant interviews conducted by Morsink: with Amy Schmidt, Director of Programme Development and Quality in Save the Children's Tanzania main office, who oversaw commissioning of the study; and with Martin Whiteside, in his first experience as lead evaluator of a QuIP study (interviews are cited as AS and MW, respectively). Both Schmidt and Whiteside also read and commented on an earlier draft of the chapter.

Save the Children and integrated approaches to improving child nutrition

Save the Children is 99 years old, encompasses 28 member organizations, operates in 120 countries, has an income of US$2.1 bn and is a globally recognized brand. Member organizations lead on activities within their home territory and work with other development programmes abroad, coordinated by a central body – Save the Children International, which was established in 2012. Save the Children's mission is to give children around the world a healthy start in life, the opportunity to learn, and protection from harm. It envisions a world where no child dies from preventable causes before their fifth birthday, all children learn from a quality basic education, and violence against children is no longer tolerated. In 2016 over 56 million children directly received support from and participated in activities run by Save the Children, or accessed services provided by its partners. Many millions more were reached indirectly through information, education, and awareness-raising activities (Save the Children, 2016).

Health and nutrition as a programme area, and East Africa as a region, are important foci for Save the Children: 43 per cent of total programme spend is devoted to health and nutrition, while East and Southern Africa receive 31 per cent of total spend (Save the Children, 2016). Save the Children estimates it helped 14.7 million children worldwide through its health and nutrition work in 2016. Child mortality figures almost halved between 1990 and 2015, yet more than 5 million children still die each year, many

from easily preventable causes such as diarrhoea and pneumonia (Save the Children, 2016; *The Economist*, 2018).

The causal pathways linking food production, food consumption, diet, and children's nutrition and health outcomes are hugely complex (von Braun, 2018; McDermott et al., 2013). They are shaped by diverse, changing, and risky environmental contexts, as well as by the intricate power and work/care relationships within families and with other actors. Child-oriented NGOs such as Save the Children have grappled with designing activities to address these issues in a holistic way: how to balance sensitivity to local variation with the benefits of scale that can be achieved through standardization, and how to weigh grass-roots ownership against the need for some top-down coordination to prevent duplication and waste (Jaenicke and Virchow, 2013). Save the Children's Harnessing Agriculture for Nutrition Outcomes (HANO) project in Tanzania is just one of many examples that could be selected. At the time HANO was being launched, Save the Children (2012) had this to say about the connection between agriculture and nutrition (emphasis added):[2]

> Making the food system work to improve nutrition means *more than simply increasing production*–more food does not automatically mean better nutrition. The real challenge is to improve the quality, availability, utilization, affordability of and access to food. However, the potential of agriculture initiatives to translate into better nutrition outcomes has been largely assumed and often overlooked. Improvements can be achieved through agriculture by considering a number of factors, including: investments in smallholder farmers, assessing the functioning of local markets and the availability of affordable nutritious food, focusing on women farmers, *boosting nutrition education*, investing in better research, considering the impact of agriculture on health (Save the Children, 2012: 49).

An important corollary of this view is the need for continuous and grounded research into what works.

> The potential of agriculture to improve the nutrition of children and their families is not yet fully explored, fulfilled or prioritised. Studies show that when improved nutrition is made an explicit objective of agricultural programmes they can lead to increases in the quantity, nutritional quality and affordability of the food families eat … More research is needed to provide evidence on the impact of agricultural interventions on rates of malnutrition, on models of best practice and on strategies for implementing programmes at scale. When agricultural interventions fail to take nutrition into account, an opportunity is lost to get the maximal return on investment (Save the Children, 2012: 49–50).

The Harnessing Agriculture for Nutrition Outcomes (HANO) project

HANO in Tanzania aimed to reduce the number of chronically malnourished children under two years old in target communities of Lindi and

Ruangwa Districts of Lindi Region by 10 per cent over five years (2012 to 2017) using an integrated package of agriculture and nutrition interventions. It was implemented across the two districts in collaboration with two district level NGOs: Lindi Support Agency for Welfare (LISAWE) and Ruangwa Organization for Poverty Alleviation (ROPA). The project aimed to reach 14,000 infants and children, along with 51,190 women of reproductive age (15–49 years). To do so, it set out to collaborate with 120 care groups (each with an average of 15 members), 80 community resource persons, 40 village-level workers (in agricultural extension, community development, and health promotion), 50 small-scale salt producers, eight district-level government staff, and five civil society organizations (CSOs). In addition, the project also targeted more than 400,000 people indirectly by disseminating information via government radio messaging on nutrition, food production, processing, and preservation.

An additional objective of the project was to build local capacity by working in an integrated and participatory way with selected government ministries, networks, the two partner NGOs, five CSOs and communities, including eight members of Partnership for Nutrition in Tanzania (PANITA) in the two districts to influence planning and budgeting processes.[3] The project also had a wider dissemination objective (see Table 7.1).

HANO adopted a peer-support approach to promote behaviour change in key infant and young child feeding practices. A mid-term assessment

Table 7.1 HANO project objectives and intended outcomes

Objective	*Intended outcomes*
To increase food and nutrient intake for infants and young children (0–23 months) and women of reproductive age	Increase the share of infants (up to six months) who are exclusively breastfed by 12%. Increase the share of toddlers (aged 6–23 months) consuming appropriate complementary food by 6%. Increase the percentage of women and men able to process and preserve foods, and to plan for the lean season to ensure consumption of sufficient quantities and diversity of foods throughout the year by 50%. Increase the diversity of foods available for children and women of reproductive age at household level throughout the year.
To increase the capacity of local district government and CSO staff to deliver nutrition-sensitive agriculture programmes	District developmental plans integrate nutrition components effectively. Local CSOs have increased capacity to deliver a package of nutrition interventions.
To document and disseminate widely project achievements among local and national stakeholders	Evidence is generated on impact of integrating nutrition and agriculture interventions. Achievements/lessons learned are disseminated to relevant government and development partners, and CSO stakeholders, throughout project duration.

Source: BSDR (2017)

was carried out in 2015, two-and-a-half years into the project, and observed sufficient progress to warrant its extension into a second phase, from 2015 to May 2017. In addition to continuing with activities deemed successful, this second phase also incorporated new activities and streamlined others. In the first phase, the project adopted the 'mother-to-mother support group' (MMSG) approach, but from April 2016, these groups were refashioned to fit the 'care group' (CG) approach, alongside establishment of new care groups. Staff regarded the CG approach as a form of action research, seeking ways to improve it as implementation proceeded.

The MMSG approach was based on training groups of 30 women, each of whom was then encouraged to reach out to another 10 women and to visit them at least four times to discuss maternal and infant and young child nutrition (MIYCN) issues. On the nutrition side, women received 12 consecutive days of two-hour classes about eating during pregnancy and while breastfeeding, meal preparation and feeding of children aged 6–24 months, sanitation and hygiene, care of a sick child, and growth monitoring. The idea was that once mothers saw evidence of weight-gain in their children after feeding them balanced meals as recommended in the 12-day training, they would not only continue these practices themselves but be motivated to support other women in adopting them. On the agriculture side, groups were given a communal solar dryer for the preservation of fruits and vegetables, and received training in cultivating kitchen gardens and fruit crops. After the 12-day training, mothers were encouraged to join a Village Community Bank (VICOBA) so they could use savings and lending to sustain the MMSGs.

The CG approach adopted from 2016 was based on training groups of up to 15 people, both women and men, each of whom again committed to visit and train at least 10 neighbouring women and their families on what they had learned during the training, using teaching aids. The original group of 15 participated in interactive training sessions, intended to be motivating and to encourage knowledge sharing. Participants made commitments regarding new behaviour they would personally adopt. The meetings were also an opportunity to share information on any births or deaths, thereby strengthening reporting on vital events to district and health authorities. As with the MMSGs, the CGs received training in maintaining individual kitchen gardens and communal gardens, as well as rearing small livestock. Members were also encouraged to join the VICOBA.

There were two main differences between the MMSG and CG approaches. First, MMSGs included only pregnant and breastfeeding women, whereas membership of CGs was also open to men. Second, the MMSGs were not provided with any formal guidelines or tools for monitoring and follow-up, whereas the CGs used a standard training and supervision toolkit, along with guidelines for monitoring, follow-up, supervision, and reporting. The CG approach was much more structured than the MMSG approach, with additional layers of supervision and coordination. From 2016 there were 33 villages participating in the project in Lindi Region: 19 in Lindi Rural

District and 14 in Ruangwa District. Each typically had two or three CGs (though some villages had more than three, depending on demand), for a total of 87 CGs and 1,300 care group volunteers (CGVs). Each village had a promoter selected from among the CGVs, who was paid an incentive and provided with a bicycle, and who was responsible for carrying out monthly supervision of the CGVs. These 33 promoters were in turn supervised by six supervisors, three of whom were staff from LISAWE and three from ROPA. Finally, there was one coordinator each for Lindi Rural District and Ruangwa District, who prepared monthly reports on the project, based on checklists from all levels of supervision.

Choice of impact evaluation methodology

Save the Children has a strong culture of engaging consultants to conduct external evaluations, with its Monitoring, Evaluation, Accountability and Learning (MEAL) team at any time managing 20 or more projects. Schmidt explains: 'Aside from the objectivity they bring, outside evaluators are hired because we don't have the staff to analyse the quantities of data or to write up reports. It's a capacity and time issue, and it's rare for a senior full-time staff to be dedicated to a single project' (AS).

However, the decision to use the QuIP for the end-of-project evaluation of HANO represented something of an about-turn for Save the Children, as HANO had originally been designed as a randomized controlled trial (RCT), starting with a baseline survey in 2012, and with plans for an end-line survey in 2017 to measure changes in a battery of proxy indicators for project objectives. Two issues prompted the decision not to conduct the end-line survey as planned. First, the combination of high turnover in project staff, and difficulties in keeping track of which intended beneficiaries belonged to the different groups (partly as a result of redesign of the project), cast doubt over the feasibility of distinguishing between 'treatment' and 'control' samples. Second, results from the 2015 Tanzania Demographic and Health Survey (TDHS) indicated an impressive (in the order of 15 per cent) reduction in stunting across the whole of Lindi Region since the previous survey in 2010. This threw doubt on the value of another expensive survey, with limited prospects for being able to attribute falls in stunting and other indicators specifically to HANO, particularly as it was not the only nutrition-oriented programme operating in the region during this time.

Some months earlier Save the Children had hosted an internal webinar about the QuIP, and the Tanzania country office responded by proposing its use alongside the RCT in order to gain more insight into causal mechanisms underpinning the project. This was reinforced by the TDHS results, and when the complexity and budgetary cost of conducting the end-line became clear, the country team suggested abandoning it altogether and relying on a QuIP study instead. However, attitudes towards the QuIP within Save the Children were mixed. Some evaluation staff were more familiar with quantitative

research methodologies, and wary of approaches that could not offer precise estimates of impact backed up by power calculations and confidence intervals. But more senior staff emphasized the need for greater insight into causal mechanisms, and drew on experience with other qualitative approaches, including *Most Significant Change* and *Outcome Harvesting*. The decision to rely on the QuIP also received the support from a senior Irish Aid staff member in Tanzania. They had taken an active interest in the HANO project from the start and had also attended a presentation about the QuIP, in the context of a wider push within Irish Aid to encourage innovation in the way its partners reported on the impact of the projects funded. Meanwhile, the fact that the HANO project had been reformulated mid-stream and had experienced a high turnover of project managers reinforced the case for adopting a more flexible approach to its evaluation.

The QuIP study

Integrating local implementing partner perspectives into the study

Design of the QuIP was influenced by the fact that HANO's objectives combined reducing stunting at the household level with strengthening local civil society and government capacity, as indicated in Table 7.1 (see section 'The HANO project', above). This stimulated discussion between Save the Children and BSDR about how to combine assessment of both. For the institutional evaluation it was agreed to try out blindfolded interviews with a small number of CSO and local government officials, limiting the role of Save the Children to identifying relevant people to interview. All interviews were in fact conducted blindfolded: 30 individual interviews and four focus groups with intended beneficiaries, six interviews with government staff, and four interviews with civil society partners. An anecdote about blindfolding in practice is captured in Box 7.1.

Box 7.1 Feedback on the feasibility of blindfolding

There was some doubt within Save the Children that partner agency-level interviews with key informants could really be blindfolded. Just how effectively this 'veil of ignorance' was in fact achieved was demonstrated part-way through the field work. The HANO project manager hadn't fully grasped the nature of the blindfolding, and asked the local nutrition officer whether someone had met them to interview them about HANO – to which the nutrition officer replied: 'No, nobody from HANO has come to talk to me.' The project manager then escalated the matter internally (being concerned that the researchers were not following the list of respondents supplied) only to be informed that the interview had indeed already been conducted. At the 'unblindfolding' workshop both the interviewing team and institutional partners also confirmed that they had not been aware that HANO was the subject of the evaluation. The workshop was a chance for the interviewers, and agency-level respondents to meet face-to-face, and you heard exclamations like, 'Oh, it's you! You came to interview me.'

Source: Amy Schmidt, post-evaluation interview.

A second and equally important component of the institutional assessment was the 'unblindfolding workshop' referred to in Box 7.1, at which the QuIP research team could share and discuss findings with representatives of all the agencies involved in implementing the project. This was held in Lindi on 25 May 2017 and was attended by a total of 25 people, including Save the Children project and country office staff, district government officers (for health, nutrition, and agricultural extension), two people from each CSO partner, and a representative from Irish Aid. Intended beneficiaries and Village or Ward Executive Officers were not included.[4] The workshop provided a time and space to discuss the findings from the QuIP and what recommendations these could lead to. For example, Schmidt reports 'I sat in on one of the small groups during the workshop, and people were discussing the need to better integrate the agriculture side of the project. That's something we're going to try to address immediately in the next iteration of the project, which we're already starting on.'

Sampling

The sample frame for the individual interviews and focus groups comprised 6,450 individuals from Lindi Rural District and 2,325 individuals from Ruangwa District, who were registered as participants in agricultural groups, mother-to-mother support groups, care groups, or as 'neighbouring women' (NW) across the HANO villages. The budget permitted 30 interviews plus four focus groups to be conducted at the household level, split evenly between the two districts (see Table 7.3). Two villages were first purposively selected in each district: one where health workers were perceived to be more effective and one less. Otherwise the villages were understood to have broadly similar characteristics and to have received similar interventions. Within each village, participants were selected from village lists to achieve a mix of five pre-determined beneficiary groups, as indicated in Table 7.2.

Table 7.2 Household-level interview sample

District	*Ward/Village*	*CGV (Women)*	*NW (Women)*	*MMSG (Women)*	*AW (women)*	*AM (Men)*	*Total*
Lindi	Kiwalala (Mahumbika)	1	2	1	1	2	7
Lindi	Nyangamara (A)	2	1	2	2	1	8
Ruangwa	Mnacho (Manokwe)	1	2	1	1	1	6
Ruangwa	Nangumbu	2	1	2	2	2	9
	Total	6	6	6	6	6	30

Note: CGV: care group volunteer; NW: neighbouring women group member; MMSG: mother-to-mother support group member; AW: agricultural group member (women); AM: agricultural group member (men)

Table 7.3 Location and participation in the four focus group discussions

District	*Ward/Village*	*Women*	*Men*	*Total*
Lindi	Kiwalala (Mahumbika)	7	6	13
Ruangwa	Nangumbu	6	6	12
	Total	13	12	25

Four focus group discussions (FGDs) were carried out, two in each district, and two each with men and women. These were intended as a cross-check on the individual interviews. Discussions were differentiated by gender and location, and were conducted away from respondents' own homes, prompting more general responses. Indeed, the interview team remarked that the dynamics of the FGDs were quite different from those of the individual interviews: 'There was a very marked difference between the FGDs compared to the individual interviews: where a respondent raised a point, it assisted the other members to ponder, reflect and comment, hence enriching the experience.'

Household-level interviews and focus groups were supplemented with key informant interviews at the institutional level. Box 7.2 lists the institutional interviewees who participated in the individual interviews and one focus group.

A standardized interview schedule (used for household interviews, key informant interviews, and focus groups in order to facilitate integrated analysis) explored what changes respondents had experienced over the past five years across domains that corresponded to HANO's objectives and desired outcomes as set out in Table 7.1 (see section 'The HANO project', above). The domains were health, farming and income, food consumption, who eats

Box 7.2 Agency level respondents

Lindi

- Adviser to the District Executive Director and Economic Development Coordinator, with two years of experience in Lindi.
- Acting District Agricultural Officer, with 31 years of experience in Lindi.
- District Nutrition Coordinator, with nine years of experience in Lindi.
- LISAWE focus group with five experienced members of the nutrition committee.
- LISAWE Education and Finance Director.

Ruangwa

- District Executive Director, with seven months of experience in Ruangwa.
- District Agriculture Information Officer providing link between Executive Director and Extension Workers, with seven years of experience in Ruangwa.
- District Nutrition Coordinator and adviser to the District Nutrition Committee, with seven years of experience in Ruangwa.
- Member of Baraka CSO *Kilwa Njia Nane* (nutrition and family welfare group).
- ROPA Chairman, Treasurer, Coordinator, M&E officer, and Field Officer.

what and when, spending and saving, gendered family relations, community relations and development context, and overall wellbeing. Although not explicit in the project's objectives, gendered family relations was added as a domain at the request of Save the Children staff.

Findings

Household-level perceptions

Individual responses to closed questions about the overall direction of change in different domains of their life over the past five years were overwhelmingly positive (BSDR, 2017: 22). Many explanations for this were incidental to the HANO project, including general improvement in farming, income, spending, and saving. But asked to name the most important external agencies driving change, Save the Children was by far the most often named.[5] This was borne out by attribution analysis of answers to the open-ended questions, which revealed many explicit positive references to HANO, reinforced by an even larger number of implicit references (see Table 7.4). In contrast, only three causal statements were explicitly negative: two respondents mentioned groups in their village that they did not attend, and one mentioned receiving small tree seedlings that were yet to give fruit or had died.

Turning to positive drivers of change, Table 7.5 lists the inductively coded links that were most frequently cited.[6] Presenting the data in this way provides only a superficial overview of the narrative data, but it does highlight some of its general features. First, it confirms that respondents perceived nutrition-related training to have been important and consequential in improving the diet and health of them or their children: seven of the eight top-ranked claims

Table 7.4 Attribution to HANO of positive and negative changes by outcome domain

Outcome domain	*Positive changes*			*Negative changes*		
	Explicit	*Implicit*	*Other*	*Explicit*	*Implicit*	*Other*
Health	8 (1)	11 (3)	1 (1)	–	1 (0)	4 (1)
Farming and income	5 (1)	12 (3)	19 (3)	1 (0)	2 (0)	6 (0)
Food consumption	1 (0)	16 (3)	12 (2)	–	5 (2)	7 (4)
Who eats what and when?	10 (2)	18 (2)	1 (1)	–	5 (0)	1 (0)
Spending and saving	1 (1)	11 (4)	19 (3)	–	1 (0)	3 (0)
Gendered family relations	2 (0)	15 (4)	13 (0)	–	1 (0)	2 (0)
Community relations	6 (2)	25 (4)	–	2 (0)	4 (0)	1 (0)
Overall wellbeing	1 (0)	19 (4)	–	–	–	1 (0)

Source: BSDR (2017)
Note: Numbers indicate how many of 30 respondents (and in brackets, how many of four focus groups), reported at least one driver of change in the corresponding impact domain.

Table 7.5 Most frequently cited causal links between coded drivers of change and outcomes

Drivers of change	*Outcomes*	*Count*	*Rank*
Received education about nutrition for pregnant and lactating women	Broke taboos about pregnant women's eating habits	27	1
Received education about child nutrition	Improved diet of children	24	2
Received education about breastfeeding children up to 6 months	Breastfed an infant aged 0–6 months	24	3
Community members working together	Better community relations/ social cohesion	22	4
Received education about nutrition for pregnant and lactating women	Improved diet of pregnant and lactating women	21	5
Received education about diet and nutrition	Improved diet	19	6
Received education about diet and nutrition	Improved health	18	7
Received education about child nutrition	Improved health of young children	16	8
Able to preserve food for hunger season	Have food all year round	15	9
Trained/understand how to store crops properly	Have food all year round	15	10
Received more advice/training	Improved standard of life/ wellbeing	14	11
Community groups work for development	Better community relations/ social cohesion	14	12
Price paid for crop harvest increased	Increased income	14	13
Received education on gender roles	Men involved in child care	13	14
Community groups work for development	More community groups in the village	13	15
Husband and wife make decisions together	Women have more voice	13	16
Give extra food/5 nutrients to young children	Improved diet of children	12	17
Increased income	Increased ability to save	12	18
Grew more/higher yield of crops	Increased income	12	19
Started growing fruit and vegetables	Improved diet	9	20
Able to preserve food for hungry season	Increase yield of crops (to sell)	9	21

Source: BSDR (2017: 30)
Note: The 'count' indicates how many times the causal link was cited across all household-level interviews and focus group discussions.

referred to such links. Second, although less frequently cited, several statements about improvements in farming, income, and food security were made (those ranked 10, 13, 19, 20, 21). Third, another cluster of statements highlighted improved community relations (those ranked 4, 12, 15) and gender relations (14 and 16). This was consistent with a theory of change based on (a) synergy between the agriculture and nutrition activities, mediated by (b) increased

Box 7.3 Selected quotations about the 'gender relations in the family' domain

27-year-old woman and mother-to-mother group member

'We decide together on what we should cultivate, what we should buy and what we should eat in the family, unlike before when my husband was the only decision maker of these things. The main reason for this change is information and education we get in our groups. My husband helps in taking care of the child and he even takes our child to the clinic, unlike before where he didn't do this, and the reason for this change is education and these changes are good.'

Young women's focus group

'In general, before men were the ones who made decisions in the household, but nowadays due to education given, men involve their women in making decisions as there is equality, hence even in decisions on what crops to cultivate women are involved in making these decisions unlike before. There is a change in the work women and men do, as nowadays any woman or man can do any work and nowadays even men are involved in taking care of children unlike before when they didn't do this, and this is caused by education given.'

27-year-old man and agriculture group member

'There is a huge change on deciding what crop we should cultivate, as before I was the only one who decided but now I decide together with my wife. Also, nowadays I help my wife with some chores, something I didn't do before. Before men did not involve themselves in taking care of babies/children but now we take care of our children also and these changes are good'

Source: BSDR (2017: 41–2).

family and community cohesion, based on (c) a stronger sense of common purpose, and catalysed by (d) the training and group activities promoted by Save the Children, along with (e) a general improvement in wellbeing within the region.

Many detailed quotations could be drawn from the data to elaborate on this overview. For example, Box 7.3 provides three statements that highlight the importance of the gender dimension. The first two (both from women) highlight how women are having more say in crop cultivation, and attribute this explicitly to education and training provided in groups. The third confirms that the message is also being internalized by at least some men.

Agency level feedback[7]

The holistic view of a virtuous cycle of rising farm income, gender-sensitive training, community cohesion and child-focused changes in diet, could be explored further by triangulating it against data from the key informant interviews. These also confirmed an overwhelming sense of progress during the last five years, as well as a positive outlook on the future. District nutrition coordinators reported significant positive progress on stunting, hospital deliveries, countering harmful traditional beliefs, and maternal/baby/child nutrition. They highlighted community health workers, who lived in the

community and were supported by CSOs, as driving this. Wider explanations for progress included increased ownership of mobile phones, improved transport and market access, favourable weather, increased enthusiasm for learning, openness to change among the population, and closer relations between government officers and communities. There was broad consensus that cash crops and, to a slightly lesser extent, food crop quantity and diversity had increased and were continuing to increase. There was a significant increase in vegetable consumption from a low level, and improvement in feeding infants, young children, and pregnant/lactating women, although there were still some households that had not changed. Change in nutritional behaviour was primarily attributed to sustained and consistent awareness campaigns delivered by a range of organizations.

Household and agency level responses also diverged on some issues. There was more emphasis from government on the importance of livestock (from chickens to cattle), and one agency respondent highlighted the need to find solutions to conflict between farmers and pastoralists – an issue that wasn't highlighted at all in the household-level data. In contrast, the key informant interviews with agencies did not pick up on increased participation of men in relation to family nutrition. However, they did confirm the increase in community level activities. For example, one of the CSO representatives reported: 'these days men and women sit and contribute equally in village meetings so gender parity is assisting breaking down barriers which were stifling.' Key informants at the agency level made more of increased vegetable and fruit production, whereas household-level respondents confirmed increased production and consumption of vegetables, but rarely mentioned fruit. Nutrition coordinators reported progress in encouraging solar dryers for preserving vegetables (a component of the HANO programme) but dryers were not mentioned explicitly by the intended beneficiaries. In contrast, they mentioned improved pesticide use for crop storage, but this didn't come up in any of the agency level interviews.[8]

Interviews at the two levels identified broadly the same set of external agencies operating in the two districts. But the agency key informants emphasized Save the Children less. The difference may be due to more overt Save the Children branding of their work in the communities (e.g. with T-shirts worn by volunteers). The lower attribution to both LISAWE and ROPA is probably because each has a smaller geographic focus; LISAWE was mentioned by four key informants but by no individuals or focus groups at the household level, for example. It seems likely that if HANO staff were working within the LISAWE catchment area they presented themselves as Save the Children, rather than as LISAWE. The outcomes identified by institutional interviewees were generally positive. However, lack of progress was identified for conservation agriculture, which was considered to be too much effort for farmers compared with the benefits. Progress on savings was mixed, with some saving groups considered to lack capacity.

Interviews at both levels similarly highlighted the importance attached to promoting learning and behaviour change, particularly in relation to nutrition practices. There was also agreement on why this was working: openness to change within the communities, and close collaboration and consistent nutrition messaging among external agencies. The benefit of 'working together' was a recurring theme.[9] It mostly referred to government working with international agencies and CSOs, but also within each community. CSO representatives felt that they had been able to 'prove themselves' and were now more accepted as partners by government. One noted 'there has been great cooperation between us and government; it took a long time before they understood us'. A district level economic adviser noted in relation to agriculture, 'Save the Children and Aga Khan Foundation have played a significant role in initiating these changes and they have been effective as they work with government officers at the village level'. Rather than handing over to international NGOs, government consider that they are in partnership and are indeed building on what NGOs are doing: '14 project villages were taught new skills by Aga Khan Foundation but we have added six more to motivate other farmers and expand the project area and thus increase production'. Individual and focus group interviews also confirmed more positive engagement by government officers with communities. Linked to this was a shared emphasis on individual and community learning (particularly in relation to nutrition), rather than distribution of physical inputs.[10] A district community development officer commented, 'overall government messaging has had a huge impact on the way people have understood the role of government as an enabler not as a charity'.

Outcome of the unblindfolding meeting

After an initial briefing and question-and-answer session, participants broke into four topic groups to discuss working with CSOs as partners, methods for engaging with community members, combining agriculture and nutrition in one project, and working with the local government administration. Within these areas, discussion focused on looking at what had worked well, what had not worked well, and recommendations. Table 7.6 provides a summary of points that emerged.

Overall findings[11]

The QuIP study documented a very high level of self-reported improvement in knowledge and understanding of the nutritional requirements for infants (0–6 months), young children (6–24 months), pregnant women, and lactating mothers, in line with HANO's objectives, and among women and men of different ages. Respondents also reported having made corresponding changes in food preparation and consumption practices. This included a greater variety of food being eaten by all the family (including more fruit and

Table 7.6 Summary points from the HANO unblindfolding meeting

What worked well	*What did not work well*	*Recommendations*
1. The CG model was successful, enabling easy access into the community, building local participation and capacity, and fostering improved gender relations by involving both men and women 2. The link between the project and government clinics was well structured and well received by participants It increased the capacity of nurses to provide nutrition education 3. Combining nutrition and agriculture was successful, with the community able to produce what they needed for improved nutrition; and nutrition messages increased the demand for agricultural knowledge 4. Increased food preservation and nutrition knowledge was successfully adopted	1. Operating HANO in only selected villages was a disadvantage, as issues affect the wider community 2. The solar dryers were not appropriate for villages that grow vegetables all year 3. Branding and visibility was biased towards Save the Children and not the CSO implementers 4. CSOs lacked operating funds and access to independent sources of income, to reduce dependence on donors 5. Problems with the mother-to-mother approach: unclear which trainers would receive allowances; the 12-day cooking and learning sessions where participants brought food were burdensome and deterred involvement of men 6. The 'no pesticide' policy of Save the Children reduced vegetable production, with promotion of traditional control methods being introduced too late 7. Lack of participation by some extension officers in monitoring and supervising implementation 8. Poor knowledge (sometimes) of the connection between agriculture and micro-nutrients for better nutrition	1. Nutrition and agriculture activities should be integrated 2. The CG model should continue 3. It could incorporate income-generating activities through savings groups 4. District level authorities and CSOs should participate in project formulation 5. The project should extend to cover the whole district and other districts 6. Need to strengthen communication between district and local government staff 7. Bottom-up planning and follow-up can help to identify and resolve problems early (such as those encountered with the mother-to-mother approach) 8. CSOs and extension officers should be involved more closely with the comprehensive council health plans 9. CSOs need more funding to cover their costs and should be less dependent on INGOs 10. There should be improved CSO branding during projects to strengthen their status in the community beyond the project lifespan 11. Ensure messaging in the community reinforces ownership of project outcomes 12. Save the Children should review its 'no pesticide' policy

Source: BSDR (2017: 64)

vegetables), improved cleanliness of food preparation, use of multiple food groups, reduced vegetable cooking times, and use of an iodine supplement. These changes were attributed to education and training, and were confirmed by government and CSO staff. Respondents highly valued opportunities to learn, and did not mention any desire for 'hand-outs'. Save the Children (and by implication HANO) was the organization most associated with this learning, but attribution is complicated by the involvement of a wider range of organizations, including community health workers, health clinics, the Aga Khan Foundation (AKF), LISAWE, and ROPA.

In parallel with nutrition knowledge, household-level respondents cited new knowledge and changed practices in relation to agriculture. This included growing vegetables in kitchen gardens, food and cash crop diversification (particularly into sesame), cultivation of larger areas, and improvements to crop storage. These changes were linked to a demand to consume more and a greater variety of food, including having increased income to buy food. A minority of respondents reported less positive change and continued food insecurity; and few household-level respondents referred to improvements in keeping livestock or fruit tree cultivation, despite reports by some government interviewees. The main drivers of agriculture change cited were new knowledge, with a strong positive association with HANO and to a lesser extent AKF, government agricultural officers, and others. Other reported explanations included favourable weather and improved crop prices.

Respondents also reported significant changes in intra-household gender norms: more joint decision-making between husband and wife on household expenditure and crop growing; more involvement of men in childcare; and women being able to do a wider variety of activities, including involvement in saving groups. Both men and women considered these changes to be positive, and they were attributed to better education. This was not generally explicitly identified with HANO; rather it seems to have been driven by diverse and mutually reinforcing messages and examples from different sources, including role models in the HANO groups and advice relating to childcare and nutrition. Agency level respondents also reported more active involvement of women in community meetings, although this wasn't mentioned by household-level respondents.

Household-level respondents did, however, report changes in inter-household relationships: more reciprocal learning, more community groups, greater social cohesion, more community level collaboration, and confidence in being able to solve shared problems. They also noted increased advice and support from government officers. CSO and government respondents reported good working relations, with government more willing to work with them and to include education/advice activities within government budgets to enable sustainability when international funding ended.

Overall, improved social relations within and between households were contributing to a generally favourable development context, reflected in respondents reporting overall improvements in wellbeing and expectations of future improvement. This suggested a virtuous cycle, with different

components reinforcing each other. It seems likely that the various groups promoted under HANO contributed to these wider improvements, although the attribution was less explicit than changes in nutrition and agriculture. Some interventions were identified that hadn't worked, and local government staff identified organizational factors that could be improved, but overall the negative feedback was minor and heavily outweighed by the positive.

Neither household- nor agency-level respondents volunteered much information on preferred learning methodologies or community approaches (e.g. CGs vs. MMSG groups). In this respect, QuIP's blindfolded interview approach did limit the opportunity to probe more deeply into specific HANO project approaches with respondents. Care groups were, however, identified as an effective community engagement strategy by participants in the unblindfolded feedback workshop.[12] The existence of a virtuous cycle can also be positively linked with the HANO strategy of linking agriculture and nutrition, food supply and demand, most clearly in the links made between growing and eating vegetables. One important positive contributor to the synergies achieved was consistent nutrition messaging by different organizations.

Methodological reflections and conclusions

Given somewhat guarded initial expectations it is not surprising that the reaction of Save the Children staff to the study was mostly positive. Schmidt recalls, 'I was floored by the results and the changes the QuIP was able to pick up on and document, as per testimonials from the community. I was floored. Especially as I would say my expectations were quite low in terms of what impact we could hope to see, given inconsistencies in the design and implementation of the project.' With the project due to end, the study offered immediate feedback to staff in Lindi Region about activities to persist with, to review, and to stop. Save the Children also had similar projects starting up in two other regions (Dodoma and Singida Regions) and was able, according to Schmidt, to take some of the evaluation findings and apply them to the work on these new projects. 'We can understand what worked well and what did not work well with HANO, and make appropriate changes to the design of these new projects.'

Immediately following the unblindfolding workshop, Irish Aid organized a debriefing meeting about the study and the QuIP methodology in Dar es Salaam, inviting UNICEF, PANITA, and the heads of cooperation from various embassies. Hence feedback from the study fed quickly into wider debates over linking agriculture and nutrition activities in Tanzania. Schmidt acknowledges the willingness of Irish Aid staff to take a risk with a relatively unfamiliar methodology: 'I really tip my hat to Irish Aid for its interest and curiosity, and for promoting this amongst its nutrition partners as an innovation to be thinking about in the evaluation space.' At the same time, the study left Save the Children with some unanswered questions. Irish Aid had suggested inclusion of a gender domain in the interview schedules, and this had helped

draw out the unexpectedly strong and positive story of changing gender norms (illustrated by Box 7.3). Schmidt linked this to the extra challenge of conducting interviews blindfolded. 'If there are unanticipated outcomes of the project and the research team does not know that, they do not know to probe the responses further… This left us scratching our heads: it was interesting, but "how" and "why" did that happen, given that it was not a purposeful intervention on the gender front?'

A more general point is that the depth of insight that can be gained into causal drivers of change (unexpected or otherwise) depends on the experience, motivation, initiative, and training of field researchers. This is one of a number of methodological points made by Martin Whiteside in a memo reflecting on his experience leading the QuIP evaluation, written after the study was completed. Box 7.4 reproduces the relevant part of this.

Data quality assurance is an issue for all field research. But in the case of qualitative research it goes beyond being a problem of controlling the measurable quality of known outputs. With HANO there is a particularly complex causal story to unravel. Accepting for the moment that the nutrition-related practices of many intended beneficiaries of the project were indeed transformed, then the attribution challenge is to identify the bundle of necessary and sufficient conditions for achieving this. Was it a necessary condition that incomes were already rising for reasons outside of project control? How much difference did the linking of agriculture and nutrition components of the project make? How important were changing gender norms within families and village-wide improvements in social cohesion?

Box 7.4 Quality assurance: methodological reflections on the QuIP

The whole process is totally dependent on the quality of the interviews and the translated summaries entered in the database. Fortunately, in this instance the interviewers were very conscientious. However, it is a challenge to ensure that all relevant detail from the interview, particularly on causes of change, is recorded in the translated interview summary. This is vital as the summaries form the raw material of the analysis. Next time I would do more, as the lead evaluator, to satisfy myself about this. There are three key opportunities to do so. Thorough training of the interview team is the first step. Then there is the opportunity to consolidate the training in detailed feedback on pilot interviews. There may be a language challenge here. Ideally, the person training and mentoring the interviewers will have the language skills to compare the original language recordings with the English summaries, but this may not always be the case. The third opportunity for quality assurance is during the data collection period. With interview summary transcripts entered daily into web-based databases, it should be possible for even a remotely located lead evaluator to provide feedback within 24 hours, enabling the interviewers to go back to their notes, or digital recordings, while their memories are still fresh.[13] This real-time feedback could continually build the quality of interviews during the often-hectic interview schedule. In this case, I only asked the field interviewers to go back to the recorded transcripts once, and significantly more relevant information was forthcoming that wasn't in the original translated summary.

Source: Martin Whiteside, memo to BSDR reflecting on use of the QuIP (November 2017)

And if these were necessary conditions, then how far can they be attributed to project activities and how far was it serendipity that the project took place when they were happening anyway?

Within a perhaps rather idealized view of 'realist' interviewing (Manzano, 2016), such causal claims should be tested through open and equal conversation between field investigators and research subjects. The idealized transcript is then an agreed report on a jointly created account of events: one that draws on, compares, deepens, and combines their prior understandings. However, this is easier said than done. One advantage of blindfolding is that by depriving the researcher of a privileged understanding of the official project 'script' it can help to ensure they engage with respondents in a more equal and holistic way, reducing any tendency to view them more instrumentally as 'data points'.[14]

In seeking to understand what factors combine to bring about behaviour change, realist evaluators also emphasize the importance of the unobservable mechanisms that causally link context to outcomes. For example, in the case of HANO, Whiteside reflects:

> What was interesting to me in HANO, was that people were totally convinced that the dietary changes they'd made were making their children healthier. I have no idea if they were or not. And there were no anthropometric measurements made to check against the statements. But to me it was quite important that they were likely to carry on with these behaviour changes *because* they believed they were working. Evidence from elsewhere suggests that, by and large, most of those behaviour changes will in the longer term produce more healthy households, even if they do not do so in the short term. But if people believe they are having an effect, they'll continue with the good practice.

The mechanism behind this might be purely cognitive (e.g. new knowledge that is acquired about kitchen gardening or nutrition, and self-interest in applying it) or more affective (e.g. peer pressure achieved through group training with sustained and consistent 'messaging'). But there may also be a normative or cultural dimension to mechanisms that are more transformative. This would be the case, for example, if a key mechanism included opening up or reinforcing the possibility of new kinds of relationships between men and women as partners in the joint task of securing food and raising well-nourished infants: an idea and aspiration that might at the same time require overcoming more restrictive views of the roles of men and women.[15] If such mechanisms were critical to project outcomes, then identifying them would require a form of enquiry that can understand and engage with household-level respondents not just as rational maximizers of utility, or even as social animals, but as moral beings (White, 2018). In short, quality assurance in data collection hinges on the moral calibre researchers bring to their engagement with respondents at least as much as it does on methodological details such as blindfolding or data checking.

One way to build a stronger rapport with research subjects, and to achieve deeper insights into the causal mechanisms behind project outcomes, is to conduct unblindfolded follow-up interviews. In the case of HANO, this took place with local project staff. Even the brief account provided above indicates how some participants at this follow-up meeting responded to the opportunity by providing more insights, both about the energy created through effective collaboration and residual power issues (between CSO and NGO staff, for example). To gain more insight into the mechanisms (cognitive and cultural) at the core of the behaviour change achieved during the project period, follow-up interviews with a sample of household-level respondents would also be needed. Without these, uncertainty remains about precisely how and under what conditions the bundling of agricultural and nutritional interventions can achieve more than they do in isolation.

An additional line of reflection posed by the HANO project concerns the counter-factual question of what Save the Children might have learned if it had persisted with an RCT instead of switching to commissioning a QuIP study. This begs many practical questions, including how far it would have been possible to identify control group respondents. Even assuming that some robust findings could have been generated, how useful would this evidence have been? Taking a positive view, it could have generated more precise estimates of changes in indicators of the health and nutrition of infants, children, and mothers, as well as indicators of livelihood changes, including food and cash income. It might also have been possible to attribute these changes to the project on the basis of differences with observed changes in non-treatment villages, to the extent that these were not contaminated by HANO project activities (or those of similar projects), nor affected indirectly by spillover effects from project outcomes. However, it is unlikely that the RCT would have revealed very much at all about which components or combinations of project activities were most effective; still less how and why. This reflects the particular limitations RCTs face in assessing the impact of one project relative to a myriad of other project design possibilities. This problem is particularly acute for activities that entail 'bundling' together inputs and activities that interact in complex or non-linear ways, also referred to as a 'rugged design landscape' (Andrews et al., 2012).

Reliance on quantitative estimates of key impact indicators would also have failed to address the question of how important or meaningful the measured changes and impacts were to the intended respondents and beneficiaries themselves, leaving commissioners and other users of the evaluation to decide for them whether the attributed impact was sufficient on its own to justify project costs, without regard to hard-to-measure effects on such matters as gender relations or community solidarity. Box 7.5 develops this criticism of the limitations of quantitative data more generally and radically. At the very least, this suggests that commissioners who rely solely on 'lean' estimates of change in key impact indicators to justify their investments do so at their peril.

Box 7.5 Further methodological reflections arising from the QuIP HANO study

The QuIP experience helped me to recognize what has become increasingly clear to me over a number of years. Being contracted to do many evaluations, I am often expected to use conventional base-line and end-line data to judge outcomes, and I am increasingly concerned about just how adequate this data is for robust analysis. A representative sample of beneficiaries may be asked their monthly income at the start and end of a project, and conventional statistical analysis can tell us the sampling error (i.e. the confidence with which the sample replies can be assumed to be the same as the response given by all beneficiaries). But this gives us a false sense of security. This 'before' and 'after' data is often very weak for at least four reasons.

First, are respondents being truthful about their income to an unknown interviewer who turns up at their door (they may be worried about tax, or not being included in 'the project', and under-report their income). Five years later they may be proud to report to the project on how well they have done and over-estimate their income – we don't know. These differences are not trivial. I reviewed some figures in Ethiopia where livestock ownership reported from household surveys was less than half that reported from dip-tank records.

Second, even if respondents are trying to recall as accurately as they can, problems remain. Income is earned by different household members, comes in kind as well as cash, and varies between months and seasons. How good is a single 'before' and 'after' measurement at representing the five-year trend? Changes in the timing of the rains may mean that crop selling had started in one survey and not in the other even if they were carried out in the same month.

Third, how do we interpret a single measure of change for a five-year period? For example, a 30 per cent increase in income may sound good, but what inflation rate should be used in the comparison? In a remote rural programme in Kenya that was largely being judged by its outcome on income, there was a belief that local inflation was much higher than the published national rate, but there was difficulty in deciding how to estimate the local rate. Whether the project was judged a success or failure by the donor depended on the inflation rate chosen!

These problems come before addressing the thorny problems of cause and attribution. These are not new. But it is often so much easier not to question the data too deeply. Quantitative data is so seductive! It sounds much better to say: 'mean incomes increased 30 per cent plus or minus 2 per cent', rather than: 'participants reported an increase in income'. Yet the latter is probably the more accurate reply. Another thing I have noticed about how project managers use data is a reflection of human nature. If the data looks good, then they are unlikely to question it; but if it is disappointing, then they appeal to lots of valid reasons for the data not being representative.

In my experience of the QuIP and other qualitative questioning approaches, many of these problems can be overcome, with evidence still collected in a robust and systematic way. What is exciting about these qualitative approaches is that if well designed, they overcome many of the weaknesses noted above. First, interviewees are able to give 'direction of travel' on many topics that they cannot quantify. Asking 'whether you have the ability to buy more or less than five years ago' usually prompts a more thoughtful response than asking for a single income amount. Complex concepts like resilience can be explored with questions like 'is your household more or less able to cope with a severe drought than five years ago'?

Moreover, 'direction of travel' questions don't rely on a base-line, but can be done in a single end-line qualitative survey. Given the weakness of so many project base-lines, and further loss of comparability as project focus or geographical spread changes over the life of the project, this is a major advantage. And of course, they also open up the possibility of further probing on why their ability to buy necessities has changed, etc. So we start getting the really important information about context and mechanisms

(Continued)

Box 7.5 Continued

(what worked, for whom, where, when, and why) as well as an indication of the project's contribution to these. As an evaluation practitioner, I am starting to have much more confidence in listening to participants' own analysis of what has changed for them, and their reasoning about why change has happened (or not). My frustration is that in many evaluations this qualitative questioning is usually limited by time and budget to a not very systematic 'quick and dirty' add-on to complement often meaningless yet widely believed quantitative data.

Of course, some data may still be best collected quantitatively at base-line and end-line. But I believe there is a big opportunity to do so far more selectively, and to invest more in collecting representative and systematic qualitative change information instead. We can produce robust evidence that can be semi-quantified – '80 per cent of participants reported an increase in income and 20 per cent reported no change'. But we must stop being seduced by the claim that 'mean incomes increased 30 per cent plus or minus 2 per cent' actually means what it says. The QuIP is a very important step along this road.

Source: Edited transcript of an interview with Martin Whiteside, December 2017.

Notes

1. This debate has played out over the years in many contexts: in relation to integrated rural development programmes during the 1970s and 1980s, and more recently the Millennium Villages Project, for example. Bundling (and the related concept of market interlinkage) has also attracted sustained attention from institutional economists who associate it not only with jointly marketing products and services, but also with cost reduction that arises from internalizing information externalities between them.
2. A linked issue is the balance between selling crops and growing food for own consumption. The report takes the position that commercialization is neither unambiguously positive nor negative. Agriculture can benefit children's nutrition in two ways: through the type of food that families grow and raise to eat themselves, and through the crops and livestock farmers grow and raise to sell in the marketplace in order to make an income with which to buy food. Benefits to children's nutrition can be achieved by encouraging households to grow more nutritious crops, including fruits and vegetables, or to rear animals for meat, eggs or milk, in addition to the staple crops they tend to rely on. Increased production also means they can sell more, generate more income and afford to pay for foods that will make up a more nutritious and diverse diet (Save the Children, 2012: 49–50).
3. Panita is a civil society platform, representing approximately 300 Tanzanian organizations. It was established in 2014 with support from Save the Children and funding from Irish Aid.
4. To invite all those involved was ruled out by cost, and the team was unable to come up with a clear rationale in time for selecting representatives. But with hindsight they were of the view that 'if we'd had a better understanding and a bit more time to organize it, something bigger would have been possible' (AS).

5. More than a dozen organizations were mentioned, but 13 out of 30 respondents ranked Save the Children as the most important, and two also picked out ROPA (BSDR, 2017: 53).
6. Note that a 'driver' is simply the stated cause, and 'outcome' the effect. Hence it is possible for the same coded item to be both an outcome (of a previous driver) and a driver (of a following outcome).
7. This section draws heavily on Section 8 of BSDR (2017) but does not cover evidence collected from key informants about capacity building, exit, and sustainability because these issues were not covered in household-level interviewing, so making comparisons in those areas is not possible.
8. Key informants noted the importance of cashew nut production (not linked to HANO), but this did not feature much in the household-level interviews except in relation to trees being cut, without compensation, to make way for a gas pipeline. However, cashew is likely to have been implicit in comments on improved crop prices.
9. By comparison, tensions between organizations were rarely raised. One CSO felt that the HANO project had 'dumped' staff on them who were troublesome because these staff retained the mind-set of being Save the Children employees. But they also reported that the lesson had been learned, with CSOs subsequently appointing their own staff.
10. Key informants made some mention of inputs (seed, fertilizer, solar dryers, fruit tree seedlings, etc.), but not as a major limiting factor. One CSO mentioned 'the idea of giving seeds and extension services has been very successful'. However provision of inputs was hardly mentioned at the beneficiary level and did not feature as a motivating factor, and only occasionally as a constraint.
11. This section draws heavily on the concluding section of the QuIP report (BSDR, 2017), but does not cover the linked issues of capacity building, project exit/follow-up, and sustainability.
12. The care group model had also been the subject of a Save the Children study 'Implementing the Care Group approach: Tanzania case study' in October 2016.
13. Web-based uploads of interview transcripts is a feature being introduced by BSDR in 2018. However, it is likely that real-time feedback would be challenged by unreliable internet connectivity in the field and adequate time for field researchers to prepare high quality transcripts – usually undertaken on their return.
14. The underlying issue here is power, and the situation is analogous to how much of it remains with the person at the office meeting who sets the agenda and takes the minutes, regardless of the quality of the discussion. When alternative views have been exchanged, and even if there has been reciprocal illumination, whose interests (and ego) prevail in the way it is recorded?
15. The word 'views' here can be linked to social norms, shared mental models, and what Rao and Walton (2004) referred to even more dryly as 'preference constraints'.

References

Andrews, M., Pritchett, L. and Woolcock, M. (2012) *Escaping Capability Traps through Problem-driven Iterative Adaptation* [pdf], Faculty Research Working Paper Series RWP12-036, Cambridge, MA: Harvard Kennedy School <https://research.hks.harvard.edu/publications/getFile.aspx?Id=841> [accessed 22 October 2018].

Bath Social and Development Research Ltd (BSDR) (2017) *QuIP Report on Save the Children's Harnessing Agriculture for Nutrition Outcomes (HANO) Project in Tanzania*, Bath: BSDR.

Jaenicke, H. and Virchow, D. (2013) 'Entry points into a nutrition sensitive agriculture', *Food Security* 5: 679–92 <http://dx.doi.org/10.1007/s12571-013-0293-5>.

McDermott, J., Ait-Aissa, M., Morel, J. and Rapando, N. (2013) 'Agriculture and household nutrition security: development practice and research needs', *Food Security* 5: 667–78 <http://dx.doi.org/10.1007/s12571-013-0292-6>.

Manzano, A. (2016) 'The craft of interviewing in realist evaluation', *Evaluation* 22(3): 342–60 <https://doi.org/10.1177%2F1356389016638615>.

Rao, V. and Walton, M. (eds) (2004) *Culture and Public Action*, Stanford, CA: Stanford University Press with the World Bank.

Save the Children (2012) *Life Free From Hunger*, London: Save the Children.

Save the Children (2016) *Care Group Case Study & QuIP Report*, October. London: Save the Children.

The Economist (2018) 'Special report: universal health care. An affordable necessity', *The Economist*, 28 April.

Von Braun, J. (2018) *Innovations to Overcome the Increasingly Complex Problems of Hunger*, ZEF working paper, Bonn: Center for Development Research, University of Bonn.

White, S. (2018) 'Moralities of wellbeing: inaugural professorial address' [video], 25 April 2018, University of Bath <https://vimeo.com/266706372> [accessed 19 October 2018].

About the authors

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Gabby Davis, PhD, is a research fellow in the digital skills observatory of the Institute of Coding at the University of Bath. Previously, she was a Senior Project Manager at Bath Social and Development Research, where she led on training and worked on ten QuIP studies in seven countries. Her main expertise is in qualitative research, wellbeing research, socio-ecological relations and post-conflict settings.

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of QuIP across a range of contexts and countries.

Martin Whiteside is a freelance environment and development consultant with over 30 years of experience working on and evaluating bilateral, multilateral, NGO, and community-based programmes in Africa and Asia. He specializes in rural development, climate smart agriculture, and community-based climate resilience programmes. He was lead evaluator on the QuIP study of Save the Children's Harnessing Agriculture for Nutrition Outcomes (HANO) project in Tanzania.

Amy Schmidt, MPH International Public Health, is the Director of Programme Development and Quality at Save the Children Tanzania. Prior to this post, she held the same position with Save the Children in Jordan and worked in the field of monitoring and evaluation for over 10 years. She commissioned the QuIP study on the Irish Aid funded Harnessing Agriculture for Nutrition Outcomes project in Tanzania.

CHAPTER 8

Placing volunteer educators: the Global Health Service Partnership in Uganda, Tanzania, and Malawi

Marlies Morsink, James Copestake, Eva Burke, Gabby Davies and Moses Mukuru, with Clelia Anna Mannino

This chapter reports on a set of three Qualitative Impact Protocol (QuIP) studies exploring contributions of international volunteer educators to university-level nursing, medical, and midwifery training. The US-based NGO Seed Global Health (Seed) commissioned the studies in Uganda, Tanzania, and Malawi on behalf of the Global Health Service Partnership (GHSP), a collaboration between Seed, the Peace Corps, and the US President's Emergency Plan for AIDS Relief (PEPFAR). The studies relied on blindfolded interviews with students and some staff, in combination with unblindfolded interviews with university heads of department. This data supplemented routine activity reports produced by volunteer educators. One striking finding, amid respondents' rich and diverse reflection on their learning, was their appreciation of the volunteers' efforts to take training beyond the classroom and into clinical practice. This chapter is one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.

Keywords: impact evaluation, causal attribution, medical education, nursing education, international volunteering, Peace Corps

Introduction

For all the science and technology wrapped up in modern medicine, when it comes to delivering better health outcomes there is no substitute for well-trained and motivated staff. Despite some recent improvement, availability of nurses and doctors across much of Africa remains shockingly low. World Health Organization figures indicate that for the region as a whole there was one physician for every 4,545 people, and one nurse or midwife for every 855. In the UK, the corresponding ratios were 1:36 and 1:10.[1] Agbiboa (2012: 162) cites a UN claim that 'the city of Manchester holds more Malawian doctors than all of AIDS-stricken Malawi'.

How to remedy this situation? An obvious answer is to invest a lot more in training doctors, nurses, and midwives where they are most scarce, and to

http://dx.doi.org/10.3362/9781780447469.008

ensure they are sufficiently well supported to want to stay there. But this is constrained by tight health budgets and limited in-country capacity to train more staff. One option is to deploy international volunteer educators to work alongside local medical and nursing faculty to meet teaching needs and build capacity. Of course volunteers do not come entirely free; in addition to the costs of travel and living arrangements, there is the potential burden of cross-cultural adjustment. But against this there are possible intangible benefits to receiving volunteers, including transfer of new ideas and recruitment of potential allies in the longer-term political struggle for a more equitable global distribution of health resources.

This chapter reports on a set of Qualitative Impact Protocol (QuIP) studies conducted to collect qualitative evidence of the hard-to-measure effectiveness of volunteer contributions to ongoing nursing and medical training programmes in Uganda, Tanzania, and Malawi. It is particularly interesting to review what led the commissioner of the studies to adopt the approach, how far it proved possible to adapt the QuIP methodology to different contexts and roles, and the implications of the study for further evaluative research in this field of practice. In contrast, and given that the commissioner's primary goal for the studies was to support internal learning, it is beyond the scope of this chapter to draw more general conclusions about the impact of international health volunteers in the three countries.

The commissioner of the studies was a US-based NGO called Seed Global Health (Seed for short), a member of the Global Health Service Partnership (GHSP). Seed placed health professionals into medical and nursing training institutions alongside faculty counterparts. Seed commissioned one QuIP for each country to inform learning about GHSP's outcomes across multiple sites. The following section provides background on Seed and the GHSP programme. The chapter then elaborates on the logic that led the commissioner to use the QuIP and provides an overview of the studies. Next, we provide an illustrative account of the findings and reflect on methodological lessons learned from adapting the QuIP to inform lesson-learning within the very different context of an international programme to promote medical and health education.

This chapter was drafted by Morsink and Copestake, drawing on the original study reports (authored by Davies and Burke with contributions by Mannino), secondary documents, and the transcript of an interview conducted by Morsink with Clelia Anna Mannino from Seed (from which quotations labelled 'CAM' are drawn). Moses Mukuru was the lead QuIP field researcher for the study. Burke, Davies, Mannino, and Mukuru also read and commented on an initial draft.

The Global Health Service Partnership

The Global Health Service Partnership (GHSP) was a collaboration between the US President's Emergency Plan for AIDS Relief (PEPFAR), the Peace Corps, and Seed Global Health (Seed). It was established in 2012 in response to the

striking shortages of health providers in many parts of the world. GHSP hoped to increase clinical care capacity and strengthen health systems in resource-limited settings by cultivating the next generation of local doctors, nurses, and midwives. The programme placed US health professionals alongside local medical, midwifery, and nursing staff to meet the teaching needs identified at each partner institution. The Peace Corps oversaw the logistics of volunteer placement and provided in-country and programmatic support, including overseeing safety and security. PEPFAR was the main funder. Seed provided clinical expertise, programming support, and overseas monitoring, evaluation, and learning (MEL) activities. Seed team members had extensive medical and nursing experience, including specialist expertise in medical and nursing education in settings where these were scarce. Over five years, GHSP placed 186 nurse and physician volunteer educators at 27 institutions across five partner countries (Liberia, Malawi, Swaziland, Tanzania, and Uganda). The GHSP partners that participated in the QuIP studies comprised nursing and medical institutions in Malawi, Tanzania, and Uganda established between 1979 and 2015, and offering a range of diploma, undergraduate, and post-graduate nursing, midwifery, and medical courses and degrees. For evaluation of the programme, Seed's dedicated MEL team worked closely with GHSP's Programme Manager at Peace Corps: 'We always say that she is part of our MEL team. We touch base regularly, and she helps to drive the MEL process for the programme' (CAM).

The QuIP study

The commissioning process

Using an external consultancy to evaluate one of their own programmes was a departure from normal practice for both Seed and Peace Corps: neither had previously outsourced an evaluation in whole or in part. The proposal to do so came from a new member of the Seed MEL team, who saw it as a way to strengthen credibility of findings and perhaps generate unexpected insights: 'My preference is for most evaluative work to be done externally; not all MEL work, but in particular evaluation. It brings a different lens to what we're seeing' (CAM).

The suggestion prompted internal discussion, and a chance to clarify that an external evaluation could be primarily oriented towards internal learning, neither authorizing the evaluator to censure the organization if aberrations were found, nor surrendering control over the communication of findings. A third concern was that an external consultant wouldn't really understand the context or the ins-and-outs of the programme. 'I've certainly worked on external evaluations in the past where the recommendations made by the external consultants were a little "off", where it became apparent that they'd not necessarily understood the programme that well' (CAM). This last concern was addressed by agreeing that the external evaluation would be conducted in close collaboration with commissioner staff. 'There's a partnership aspect

to the work, which applies to framing the questions the evaluation aims to answer, creating the content of the questionnaires, all the way through to how findings are communicated' (CAM). This discussion helped to ensure that the commissioner was closely engaged in specifying not only the purpose of the study but also the methodology.

> Seed knows there's a lot of nuance in what our volunteer educators do, and nuance in what we'll see as impact. Given this, there was support across the board at Seed for qualitative methods, and recognition that evidence of impact doesn't necessarily have to be quantitative. The challenge in examining impact on metrics like student grades or test scores is that there are so many other factors that influence those results (CAM).

This consensus over methods within Seed's MEL team was also achieved through discussion of why using a randomized controlled trial (RCT) approach would not work:

> We've discussed whether to use an RCT and, to date, that methodology just does not work in our context. For one, it's hard to determine an appropriate control group; since GHSP is implemented at universities, we enter into a pre-functioning system. In the spirit of our partnership with sites, it is not appropriate for us to compare students who have been taught by GHSP volunteers with those who have not. Also, there is no one uniform intervention that you can measure: each GHSP volunteer approaches their work differently, has different interactions with students and colleagues, and has a scope of work tailored to their context (CM).

Seed was already generating a lot of quantitative data through monitoring volunteers' activities and the projects they were working on. The knowledge gap that Seed identified was between this measurable evidence of activity from monitoring, and evidence of impact in the complex setting of a university programme with multiple inputs. This is what prompted Seed to seek a suitable qualitative approach. Box 8.1 explains how the commissioner identified the QuIP and what prompted Seed to select it for the evaluation.

Adapting the methodology: dilemmas over blindfolding

While acknowledging that blindfolding would add to the credibility of the impact evaluation, Seed and staff from Bath Social and Development Research Ltd (BSDR) also recognized the difficulties and disadvantages arising from it, including negotiating access and limiting the ability of interviewers to enquire into detailed aspects of the activity being assessed. They addressed this dilemma by supplementing the blindfolded interviews conducted with all students and some staff, with un-blindfolded interviews conducted with department heads. GHSP also fully briefed university deans over the

Box 8.1 Commissioner rationale for selecting the QuIP

'Once we'd decided to do an external evaluation with mixed methods, we were searching for ways to find an appropriate approach. One way would have been to do an open call for proposals, inviting consultants to propose different methodologies. But around that time I came across the QuIP on the Pelican Listserv.[2] The more I read about it, the more I thought it would really fit with what we were looking for. It promised a rigorous approach to qualitative work for practitioners whose programming just doesn't lend itself to quantitative assessment. I liked that the QuIP bridged the gap between certain quantitative methodologies and the more qualitative lens we needed to use for this programme. The QuIP offered an interesting combination of some of those principles, by adding in both attribution analysis and blindfolding, which added a unique and robust dimension to a qualitative approach. Another piece of the QuIP that was compelling for us and helped our internal conversation around adopting the methodology was the QuIP's approach to reporting on findings. In the QuIP, the findings of the study are intended to serve as the starting point for conversations within the organization around recommendations and how those findings will be used. The QuIP provides the methodology and the findings, but then it's up to us as an organization to figure out what it means for us and what we want to do with the information'.

Source: Interview with CAM

proposed methodology, and sought their approval to use it. While this strategy was based on ethical principles, the double-barrelled approach also made it possible to triangulate the unblindfolded observations of senior staff with findings from the blindfolded interviews, potentially leading to additional insights.

Initial discussions between Seed and BSDR also focused on how to ensure blindfolding did not result in volunteers' contributions being missed, because student testimony might stop at referencing activities but without specifying who was behind them. A key point here was the availability of secondary data from volunteers' reports (covering classes they taught, projects on which they worked, and so on) to assist in piecing together attribution stories. 'Having good monitoring data helped a lot with attribution: if a particular activity or intervention was mentioned in an interview, we were able to triangulate with volunteers' reports to see if we could pull that information' (CAM). By providing the QuIP analysts with details of volunteers' activities both from the reports and their own programme knowledge, the Seed team felt more confident that where there was perceived change, the evaluation would be able to capture what input had come from GHSP volunteers.

Adapting the methodology: defining domains

The diversity of volunteers' activities and institutional settings also posed a challenge when it came to designing data collection instruments. This was addressed by referring to the programme's theory of engagement, as well as outcome areas identified by an internal qualitative study during the programme's early years, reproduced as Figure 8.1.

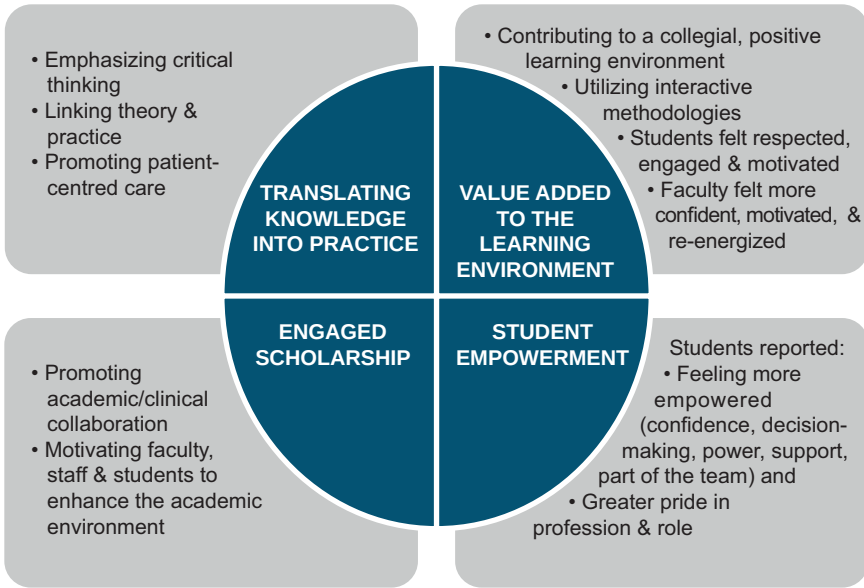


Figure 8.1 Preliminary outcome areas

Considerable iteration was needed between the commissioner, evaluation, and field teams to come up with a workable domain structure and a set of questions for each domain. This provided a useful reality check in itself on the feasibility of aligning programme goals with questions about change that would make sense to staff and student respondents. Ultimately four open-ended domains were decided upon: learning environment; clinical practice; collaboration and communication; aspirations and plans for the future. Respondents were asked to comment on change over the previous three years across each domain, and each section of the questionnaire ended with one or two closed questions. A final section asked respondents to list, rank, and assess their experiences with teachers from abroad – thereby further aiding attribution. The appendix to this chapter outlines the final version of the questionnaire used for both student individual interviews and focus group discussions (FGDs).

Going through this process jointly with the evaluator provided the commissioner with added reassurance that respondents' experience with volunteers would be discussed, despite the blindfolding. This illustrates the important general point that engagement by the commissioner in the design of an external evaluation can help create a sense of ownership in the findings, and increase prospects that findings will feed through into action.

The robust domain structure informed deductive coding and confirmatory analysis. Delimiting data collection by using domains sped up and simplified the thematic coding and reporting of data. While the domains framed the scope of interest of the interview, the respondents remained free to identify

any number of specific outcomes in responding to the questions in each domain. To the extent respondents mentioned unanticipated outcomes, coding became inductive (i.e. *ex post* and more exploratory). This ensured that the quality of findings, while delimited by the domain structure, was not constrained by it.

Illustrative findings

Securing agreement of department heads to the blindfolded approach did not turn out to be an obstacle, in part because they were familiar with its use in research in other clinical contexts. Nevertheless it was important to explain fully the rationale for blindfolding, and also commit to sharing findings with them at the end of the study. The challenges the field teams did encounter were more prosaic: staff were busy, and in some cases weary of being interviewed. With persistence, however, the sampling quotas were achieved – apart from two staff declining to be interviewed in Uganda, and one FGD having to be cancelled in Malawi.

Table 8.1 provides a summary of the interviews and FGDs completed across all three countries. The table divides samples between medical and nursing/midwifery programmes in order to illustrate the range of respondents interviewed, including their gender composition, with a higher proportion of men in medical programmes, and women in nursing and midwifery programmes. All students interviewed were in either their final or penultimate year of study.

Answers to the closed questions (addressed to each individual respondent and FGD participant) were almost universally positive across all four domains, with no negative replies from 71 respondents in Tanzania, only five (two for change in teaching methods, and three for experience of clinical practice) from 70 respondents in Uganda, and 13 negative responses in Malawi (two for clinical practice, five for staff/student relationships, and six for confidence in the future). While this could reflect an objective

Table 8.1 GHSP interviews and focus group discussions

	Uganda		*Tanzania*		*Malawi*	
	Medicine	*Nursing & midwifery*	*Medicine*	*Nursing & midwifery*	*Medicine*	*Nursing & midwifery*
Number of sites	2	2	2	2	2	2
Student interviews	6	6	6	6	7	6
(of which women)	(0)	(6)	(0)	(2)	(3)	(4)
Student FGDs	4	4	4	4	3	4
(total participants)	(23)	(24)	(23)	(24)	(16)	(24)
Staff interviews[3]	5	6	6	6	6	6
(of which women)	(0)	(6)	(0)	(3)	(2)	(4)

Source: BSDR (2017a,b,c)

improvement in training across the three countries in the last three years, it may also partly reflect growth in students' own confidence and competence as their training had proceeded. Either way, answers to these questions did not reveal any significant differences in perception of change between countries, institutions, or types of respondent (staff, medical student, nursing or midwifery student).

Turning to answers to the open-ended questions, Table 8.2 shows the frequency counts of positive and negative causal claims coded as 'explicit', 'implicit', and 'incidental' relative to GHSP's theory of change.

Four observations can be drawn from this table. First, the highest frequency of coding was for incidental or 'other' drivers of change that alluded neither explicitly to GHSP nor implicitly to the presence of volunteer educators from abroad. This is not surprising, given the many other factors affecting students' learning experience. That said, the interviews and FGDs were also successful in collecting a lot of explicit or implicit attribution evidence, with positive causal claims greatly outnumbering negative claims across all countries and domains – a finding that was consistent with the overwhelmingly positive

Table 8.2 Frequency counts of coded causal statements for all programmes by country, domain, and attribution tag (across all interviews and FGDs)

	Positive			*Negative*		
	Explicit	*Implicit*	*Other*	*Explicit*	*Implicit*	*Other*
Uganda						
Learning environment	11 (4)	11 (5)	23 (7)	–	– (2)	10 (6)
Clinical practice	10 (3)	13 (7)	22 (7)	–	2 (–)	16 (7)
Collaboration and communication	4 (–)	9 (1)	19 (5)	–	3 (–)	14 (6)
Aspirations and future plans	1 (–)	6 (–)	19 (7)	–	3 (2)	9 (5)
Tanzania						
Learning environment	11 (6)	16 (7)	24 (6)	–	2 (–)	7 (3)
Clinical practice	7 (2)	19 (7)	24 (6)	–	–	11 (6)
Collaboration and communication	6 (2)	12 (8)	23 (8)	–	–	16 (5)
Aspirations and future plans	8 (1)	3 (1)	23 (8)	–	–	10 (5)
Malawi						
Learning environment	3 (1)	11 (4)	19 (6)	1 (–)	3 (1)	11 (4)
Clinical practice	1 (–)	14 (6)	25 (7)	2 (–)	3 (–)	21 (6)
Collaboration and communication	1 (–)	4 (2)	22 (7)	1 (–)	–	16 (7)
Aspirations and future plans	– (1)	5 (–)	25 (7)	1 (–)	–	14 (4)

Source: BSDR (2017a,b,c)
Note: First number refers to individual interviews, and the second (in parenthesis) to focus group discussions. Not shown, but also included in the original data, were coded outcome claims made without a causal explanation.

response to closed questions. Third, explicit and implicit positive attribution most often referred to effects on the learning environment and clinical practice. Fourth, across 72 interviews and 23 FGDs only five statements were coded as explicitly negative, all of them from Malawi where positive explicit statements were also less common than in the other two countries.

Of course, this table provides only a very superficial view of the attribution evidence collected. The QuIP reports for each country explored the evidence in far greater depth; and by showing the codes of the interviewees in each cell (rather than just indicating how many there were) the reports also made it easy to pull out the coded text underpinning the numbers. To illustrate the process of delving into the qualitative data, Box 8.2 provides two examples of positive change explicitly attributed to GHSP volunteer educators working in Uganda.

The quotations in Box 8.2 suggest that one potentially far-reaching contribution of the GHSP volunteer educators was to strengthen the linkage between class-based teaching and learning through supported clinical

Box 8.2 Illustrative causal claims from Uganda coded positive explicit

Female midwifery student

'The other important thing that I have liked is that we have what they call the Peace Corps and the Seed Global Health and they have been mainly with the nurses. They teach you the theory and they take you into the practicals to see whether you mastered what they have been teaching you. It's the practical aspect and following up of students in the wards that makes them stand out so much because most of the lecturers, be it in nursing or pharmacy, when they teach you in class they leave you there and when it reaches time for going on the ward, they leave you to go on your own to implement what you learnt from theory which is too, too hard because some of the things are too complex for a student to implement on their own. But these guys after teaching, they go with you to the wards … the Peace Corps who come in as volunteers … are more in the wards. Because they are more clinical, it's an advantage to us because nursing is a clinical course. The way they have been doing their clinical is different from how we have been doing it. We used to identify a case, we go through it and like that but with the way they are doing it, we discuss the patient there and then, you even do a ward round; it's more nurse-centred. Before the doctors owned the patients but with the current trend we encourage interdisciplinary teaching. When I am on the ward round and all other students are there, I can invite all of them and we discuss the patient. We also encourage the nurses to speak during the ward round. This is what we call team work and at the end of the day the students learn to appreciate one another unlike before.'

Male medical student

'Every year we receive these people from outside called the Peace Corps; they come directly to the wards because that is where most of their interest is. They come as volunteers depending on I think which group is available, because last year we had paediatrics and surgery but this year we have only received 2 for paediatric medicine. In our 3rd year, we had a lady from the Peace Corps … she was in paediatrics but she helped teach most of us during that rotation. In fact, about 70% of the paediatrics that I know and I think most of what students learnt at that time in paediatrics is because of her.'

Source: BSDR (2017a).

practice. This in turn raises further questions about why this may be the case, as well as how far such a change is or could be sustained without the presence of the volunteers.

The QuIP's systematic approach to interrogating the data involved supplementing attribution tags by (a) identifying more specific drivers of change, linked both to the activities being evaluated and those incidental to them, and (b) identifying more specific outcomes than the broad domains used to structure data collection. By viewing causal claims in the context of the whole interview, the analyst could usually also classify them as either positive or negative from the respondents' perspective. To illustrate, Tables 8.3 and 8.4 reproduce data for the frequency count of causal claims, extracted from the Uganda and Malawi datasets, respectively. Each row corresponds to a different driver of change and each column to a different outcome, while the numbers refer to how many times the analyst recorded a causal link from the first to the second, as made by a respondent. The frequency counts include multiple citations that arise when the respondent or focus group made the same link more than once. While there is an element of subjectivity about the frequency counts, arising from how the analyst chose to label different statements, the table does nevertheless provide an initial overview of respondents' collective perception of what changes are taking place and why.

The most widely cited outcome in the Uganda data was that 'students practise and improve their clinical skills' (frequency count of 48), and the highest ranked driver of this was 'increased practise, assessment, and supervision of clinical skills' (16 counts). These citations did not necessarily relate to GHSP, but in other cases they were more likely to do so. For example, 'visiting professors teach students' was the fourth most cited driver of change overall (23 counts), and could be linked to 'improved management and quality of teaching' (12 counts) and 'students practise and improve clinical skills' (8 counts). This is supported by the illustrative quotes in Box 8.2.

Table 8.4 turns to coded evidence on negative change, which was smaller in number than for positive change (71 compared with 233), and again did not necessarily relate to GHSP. The most cited negative outcome was 'difficult relationship between students and staff' (15 counts). Not surprisingly, many of the drivers overall concerned lack of material resources; yet the most frequently mentioned overall driver referred to attitudes of 'unwilling/unsupportive clinical staff' (15 counts), leading among other things to 'difficult relationships between students and staff' (7 counts). While these comments were mostly *not* explicitly linked to GHSP they did reinforce the need to mobilize and motivate more staff. The small number of negative explicit comments about GHSP related to the limited duration of volunteer educators' stay, as illustrated by the quotations in Box 8.3.

While it is not possible from this data alone to assess whether these criticisms are fair or not, it is an important fact that students felt this way. This suggests scope for further analysis of a complex negative dynamic (with drivers and outcomes caught in two way relationships), very possibly reinforced

Table 8.3 Frequency count of positive causal claims identified within the Uganda data

Driver of change	*Change or outcome*									
	Students practise and improve clinical skills	*Students more confident in their abilities*	*Improved management and quality of teaching*	*Improved rapport and relationship between students and patients*	*Students feel more supported*	*More able to put theory into practice*	*Students more engaged in learning*	*Students plan to specialize/ continue with PG study*	*Improved relationships between staff and students*	*Total citations (can be more than one per respondent)*
Increased practice, assessment. and supervision of clinical skills	16	7	3		10	3	2		1	42
Increased medical knowledge and skills	2	20		1		9			1	33
More interactive and student-centred teaching	1	1	8				14			24
Visiting professors teach students	8		12		1		1		1	23
Increased focus on practical training	4	4				4	2			14
Staff are more approachable and accessible	1				6				7	14
Students have higher aspirations								13		13
Students spend more time interacting with patients				12						12
Communication skills and language learning				11					1	12

(Continued)

Table 8.3 Continued

Driver of change	*Change or outcome*									
	Students practise and improve clinical skills	*Students more confident in their abilities*	*Improved management and quality of teaching*	*Improved rapport and relationship between students and patients*	*Students feel more supported*	*More able to put theory into practice*	*Students more engaged in learning*	*Students plan to specialize/ continue with PG study*	*Improved relationships between staff and students*	*Total citations (can be more than one per respondent)*
Increased availability of medical equipment	8		2		1					11
More well-qualified staff employed	2		4		3				2	11
Using skills lab to demonstrate and practice clinical skills	5	1				4				10
Improved planning and management of student learning	1		4				1		2	8
Increase in student led tutorials							6			6
Totals	48	33	33	24	21	20	26	13	15	233

Source: Adapted from BSDR (2017a).
Note: This table classifies all coded statements – explicit, implicit, and incidental. In other words not all causal links shown are directly linked to the GHSP programme

Table 8.4 Frequency count of negative causal claims identified within the Malawi data

Driver	*Outcome*									
	Reduced supervision of students in a clinical setting	*Staff are not approachable*	*Lack of collaboration between students and clinical staff*	*Gap in knowledge between theory and practice*	*Students do not feel confident in their abilities*	*Difficult relationship between students and staff*	*Insufficient support for students*	*Students unable to learn effectively/ insufficient training*	*Increased demotivation and demoralization*	*Total citations*
Fewer practicals/ supervisions assisted by faculty staff					1		4	1		6
Students self-learn/ left to find balance between theory and practice				1	2					3
Difficulty for students to speak up on wards						1	1	1		3
Preceptors not supportive of certain students						1	3			4
Understaffing/staff shortages	2						2			4
Gap in knowledge between theory and practice					3					3
Lack of equipment/ facilities/resources				4	1			3	1	9
Absence of clinical staff to supervise students		1				1	1			3

(Continued)

Table 8.4 Continued

Driver	*Outcome*									
	Reduced supervision of students in a clinical setting	*Staff are not approachable*	*Lack of collaboration between students and clinical staff*	*Gap in knowledge between theory and practice*	*Students do not feel confident in their abilities*	*Difficult relationship between students and staff*	*Insufficient support for students*	*Students unable to learn effectively/ insufficient training*	*Increased demotivation and demoralization*	*Total citations*
Unwilling/unsupportive clinical staff		2	4			7	1	1		15
Different education levels between clinical staff and students		1	1			4				6
Increase in student numbers	2						1	1		4
Clinical staff do not teach, but use students						1	1	1		3
Increased competition for graduate jobs/ lack of employment opportunities									5	5
Class sizes too large for effective learning								3		3
Total citations	4	4	5	5	7	15	14	11	6	71

Source: Adapted from BSDR (2017b).
Note: This table classifies all coded statements – explicit, implicit, and incidental. In other words not all causal links shown are directly linked to the GHSP programme.

Box 8.3. Illustrative causal claims from Malawi medical students coded negative explicit

Male medical student

'Sometimes there is no one to help with some theatre procedures, as compared to office procedures. Not all supervisors from abroad are conversant with theatre procedures.'

'The length of stay of volunteers is too short. The training is different and context based for family medicine, volunteers have to learn about the local context when they arrive and when they finally get to do that, it will be time for them to leave.'

Female medical student

'We have supervisors from America [Peace Corps] who come for twelve months and then return and another team comes. They are very helpful. The first cohort knew the portfolio very well. They took us through all relevant procedures we needed to acquire. The second cohort was more office based. I have never seen one of them in theatre that could guide you how to do a procedure as compared to the first cohort.' '… We would have loved it if volunteers stayed one year longer since they leave just after they have settled or adapted. Otherwise they always bring rich experience.'

Source: BSDR (2017b)

by the effect of lack of resources on staff attitudes. Note that while none of the negative drivers listed in Table 8.4 self-evidently refers to the presence of volunteer educators, the data may be relevant to analysis of arguments for and against supporting medical and nursing/midwifery training in this way. In a context of generalized resource scarcity it would not be surprising if respondents were reluctant to criticize any support they did receive on the basis of 'not looking a gift horse in the mouth'. Hence the fact that interviews were blindfolded does add to the credibility of the claim that volunteers were not perceived to be a hindrance. That said, one of the most explicit negative statements about the effect of volunteers came from an un-blindfolded interview with a department head:

> One of the biggest challenges we have is that being a teaching medical institution we have had some of our volunteers who have not been very flexible when it comes to seeing patients. They know their primary role is to teach students both undergraduate and post graduate, so when you ask them to see patients, they feel it is not their role. Personalities vary, we have had very good volunteers and some not very good ones, but that has not been a very big challenge. The older and super specialised volunteers are rather rigid and would rather do what they were trained to do. The younger ones are flexible.

Reflections

Building on findings: potential for further analysis and application

Overall, the three studies provided substantive evidence that the GHSP contributed positively to medical and nursing training in each country, particularly to students' learning environment and their opportunity to

learn through clinical practice. This is illustrated by excerpts from the Uganda report in Box 8.4. The three study reports also indicated that while the presence of volunteer educators had an important influence on students' experiences, it was not the most important.

Looking across groups, the data from individual interviews and FGDs in Uganda also suggested that GHSP volunteer educators had more positive

Box 8.4 Illustrative summary of conclusions from the Uganda report

Translating knowledge into practice effectively

This was the area of most positive change, and the one that was most explicitly attributed to GHSP, especially regarding the positive change 'students practise and improve clinical skills'. Major drivers contributing to this positive change were the 'increased practice, assessment, and supervision of clinical skills', 'visiting professors', 'the availability of medical equipment', and 'more practical training sessions'. All of these factors were *explicitly* attributed to the GHSP to some extent, with more explicit attribution from the nursing departments at two locations.

Promoting patient-centred care

Although many respondents reported positive changes in this area, such as students spending more time with patients and improvements in communication, these were generally not explicitly linked to GHSP. However, key informant interviews confirmed the positive contributions of the GHSP to clinical practice, with the volunteers contributing to improvements in course structure, objectives, hands-on training and supervision, and being good role models. Increased supervision on wards was particularly highly regarded and frequently alluded to throughout the study by many respondents.

Learning environment

The QuIP study revealed success in this area, with positive changes reported in the 'management and quality of teaching' and 'student engagement in learning'. 'Teaching methods' was reported as an area of positive change by staff interviews, a finding supported by all the department heads. Drilling down revealed that specific drivers could be attributed to the GHSP, including 'visiting professors teaching students', 'more interactive and student-centred teaching', and 'improved planning and management of student learning'. Respondents spoke highly of the demonstrations and practical sessions that complemented the theoretical sessions, and of more student-centred ways of teaching. Respondents in the nursing and midwifery programmes explicitly attributed this to the GHSP, more than did medical students and staff. Department heads supported these findings and confirmed that new, more interactive teaching methods had been introduced, with volunteer educators also bringing new ideas and innovation. One department head particularly appreciated the introduction of feedback and evaluation mechanisms for staff and students.

Enhancing the academic environment and promoting academic/clinical collaboration

The QuIP study yielded some evidence of staff being more approachable and accessible. However, few blindfolded respondents attributed these changes explicitly to GHSP. This contrasted with department heads, who especially linked GHSP to creating a culture of feedback which helped shift the staff–student hierarchy to a more open and equal footing. While they did not attribute many changes to GHSP in terms of student/clinical staff relationships, all of them also reported improvements in cross-departmental collaboration. There was little improvement in the relationships between different types of students, with nursing and medical students doing ward rounds together but with minimal interaction.

Source: BSDR (2017a)

impact on nursing students and staff than on medical students. However, this was not so evident from the studies in Malawi and Tanzania: further research would be needed to identify the diverging positive and negative outcomes reported by different disciplines, and also by students and faculty/clinical staff. The same would be true for a disaggregated analysis of variation in drivers of change (by country, site, course, gender), unrelated to GHSP or during the previous three years.

One potential use of the evidence generated would be for induction and training of new volunteer educators, both in highlighting obstacles to learning that were likely to differ from those that they had encountered personally, and factors that could contribute to their own effectiveness. Other volunteer programmes and/or collaborations with institutions were widely mentioned, including volunteer visitors from Cuba, Sweden, Germany, UK, Belgium, and Italy. This suggested scope for analysing not only the absolute impact of volunteers but also perceptions of their relative effectiveness, drawing on the multiple exposures of students and staff to a wide variety of volunteers. It would be useful to analyse how volunteers on other programmes operated, adapted, and contributed to learning. A key issue here is also sustainability: how to ensure teaching programmes continue to be more student-centred, interactive, and practically focused after volunteer educators have left, for example. This reflects the underlying problem of lack of resources, and its effect not only on the ability of permanent staff to practise effectively, but also on their energy and enthusiasm to support students in doing so.

Methodological lessons

Representing the commissioner perspective, Clelia Anna Mannino (CAM) suggested that the QuIP did generate sufficiently detailed and nuanced evidence to support programme learning, but also emphasized that where programming was relatively fluid and differed between implementation sites and individuals, then triangulation against activity monitoring data was particularly important (see Box 8.5). The studies also illustrated the benefits that could be gained from combining blindfolded feedback with fully informed feedback from key informants, particularly for programmes such as GHSP where change due to the programme might have been a relatively minor driver of overall change in a complex context. Close communication between the analysts and the commissioner was an important ingredient for ensuring accurate attribution, supplementing field data with access to the volunteer educators' own feedback reports. Since volunteers have the potential to affect students' learning in many ways, sometimes unexpectedly, it was also important for the analyst to be able to conduct *ex post* inductive coding rather than relying only on the domain structure used in interviews.

Another methodological issue illustrated by the study was the challenge of dealing with a high degree of heterogeneity in the analysis stage. The analysts were confronted with understanding the dynamics of three institutions in

Box 8.5 Reflections of the commissioner on the QuIP reports

For us, because we work in the context we do, it was first of all interesting to learn about the general landscape of positive and negative changes – even before we got to the attribution of changes to our programme. Just seeing this landscape was incredibly valuable for us, helping us to better understand the institutions with whom we partner and how we could think about supporting them in the future … It was especially exciting that the data pointed to evidence of impact, particularly given our complex programming and the unique differences between volunteers, between sites, etc. The QuIP approach really worked well for us to see both GHSP impact and get a more general sense of the landscape in which we work … I really appreciated the QuIP's approach where no value judgements are placed on findings. A QuIP report really is a "report of findings", leaving us as a team to think about what exactly it tells us, what is surprising, and how we will translate findings into programmatic actions … Given the complexity and uniqueness of our programming and the challenges of capturing impact quantitatively, the ability to attribute impact depends on solid monitoring data gathered throughout the course of the project. The work of our volunteers is nuanced and not all elements are captured within their monthly activity reports; however, between the volunteer activities and the QuIP analysts' work, we were able to cross-reference and build a solid base for attribution analysis.'

Source: Interview with CAM

three countries, spanning medical, nursing, and midwifery programmes; as well as the gender balance across them among both staff and students. This chapter illustrated the range of options for interrogating the data, but it was beyond its scope to present a systematic analysis across all the cases. More ambitious research could extend the analysis to address variation in the impact of volunteer programmes of different countries/agencies, including a time dimension. Nevertheless, the study did break new ground for the QuIP by adapting it to the very different context of health students' learning experiences.

Appendix: questionnaire outline for student interviews and focus group discussions

A. Respondent details
B. Setting the context

B1. How long have you been studying? This study is looking at your experience since you started your programme of study, particularly around the last three years.
B2. What motivated you to choose this course of study?

C. Learning environment

C1. Thinking back, have there been any changes in the way you are taught over the last few years? How do you find the teaching methods at the college/university differ from the teaching methods you've encountered previously? What difference does this make to your learning experience?
C2. Are there any lectures or rotations in which teaching methods have changed since your programme of study began? Which methods of teaching/activities do you find work best for you? Which methods do you find are least good for learning? Why?

C3. How confident are you that what you were taught (through the content of the lectures) is giving you the skills and knowledge you need to practise? Please explain.
C4. How confident are you that the way you are taught is giving you the skills and knowledge you need to practise?
C5. Overall, compared with when you started your course do you feel that teaching methods are: better/the same/worse?

D. Clinical practice

D1. What over the past three years has made a difference to your experience of clinical practice?
D2. Do you feel that you are working in a collaborative team environment when undertaking clinical practice? Has anything changed in the relationships between students and medical staff? Do you feel confident/empowered to speak up in clinical situations?
D3. How about the relationship between you (students) and patients? Has anything changed in the past three years which has changed the way you relate to patients? If so can you say what caused this change?
D4. How do you feel about the level of training and support you receive while on your clinical rotations? Do you have someone you can go to if you have a question or if there is a problem? Is there someone available who would help you and demonstrate a procedure if needed?
D5. How do you feel about your ability to take what you've learned in class and use it in your clinical practice? What has helped you? What makes that difficult? Has that changed at all in the last 3 years?
D6. Overall, compared with when you started your course do you feel that your experience of clinical practice is: better/the same/worse?

E. Collaboration and communication

E1. How accessible and easy to approach are your university/college [insert name of institution] teachers? Has this changed at all since you started your programme of study?
E2. How about clinical staff?
E3. What are relationships like between students, including between medicine, midwifery, and nursing students?
E4. Overall, compared with when you started your course, do you feel that relationships between staff and students have: improved/stayed the same/got worse?

F. Aspirations and plans for the future

F1. What plans do you have for the future when you finish your course? Have these plans changed since you started studying? If so, why?
F2. Thinking about when you first started your course of study, how have your feelings towards becoming a nurse/midwife/doctor changed? Why?
F3. How has your level of confidence in your ability as a nurse/midwife/doctor changed in the past three years?
F4. Do you plan to continue adding to your medical, midwifery or nursing knowledge once you are working? If so, in which areas?
F5. Overall, compared with when you started your course, how positive/confident do you feel about your future as a physician, midwife or nurse? Much more confident/a little more confident/much less confident.

G. Any other comments about your experience as a student?

Source: BSDR (2017a)

Notes

1. See https://afro.who.int/countries/ for the Africa region (figures for 2005) and http://apps.who.int/gho/data/node.main.A1444 for the UK (figures for 2010).
2. The Pelican Initiative is an online platform for evidence-based learning and communication for social change (https://dgroups.org/groups/pelican)
3. This excludes interviews with department heads, which were not blindfolded.

References

Agbiboa, D. (2012) 'Offsetting the development costs? Brain drain and the role of training and remittances', *Third World Quarterly* 33(9): 1669–83 <https://doi.org/10.1080/01436597.2012.720847>.

Bath Social and Development Research Ltd (BSDR) (2017a) *QuIP Report on the Global Health Service Partnership (GHSP) Program in Tanzania*, July 2017, Bath, UK: BSDR.

BSDR (2017b) *QuIP Report on the Global Health Service Partnership (GHSP) Program in Tanzania*, August 2017, Bath, UK: BSDR.

BSDR (2017c) *QuIP Report on the Global Health Service Partnership (GHSP) Program in Malawi*, September 2017, Bath, UK: BSDR.

About the authors

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of QuIP across a range of contexts and countries.

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Eva Burke, MSc Reproductive and Sexual Health Research, is an independent consultant specializing in sexual and reproductive health (SRH). Her main expertise includes the programming, research, and monitoring and evaluation of family planning programmes and access to SRH services in sub-Saharan Africa and Latin America. She has worked on two QuIP studies: one on the impact of the Global Health Service Partnership programme in three East African

countries; and the other on the impact of a training programme on the quality of life of midwives in Uganda.

Gabby Davies, PhD, is a research fellow in the digital skills observatory of the Institute of Coding at the University of Bath. Previously, she was a Senior Project Manager at Bath Social and Development Research, where she led on training and worked on ten QuIP studies in seven countries. Her main expertise is in qualitative research, wellbeing research, socio-ecological relations and post-conflict settings.

Moses Mukuru, MA Social Sector Planning and Management, is a Research Fellow at Makerere University with 13 years' experience in programme management, evaluation, and research on social development programmes. His expertise is in programme design, evaluation, and public policy research. He is a trained QuIP researcher and has supervised four QuIP studies in Uganda focusing on food security; two studies on university-level training programmes for health workers; and a study on poverty reduction and welfare improvement in Uganda.

Clelia Anna Mannino, PhD, is Director of Monitoring, Evaluation & Learning (MEL) at Seed Global Health, where she oversees the MEL portfolio in all countries of operation. She was responsible for commissioning and overseeing the QuIP studies carried out on behalf of the Global Health Service Partnership in Uganda, Tanzania, and Malawi.

CHAPTER 9

Adapting the QuIP for use with local authorities in England: bending but not breaking

Marlies Morsink and Fiona Remnant

This chapter reports on the first Qualitative Impact Protocol (QuIP) pilot studies in the UK, conducted for local government authorities in the south-west region of England (Bristol and Frome). This shift in context entailed a range of methodological adaptations, not least in response to the chronic lack of funding for impact evaluation in the UK local government sector, when compared with international development. The QuIP study in Bristol broke new ground by analysing drivers of change at an organizational level in assessing the impact of Bristol City Council's support for community level organizations through its civil society support partner Voscur. The exploration into the effects of Frome Town Council's interventions in public green spaces, meanwhile, led to a re-framing of impact in terms of citizens' choice architecture. This is one of seven case studies exploring how the QuIP was used in specific contexts during 2016 and 2017.

Keywords: social impact evaluation, UK, local government, green spaces, voluntary sector, social innovation

Introduction

Central government in the UK has for several years been devolving responsibility for social services to lower tiers of government, making itself responsible for less, and local government responsible for more. At the same time, it has reduced the amount of money redistributed via central government to the local level (Sutaria et al., 2017). Increasingly local authorities, including city and town councils, are being called upon to help solve the problems created by budget cuts to local services previously prescribed and paid for by upper tiers of government.

Within UK local government the culture of impact evaluation based on data collection involving intended beneficiaries is much weaker than it is in the field of international development. Town councils, for example, still receive a significant amount of their funding from council tax paid by the town's residents, and their primary line of accountability is to them.[1] Councils are required to prepare externally audited reports, and to make

http://dx.doi.org/10.3362/9781780447469.009

them publicly available (Howes et al., 2013), but these typically do not play an important role in voter decisions at election time. However, city and town councils in the UK have recently begun to take more interest in demonstrating social impact, partly in anticipation that this may help them access non-governmental sources of social and community finance.

The field of international development is one potential source of ideas about how local governments might conduct impact evaluation studies. Abundant research indicates that such institutional transfer of knowledge is not straightforward (e.g. Dolowitz and Marsh, 2000: 9). First attempts to apply the QuIP in the UK were modest in scope and undertaken as pilots, and had a lot of latitude for learning and adjustment. They entailed making a large leap, but also from the voluntary to the public sector, and into a very different institutional context.

The first QuIP studies in the UK were conducted for local authorities in the south-west of England: the first for Voscur, an organization partnered with Bristol City Council (BCC); and the second for Frome Town Council (FTC). The adaptations made to the QuIP approach for both these studies were extensive and very different. In Bristol, an initial pilot was carried out and fed into an ongoing three year study. In Frome, the town council supported a Master's dissertation to investigate how FTC could evaluate the impact of its activities in parks and green spaces.

This chapter was finalized by Remnant from a draft produced by Morsink, who conducted the QuIP pilot in Frome (Morsink, 2017). James Copestake and Ed Howarth (who conducted the QuIP pilot for Voscur, while employed there as a member of staff) also provided advice and input into the final version. Members of Frome Town Council provided invaluable support for the Frome study but have not contributed to the writing of this chapter. The following section outlines the different motivations and needs of Voscur and Frome Town Council that led to them piloting QuIP studies. The chapter then considers the adaptations made to the QuIP approach for each and reflects on the lessons learned, and what the case studies revealed about the adaptability of the QuIP to different contexts.

The two QuIP pilot studies: selection of approach and scope of study

Both Voscur and FTC are small organizations, and in 2016 when work started on designing the QuIP studies neither had built-in monitoring, evaluation, and learning (MEL) functions.[2] Voscur monitored certain indicators regularly across the whole organization and reviewed changes over time using customer relationship management (CRM) software. Both organizations conducted more elaborate evaluations of selected programmes internally, and these were typically led by employees directly involved with the programmes. Both trialled QuIP in 2017, looking for a more robust way to demonstrate the impact of their organizations on their 'intended beneficiaries', namely the voluntary, community, and social enterprise (VCSE) sector in Bristol, and the town's population in Frome.

Voscur is Bristol City Council's main partner in supporting the VCSE sector in the city. Bristol City Council (BCC) is one of England's 56 Unitary Authorities, responsible for an urban area with a population of about 600,000. Frome Town Council (FTC) represents the most local tier of government in England and serves a population of about 27,000; it is one of around 10,500 similar councils in England. These used to play a much larger role, but now have more limited formal functions such as providing village halls, leisure facilities, playgrounds, and cemeteries; maintaining public footpaths; and funding cultural projects, community transport initiatives, and crime-prevention equipment. In addition, they must be notified of all planning applications and consulted on the making of certain by-laws. Figure 9.1 shows the administrative structure of England and the positions of both Unitary Authorities and Town Councils.

The QuIP came to Voscur and FTC's attention through personal contacts. Ed Howarth, who commissioned the QuIP pilot at Voscur, had a background in international development and learned about the QuIP through his

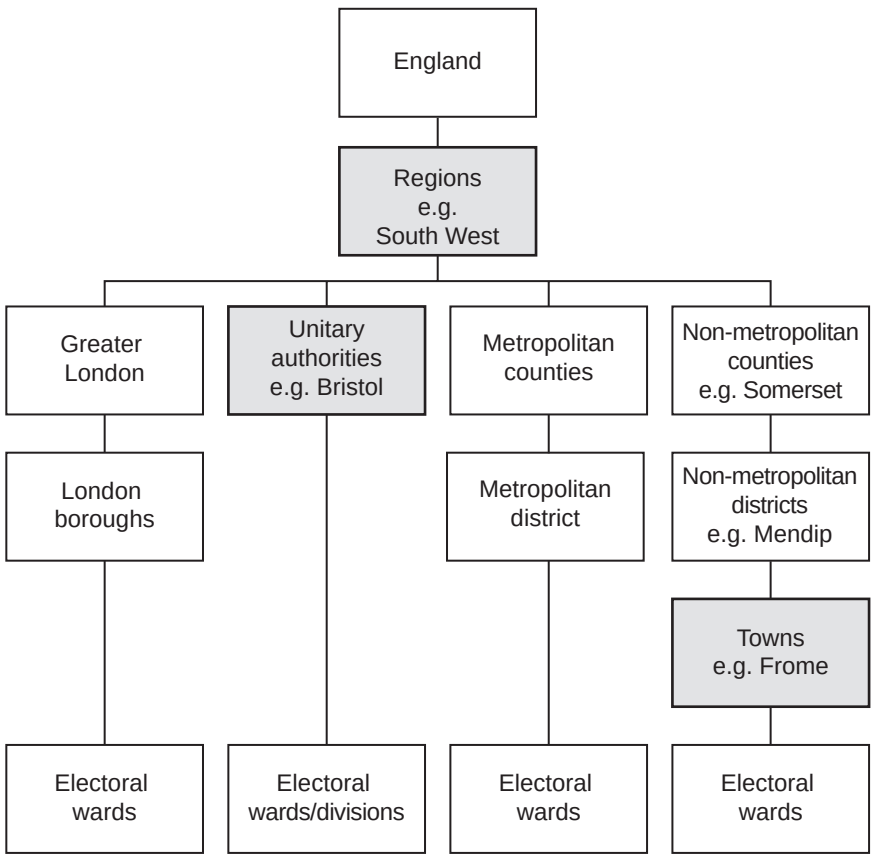


Figure 9.1 Tiers of government in England

involvement in the South West International Development Network (SWIDN). In Frome, the town councillors engaged a Master's student (Morsink) enrolled at the University of Bath to use her dissertation to explore the question of how they could assess their social impact as a council. Morsink learned about the QuIP through colleagues at the University of Bath, researched the Voscur pilot, and then proposed using the QuIP as the basis for a pilot study in Frome. Both organizations demonstrated entrepreneurial mettle through their willingness to trial the QuIP in quite radically different contexts from those in which it had hitherto been employed.

Voscur, along with its client organizations in Bristol's VCSE sector, was under increasing pressure to demonstrate what contribution it was making to improving the wellbeing of the city. In 2011, BCC rationalized its funding for VCSE infrastructure in Bristol, consolidating funding streams to five different organizations into a single grant provided to Voscur. This change was driven by feedback from the VCSE sector that organizations didn't know where to go for different kinds of support, and by BCC's need to have better insight into the effectiveness and efficiency of its support for the sector. The first cycle of funding for Voscur covered 2012 to 2016, with the second cycle covering 2016 to 2020. In 2015, BCC changed its commissioning guidelines for council grant funding to the VCSE sector with the aim of clarifying its key priorities and reviewing what the funding was expected to achieve (Bristol City Council, 2015). While Voscur had a fairly detailed approach to monitoring change, it was seeking new approaches to generate evidence of contribution and attribution. This was in direct response to new commissioning guidelines stipulating that applicants for BCC funding needed to be able to demonstrate what changes their interventions had brought about and how. Figure 9.2 depicts Voscur's own theory of how the QuIP could contribute to its MEL system. This illustrates how it aimed to use qualitative interviews to find out more about its impact on the capacity of VCSE organizations in the city, as well as to explore a range of other drivers of change.

The change domains selected for the study reflected the VCSE activities eligible for Voscur support. These were service delivery, access to funding, access to support services, people and skills, governance and leadership, and voice and influence. A major adaptation of the QuIP was that interviews would be conducted exclusively at the level of VCSE organizations, not at the individual level – moving data collection one level up the funding chain. No interviews were conducted with individual intended beneficiaries of the VCSEs supported by Voscur. This reflects the fact that for Voscur (and BCC), the goal of the study was to assess success in strengthening the institutional capacity of the VCSEs, not to assess their impact at the community level. Informant interviews with institutional partners had previously been included in QuIP studies (see Chapter 7, for example), but the Bristol pilot was the first to focus *only* on this level. Additionally, Voscur quickly decided not to attempt blind-folding, because lack of funds meant its own staff would have to conduct interviews, and do so in the context of ongoing institutional partnerships.

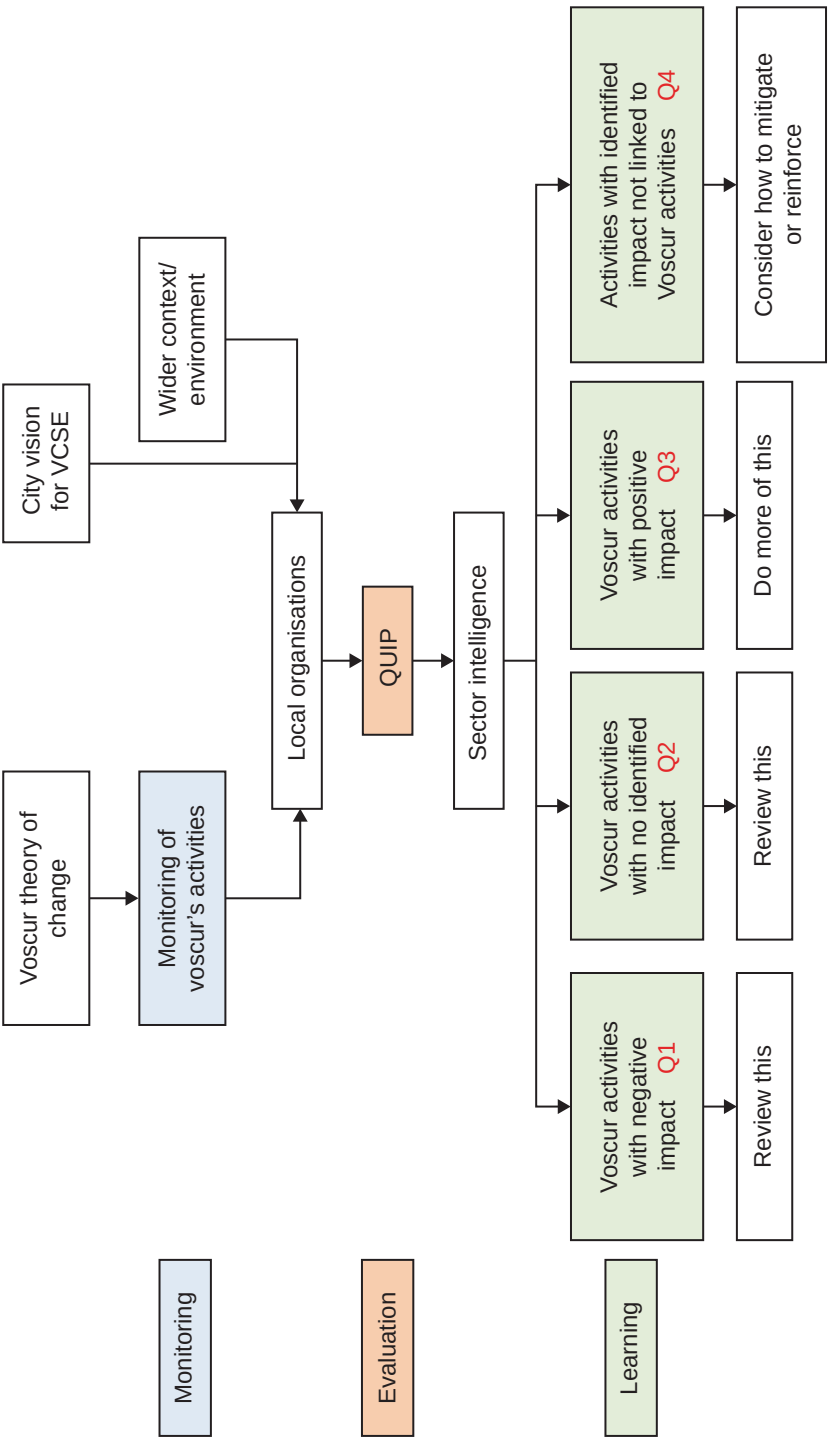


Figure 9.2 Voscur's framework for learning through use of the QuIP
Source: Voscur (2017)

The Frome QuIP pilot was an exploration of the social impact of a town council, through narratives of change recounted by the town's residents. Relative to the Voscur study, the Frome pilot was more closely aligned to previous QuIP studies, as respondents were individuals (rather than institutions), and a semi-blindfolded approach was used (respondents being unaware that the study was particularly interested in the impact of town council initiatives). Despite these overlaps the FTC pilot also departed significantly from the standard QuIP in other ways, as explained below.

Town councils in England have historically had limited responsibilities, and commensurately few formal channels through which to effect change in their communities. However, in 2011 the Localism Act was passed, endowing town and parish councils with 'powers of general competence' which formally enabled them to do whatever an individual is entitled to do providing they have a suitably qualified officer (usually the town clerk) in charge. This has freed up forward-looking town and parish councils to review their activities and venture into new territory. Elected councillors have always been able to act as statutory consultees on planning applications, representing the interests of the residents to higher tiers of government such as the district council. However, the 2011 Act empowered town councils to prepare a neighbourhood development plan which would, if adopted by referendum, become part of the district council's development plan for the area and would have to be used as a basis for making decisions on planning at upper tiers of government. In the wake of these regulatory changes, what a particular town council *actually* does came to depend to a larger degree on the energies and interests of individual elected officials, influenced also by their political party affiliations, and the competence of council staff.

Around this time, the town council in Frome was beginning to attract national and international media attention for its novel approach to local governance. Independents for Frome (IfF), a new grouping that eschewed conventional party politics, took control of the council. Established by a group of enterprising Frome citizens, it won 10 of the 17 council seats in the municipal election of 2011, going on to win all 17 seats four years later. At the time of the QuIP pilot in 2017, IfF had held a majority on the council for a total of six years.

IfF is a party without an overt political ideology. It operates primarily at election times as a way to enable candidates to stand, who do not wish to be affiliated with existing UK political parties (e.g. Conservative, Liberal Democrat, Labour, Green, UKIP). Once elected, IfF councillors are not subject to any formal leadership or 'party whip', and are free to work unconstrained by party ideology. IfF does, however, adhere to certain 'ways of working' – set out by two-term councillor Peter Macfadyen (2014) in his book *Flatpack Democracy: A DIY Guide to Creating Independent Politics*, as well as in official documents. These emphasize the values of independence, integrity, positivity, creativity, and respect. Both the selection process for candidates wishing to stand under the IfF banner, and councillor decisions when in office, are grounded in these values.

At the time of the QuIP pilot, there were 17 elected IfF councillors providing the strategic direction for the council, and 22 council staff under the leadership of the Town Clerk, responsible for implementation. There was

Box 9.1 Excerpt from Frome Town Council Strategy 2016–2020 (emphasis added)

Section 2.3: The core of our strategy

The central theme underpinning the council's approach will remain a focus on developing a sustainable town, but we have expanded what we mean by that. Everything we do and support will fall into three areas:

- Wellbeing: a flourishing and active community of people and organizations working together.
- Prosperity: a thriving business community, connected with each other and with the town, providing employment and prosperity.
- *Environmental sustainability: covering the attractiveness, variety and accessibility of the town's green spaces* and an increased focus on renewable energy, energy efficiency, waste reduction, and community transport.

Wellbeing, prosperity and environment are intrinsically interlinked. For example: we will look to focus business support in ways that enhance ethical, environmentally-sensitive business practice – strengthening business, wellbeing and environmental sustainability together. Similarly, many projects which enhance wellbeing also enhance green spaces, and a focus on green energy not only reduces emissions but also reduces costs and sustains regional economies.

Source: Frome Town Council Strategy 2016–2020

a high degree of collaboration and sharing of responsibilities between the two groups. The council produced documents describing its vision and approach, including a strategy document for 2016–2020 (see excerpt in Box 9.1), and annually updated work programmes and budgets.

FTC planned and implemented a wide range of interventions across the three core areas of wellbeing, prosperity, and environmental sustainability. To keep it manageable, the scope of the QuIP pilot was restricted to the social impact of the council team taking the lead on environmental issues, and more narrowly, the impact of this team's interventions relating to the town's green spaces. The council team shared a belief in the benefits to health and wellbeing of spending time outdoors in green spaces – whether relaxing, exercising or socializing; and this was reflected in a range of council activities intended to increase the variety, attractiveness, and accessibility of green spaces in the town. These included hiring rangers to look after all the parks day-to-day and to be on hand to help visitors; assisting and supporting community groups to improve local green spaces (e.g. building the Roundhouse on the Otherside, clearing access to and through The Dippy, saving Whatcombe Fields from sale to a developer); funding play and exercise equipment in parks (e.g. a toddler play area in Victoria Park and adult exercise equipment in the Old Showfield); promoting wildlife diversity and wildlife corridors (e.g. managing open spaces like Rodden Meadow); and organizing and supporting events in open spaces (e.g. fun days for children, and markets for all ages).

In the short to medium term, the FTC expected residents to spend more time outdoors in the fresh air, be around flora and fauna more, get more exercise, and have more opportunity to meet other residents. In the longer

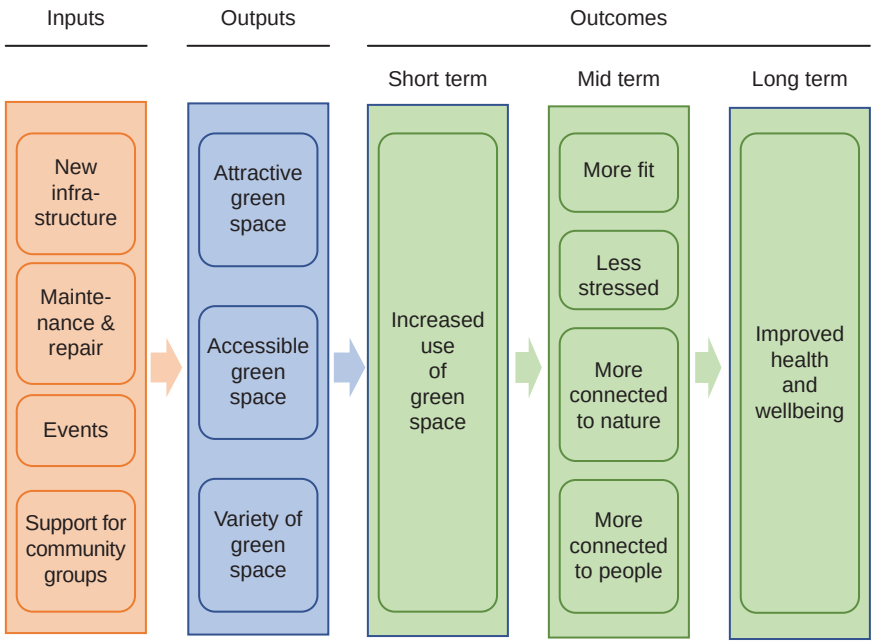


Figure 9.3 Frome Town Council theory of change for green spaces

term, this would – it was hoped – lead to a happier, healthier, and more interconnected population. Figure 9.3 summarizes this theory of change (ToC), as sketched by Morsink in conducting her research.

Methodological adaptations of the QuIP in Bristol and Frome

As the Voscur and Frome studies were conducted in such different settings compared with other QuIP studies, they entailed significant departures from the standard guidelines. This section explores five issues that influenced these adaptations:

- the focus on institutions (rather than households or individuals) as a point of reference;
- how power dynamics affected setting up interviews (including blindfolding);
- how the extent of monitoring data influenced sample selection;
- the relative significance of drivers of change as judged by respondents and by the ToC;
- challenges in defining and understanding what 'impact' looks like.

Institution-level interviews

The Voscur QuIP pilot broke new ground in data collection for the QuIP by conducting interviews exclusively at the institutional level. This demanded

adapting the design of the questionnaire, as well as the selection of interviewees to ensure that they would be able to respond on behalf of the organization. There were challenges associated with this: organizational change can be harder for a respondent to assess than change linked to their personal experience; and the respondent's assessment of change might be influenced by their personal position within the workplace and ongoing work-based politics. Staff turnover might also limit the ability of key individuals to assess change over time. Voscur's knowledge of the selected organizations enabled it to mitigate this issue by selecting respondents who had been in senior enough positions for long enough to enable them to discuss changes. But, as with household level QuIP interviews, there was debate over the potential benefits of interviewing more than one staff member. Time and budget constraints pre-empted this, as well as issues around buy-in from the organizations that would have had to commit more staff time. For household level QuIP studies this limitation is offset by complementing individual interviews with focus groups. With adequate funds this is also an option for organizational level QuIP studies, as the Global Health Service Partnership studies in Africa illustrated (see Chapter 8).

In a 2018 full-fledged follow-on study to this QuIP pilot, feedback from the independent research team suggested that not blindfolding respondents at this level could help with buy-in: interviewees were reported to be happy to take part, as they believed the research was important and were pleased it was being done. This was the case even for staff of very small organizations who faced serious time pressures. Some respondents also reported finding the process of discussing evolving challenges and opportunities beneficial, and were positive about the experience in general (corroborating similar reports from other QuIP studies).

As with other QuIP studies, the domains were determined by the organization's theory of change. There was a cost to using one standard questionnaire to cater for multiple respondents, given that organizations received different types and levels of support. However, asking all organizations about all areas of potential change helped to build up a broader picture of changes across the VCSE sector, whether driven by Voscur's activities or other drivers, and was consistent with the QuIP's exploratory potential. Indeed, while the QuIP study was useful for internal learning at Voscur, it was particularly useful in providing evidence of wider changes and influences in the sector for BCC. This was a factor in prompting BCC to ask Voscur to continue using the QuIP over the next three annual reporting cycles. Voscur plans to use the evidence to inform all its funders (not just BCC) about their impact, as well as about the key drivers influencing a vulnerable sector undergoing rapid change; and potentially contribute to the design of more effective support strategies.

Securing interviews and blindfolding

The Voscur pilot in 2017 was based on interviews conducted by staff of the commissioning organization itself (namely two Voscur development officers).

This was dictated mainly by tight time and budget constraints. The follow-up study in 2018 used a team of external interviewers, but again the organizations being interviewed were aware of who had commissioned the interviews. For the pilot, a small sample of seven organizations was purposively selected based on organization type, level of support received, availability, and willingness to be interviewed. The follow-up QuIP in 2018 used the standard sample size of 24, purposively selected to ensure a spread of organization types and exposure to support from Voscur.

In both the 2017 and 2018 studies, the organizations that took part in the interviews not only sought but expected to be fully informed of the purpose of the interviews, in line with professional norms governing transparency within the sector. The QuIP study was pitched as a 'state of the sector' survey and framed the questions broadly, rather than focusing narrowly on Voscur's own activities. As Voscur is of a comparable size and status within the VCSE sector to the respondent organizations, it was not appropriate to conduct the interviews blindfolded. Temporary blindfolding might have been possible if the study had been commissioned by a larger organization through an independent evaluator, as was the case with institutional level interviews conducted as part of Save the Children's HANO project in Tanzania (see Chapter 7), but recruiting an independent organization to conduct the study was not feasible within the budget available to Voscur for the studies.

Like the Voscur pilot, the FTC pilot was conducted with the same person (Morsink) performing the roles of field researcher, data analyst, and lead evaluator. This precluded blindfolding the interviewer, hence interviews were conducted with only the respondent partly blindfolded. The respondent was informed that the interviewer was affiliated with a local university, and that the purpose of the research was to fulfil academic requirements as well as to find out more about green spaces in Frome. FTC was not named as the commissioner of the study, but the interviewer did explicitly state that the research would be shared with FTC.

In Frome, the length of the interview had to be shortened. Interviewees recruited ad hoc in green spaces would not commit to an hour-long interview, the median duration of interviews in previous studies. Instead they offered 10–15 minutes of their time. In practice, once respondents engaged with the topic, and with the interviewer, then many interviews lasted longer, some for up to an hour. However, setting a short time limit initially did interfere with establishing rapport with the interviewee. Indeed, the time limit seemed to serve as a protective strategy for some respondents who were uncomfortable, unwilling, or unable to share more detail about their lives and choices. Some questions probing for details about changes to exercise patterns, health status, and social connections over the past two to three years, were possibly perceived as overly personal and met with non-committal or vague answers.

Although blindfolding can be a useful, it is not the only way that QuIP data collection adds value. The systematic exploratory structuring of questions in

the QuIP, and the training of interviewers to ask questions in a way which gives the respondent the maximum opportunity to tell an accurate story in their own words, are just as important. The QuIP's main purpose in employing blindfolding when setting up interviews is to enhance the credibility of the data collected. Confirmation bias is widely viewed as a weakness of self-reported impact attribution, and blindfolding is a credible way to reduce the perceived risks of project-related confirmation bias. However, the nature and extent of confirmation bias remains unknown (Copestake et al., 2018: 6), and it is a logical fallacy to infer that because blindfolding can mitigate against confirmation bias, confirmation bias will necessarily increase without blindfolding. Both the Voscur and FTC QuIP studies enabled respondents to tell their stories in their own words, and demonstrated the QuIP's scope for flexibility in data collection.

Monitoring data and sample selection

Improving a population's wellbeing by making a product or service available and accessible is a common aspiration of development practitioners. Such approaches assume that the intended beneficiaries actually use the product or service, and that benefits accrue from this use. To evidence this use, programme implementers can monitor and keep records of usage. For instance, microfinance institutions (MFIs) have information systems to record clients by name and monitor their financial transactions (see Chapter 5 for a case study of the social impact of housing improvement loans in India). In the case of Frome's green spaces, however, FTC did not monitor who used the parks and the other land it managed, and the absence of such data had consequences for establishing a sampling frame for the Frome QuIP pilot. Frome has a population of about 27,000, all of whom are 'intended beneficiaries' (broadly defined) of FTC's activities. For systematic monitoring of the town's population, the council turns to government data collected at either the national level (e.g. census data), or at the county and district levels (e.g. data sets maintained by the Somerset Intelligence Unit). The town council itself does not have the capacity to do systematic longitudinal monitoring. With no recorded intended beneficiaries of FTC's interventions relating to the town's green spaces, there was no ready sampling frame, such as is available for most QuIP studies.

Without such data, the Frome QuIP pilot identified users of green spaces 'at-point-of-use', interviewing respondents outdoors in parks, meadowlands, and playing fields. This approach ensured that a respondent was indeed a user of at least one green space and had direct experience of the physical interventions made there by FTC. This approach also allowed for at least approximate classification of respondents into sub-categories based on the location of the interview (by neighbourhood), and visible personal attributes (such as age and dog ownership). It was not possible, however, to purposively select participants according to other characteristics potentially of interest to the

Table 9.1 Breakdown of interviews by green space

Green space	*Areas included*	*Interviews*
Victoria Park	Victoria Park field, toddler playground, Mary Baily fields	7
Welshmill	Welshmill playground, BMX pump track, Other Side Roundhouse	5
Old Showfield	Children's playground, adult exercise equipment, meadows	7
Rodden Meadow	Meadow and river, Millennium Green, New Road playground	7
The Dippy	Pathway and culvert	6
Total		32

Table 9.2 Breakdown of interviews by respondent characteristics

Gender		*Age (years)*		*Years in Frome*		*Dependants*	
Female	18	≤24	5	≤20	11	Child <10 years	15
Male	14	>24	22	≤20 but entire life	5		
		>67	5	>20	16	Dog	13
						Both	7
						Neither	11
Total	32		32		32		NA

commissioner, such as length of residency in Frome, the proximity of green spaces to a respondent's home, or pre-existing levels of health and wellbeing.

For the pilot, 32 individual interviews were conducted in five green spaces managed by FTC.[3] Sub-categories of users were created first by spreading the interviews across the different green spaces, and secondly by targeting respondents on the basis of their perceived age, gender, and guardianship of a young child or dog. The sample is described in Tables 9.1 and 9.2.

The interviews were conducted over the course of a week, with the interviewer in the field for eight hours each day. This approach precluded respondents who might have used a given green space at a different time from that at which the interviews were conducted, for example in the evening rather than the morning, or at the weekend rather than during the week.[4]

Relative significance of drivers of change to respondents and to the ToC

Crafting the questionnaire and interviewing respondents in Frome posed further adaptation challenges.[5] The questionnaire was trialled with potential respondents and went through several iterations, as is common research practice. However, the original QuIP questionnaire format required a full overhaul rather than mere tweaks, raising the question of whether the final version was an adaptation of the QuIP or something inspired by it but entirely different.

The pilot questionnaire was reconfigured twice before being rolled out. Initially the change domains were based on the mid-term outcomes in the ToC for green spaces (see Figure 9.3), taking a typically open and broad approach to questioning. However, the first set of trial interviews produced few mentions of green spaces, and no comments on any change in use of green spaces. Respondents attributed any increase in fitness or reduction in stress to other drivers, including gym membership and conditions at work or at home. Asking about connections to nature elicited narratives of country walks and visits to national parks; connections to other people prompted general comments about family and friendship circles. Next, the open questions were narrowed on the basis of using the short-term outcomes in the ToC: namely changes in use of green spaces – whether for activity, rest, or socializing. However, respondents reported that they had not perceptibly changed their use of spaces over the last three years, except in cases where a green space no longer existed (because it had been bought for housing development, for example).

With no significant change to report, expected drivers of change which might have been linked back to FTC were not mentioned (for example: the council's improvements to play areas, fun days in the parks or protection of wilderness areas). The final iteration of the questionnaire asked respondents to describe their use of, and reasons for using, green spaces in Frome, together with any changes they might have noticed to green spaces. The interview incorporated prompts in the form of lists of green spaces in Frome (whether managed by FTC, the district council, or privately owned); lists of events and activities regularly held in green spaces (by a variety of local groups); and lists of government, community, and volunteer groups (among which was FTC) involved in using or looking after green spaces.

Respondents were on the whole knowledgeable about the availability of green spaces in Frome, and very aware of physical interventions that had been made to those that they frequented. They were far less au fait with the range of activities or events available at different times of year, or with the various community groups organizing them. When it came to attribution, respondents generally hesitated to attribute interventions with any confidence to FTC or any other parties, except on those occasions where the respondent had been directly involved with the organizers, or personally knew one of them. Respondents often said they assumed FTC had done things, not on the basis of specific knowledge but because they thought those were the 'kinds of things' that councils had authority over, or because they had seen park ranger vehicles with FTC markings in the vicinity of the green space. Knowledge regarding what was the town rather than district council's responsibility was mostly hazy or incorrect. This reflected in part the level of citizen engagement with local politics, but also that FTC did not generally advertise its involvement in projects by marking infrastructure or events with plaques or the FTC logo.

Defining and understanding 'impact'

This extensive adaptation of the Frome QuIP pilot questionnaire meant that data was no longer being gathered on respondents' experiences of change in an open-ended manner. Analysis therefore could not elucidate causal pathways of change, or attribute change. What the questionnaire *did* elicit were *narratives of choice*: why a respondent chose (or would have chosen) to use one green space or another. The analysis of the data shifted from identifying drivers of change in behaviour, to identifying inputs to choice.

Attribution remained a useful concept in this analysis, defined as who or what was creating the possibility of that choice. For example, by building a toddler playground, FTC created the possibility for residents to take a toddler to play in a safe and appropriate green space (whether or not this happened in practice), and this had an impact on perceptions of wellbeing among residents. If the toddler playground hadn't existed, residents wouldn't have had *that* choice – and it was an option that existed thanks to an FTC intervention. This adaptation of the QuIP to look at choice architecture rather than more tangible short or mid-term impacts as described in the ToC, was useful in providing evidence of potential mechanisms linking changing context with changing outcomes, even in a case where pathways of change could not be more explicitly identified.

Conclusions

This chapter reported on two pilots of the QuIP in the UK, and explained how these demanded a significant amount of adaptation. In Bristol, Voscur used unblindfolded interviews to investigate its role in building capacity of voluntary and community social enterprises, and generated credible findings, not least from the perspective of Bristol City Council which authorized a larger follow-up study. That pilot also demonstrated that the QuIP's systematic approach to domain selection, coding, and data analysis could still be used effectively with the data collected. In Frome, outcomes were explored that were of a different calibre and more marginal magnitude than those the QuIP had investigated previously. The pilot responded by shifting focus from pathways of change to architectures of choice, but retained a focus on attribution.

The rest of this section focuses on what these experiences tell us about when a QuIP is appropriate and when it isn't, and to what extent it needs to be adapted. The experiences from these two UK pilots highlight three important questions which should be asked at the outset of any study:

- How much prior knowledge does the commissioner already have?
- Given known characteristics of intended beneficiaries, what form of interviewer–interviewee relationship is appropriate?
- What does positive change look like for the project?

How much prior information is required?

The Frome pilot highlighted two factors (in absentia) that ordinarily serve to enhance the relevance and impact of the QuIP. The pilot found itself in unfamiliar terrain, without monitoring data or a theory of change tailored to the beneficiary population. Although this demonstrates how a ToC may not be indispensable to conducting a QuIP, having a clearly articulated ToC does facilitate the definition of questionnaire impact domains and the coding of implicit attribution. Similarly, descriptions or reports of interventions are indispensable to attribution of causal claims when the commissioner isn't identified by name, and for spotting 'missing' narratives where reported change would have been expected. As a result, this QuIP study was necessarily mostly exploratory. How useful a QuIP is to a commissioner depends on the commissioner's expectation of the balance between confirmatory and exploratory findings, which in turn is related to the extent of the ToC and programme monitoring data.

The importance of these two 'anchors' to a QuIP study further highlights the importance of planning the programme cycle holistically, in particular thinking about evaluation and impact assessment *before* implementation starts, not once it has already taken place. Thinking about how to change and improve a programme can only happen if there is an understanding of the current state of the intervention. Some commissioners are keen to jump ahead to the question, 'How did we impact people's lives?' and pass over crucial interim questions, including: 'Did we actually do what we said we were going to? Who did we engage with?'

Comparing the experiences of these QuIP pilots to the other case studies in this book leads to the conclusion that the QuIP works best in contexts where:

- there is a defined population of intended beneficiaries;
- there is monitoring data on population attributes relevant to the commissioner;
- there are records of population exposure to the commissioner's intervention;
- the intervention is defined and evidence exists to show the intervention has been implemented as intended;
- there is some reason to believe change has occurred, ideally backed up by some form of monitoring data to substantiate that change has occurred.

The QuIP then comes into its own, offering insight into the how and why of the change; and who or what was at the root of the change. These criteria are not all necessary, but they are important, and where any are lacking it requires more adaptation to accommodate a QuIP to the context. Much of the need for the adaptations of the QuIP described in this chapter was due less to differences between the UK and other countries where it had already been used, and more to the lack of this kind of information.

It is possible to think of examples of activities in the UK where conducting a QuIP would have been easier. Indeed, in Frome just such an example came to national attention in early 2018, sadly too late for the research undertaken in 2017 (Sutaria et al., 2017; Monbiot, 2018). The Compassionate Frome project was launched in 2013 by the medical practice in Frome, in collaboration with the NHS group Health Connections Mendip, and supported by FTC strategically and financially. This community-based welfare scheme linked the health centre, the community hospital, and social services with local charities and other voluntary groups providing care. In addition to creating a directory of services, it employed 'health connectors' and trained 'community connectors', to help Frome residents plan their own care and find the support they needed. Alongside improved access to information, volunteers offered health and non-health related help, such as transportation and mobility, home-care, grocery shopping, and support in joining social activities. As an indicator of the success of this approach to health provision, champions of the programme pointed to a drop in the number of emergency hospital admissions: in Frome, these fell by 17 per cent over three years, while across the whole of the county of Somerset, they rose over the same period by 29 per cent.

The correlation between the programme and the relative changes in emergency hospital admissions in the area is astonishing, and claims to causal links are indeed very tempting to make. Yet how and why the drop in emergency hospital admissions occurred has yet to be explored, and the link to the Compassionate Frome interventions has yet to be substantiated. Systematically collecting and analysing narratives of change from the population concerned would vastly increase understanding of how and why the programme worked and what the causal pathways were.

What form of interviewer–interviewee relationship is appropriate?

Rapport between the field researcher and respondent is important in all qualitative research, but in particular when asking open questions requiring respondents to reflect on and divulge personal experiences of change and reasons for change. Such rapport is affected by the balance of power between field researchers (and the commissioners on whose behalf they are working) and respondents, and pertinent to the choice to blindfold.

If the commissioner has control over resources that the intended beneficiary wants or needs, then blindfolding respondents reduces their incentive to seek advantage by telling the interviewer what they think the interviewer wants to hear. Blindfolding both the respondent and the interviewer goes even further: by placing interviewer and respondent on a more equal footing it can encourage interviews to be conducted in a more equal and reciprocal way.[6]

In the Voscur pilot, there was no significant power difference between interviewer and interviewee; and there was no obvious power advantage to be gained by the interviewee, as an institutional representative, from reporting that Voscur was contributing more to the functioning of their organization

than they knew to be the case (indeed there might have been disadvantages associated with doing so). As a private individual, the interviewee might be eager to please the interviewer, but this would be offset by their professional interest in doing what was required by their organizational role. The case for blindfolding was correspondingly weaker. Hence the Voscur experience demonstrates that there is no one answer to the use of blindfolding in QuIP data collection.

The time limits enforced by some Frome respondents on interviews may have curtailed rapport-building in some cases, and it is possible there is a socio-cultural dimension to reticence in the UK. Where answers from Frome respondents were vague or non-committal, it was not always clear whether this was a function of respondent reticence or poor recall. It is also possible that respondents were flummoxed by the invitation to converse openly with a total stranger, being culturally conditioned to respond to surveys with tick boxes and rating scales. It is also conceivable that in a written rather than oral culture, people have difficulty remembering details of their day-to-day experiences without aides-memoires in the form of diaries or documentation.

What does positive change look like?

There are three dimensions to this topic: whether maintaining the status quo counts as 'change'; the importance of the drivers of interest to the commissioner (typically based on their own interventions) *relative* to other drivers affecting the hoped-for change; and what role non-observable mental changes, for example changes to a respondent's choice architecture, may have on wellbeing.

Both pilots raised the question of whether there can be impact without change. Even if we can agree that change indicates impact, it is a mistake to conclude that an absence of change necessarily indicates an absence of impact. Maintaining the status quo can require intensive and ongoing intervention. In such situations, lack of change can be an indicator of impact: interventions have had the desired impact of preventing a deterioration of the status quo. Local authorities can have a large impact just by ensuring there is no decline in the quality of life in a town or city despite budget cuts and other shocks; under such circumstances they might be doing a good job if they keep things running smoothly, possibly making improvements only at the margins.

Similarly, other QuIP studies looking at rural development projects included areas which were negatively affected by climate change. Project theories of change may have included efforts to mitigate the worst effects of climate change for subsistence farmers. Outcomes which then speak of 'no change' in the context of a deteriorating environment may in fact indicate success; the latent counterfactual here is how much worse outcomes could have been in the absence of any action, rather than how much improvement has been recorded. It is important that the lead evaluator and commissioner consider this from the outset (whether or not it is apparent from the theory of change

and any monitoring data available) to ensure a suitable approach to coding is adopted when it comes to the analysis.

Second, the Frome study dealt with interventions whose short-term outcomes did not loom so large in people's day-to-day lived experiences as those featured in other case studies in this book, even though all interventions were concerned with improving the wellbeing of intended beneficiaries. This highlighted that every commissioner considering a QuIP should first review how their intervention ranks relative to other factors contributing to outcomes envisaged in the ToC. In the case of Frome, when it comes to increasing the use of green spaces, this would entail considering how enhancements to vegetation might rank relative to installing exercise equipment or providing outdoor shelters. And when it comes to improving health and wellbeing, how does increased use of green spaces rank relative to going to the gym, eating organic food or decreasing stress at work? If those other factors dominate, a respondent in Frome may not mention the council's interventions at all. If a commissioner believes that their intervention is relatively marginal compared with other drivers of change, and that people are unlikely to refer to it when asked broad questions about change in a particular domain, then a standard QuIP approach is probably not suitable.

This notion of *relative* importance can be controlled for by broadening or narrowing the scope of the impact domains in the questionnaire. For instance, the data collected using a narrower domain (e.g. physical health or tertiary education) will differ from that collected using a broader domain (e.g. overall wellbeing); narrowing the domain could avoid letting the intervention of interest get drowned out by other drivers. However, as with the decision over whether to blindfold researchers and respondents, deciding to make domains or questions more specific means accepting a trade-off. At what point does the domain become so narrow that there is only one answer possible? At what point may the information collected no longer be considered unprompted or unbiased? This trade-off should be factored into the evaluation design, and weighed alongside the value added by the coding and analysis aspects of the QuIP approach.

Lastly, just knowing (or believing) that one has a choice can arguably have a big impact on wellbeing, even if one never exercises that choice (i.e. even if one doesn't make or experience any observable changes). For example, some Voscur-supported institutions reported that knowing that there was support and advice available if they needed it contributed to their improved confidence. Various Frome respondents commented that although they had not been to a particular green space for years, if ever, it made them happy to know that those spaces existed. A more traditional approach to impact evaluation focussed on measurable change would struggle to reveal the benefits of changes to architectures of choice or belief, though such mental structures might be significantly linked to levels of wellbeing. The Frome pilot showed that the QuIP can be adapted to attribute pathways of choice rather than pathways of change – but the QuIP approach is better suited

to situations where respondents provide narratives containing observable markers and milestones.

This chapter set out to reflect on the question of whether using the QuIP in the UK might change it beyond recognition. Certainly, both pilots encountered significant contextual differences in power dynamics, gatekeeping issues, reticence, and recall; as well as differences in prior MEL activity, funding chains, and conceptualization of beneficiary populations – and consequently adaptation was more radical than in previous studies. The QuIP nevertheless proved sufficiently malleable to allow responses to these challenges, demonstrating how it offers a useful starting point for agile impact assessment and attribution – an approach that can bend without breaking. That said, the insights that can be gained from using the QuIP are much strengthened through more cogent upfront planning and integration into wider MEL activities over the course of the activity being assessed.

Notes

1. Town councils are funded in a first instance by council tax paid as a percentage – or 'precept' – of the amount collected by the district council. Town councils do not receive any government funding or income from business rates, but their budgets may significantly expand through gifts and donations, loans and grants, and other income.
2. Voscur had 27 employees, and Frome Town Council consisted of 17 elected councillors, plus 22 full-time and part-time staff.
3. No focus groups were held owing to the tight time frame for the pilot, and because of delays resulting from the sampling issues and questionnaire adaptation.
4. Ways around this limitation in future, short of monitoring park usage on a regular basis, include hiring a large team of researchers to interview in all the green spaces around the clock, or conducting an online usage survey ahead of conducting the QuIP.
5. Adaptation of interviews is a feature of all QuIPs, but seems to have been particularly challenging in Frome.
6. This is not to suggest that blindfolding alone is sufficient to guarantee appropriate conduct in interviews where such power distance is large. Careful selection and training of field researchers is also critical to ensuring interviews are respectful and based on developing a good rapport (Copestake et al., 2018). Being 'local' to the area can help – not least with respect to language – but is not itself sufficient to ensure this.

References

Bristol City Council Investment and Grants Team (2015) *The VCS Grants Prospectus: A Proposal for a New Approach to Voluntary and Community Sector Grants* [online] <https://www.bristol.gov.uk/documents/20182/303221/Grants+consultation+prospectus+proposal/920b9299-7dcb-4e5f-a572-07593f1c5a61> [accessed 5 November 2018].

Copestake, J., Remnant, F., Allan, C., van Bekkum, W., Belay, M., Goshu, T., Mvula, P., Remnant, F., Thomas, E. and Zerahun, Z. (2018) 'Managing relationships in qualitative impact evaluation of international development practice: QuIP choreography as a case study', *Evaluation* 24(2): 169–84 <http://dx.doi.org/10.1177/1356389018763243>.

Dolowitz, D.P. and Marsh, D. (2000) 'Learning from abroad: the role of policy transfer in contemporary policy-making', *Governance* 13: 5–23.

Frome Town Council (2015) *Frome Town Council Strategy 2016–2020* [pdf] <http://www.frometowncouncil.gov.uk/wp-content/uploads/2014/04/Frome-Town-Council-Strategy-2016-20.pdf> [accessed 5 November 2018].

Howes, L., Skinner, E. and Derounian, J. (2013) *The Good Councillor's Guide: Essential Guidance for Parish and Town Councillors* [online], National Association of Local Councils. <https://www.nalc.gov.uk/publications> [accessed 19 December 2018].

Macfadyen, P. (2014) *Flatpack Democracy: A Guide to Creating Independent Politics*, Bath, UK: Eco-Logic Books.

Monbiot, G. (2018) 'The town that's found a potent cure for illness – community', *The Guardian*, 21 February <https://www.theguardian.com/commentisfree/2018/feb/21/town-cure-illness-community-frome-somerset-isolation>.

Morsink, M. (2017) *An Exploration into How to Assess Impact for UK Town Councils: Adapting the Qualitative Impact Protocol for Frome*, MRes Dissertation, Department of Social and Policy Sciences, University of Bath.

Sutaria, S., Roderick, P. and Pollock, A.M. (2017) 'Are radical changes to health and social care paving the way for fewer services and new user charges?' *BMJ* 358 <http://dx.doi.org/10.1136/bmj.j4279>.

Voscur (2017) Measuring the changes affecting Bristol's VCSE and the impact of Voscur's Infrastructure Support Service

About the authors

Marlies Morsink, MRes Social and Policy Sciences and MBA Finance, is Project Manager at Bath Social and Development Research Ltd. She previously worked in business consultancy, finance, journalism, and community development. She adapted the QuIP for use by a town council in the UK, and has researched stakeholder experiences of QuIP across a range of contexts and countries.

Fiona Remnant, MSc International Policy Analysis, is Managing Director of Bath Social and Development Research (BSDR), and has worked in development for over a decade, specializing in the application and communication of academic research to practitioners and policymakers. She has worked for the Centre for Poverty Analysis in Sri Lanka, Oxfam in the UK, and the Centre for Development Studies at the University of Bath. She collaborated with James Copestake on the Assessing Rural Transformations action research project at the University of Bath between 2012 and 2016 which culminated in the development of the QuIP and the creation of BSDR.

CHAPTER 10

Analysis and conclusions

James Copestake and Fiona Remnant

This chapter draws practical conclusions from 10 case studies of using the Qualitative Impact Protocol (QuIP) in diverse contexts during 2016 and 2017. To do so it reports on a comparative thematic analysis based on the stages of each study, as follows: (1) scoping, especially the balance between exploratory/confirmatory goals and internal/external audiences; (2) detailed design, especially the influence of prior theory, availability of potentially complementary quantitative data and options for combining a QuIP study with other evaluation activities; (3) data collection, especially the importance of social relationships in the field and careful coordination of different contributors to the study; (4) analysis, including scope for more systematic and transparent coding and visualization of data; and (5) use, including willingness to engage with researchers and other stakeholders in collaborative and creative interpretation of findings. More generally, the case studies illustrate the ultimately political nature of impact evaluation as a device for structuring deliberation over what is working, how, for whom, and why. The chapter concludes that by unpacking the stages of social research and fostering an embedded, practical, and flexible approach to tackling the underlying attribution challenge, the QuIP can contribute to a more agile, adaptive, effective, and meaningful approach to doing development.

Keywords: impact evaluation, causal attribution, qualitative research methods, international development, adaptive management, mixed methods

Introduction

This chapter draws on the case study material presented in the book to reflect on the Qualitative Impact Protocol's (QuIP) relative strengths, weaknesses, complementarities, and potential to contribute to more agile evaluation and international development practice. In doing so, it is important to emphasize the primacy of our practical over our academic intent. Judgements about impact evaluation can be informed by many abstract criteria: validity, credibility, timeliness, reliability, sufficiency, and so on (see Chapter 2). Here we reflect on how the QuIP was able both to meet the expectations of commissioners and other users, and to challenge them. This means reflecting on the QuIP studies in a way that takes into account users' prior knowledge, resource constraints, and operating environments. It also means reflecting on how the QuIP addresses the attribution challenge not in isolation, but in the context of

http://dx.doi.org/10.3362/9781780447469.010

the four challenges facing development organizations (the other three listed in Chapter 1 being goal formulation and planning, change monitoring, and adaptive management). To view impact evaluation in this wider perspective is to emphasize the need to manage trade-offs between practical usefulness and the highest standards of scientific rigour.

Much discussion of attribution and of impact evaluation takes a more abstract and academic view. And of course, academic specialists have an important role to play in informing impact evaluation and development practice, including as independent commentators, peer reviewers, specialist advisers, and critical friends. But the intermediate feedback loop to which impact evaluation contributes has a more direct audience and practical purpose than academic research (as discussed in Chapter 2). In an international development context, it includes enabling organizations to be not just more adaptive, but also more responsive to the voices of their intended beneficiaries. Assessing the authenticity of claims to be so entails looking at studies as actually used and in relation to what is possible, rather than holding them up against abstract criteria.[1]

The comparative analytical part of this chapter relies mostly on thematic analysis of the seven case studies presented in Chapters 3 to 9, with some reference also to other QuIP experiences. The five themes considered in turn in this chapter are as follows:

- QuIP commissioners: purpose, priors, and priorities;
- Reasons for using the QuIP and links to other sources of evidence;
- Designing QuIP studies: timing, scope, and sampling;
- Implementing QuIP studies: data collection and analysis;
- From evidence to use: workshops, decisions, and dissemination.

Comparative thematic tables which served as an intermediate step in this analysis are reproduced in the appendix to this chapter. The generalizations this analysis provides remain highly subjective, and may not reflect the views of other contributors to the book. The same applies (even more so) to 10 'takeaway' lessons on how to do qualitative impact evaluation more effectively, listed below.

1. *Use benchmarks flexibly*. Inflexible attempts to transfer evaluation blueprints or benchmarks to new contexts often fail. But more flexibly used standards or benchmarks (such as the QuIP) can be useful as a starting point and common reference for discussion and adaptation.
2. *Learn incrementally*. The most cost-effective impact evaluation does not start by assuming a blank slate nor by assuming the evaluation can answer all possible questions. Rather, it builds critically on prior knowledge and understanding. This entails making trade-offs, e.g. between generating more self-contained, certain, and precise attribution claims; and claims that test prior knowledge are more comprehensive, context contingent, and timely.
3. *Combine confirmatory and exploratory goals*. Investment in independent impact evaluation can most easily be justified when it has the potential

both to confirm/challenge prior evidence and theories of change *and* to explore unintended/unexpected changes and their causes.

4. *Deliberate early*. The scope for an impact evaluation study to contribute to challenging and changing understanding within a commissioning organization is linked to how actively and widely potential users engage in designing the study. This includes elements such as the extent to which stakeholders identify with the study, trust the evaluation team, manage expectations in relation to results (see point 2), and are able to apply findings to live questions.
5. *Address possible interviewing biases*. The risks of biased self-reported attribution in qualitative evaluations can be reduced in many ways, including: blindfolding, distancing field staff from implementing agencies, and data collection around wellbeing outcomes rather than interventions (i.e. working back from outcomes to causes in an open-ended way).
6. *Integrate qualitative and quantitative methods*. Quantitative monitoring of change in key indicators, and qualitative enquiry into causal drivers of that change, are highly complementary. Qualitative data can also usefully be summarized quantitatively in tables and charts, using labels and dashboards to enable rapid reference back to the original text so as not to hide the people and the narrative text underpinning the visualizations.
7. *Address contextual complexity*. Attempts to quantify impact using statistical inference are likely to be poor value for money, and can be highly misleading, unless sufficiently informed by prior qualitative research and elaboration of theory concerning alternative possible causal explanations for observed or expected changes.
8. *Follow through*. Producing and delivering a written report is rarely sufficient: involving staff in analysis (including training in use of data dashboards), sense-making workshops (including with intended beneficiaries), and joint presentation and publication of findings, can add substantially to impact.
9. *Contribute to middle range theory*. While a study may contribute to specific operational decisions (e.g. to close or scale-up a project) this is hard to document. The contribution that studies can make to the evolution of an organization's broader understanding of how change happens (and its own role in these processes) is also hard to establish, but may ultimately be more important.
10. *See evaluation as a political, social and moral process*. Professional understanding of impact evaluation theory and methodology is a necessary condition for effective design and execution of qualitative, quantitative, and mixed method studies. But conducting effective studies is also a political process constrained by power relations between stakeholders, and a social process that hinges on investing time in building mutual understanding, trust, and respect. This includes a moral obligation to involve intended beneficiaries in the outcomes of the study in whatever way is possible.

The book concludes with reflections on the process of developing the QuIP as a case study of institutional innovation, and on how the book contributes to thinking about the wider challenge of 'doing' development in ways that are more agile, responsive, and effective.

QuIP commissioners: purpose, priors, and priorities

The commissioners of the QuIP studies featured in this book ranged from a multinational corporation to a town council, and included international NGOs and philanthropic organizations. They are all listed in the Appendix in Table A10.1, along with information about additional studies for two other international NGOs, Oxfam and Self Help Africa (SHA) formally known as Gorta Self Help Africa (GSHA), and an impact investor, Acumen (these studies were briefly introduced in Chapter 1, Box 1.4).

The Save the Children and SHA case studies conformed to the model of time-bound projects that link demand for impact evaluation to the requirements of official donors, as well as to decisions over whether to continue, replicate, adapt or close down specified projects (this was also the case for Seed and C&A Foundation). But it is striking that other case studies concerned investments that were less rigidly tied to official donor project cycles, being based on activities financed by private supporters (Tearfund), foundation income (C&A Foundation, Terwilliger Center), corporate social responsibility budgets (Diageo), private investors (Acumen), and tax payers (Frome Town Council). In this sense, the case studies reflect the diversity of development finance, if not quite the magnitude of the 'philanthro-capitalist turn' recorded by McGoey (2014).

Variation in the primary audience of the study helps to explain commissioners' different attribution priorities. Seed was the strictest in viewing the studies commissioned as an investment in internal learning, giving a low priority to external publication. For others, generating evidence that could be shared with supporters and could help to back up their claims to positive social impact emerged as the main priority, even when this was not explicit at the outset (see section headed 'Wider dissemination', below).

The extent to which the QuIP studies were able to draw upon an agreed theory of change (ToC) for the intervention also varied widely. Theories of change were often held tacitly within commissioning and implementing organizations rather than being explicitly written out and agreed. Indeed an unanticipated effect of several QuIP studies was to prompt and assist the commissioner to develop the ToC underpinning their project more explicitly. In the case of C&A Foundation in Mexico, the QuIP study served partly to stimulate debate over how far theory underlying the project was congruent with a revised set of ToCs for C&A Foundation's programmes at the global level. Other evaluation consultants have commented that clarifying the theory behind a project is often an under-recognized and under-budgeted part of the contribution they find themselves having to make.

Formal terms of reference for QuIP studies generally focused on procedures and deliverables, without specifying whether they should confirm or refute specific causal claims.[2] Indeed, most commissioners emphasized that the exploratory potential of the QuIP – i.e. the potential to identify unanticipated outcomes and drivers, as well as the ability to confirm or refute prior expectations – was an important factor in motivating them to use it.

Several commissioning organizations, in addition to having prior theories of change, also drew on explicit normative frameworks to guide selection of the domains within which positive impact was anticipated. Tearfund's 'Light Wheel' provides the clearest example, while Yo Quiero Yo Puedo's (YQYP) psycho-social framework for thinking about empowerment was one of the more sophisticated. These frameworks facilitated evaluation by making it easier to modify the domain structure of data collection schedules to align with the goals of the evaluation. Adapting the QuIP to suit the values and priorities of the commissioner in this way also helped to build a sense of ownership in the study, and confidence that blindfolded interviews would be a fair test of whether desired outcomes were being achieved. The Terwilliger Center, C&A Foundation, and Acumen were most proactive in tailoring questions to cover areas of interest, to the point of limiting the exploratory function of the study and compromising blindfolding. But doing so also accentuated the potential power of null returns (i.e. failures to report attributable impact in specified domains), and the commissioner's commitment to learning from the findings.

To sum up, commissioners often sought ways of investing in impact evaluation that could serve multiple purposes, not all of which were fully explicit: e.g. to generate findings suited to external audiences as well as to internal learning. Although often under-budgeted, the time spent discussing how data collection could be tailored to address questions and cover domains that were important to the commissioner was valuable in building commitment. The relationship between outcomes, prior relations (e.g. between commissioner and evaluator), and negotiation over the scope of impact evaluation as an 'invited space' for learning, is a subject worthy of further applied research (cf. Stevens et al., 2013; van Tulder et al., 2016).

Reasons for using the QuIP and its links with other sources of evidence

The decision to commission a QuIP, rather than to use some other approach, was based on discussions that ranged from relatively short conversations (e.g. led by one senior manager) to protracted debates among M&E and operational staff. Looking across the 10 case studies (see Table A10.2), three main reasons can be distinguished:

- First, there was congruence or fit with core values. For example, in selecting the QuIP to evaluate the YQYP project, the project implementer picked up on the resonance between the QuIP's emphasis on self-reported attribution and YQYP's focus on transforming intended beneficiaries'

thinking and motivation to act. This was also a factor for Tearfund, Seed, Save the Children, and SHA.

- Second, most organizations were specifically looking for an approach with both exploratory and confirmatory potential, able to pick up on both unexpected and expected but hard-to-measure outcomes, and to yield evidence of the causal mechanisms behind these. The importance of the exploratory dimension included seeing the QuIP as a way of assessing social risks as well as collecting evidence of positive social impact.
- Third, commissioners were often seeking more cost-effective and flexible alternatives to quantitative impact assessment. For Save the Children and SHA this was partly driven by the need to report to an external donor, whereas for Acumen the issue was how to provide their investors with evidence of positive social impact in a routine and affordable way.

For some organizations, an additional selling point for the QuIP was that it could be conducted without the need for a baseline or reference to other data: the Tearfund study in Uganda being one example. However, it proved easier both to design studies and to interpret findings when these were complemented by at least some baseline survey data offering details about the wider reference population, as illustrated by the Diageo study in Ethiopia. Table A10.2 also illustrates the use of a QuIP study as a follow-up to survey-based impact evaluation (e.g. the Oxfam study in Ethiopia), as a substitute for it (e.g. Save the Children in Tanzania), or in parallel with it (e.g. C&A Foundation in Mexico). This is consistent with wider discussion of mixed method approaches in research and impact evaluation (e.g. see Jimenez et al., 2018). However, experience to date of effectively combining the QuIP with quantitative studies is limited, partly for lack of clear advance planning about how they would work together. Many of the staff who commissioned the QuIP were familiar and comfortable with both quantitative and qualitative 'cultures' (Goertz and Mahoney, 2012). However, they did mention having to overcome resistance among colleagues to taking a more qualitative approach, particularly from non-specialist staff.

Missing from the case studies, but evident from many discussions between the authors and potential commissioners of QuIPs, are the reasons for *not* using the QuIP. Lack of money to invest in impact evaluation of any kind was a common reason, particularly in the UK, partly perhaps because prospects for obtaining relevant independent research (or the long feedback loop, as discussed in Chapter 2) are better. A second consideration was that the QuIP would not on its own permit impact to be quantified, or (more vaguely) did not fit with a simplistic view of research and evaluation that conflates credibility with measurability.[3]

What the case studies do reveal are a wide range of options for combining QuIP studies of household level impact with organizational and community level data, both to extend analysis of impact to that level, and to combine

impact evaluation with process evaluation. For example, the Save the Children study in Tanzania used key informant interviews and workshops to throw light on partnership arrangements and achievement of capacity building goals. The study of housing microfinance in India combined use of the QuIP for social impact assessment of loans at the client level, with portfolio analysis and financial performance assessment of the selected microfinance institutions. This illustrates how the scope for combining different evaluation approaches in practice extends beyond the much narrower issue of how best to combine quantitative and qualitative approaches and methods.

Designing QuIP studies: timing, scope, and sampling

Deliberating over the purpose and design of a study is essential, but can also be time consuming. When the QuIP was first trialled (during the Assessing Rural Transformations project), the aim was to be able to deliver a full study in 4–6 weeks. In theory this is possible: once designed, fieldwork for a study can be completed in around 10 days, and coding and analysis in a similar amount of time. However, the typical time required for a QuIP (from design to reporting) over the period covered by this book was closer to three months. Figure 10.1 identifies potential crunch points that affect the timeliness with which data can be delivered once a commissioner decides to go ahead with a study.

Once a study is agreed in principle, delay often arises from the need to obtain project level information from which to sample respondents, particularly when there is a significant gap (and principal-agency problem) between the commissioner of the study and staff actually implementing the project being evaluated. Of course, these and other delays also arise from factors beyond the control of those involved. For example, delays in securing lists of borrowers from the microfinance institutions in India arose in part from the pressure they were put under by Prime Minister Modi's demonetization initiative in November 2016.

Having identified an attribution challenge and opted to address it using the QuIP, each study also presents a different resource-constrained design problem that includes identifying a suitable sampling frame, and deciding on a sample size and selection strategy. Scoping decisions also include the timing of data collection relative to the activities being assessed, including choice of the reference period over which respondents are asked to recall major changes and causal processes.[4]

One general observation is that sampling choices were influenced less than expected by whether the purpose of the study was primarily confirmatory or exploratory, principally because most studies aimed to combine both.[5] In practice, sample selection was more heavily influenced by two other factors: (a) the size of the commissioner's budget, and (b) the availability of secondary data to permit more informed purposive sampling. These are discussed in turn.

Discrete projects funded by an external donor generally had a predetermined budget allocation for monitoring and evaluation, with funds being

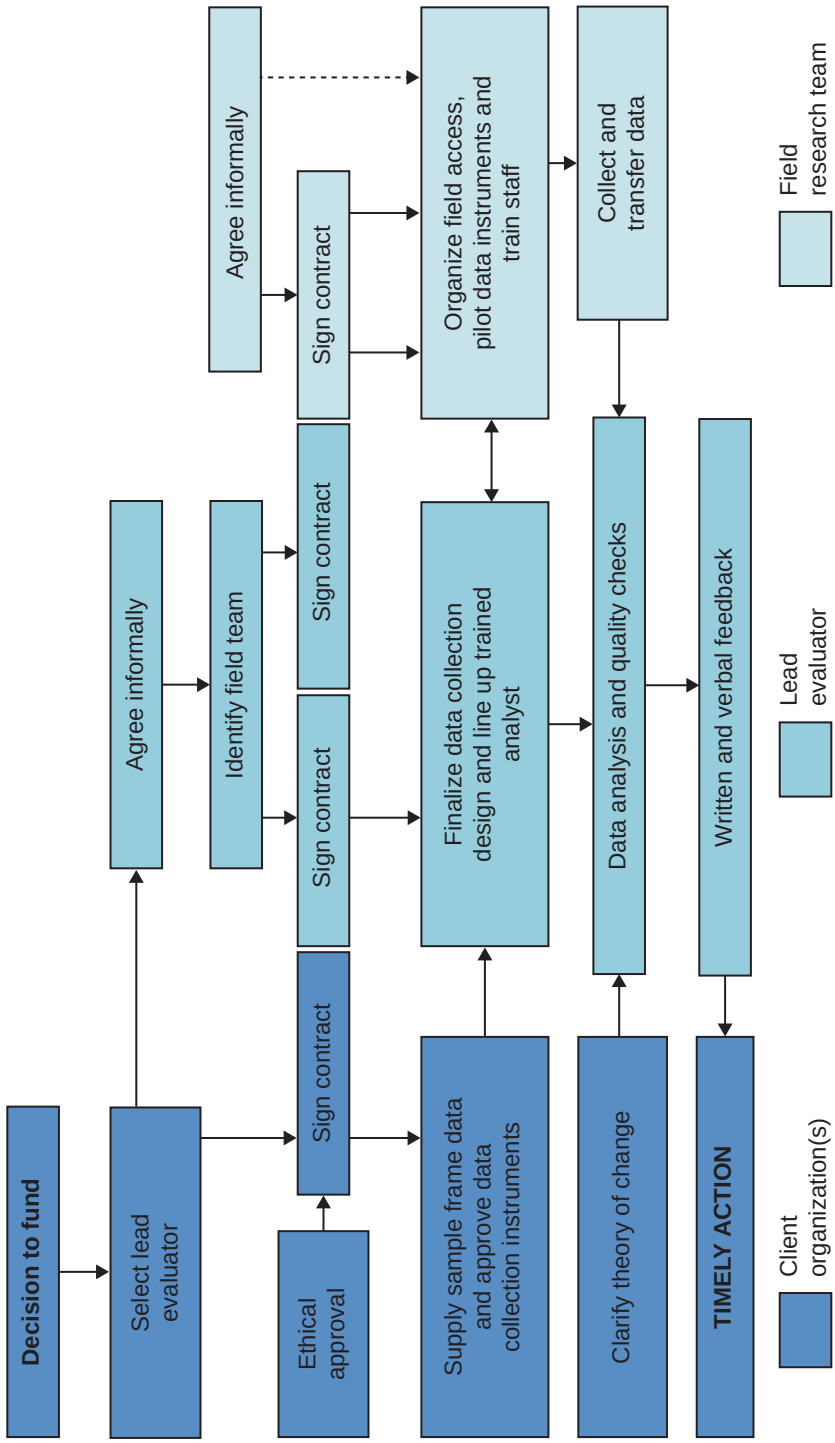


Figure 10.1 Determinants of timely feedback using the QuIP

divided between the QuIP and other studies.[6] For studies commissioned by specialist units there was more discretion, with QuIP studies contributing to a programme of evaluation activities. Within these constraints a common point of discussion was whether a 'single' or 'double' QuIP would be sufficient to support credible generalization, given the diversity of the intended beneficiary group being sampled and taking into account initial costing of each. Eight of the studies completed by Bath Social and Development Research Ltd (BSDR) in 2016 and 2017 were at least double QuIPs; one (GHSP) comprised three linked single QuIPs in adjacent countries; and four were single QuIPs (see Table 1.3 in Chapter 1). There is scope for more research into mental models governing the funding of social impact evaluation, including how this compares with conventions for financial performance assessment. Grants from the UK's Department for International Development (DFID) to NGOs have in the past suggested 5 per cent of funds should be allocated to evaluation, although such guidelines have not been monitored or enforced (Copestake et al., 2016).

The quality and availability of monitoring data influences budget discussions to the extent that it permits stronger claims to be made from smaller but better understood samples, and also because it influences the time and cost of locating and interviewing people. The Tearfund and Frome Town Council studies were extreme examples of studies for which very little data was available to inform sample selection. Despite this, the field researchers were able to construct samples of respondents likely to have been affected by the activities being evaluated by virtue of their physical location. At the opposite extreme, studies that could draw upon richer seams of secondary data were able to make cluster selections that revealed insightful heterogeneity in impact – e.g. between rural and urban takers of housing loans in India. The best example is the Diageo study in Ethiopia, partly because cluster selection was influenced by consistent data across the full population of farmers from whom the company had purchased barley. Here the major weakness was having only cross-sectional data, and hence not being able to gain a better understanding of livelihood dynamics over the longer term.[7]

While most studies succeeded in capturing some insightful variation in impact, a lack of complementary monitoring data often greatly restricted the scope for interpreting how this compared with typical experiences across the entire reference population. This was particularly the case where the scope for sampling was skewed by variation in the willingness of key gatekeepers to cooperate (senior factory managers in the case of the C&A Foundation study in Mexico, for example).

Access to consistent change data from which to draw samples and against which to triangulate findings was mostly limited by simple non-availability rather than confidentiality and data protection regulations. Obtaining and maintaining good records of intended beneficiaries over time is rarely as simple as might be expected. Seasonality, migration, and the fluidity of household structures all complicated maintenance of good records in rural Africa; and even microfinance institutions in India struggled to produce

systematic records of clients going back a few years.[8] Looking to the future, digitization, near universal mobile phone coverage, more accurate GPS systems, and strengthened national registration systems should all help to reduce this problem. But at the same time, data protection legislation is already becoming a bigger obstacle, requiring organizations to seek the permission of clients and other intended beneficiaries to release their personal data to independent evaluators.

Implementing QuIP studies: data collection and analysis

Once an impact study has been designed and commissioned, the challenge remains to make it happen. Answers to 'what' and 'why' questions about impact evaluation mean very little if not combined with answers to questions about 'how' and 'who'. This justifies why the QuIP was first tested through action research under the ART project, and has since been further tested and refined by BSDR. It also explains publication of revised guidelines in this book's Annex, alongside case study evidence and reflection based on its actual use. This section offers additional reflections on data quality and how research tasks are shared out, the role of the analyst, and data synthesis. Table A10.4 summarizes some of the case material that informs this discussion.

Top of the list of necessary conditions for producing an effective QuIP study is a field team which can conduct high quality interviews and focus groups, and also write them up to a high standard. The task of recruiting well-qualified and experienced field teams, and of ensuring they are fully briefed, trained, and motivated, adds significantly to the cost of a study but is critically important. The use of blindfolding places an added burden on the field team by leaving them to find respondents, and to set up interviews without the support of project implementation staff. Overcoming the inevitable difficulties that this brings up hinges not only on the field team's problem-solving abilities, but also on their understanding of why it is worth persisting.

In recognition of the challenge that blindfolded data collection poses, the experience reported here reinforces the case (set out in the guidelines in the Annex) for recruiting a senior field researcher to lead the team, ideally from an established local academic institution or consultancy. This person takes on the role of being the main point of contact with the lead evaluator, and bears overall responsibility for all aspects of data collection. In some cases (e.g. the Seed, Oxfam, and Terwilliger studies) the lead researcher also undertook field work, but this is optional. More important is their close interaction with the field team and the data, as emphasized particularly by lead evaluator Martin Whiteside (see Chapter 7, Box 7.4). Challenges in fieldwork are inevitable: weather, transport, physical access to communities, securing permission from local authorities, illness, strikes, unexpected elections, elusive or survey-fatigued respondents, individually or collectively affected most of the QuIP studies featured in this book, and their successful completion hinged on resourceful, on-the-ground management and leadership. Working with

researchers connected to established local academic institutions (in Ethiopia, Uganda, Malawi, UK, and Zambia) or consultancies (in Ghana, India, and Tanzania) facilitated introductions in the field, and enabled studies to draw upon graduate researchers with the personal backgrounds, language skills, and enthusiasm that facilitated building a strong rapport with respondents.

While practically convenient, separating the field team from the commissioner required a leap of faith for some commissioners. Several of those involved in this book admitted to having been concerned about their lack of direct control over fieldwork compared with other studies they had commissioned. In the C&A Foundation study, for example, this tension was addressed by allowing local staff to participate in some focus groups, thereby partially unblindfolding field researchers and respondents (see Chapter 4). A particular concern linked to blindfolding was that interviews would not generate sufficient narrative evidence about the project. However, this concern almost always turned out to be unwarranted, and with the benefit of hindsight, most commissioners agreed that the credibility of evidence generated through the QuIP was improved by separating these responsibilities.

The next vital step in producing high quality evidence is engaging the best possible analyst. This brings us back to the issue of 'positionality' of the analyst, discussed in the section 'Analysing and presenting data' in Chapter 1. Most of the studies in this book relied on independent researchers who were employed directly by the lead evaluator rather than the commissioning organization. The Tearfund study was an exception, as the analyst was a former employee, able to draw on prior experience of the programme and of Tearfund's monitoring, evaluation, and learning culture (see Chapter 6). This illustrates the need for flexibility in managing a trade-off. An analyst who is closer to the project may be better placed to conduct the attribution coding and to ensure findings are written up in a way that connects strongly with the commissioner and project staff. But from an outsider's perspective this introduces additional risks of bias in coding, e.g. if the analyst makes assumptions about what the respondent might have 'meant to say'. For this reason, the best choice also depends on who the priority users of the study are and what they most want to know (as discussed in the section 'Choosing between approaches to impact evaluation' in Chapter 2). But in general, the QuIP studies described in this book positioned the analyst outside the project, encouraging them to code drivers and outcomes inductively. And by delegating back to commissioners the primary responsibility for generating recommendations, analysts focused on interpreting the data rather than working out its practical implications.

From evidence to use: workshops, decisions, and dissemination

The final and critical link in a QuIP study is how the commissioner chooses to use the evidence it generates. This links back to the discussion above in the section 'QuIP commissioners' about the intended purpose of the

study, but with the added interest of finding out more about what actually happened.

The focus of the QuIP is to assess impact on the intended beneficiaries of a project, rather than to offer a process evaluation, conduct a full stakeholder analysis, weigh up costs and benefits, or engage in scenario planning. A QuIP study is therefore usually insufficient to generate clear recommendations for action on its own, and this is consistent also with our emphasis on the way QuIP studies enhance what users already know, and can complement evidence generated through other feedback processes – short, intermediate, and long. For this reason, the QuIP guidelines emphasize the importance of passing on the baton of evidence to the commissioner, and to other potential users, in ways that can support its effective use.

The studies covered in this book were also conducted under commercial conditions, rather than as part of an explicitly designated action research project. Hence while key informant interviews with commissioners did cover dissemination and use of studies, the commissioners were under no obligation to share information about this with us or anybody else. Indeed, discussion of earlier drafts of four of the case study chapters prompted requests from commissioners to cut out descriptions of operational decisions made in light of the QuIP, on the grounds that it was not appropriate to share this information publicly. While frustrating for us as co-authors, this was entirely appropriate given the contractual basis on which the original studies were conducted. Nevertheless, some information and insight into the crucial last step from evidence to use remains, as summarized in Table A10.5. This is discussed under four headings: follow-up workshops, influencing operational decisions, dissemination to wider audiences, and follow-up studies; leading in turn into a discussion of the institutionalization of impact evaluation.

Follow-up workshops

Despite being strongly encouraged in the QuIP guidelines, there were few follow-up workshops in the case studies covered by this book. This partly reflects their cost (particularly when planned to involve intended beneficiaries and project staff), but also perhaps the lack of experience with feeding QuIP data into participatory planning, and the need to prioritize upward accountability to project funders over downward and peer accountability.

Two exceptions stand out: Tearfund's follow-up focus groups in Uganda, and Save the Children's stakeholder workshops in Tanzania, as documented in Chapters 6 and 7, respectively. Both promoted learning and strengthened links with local partners of the commissioning organization. They also demonstrated that if staff are committed to the principle of doing this, then the logistical challenges and costs of doing so are not insurmountable. However, institutionalizing such follow-up activity will become more common only if it comes to depend less on the enthusiasm of particular staff and more on political pressure on commissioners to be seen to report back to local

stakeholders in this way. Better documentation of the ways in which such workshops can correct and augment findings would also help such events become the norm.

Some internal and follow-up triangulation of data is also possible without organizing focus groups or workshops. Examples include the unblindfolded key informant interviews conducted in the GHSP study with senior medical staff, and ongoing consultation between the lead researcher and members of Frome Town Council in the UK. Key informant interviews with gatekeepers often entail making vague promises to feed back findings; this is better seen as an opportunity to learn more and to deepen relationships, rather than as a duty and a chore.

Influencing operational decisions

How impact evaluation studies (whether based on the QuIP or another approach) are used by commissioning organizations to inform operating decisions is not easy to document for both methodological and political reasons. With respect to the former, attributing a specific decision (as an outcome) to a specific driver (such as a QuIP study) presents its own attribution challenge, particularly given the likelihood that most major decisions are the consequence of multiple drivers, triggers, and contextual factors. Fear that explanations of how decisions are made will not do justice to this complexity, and that being more outwardly transparent can be demanding of staff time, partly explains the political reluctance of organizations to be more open about decision making processes.

Nevertheless, the processes of conducting a QuIP study and (particularly) of interviewing commissioning staff one or two years after it was completed, did yield interesting insights into decision making. Several chapters in the book include specific reflections on conducting the study and what it meant to the commissioner. But other reflections and insights can only be shared anonymously. The wider operational context for these more critical reflections is one of structural tensions: (a) among global, regional, and country level staff within the same organization; (b) between commissioning organizations and official donors funding them; and (c) between commissioning organizations and locally contracted 'partner agencies' implementing projects. Such tensions will also be familiar to those who have conducted evaluations based on other methodological approaches, including 'aidnography'.

A stylized fact is that more power over project resources resides at the global level, although this was reduced to some extent where opportunities existed to raise funds at country level. Either way, gaps often emerged between: (a) the aspirations and discourse that attached to funding proposals and commitments (to achieve 'sustainability', 'transformation', 'graduation from poverty', etc.); and (b) what staff deemed possible to actually achieve on the ground within agreed project budgets and time frames. The arrival of a QuIP study (or generation of any other empirical evidence on actual impact)

threatened to expose such gaps to a wider audience, as well as gaps between the organization's official 'script' and the perceptions of intended beneficiaries (Copestake, 2011). This threat presented organizations and their staff with a political difficulty; one that they in part sought to manage through rhetorical ambiguity and/or silence (cf. Mosse and Lewis, 2006).

Difficulties of this kind emerged to varying degrees in finalizing the drafts of some QuIP reports prior to these being accepted. From the lead evaluator's point of view, detailed objections to draft text (and to the analysis underpinning it) can come across as a defensive tactic, seeking to soften or bury criticism. But for the staff themselves the QuIP evaluator's observations perhaps evoke Alexander Pope's 300-year-old remark that 'fools rush in where angels fear to tread'. Either way, a key learning point for users of the QuIP is to ensure that all stakeholders have a good understanding of what to expect from a QuIP study as well as how the report is to be used, *before* they receive draft findings. The chance to discuss findings verbally can also help to avoid costly struggles over what is officially written and reported.

The key point about these drafting battles is that they can be a window onto more substantive debates over operational decisions, including whether to cut, continue, increase or alter the terms of funding of different activities. Without being able to openly identify the specific impact of the different QuIP studies reviewed here it was clear in more than one case that the studies did influence specific decisions of this kind, not only on the basis of evidence of impact but also by clarifying how project theory was understood by different stakeholders. Pushing the point further, instances arose where the process of drafting and commenting on drafts could itself be viewed as the mechanism by which internal deliberation occurred, with externally led studies being a relatively neutral artefact for triggering the process.

Wider dissemination

At the time of writing, half the commissioners of the 10 studies listed in Table A10.1 had published full or abbreviated versions of QuIP reports (Diageo, C&A Foundation, Terwilliger Center, Tearfund, and Oxfam).[9] While carefully edited, these mostly reflected both positive and negative findings, although it can be argued that this was made easier by the fact that positive clearly outweighed negative findings for all these studies. The editing for wider publication was not oriented towards putting a positive spin on findings. Rather it mostly focused on shortening, simplifying, and polishing presentation, checking facts and ensuring that the descriptive text was consistent with other material published by the organization.

The time lag from finalization of reports to wider publication of edited versions varied from a few months to years. For example, Oxfam's research note (Mager et al., 2017) came out several years after the project itself finished, and 18 months after the QuIP report was submitted. This might be viewed by some observers as an indication of the rather relaxed lines

Table 10.1 Middle range theory behind selected commissioners' missions

Study commissioner	*Theory*
Diageo (Ethiopia)	Purchasing barley as a cash crop from small-scale farmers does not have unintended negative social consequences
C&A Foundation (Mexico)	Garment factories can be used as an entry point to strengthen the communities and intra-household relations of their employees
Habitat for Humanity International (India)	Incremental home improvement funded by commercially self-sustainable housing microcredit has a positive social impact on borrowers
Tearfund (Uganda)	Faith-based community development can have a positive transformative effect, even when not linked to material transfers
Oxfam (Ethiopia)	Promoting fair trade coffee as a cash crop does not have adverse effects on the wellbeing of women by increasing their work burdens

of accountability to funders that some international agencies enjoy. More importantly, however, it indicates that the demand for more evidence of impact did not primarily arise from gaps in the organizations' short feedback loops, but from a perceived need to strengthen the middle range theory of change underpinning the organizations' public missions. Examples are summarized in Table 10.1.

By comparison, it is interesting to speculate on why some of the other commissioning agencies chose not to invest in similar publications. For example, while Save the Children and SHA were more focused on producing evidence for an official donor, both would also benefit from convincing a wider public that integrating climate smart agricultural innovation, nutrition education, and community development can have positive effects. However, it can perhaps also be argued that independent research (via the long feedback loop) is a more effective route to demonstrating this than more modest impact evaluation studies, such as the QuIP. Likewise, as a leading impact investor, Acumen has a strong interest in strengthening the evidence base that demonstrates the positive social impact of private sector development. But, while keen to strengthen its short feedback loop, it can be argued that independent academic research linking private enterprise, economic growth, and poverty reduction provides them with a sufficient evidence base. Conversely, critics of impact investors would argue that their future prospects rely more on using broad social impact claims to bolster a benign view of neo-liberalism (i.e. a market-led vision of development) than on evidence-based public policy (McGoey, 2014).

Follow-up studies

Given that the potential demand for evidence of development impact flows through time (and hence is unlikely to be satisfied by a single study), the usefulness to commissioners of QuIP studies can also be assessed by whether they are willing to invest in follow-on studies. Here the evidence is fairly positive.

Diageo followed up the Ethiopia study with a second study in Uganda in 2017, and plans a third in 2019. The Terwilliger Center commissioned a second QuIP study in 2018 of its support for Mibanco's housing microfinance products in Peru. Tearfund saw the Uganda study of CCM as a pilot, and committed to following it up with four more studies in other countries. Save the Children commissioned a second QuIP study in Ethiopia and is exploring options for more. And the Zambia study is one of three commissioned by SHA since the end of the ART project (the others having been in Kenya and Burkina Faso). In contrast, Farm Africa has not commissioned a single study since its involvement in the original ART project. Oxfam has also only undertaken the one study, as has C&A Foundation. Acumen, in contrast, has pursued a distinctive strategy based on internalizing lessons learned from QuIP studies into its own programme of 'lean' social impact evaluation. Likewise the Aga Khan Development Network has adopted the strategy of investing in the capacity of its own staff to conduct QuIP studies. This quick review raises a bigger question: how can and should the QuIP contribute to institutionalizing qualitative impact evaluation beyond the level of discrete organizations? This is addressed in the next section.

The QuIP as a case of institutional innovation

To briefly recap, the last section of Chapter 1 outlined 'The backstory of the QuIP', including drafting and piloting QuIP guidelines under the ART project prior to 2016, and mainstreaming its use since. Chapter 2 then examined more fully how it compares with other approaches to impact evaluation, and subsequent chapters have reviewed examples of QuIP studies delivered under contract to a wide range of commissioning organizations. Chapter 9 also opened discussion of how radically it can and could be adapted to different contexts and purposes. One model for the dissemination of the QuIP, as the product of applied research, would have been to publish the guidelines and then wait to see what pattern of diffusion and adoption occurred, including how steep and long the 'S curve' of references to the QuIP label turned out to be. Instead this book reports on a more active strategy of promoting its use by setting up a social enterprise, Bath Social and Development Research Ltd, with the explicit goal of doing so.[10] BSDR benefitted indirectly from grant funding to the University of Bath for initial dissemination of the QuIP and market research into potential demand for it. But it has subsequently operated out of income earned commercially.

The creation of BSDR in early 2016 and its subsequent development of the QuIP can be considered an ongoing action research project in itself. It was motivated by the judgement that potential for methodological innovation and market 'disruption' using the QuIP was far from complete, and would be better enhanced by promoting its commercial use, rather than relying on more grant funded action research. A more ambitious business plan for BSDR expansion was developed and submitted for funding to a government

sponsored scheme to promote university start-ups. This was unsuccessful, and gave way instead to a strategy of financing growth through reinvestment.[11] Contracts to deliver QuIP studies from Self Help Africa, Oxfam GB, and Diageo were all instrumental in enabling BSDR to meet its start-up costs.

Since promoting the QuIP as a 'project' with development goals in its own right, it is appropriate that BSDR should have its own theory of change. This is reproduced as Figure 10.2.

The ultimate goal (far left of Figure 10.2) is generation of more and better evidence of social impact as a contribution to better development practice. Three intermediate goals are to promote more and better understanding of how to produce good social impact evidence, wider demand for this, and increased capacity to supply it. BSDR aims to deliver this through QuIP related research collaboration, conducting more QuIP studies, and enabling others to do so (including through training, networking, and dissemination). This triple strategy of researching, doing, and enabling QuIP studies is premised on the existence of strong complementarities between the three.

It is too soon to evaluate the effectiveness of this strategy for opening up the range of possibilities for effective social impact evaluation within the international development field, and still less in other areas. However, one key tension or dilemma within the model is worth highlighting, and this concerns intellectual property and branding.

Chapter 1 suggested that an obstacle to doing more applied impact evaluation using qualitative social research remains the uncertainty in the minds of many commissioners about variation in the rigour of findings, and how to go about assessing this. The way the perceived 'black box' between collection of data and generation of findings fuels doubts about credibility was likened to the 'lemon problem' made famous by Akerlof (1970). This underpinned the decision to invest in the QuIP as a label for a distinctive and transparent approach to impact evaluation, to facilitate clear comparison with other approaches – as outlined in Chapter 2. The dilemma that remains is how liberal to be with the QuIP label and promotion of its use.

One strategy is for BSDR to openly share everything it has learned. This appears consistent with its social mission, and with the ethos of having utilized public funding to develop the QuIP this far. But to give away every detail of how to conduct a QuIP study (including expensively refined and tested details of how to conduct the data analysis) risks losing all influence over how well QuIP studies are conducted. It also potentially undermines BSDR's own competitiveness as a young and small social enterprise, as well as its capacity to coordinate learning and further methodological development.

The contrasting strategy is to narrowly define the QuIP label, and jealously defend it. Doing so can help to promote methodological clarity and precision. Given that the QuIP is not a straightforward piece of 'technology' to use well, there is also a case for thereby seeking to differentiate between bona fide or 'accredited' QuIPs, and those conducted by evaluators without any specific

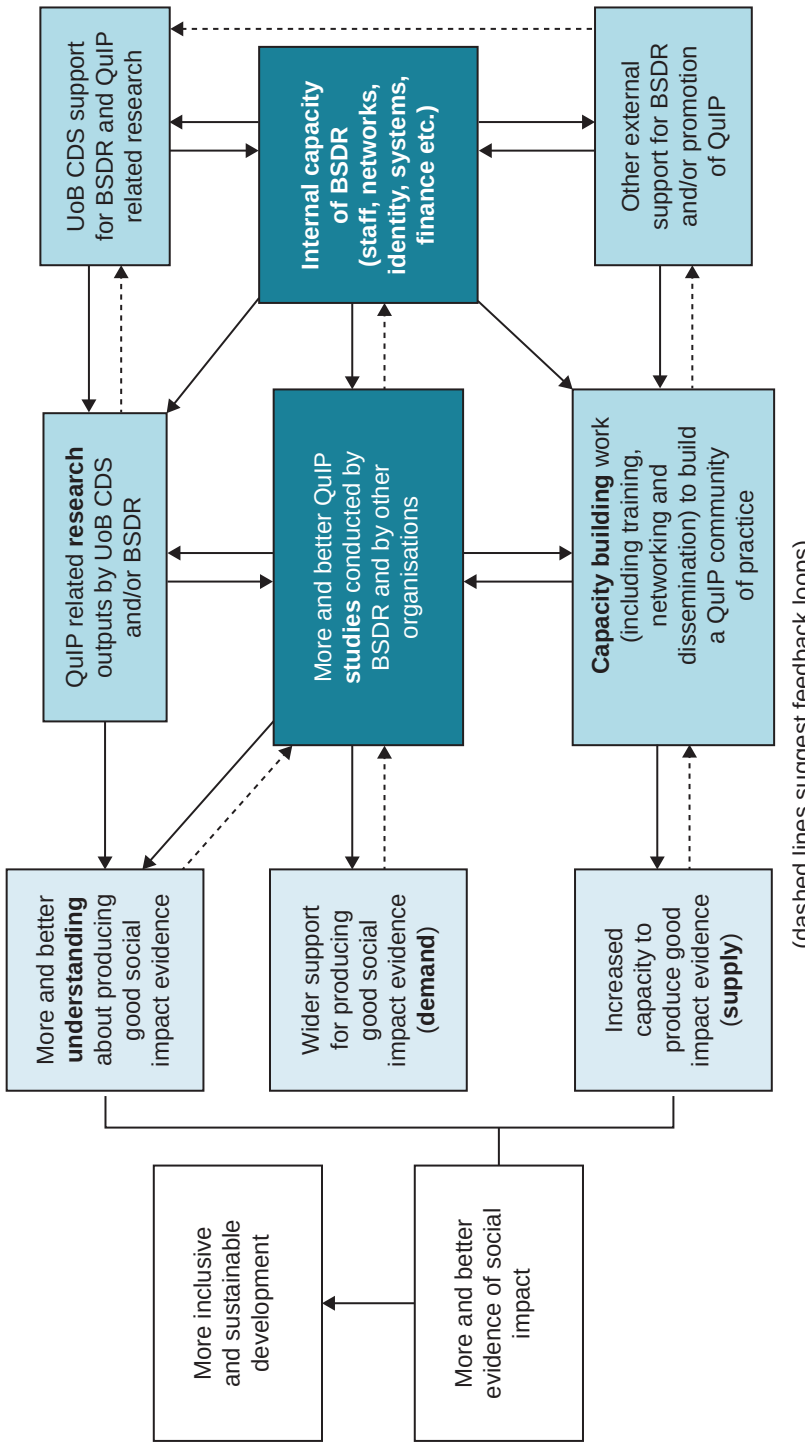


Figure 10.2 BSDR's theory of change
Note: UoB, University of Bath; CDS, Centre for Development Studies at UoB

training or support – particularly in using software to facilitate transparent data analysis and visualization.

At the time of writing, this tension has been resolved by being as open as possible, subject to the term 'QuIP' having been registered as a trademark by the University of Bath and licensed exclusively to BSDR along with the right to sub-license its use to accredited practitioners. This represents a modest step towards being able to exert some quality assurance and coordination over how QuIP studies are conducted, without limiting how ideas and practices for better evaluation promoted through the QuIP are picked up, utilized, and further adapted by others. Aspirations to build a network of accredited and trusted QuIP evaluators, practitioners, and commissioners with strong shared skills and understanding can thereby proceed without limiting wider and looser diffusion: caveat emptor!

Towards more agile evaluation and adaptive development practice

This book started by noting that anybody aiming to bring about positive social change eventually confronts the problem of how best to check whether they are really achieving what they hoped. Given the immense complexity of this problem, it further suggested that addressing the attribution challenge is central to wider international development practice. Closing the feedback loop, speaking truth to power, and learning to be a reflective practitioner may not be sufficient conditions for better development practice, but they are necessary ones. This final section draws on the case study material reviewed in the book to reflect further on how far small investments in impact evaluation can foster better development practice.

With the benefit of hindsight, taking the QuIP as an entry point for reflecting on development practice is awkward in at least three ways. First, while primarily a qualitative approach to impact evaluation it goes further than many qualitative social researchers may be comfortable with in utilizing numbers as well as words to summarize findings. And while we have emphasized its role in generating credible evidence (rather than contributing to absolute truths) many qualitative social researchers may also be uncomfortable with its realist roots – wary of being able to deliver universal truths about change processes in complex social worlds, but neither being willing to abandon the quest to produce a better overall understanding of causal processes.

Second, the intermediate feedback loops that the QuIP seeks to facilitate (identified and explored in Chapter 2) sit awkwardly between the more familiar short feedback loops of performance management and the longer feedback loops dominated by academic knowledge communities. Researching precisely how far the QuIP was effective in augmenting or correcting short feedback loop understanding has proven to be difficult, given both the complexity of operational decision making, and organizations' understandable reluctance to discuss it more openly. In contrast, several case studies suggested that the QuIP played a stronger than anticipated role in testing and supporting

middle range theory to augment theory and evidence generated through the longer feedback loop. This raises difficult questions about scale. On the one hand, there is a danger that the QuIP could be used to generate evidence to buttress the social claims of development organizations more cheaply (and with more corporate control) than could be done by commissioning and using independent academic research. But on the other hand, the QuIP can be a cost-effective way of producing good-enough evidence of the impact of highly context-specific development activities and strategies.

Third, and even more fundamentally, the QuIP can be viewed as a reformist strategy for seeking improvement in development practice. This will be of relatively little interest to those who have more radical criticisms of the whole idea of international development based on transfer of resources and strong upward accountability to richer and more powerful actors, whether in the public or private sector. While we have emphasized the importance of 'giving voice' to 'intended beneficiaries' we have not spent so much time considering how they can *seize* that influence, not as intended beneficiaries but as active political actors. Building on a three way distinction suggested by Gulrajani (2010), the QuIP can be viewed as a 'romantic' contribution to development management, that gives weight to the potential for better voluntary collaboration, and lies between a more managerial and top-down 'reformist' approach, and a more politically ambitious and aggressive 'radical' approach.

Our defence for inhabiting the awkward middle space between qualitative versus quantitative cultures, short versus long feedback loops, and radical versus reformist management perspectives is a pragmatic one. Work on the QuIP started through dialogue with individual practitioners and organizations seeking better solutions to immediate problems, and it has sought intermediate solutions for them. And of course, we are not alone in seeking new and more adapted approaches to development management, better suited for complex and rapidly changing contexts. We are not alone in emphasizing that catastrophes happen, projects go awry, black swans appear, the best laid plans are disrupted, and even the most rigorous evaluations go awry or leave important questions unanswered. In complex, rapidly changing, and uncertain development contexts we conclude by reaffirming the need to find faster and more flexible 'reality checks' and 'deep dives' to gather evidence on what is happening along what are often long and complex financing chains. This can be true for very large programmes prone to local deviance, and for very small projects with potential to generate lessons of far wider significance.

The quest for more agile evaluation is also a necessary element of attempts to 'do development differently', to engage in 'problem driven iterative adaptation' and adaptive management (e.g. Andrews et al., 2012, 2017). The case studies in this book confirm that approaches like the QuIP can be useful in scoping, framing, and prioritizing where more precise or generalizable evidence is needed across 'rugged design landscapes' that present a myriad of options for both programme developers and evaluators. They also demonstrate the scope for operating within tighter budgets, shorter time scales, and more localized

circuits of governance. At the same time, we recognize that scope for further research and innovation in this space remains large. While we hope that this book does help to promote more use of the QuIP and adaptation of the ideas incorporated into it, the larger and more important goal of the book is to promote a broader and more plural approach to impact evaluation in pursuit of more effective development.

Appendix: case study themes

Table A10.1 QuIP commissioners: purpose, priors, and priorities

Commissioner (and book chapter)	*Project activity or 'evaluand'*	*Purpose of the study*	*Prior knowledge, including theory of change (ToC)*	*Confirm and/or explore?*	*Orientation (internal or external)*
Diageo (Ch. 3)	Sourcing for Growth (S4G); Ethiopia	Assess the social impact of local malt barley procurement	Detailed crop procurement data; weak ToC	Mostly to explore	Both
C&A Foundation (Ch. 4)	YQYP training for factory workers; Mexico	Explore workers' knowledge, attitude and agency	Strong ToC, but relatively weak data on project implementation	Mostly to confirm	Both
Terwilliger Center of Habitat for Humanity (Ch. 5)	Housing microfinance; India	Reveal the social impact of housing microfinance	Strong ToC; loan portfolio and data on MFIs' financial performance	Both	Both
Tearfund (Ch. 6)	Church and Community Mobilisation (CCM); Uganda	Gather evidence on the potentially transformational mechanisms set out in the ToC	Strong normative framework; open-ended ToC	Explore	Both
Save the Children (Ch. 7)	Harnessing Agriculture for Nutritional Outcomes (HANO); Tanzania	Assess impact on completion of project, learn lessons, and report to donor	Basic ToC; some prior process studies, weak monitoring data	Both	Both
Seed and Peace Corps (Ch. 8)	Global Health Service Partnership (GHSP); Malawi, Tanzania & Uganda	Draw cross-country lessons from placement of volunteer educators	No explicit ToC; good monitoring data, including written reports from volunteer educators	Both	Internal
Frome Town Council (Ch. 9)	Council support for green spaces; England	Secure feedback on impact of a variety of initiatives	Weak ToC and monitoring data	Explore	Internal

(Continued)

Table A10.1 Continued

Commissioner (and book chapter)	*Project activity or 'evaluand'*	*Purpose of the study*	*Prior knowledge, including theory of change (ToC)*	*Confirm and/or explore?*	*Orientation (internal or external)*
Oxfam (Ch. 1, Box 1.4)	Coffee value chain upgrading; Ethiopia	Assess gender impact on caring responsibilities	Strong ToC; prior quantitative impact survey	Confirm	External
Acumen (Ch. 1, Box 1.4)	Impact investment; India	Assess client impact and satisfaction with services	Basic ToC and commercial data	Confirm	For investors and investee
Self Help Africa (SHA) (Ch. 1, Box 1.4)	Integrated area development project (IADP); Zambia	Assess impact on completion of project, learn lessons, and report to donor	ToC and data from prior and complementary studies	Both	Both

Table A10.2 Reasons for using the QuIP and links to other sources of evidence

Case study	*Reasons for choosing the QuIP*	*Linked data collection activities*
Diageo; malt barley promotion; Ethiopia	Looking for an exploratory approach, also credible enough to support a web publication	A complement to commercial procurement operations and systems
C&A Foundation; garment worker training; Mexico	Good fit with the project's goal to empower intended beneficiaries and strengthen their voice	Difference-in-difference impact based on psychometric scales of workers' capabilities
Terwilliger Center; housing microfinance; India	Seeking evidence on a wide range of potential social impacts, and mechanisms linking them to improved access to finance	Financial performance assessment, including portfolio quality analysis of selected microfinance partners
Tearfund; Church and Community Mobilisation; Uganda	Good fit with the project's emphasis on empowerment and voice. Seeking alternatives to a quantitative approach.	A second round of unblindfolded focus group discussions of findings in selected communities
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania	Seeking more cost-effective alternatives to experimental impact evaluation approaches	Process evaluation based on key informant interviews with implementing staff. An earlier plan to do an RCT was abandoned.
Seed and Peace Corps; Global Health Service Partnership; Malawi, Tanzania & Uganda	Seeking an approach that could credibly capture evidence of diverse and unexpected outcomes	Good programme monitoring, including self-evaluations by volunteer educators

Case study	*Reasons for choosing the QuIP*	*Linked data collection activities*
Frome Town Council; promoting use of green spaces; England	Innovative town council was seeking inexpensive ways to reflect whether and how it was making a difference	Direct feedback from the town's citizens, not least through local elections
Oxfam; producing fairtrade coffee; Ethiopia	One of several qualitative follow-ups to a programme of difference-in-difference impact assessments	Based on a sub-sample of interviews conducted as part of a difference-in-difference evaluation
Acumen; impact investment; India	Seeking a low cost approach to assessing social impact of investments alongside financial performance assessment	Financial assessment of investees. The lean QuIPs were part of a series of lean studies.
SHA; integrated area development; Zambia	Seeking alternatives to experimental impact evaluation approaches for assessing contribution	Complementing larger nutrition surveys and village studies of changing incomes using the individual household method

Table A10.3 Designing QuIP studies: scope, sampling, and time frame

Case study	*One strength*	*One weakness*
Diageo; malt barley promotion; Ethiopia	The sample purposively selected contrasting clusters, based on monitoring data: this picked up sharp differences in impact	The study only generated evidence on change over a couple of seasons
C&A Foundation; garment worker training; Mexico	The sample generated strong evidence of positive programme impact	The sample was both small relative to the population and skewed towards factories willing to cooperate
Terwilliger Center; housing microfinance; India	Use of portfolio data permitted stratification of the sample, and revealed heterogeneity in impact, including between rural and urban borrowers	Scope for assessing how typical the sample was relative to the wider client population was limited by lack of portfolio data
Tearfund; Church and Community Mobilisation; Uganda	Strong attribution of impact emerged despite starting with weak data on who participated in the project	It was not possible to assess how typical the experience of the four selected communities was compared with the wider population
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania	The scope was usefully expanded to include local implementing partners and thus permitted triangulation of findings between two levels of data	It was not possible to assess how typical the experience of the four selected communities was compared with the wider population involved
Seed and Peace Corps; Global Health Service Partnership; Malawi, Tanzania & Uganda	Data was obtained from students and staff across a wide range of courses and institutions, opening up scope for interesting comparisons	It was hard to unravel students' observations of changes in clinical training from their personal progress through the training

(Continued)

Table A10.3 Continued

Case study	*One strength*	*One weakness*
Frome Town Council; promoting use of green spaces; England	Rich qualitative evidence was collected despite the lack of a population frame from which to draw the sample	The lack of monitoring data with which to contextualize and interpret the qualitative data
Oxfam; producing fairtrade coffee; Ethiopia	Successful in filling a gap (concerning impact of the project on women's time) unanswered by the prior quantitative impact study	The long time lag from project to study limited the scope for exploring causal mechanisms in more detail
Acumen; impact investment; India	Narrow scope and use of telephone interviewing permitted coverage of a larger sample at lower cost	Danger of falling somewhere between being a quantitative survey and a qualitative enquiry
SHA; integrated area development; Zambia	The sample purposively selected contrasting clusters, based on monitoring data, and this picked up sharp differences in impact	Weak integration of cluster selection with monitoring data made it hard to assess how the experience of the four selected communities compared with that of the wider population involved

Table A10.4 Implementing QuIP studies: data collection and analysis

Case study	*Enabling factors*	*Constraining factors*
Diageo; malt barley promotion; Ethiopia	Built on established relationship with an experienced lead field researcher	Principal-agency issues between commissioner and within-country staff over release of data
C&A Foundation; garment worker training; Mexico	Good collaboration with the implementing agency; extensive prior discussion of the study	Variation in willingness of factory management to collaborate; some reluctance to go ahead with blindfolded interviews
Terwilliger Center; housing microfinance; India	Collaboration with a highly experienced Indian consultancy permitted integration of the QuIP with financial assessment	The gap between commissioner and selected MFIs affected timeliness of data collection, particularly given external shocks affecting the MFIs at the time (principally demonetization)
Tearfund; Church and Community Mobilisation; Uganda	Field team links with Makerere University; analyst a former employee of Tearfund	Lack of a clear sample frame within selected villages, limited possibility of wider comparisons
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania	Strong commitment to the study within the country office	Some delay arising from the need to identify a field research team in Tanzania from scratch

Case study	*Enabling factors*	*Constraining factors*
Seed and Peace Corps; Global Health Service Partnership; Malawi, Tanzania & Uganda	Strong support from within the commissioning organization	Time required to secure cooperation of multiple stakeholders across several implementing institutions, gain access to students, and agree to blindfolded interviewing
Frome Town Council; promoting use of green spaces; England	Data collection and analysis by the same researcher	Lack of a clear sample frame and of monitoring data complicated the task of identifying respondents
Oxfam; producing fairtrade coffee; Ethiopia	Sampling off the back of a prior survey	Long gap between project and study complicated the task of locating some respondents
Acumen; impact investment; India	Commissioner strongly committed to internalizing use of an adapted QuIP	Demand for more detailed evidence on specific outcomes required revisions to QuIP data analysis protocols
SHA; integrated area development; Zambia	Field team links with University of Zambia	Large distances between clusters. Blindfolding didn't work in very remote sites with very limited presence of external agencies

Table A10.5 From evidence to use: workshops, decisions, and dissemination

Case study	*Main applications*
Diageo; malt barley promotion; Ethiopia	An edited version of the study was published on the company website. There was some follow-up discussion of operational implications, but full details not known.
C&A Foundation; garment worker training; Mexico	The QuIP report was made available publicly via the web. It also influenced operational decisions and the commissioner decided that it was not appropriate to share details of how it did so publicly.
Terwilliger Center; housing microfinance; India	An edited version of the study was produced for public dissemination via the Center's website, the lead evaluator was invited to participate in a regional conference on housing microfinance, and a second QuIP was commissioned (in Peru). The commissioner decided that details of operational follow-up with selected MFIs was not appropriate to share publicly.
Tearfund; Church and Community Mobilisation; Uganda	Findings were shared with respondents through follow-up village meetings. An edited version of the report was published by Tearfund, and findings were also shared through various public events with supporters.

(Continued)

Table A10.3 Continued

Case study	*Main applications*
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania	Findings were shared through multi-stakeholder meetings in the project area and in Dar Es Salaam soon after completion of the study (attended by the lead evaluator and field research staff). Various wider dissemination meetings took place and are still planned. The commissioner reports that the study was timely and relevant to decisions on other projects in the country, as well as being fed back to the official donor of the project.
Seed and Peace Corps; Global Health Service Partnership; Malawi, Tanzania & Uganda	The study fed into internal discussions about GHSP but this was curtailed by closure of the whole programme (for reasons unrelated to the QuIP study or programme characteristics of the three countries covered by it). However, the evidence continued to inform Seed's other activities promoting volunteer educators.
Frome Town Council; promoting use of green spaces; England	Detailed findings were presented to the Town Council.
Oxfam; producing fairtrade coffee; Ethiopia	The study was published on Oxfam's website along with an edited report on cash cropping and gender relations. It also contributed to a wider programme of reviewing the issue of unpaid care work and agricultural commercialization (Mager et al., 2017).
Acumen; impact investment; India	Findings were reported back to the companies studied, and contributed to Acumen's ongoing programme of generating 'lean data' about the social impact of their investments.
SHA; integrated area development; Zambia	Findings from the study were combined with those from other surveys into a final report for the official donor of the project.

Notes

1. See Johnson and Rasulova (2016) for a fuller discussion of authenticity as a criterion of rigour in development impact assessment.
2. The Oxfam study came closest to having a more precise confirmatory purpose: to provide added reassurance that the coffee project had not exacerbated gender inequality in work allocation by adding paid work onto women's unequal responsibility for household reproduction. The study indeed found that this was not the case (Mager et al., 2017).
3. No examples were available for this book of how QuIP data can contribute to building simple simulation models (calibrated using monitoring data) to estimate magnitudes of impact; hence this remains an unrealized opportunity for further research and development.
4. As mentioned in the last section, the scope of each study was also further limited by the domain structure and choice of interview questions.
5. These generalizations draw on Table A10.3, which highlights one design strength and one weakness from each case study. The tables are based on a subjective assessment of the authors, with the benefit of hindsight.

6. The cost of randomized controlled trials (RCTs) and other survey-based impact assessment is strongly influenced by 'power' estimates of the minimum sample size required to generate findings with desired statistical significance. Qualitative approaches such as the QuIP are less prescriptive about what constitutes a minimum sample size, and as a result can easily find themselves having to make do with whatever residual remains in the allocated budget: 'double or quit' being the limited alternative to doing a single QuIP.
7. A second study for Diageo in Uganda was not so well served, as very few of the private traders supplying the company were able to furnish lists of farmers they purchased from to match the quality of those made available by producer cooperatives in Ethiopia.
8. Visiting one NGO office revealed another reason for the problem. Having endured delays in authorization of donor funding, it confronted the need to spend funds on stipulated activities within the short period remaining before the end of the financial year. Maintenance of reliable records of which individuals participated in different training events, where and when, was one casualty of the ensuing rush to spend.
9. Of the other five, two (for Save the Children and GHSA) were primarily intended for consumption by an official donor; the study for GHSP was always intended for internal use only; the Frome Town Council study was a pilot; and the Acumen studies were oriented towards feeding back to private impact investors. If anything this limited evidence rebuffs the assertion that publication of impact evaluations tends to be biased towards studies with positive findings.
10. BSDR is a company limited by guarantee, with a non-distribution clause. Profits surpluses are used to fund research and development.
11. The unsuccessful bid for capital funding was made to Innovate UK, funded by the Department for Business, Energy and Industrial Strategy (BEIS). It followed on from participation in the ICURe programme (Innovation to Commercialisation of University Research) run by the SETsquared Partnership. This funded an initial round of market research, intended to enable university researchers to validate potential business ideas, and to 'pivot' these towards promising potential customer segments (see http://www.setsquared.co.uk/research-commercialisation).

References

Akerlof, G. (1970) 'The market for "lemons": quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* 84(3): 488–500.

Andrews, M., Pritchett, L. and Woolcock, M. (2012) *Escaping Capability Traps through Problem-driven Iterative Adaptation* (PDIA), Paper 299, Washington, DC: Centre for Global Development.

Andrews, M., Pritchett, L. and Woolcock, M. (2017) *Building State Capability: Evidence, Analysis, Action*, Oxford: Oxford University Press.

Copestake, J. (2011) 'Well-being in development: comparing global designs and local views in Peru', *European Journal of Development Research* 23(1): 94–110 <https://doi.org/10.1057/ejdr.2010.45>.

Copestake, J., O'Riordan, A-M. and Telford, M. (2016) 'Justifying development financing of small NGOs: impact evidence, political expedience and the case of the UK Civil Society Challenge Fund', *Journal of Development Effectiveness* 8(2): 157–70 <https://doi.org/10.1080/19439342.2016.1150317>.

Goertz, G. and Mahoney, J. (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*, Princeton, NJ: Princeton University Press.

Gulrajani, N. (2010) 'New vistas for development management: examining radical-reformist possibilities and potential', *Public Administration and Development* 30(2): 136–48 <https://doi.org/10.1002/pad.569>.

Jimenez, E., Waddington, H., Goel, N., Prost, A., Pullin, H., White, H., Lahiri, S. and Narain, A. (2018) 'Mixing and matching: using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes', *Journal of Development Effectiveness*, forthcoming.

Johnson, S. and Rasulova, S. (2016) 'Qualitative research and the evaluation of development impact: incorporating authenticity into the assessment of rigour'. *Journal of Development Effectiveness* 9(2): 263–76 <https://doi.org/10.1080/19439342.2017.1306577>.

Mager, F., Walsh, M. and Remnant, F. (2017) *Cash Cropping and Care: How Cash Crop Development is Changing Gender Relations and Unpaid Care Work in Oromia, Ethiopia*, Research Report, Oxford: Oxfam GB <http://dx.doi.org/10.21201/2017.1329>.

Mosse, D. and Lewis, D. (2006) 'Theoretical approaches to brokerage and translation in development', in D. Lewis and D. Mosse (eds), *Development Brokers and Translators: The Ethnography of Aid and Agencies*, Bloomfield, CT: Kumarian Press.

McGoey, L. (2014) 'The philanthropic state: market-state hybrids in the philanthrocapitalist turn', *Third World Quarterly* 35(1): 109–25 <https://doi.org/10.1080/01436597.2014.868989>.

Stevens, D., Hayman, R. and Mdee, A. (2013) '"Cracking collaboration" between NGOs and academics in development research', *Development in Practice* 23: 1071–7 <https://doi.org/10.1080/09614524.2013.840266>.

van Tulder, R., Seitanidi, M., Crane, A. and Brammer, S. (2016) 'Enhancing the impact of cross-sector partnerships: four impact loops for channelling partnership studies', *Journal of Business Ethics* 135: 1–17 <https://doi.org/10.1007/s10551-015-2756-4>.

About the authors

James Copestake is Professor of International Development at the University of Bath, and has a particular interest in modalities of development finance and its evaluation. His recent research and publications have covered contested perceptions of wellbeing in Peru, microfinance in India, the relationship between social policy and development studies, the design of challenge funds, use of political economy analysis in aid management, and qualitative impact evaluation.

Fiona Remnant, MSc International Policy Analysis, is Managing Director of Bath Social and Development Research (BSDR) and has worked in development for over a decade, specializing in the application and communication of academic research to practitioners and policymakers. She has worked for the Centre for Poverty Analysis in Sri Lanka, Oxfam in the UK, and the Centre for Development Studies at the University of Bath. She collaborated with James Copestake on the Assessing Rural Transformations action research project at the University of Bath between 2012 and 2016 which culminated in the development of the QuIP and the creation of BSDR.

ANNEX

Qualitative Impact Protocol (QuIP): guidelines

Bath Social and Development Research

Introduction

The Qualitative Impact Protocol (QuIP) was developed at the University of Bath in the UK to address the challenge of assessing the impact of multi-faceted interventions in complex and/or rapidly changing contexts in a way that is credible, timely, and cost-effective. The QuIP relies on narrative evidence of the causal drivers of change obtained through semi-structured interviews and focus group discussions. It has been designed and tested to mitigate against potential response bias and to address the challenges associated with analysis of qualitative data of this kind.

These guidelines are published in conjunction with a book of case studies and reflections on use of the QuIP during 2016 and 2017. They aim to be of practical assistance to anyone planning to undertake a QuIP study. Section 2 provides a brief overview, including a discussion on how to decide whether the QuIP is appropriate to a particular context, taking into account the type of evidence being sought. Sections 3, 4 and 5 then address design, data collection, and analysis in more detail. Section 6 provides a glossary of key terms. For example, the word project is used very broadly as shorthand for the specific activities, 'treatments', interventions or investments being evaluated by a study, while the term intended beneficiaries refers to the people that a project explicitly and primarily sets out to benefit.

Overview

The main purpose of the QuIP is to collect rich and credible evidence of the causal links between project activities, and changes in the self-perceived wellbeing of intended beneficiaries. It does this by providing intended beneficiaries with an opportunity to describe their experiences in an open-ended way, placing a high value on their personal perceptions and priorities.

QuIP studies generally rely on a mixture of semi-structured interviews with individuals at the household level and focus group discussions at the neighbourhood level. Possible extensions to the approach (not covered here)

http://dx.doi.org/10.3362/9781780447469.011

include incorporating key informant interviews and focusing on drivers of change at the organizational level.

Data collection is carried out by independent researchers located close to the study area, who are informed as little as possible about the project being assessed or the organization responsible for it. The purpose of this blindfolding is primarily to reduce the potential for pro-project bias on the part of respondents, including their response to cues from the researchers. Individual respondents and focus group participants are asked a series of open-ended, non-project specific questions about changes they have experienced within a specified period of time, organized according to selected domains of their lives, livelihoods, and/or wellbeing. These domains depend on the type of project being implemented. Most questions are open-ended, aiming to elicit respondents' own account of both what has changed in each domain and why. Discussion of drivers of change in each domain ends with one or more closed questions to clearly establish the respondent's own view about how this domain of their life has changed overall during the specified time period. This provides a useful snapshot of respondents' overall experience of change and helps to close each section of the interview or focus group discussion prior to moving on to discuss another domain.

The QuIP approach also includes guidelines on how to analyse the qualitative data, assessing how the data relates to the project's theory of change by systematically identifying cause-and-effect statements embedded in it according to whether they (a) *explicitly* attribute impact to project activities, (b) make statements that are *implicitly* consistent with the project's theory of change, or (c) refer to drivers of change that are *incidental* to project activities. Coding for positive or negative attribution, as well as drivers and outcomes, enables the creation of summary tables and diagrams to illustrate how far the data collectively confirms or challenges the theory behind the project.

Asking intended beneficiaries directly about project impact seems both common sense and ethically correct, but doing so in a credible way is not easy. One challenge is to minimize possible sources of bias in the evidence offered, recorded, and shared – e.g. bias caused by respondents saying what they think researchers want to hear. A second is to be clear whose voices are being heard, how typical they are, how they differ, and why. A third is to avoid highlighting the impact of a project in isolation from other factors contributing to changes in selected domains. An additional challenge is to ensure the evidence is not only credible but also relevant, sufficient, affordable, and timely to meet the needs of those using it. QuIP's potential to add value to an evaluation is based on balanced responses to these multiple challenges and the tensions between them.

Key design issues

When designing a QuIP study there are a number of different elements that can be altered in the methodology to meet the specific requirements of the project being assessed. The first issue to address is *who needs/wants a QuIP study*

and why? It is important to give sufficient time at an early stage to examining why a QuIP study is being considered and by whom, and how they will use the evidence it generates alongside information from other sources. This will influence what other data might be needed, how the timing and sampling strategies will overlap, and who will be involved in each stage. Once a clear set of objectives for a QuIP study are agreed then the following checklist can help in working through four key questions: when to conduct the study, how to select a sample, to what extent the researchers should be blindfolded, and who should be involved.

When to carry out a QuIP? Deciding when to schedule a QuIP depends in large part on its relationship to the project being assessed.

- Early in the design phase, as a diagnostic tool for identifying drivers of change or testing the theory behind a proposed project.
- Early on or mid-way through a project, as a 'deep-dive' or 'reality check' to find out what intended beneficiaries think is happening, with time for course correction based on information gleaned.
- After, or at the end of a project, to inform reflection on what worked and why (including the relevance, sufficiency, and reliability of assumptions and theory underpinning the project), even when there isn't a baseline or control group to aid impact evaluation through statistical comparisons.

How to select a sample? Section 3 provides further guidance on sampling strategy, and answers to the question below.

- Is it more important to assess the *typical* experience of intended beneficiaries – or to focus on the diverse experiences of more narrowly defined socio-economic groups, or those exposed to different 'treatments', or who appear from monitoring data to be doing significantly better or worse than others?
- Is overlap with samples used for other studies useful? Or is it important to avoid intended beneficiaries who have already been interviewed under other studies to avoid survey fatigue?
- Is it useful to collect information from individuals or groups who were not intended beneficiaries (e.g. those who may benefit or be adversely affected indirectly)?

To what extent will the researchers need to be blindfolded? Blindfolding – including double blindfolding – can help to reduce the risk of pro-project bias and hence enhance the credibility of findings. But the extent to which the researchers are blindfolded will depend on the aims and the context of the study.

- Double blindfolding is only possible through involvement of a third party, in order that the research team can be recruited, trained, and supported

without knowing the identity of the organization implementing the project or commissioning the study.
- Partial blindfolding may be more appropriate – e.g. a trusted team of researchers might be recruited directly by a commissioner, but without being given information about the project being assessed.
- By not blindfolding them, a trusted team of researchers may be able to obtain more detailed and relevant information about the project; their professional expertise and integrity may be more than sufficient to ensure they are impartial and do not prompt respondents to respond to questions in line with prior understanding and interests.
- No blindfolding may be appropriate if it is impractical, unethical or dangerous to blindfold either interviewers or respondents. It is still possible to focus instead on designing an open-ended and exploratory questionnaire, positioning the study in a broader context, and encouraging respondents to refer to this broader context when thinking about drivers of change.

Who will be involved in carrying out the study? There are three main roles in a QuIP study:

- The lead evaluator is responsible for working with the commissioner, designing and managing the study, commissioning data collection from a research team, and overseeing analysis and reporting.
- The lead researcher is responsible for recruiting and training the field research team and overseeing the collection of data.
- The analyst is responsible for coding and analysing the data collected by the team. They work closely with the lead evaluator to produce the final report. This role can also be carried out by the lead evaluator if they have the appropriate skills.

Deciding who will be involved at each stage of the QuIP study depends in part on the answers to the previous questions. If you are a consultant then you need to decide how much of the work you are able to complete yourself, including the qualitative data analysis, which is a time consuming, but very rewarding, process. If you are a commissioner and are concerned about blindfolding but wish to keep the process in-house, there may be an option to delegate recruitment and even analysis of the data to internal staff from another project, or even another country. Once the data has been analysed you then need to decide how best to involve internal project staff, the QuIP researchers, and, where possible, the respondents, in the process of interpreting the findings.

To QuIP or not to QuIP?

The QuIP offers one solution to the attribution challenge; but it isn't appropriate in all situations and is often best combined with other methods to generate all

the evidence that may be expected of an evaluation. It is important to manage the expectations of all involved about both its potential to add value and also its limitations.

The QuIP **does** the following:

- Generate insights into intended beneficiaries' *perceptions* of change and their understanding of why these changes have happened.
- Throw light on sources of variation in change within a population of intended beneficiaries and the reasons for these.
- Assist in confirming or refuting the theory (of change) behind a project in relation to specific intended beneficiary groups and areas sampled.
- Generate such data in a more credible way by reducing the risk of pro-project bias, through incorporation of an appropriate level of blindfolding.
- Use a bespoke qualitative questionnaire developed with the commissioner to explore perceived changes across a variety of livelihood and wellbeing domains. This includes confirming how far the project itself is or is not contributing to wellbeing change in the way expected. It also includes identifying other (perhaps unexpected) drivers of change and unintended consequences of the project.
- Employ experienced and skilled local researchers, who conduct interviews with intended beneficiaries in an appropriate local language.
- Code and analyse interview data in a transparent, systematic, and rigorous way using flexible thematic coding (for identifying different drivers of change, outcomes at different levels, and the degree to which these can be attributed to the project).
- Enable and encourage users to refer back to source text data, by providing an annotated annex of all coded interview data and/or access to this digitally through a dashboard.
- Generate data that can be used in a wide range of stakeholder and 'sense-making' meetings, including with project staff and intended beneficiaries.

The QuIP **does not** do the following:

- Provide results that are statistically representative of all intended beneficiaries. QuIP studies are designed to gain a deeper insight into changes occurring in purposively selected communities or sub-groups, and to permit cautious generalization across the wider population.
- Guarantee to answer very specific questions about the impact of certain project activities. If the activity is considered important by respondents in a wellbeing domain covered in the interview (and not simply taken for granted) then the QuIP should pick up unprompted references to these project-related drivers. However, if project activities are relatively marginal to respondents' lives then a more direct and targeted line of questioning is

required. However, gaining a better understanding of the broader context of change (including factors that contribute to or mitigate the success or failure of the project) may still be useful.

- Measure the magnitude of impacts or provide detailed quantitative data. The QuIP focuses on the nature of impact rather than its magnitude. Some quantification of drivers of change and outcomes can be generated to summarize and visualize patterns and themes across the sample, but the data is not statistically representative. It may be useful to inform modelling that can simulate the magnitude of change, but other data will be needed with which to calibrate such models.
- Score or weight the overall success or failure of a project. While the visualization of coded qualitative data can make the evidence easier to digest and highlight patterns and outliers, commissioners need to be prepared to engage with the data, and where possible triangulate with evidence from other sources to make an overall assessment of the project and draw out recommendations for future action.
- Directly promote a more participatory approach to development, although findings can be used to promote reflection and learning among intended beneficiaries, and some respondents have also reported finding the interviews and focus groups to be useful and/or enjoyable opportunities for self-reflection.

Ethics

Any research involving people as participants or respondents must be based on ethical principles. Three basic principles of research ethics apply:

- *Respect*. The researcher should recognize the rights and capacity of all individuals to make their own choices and decisions.
- *Beneficence*. The researcher's primary responsibility is to avoid harm and protect the physical, mental, and social wellbeing of all participants.
- *Justice*. The researcher should ensure that the potential benefits for participants are more than sufficient to offset costs and risks.

Blindfolding respondents raises particular ethical questions that need to be carefully assessed prior to each study. Blindfolding does not have to be complete or permanent; temporary blindfolding as an appropriate means to a beneficial end is also possible. Organizations commissioning QuIP studies are encouraged to include triangulation, feedback, and 'unblindfolding' workshops to which staff and respondents can be invited once the data has been collected and analysed. Decisions about precisely how much detail will be hidden and how much revealed can be decided at the design stage, along with agreement on ethical principles and procedures concerning confidentiality and anonymity.

Designing a study

Roles and responsibilities of QuIP team members

This section of the guidelines focuses particularly on the role of the commissioner and lead evaluator in designing and scoping out the QuIP study. First, it is useful to review the roles of different participants in a QuIP study and how they relate to each other, as illustrated by Figure A.1.

Commissioner. The commissioner is the primary consumer of evidence to be collected, and responsibility rests with them to decide what sort of evidence they want, as well as when, where, how, and why to collect it. The QuIP is designed to minimize the amount of time that both the commissioner and project staff need to devote to the study. This helps to reduce potential bias, but also avoids distracting them too much from their other operational responsibilities. However, some involvement of the commissioner is important, particularly at the beginning and end.

Their main responsibilities are:

- to mobilize and provide necessary funds to permit the study to be completed to an appropriate level;
- to confirm the scope of the study;
- to recruit the lead evaluator (and possibly also the analyst, if different);
- to agree on the sampling strategy;
- to provide relevant project documentation and respondent details to enable sample selection;
- to oversee and support appropriate dissemination and use of findings, including ensuring the interpretation of QuIP data is integrated with evidence generated in other ways;
- to contribute to the overall quality of the study.

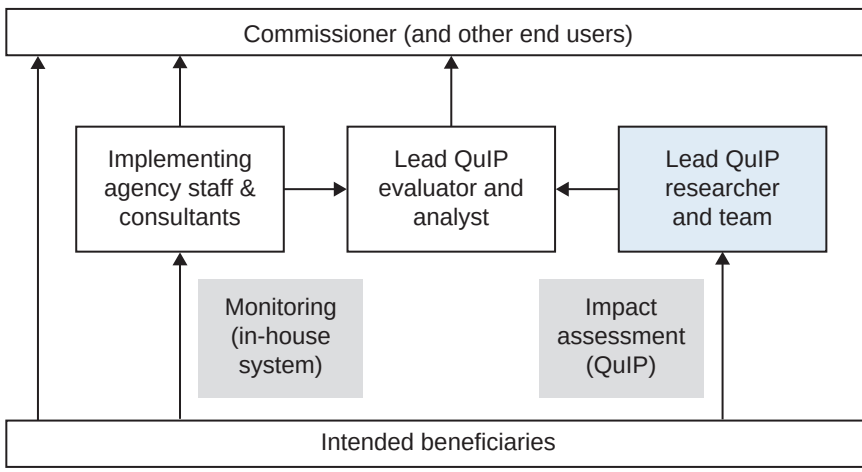


Figure A.1 Roles in a QuIP study

Lead evaluator. The lead evaluator is responsible for working with the commissioner, designing and managing the study, commissioning data collection from the research team, and overseeing analysis and reporting. The lead evaluator may be an employee of the same organization that is implementing the project (so long as they are not directly involved in management of the project). Contracting someone from outside the organization to perform the role is likely to strengthen the credibility of the evidence produced.

The main responsibilities of the lead evaluator are:

- to recruit the lead researcher;
- to refine data collection instruments;
- to brief and if necessary train the lead researcher;
- to provide the research team with the introductions needed to access communities and arrange interviews;
- to oversee data quality control, including data cleaning, briefing, and de-briefing of other staff, including the analyst;
- to produce a synthesis report;
- to support use of the findings, including through participation in additional meetings.

These tasks should be familiar to any experienced qualitative researcher or consultant, but additional training specifically in the QuIP is also likely to be useful. Experience in the selected country and region is also important, not least by allowing closer interaction with the lead researcher, and to inform oversight of data analysis and writing up. The lead evaluator will need to be familiar with the principles of qualitative data analysis and must also be in a position to manage the sub-contracting of the researchers. Other important considerations are integrity, reputation, availability, and cost.

While responsibility for selecting the lead researcher can be left to the lead evaluator, the commissioner can also participate in their identification and selection. However, they should not be in direct communication with them as this would undermine the blindfolding process.

One exercise that can inform selection of the lead evaluator is to share these guidelines with candidates and invite comments on them. This initiates dialogue with the lead evaluator at an early stage. Likewise the lead evaluator, once recruited, can share and discuss these guidelines with potential lead researchers.

Lead researcher. The lead researcher plays a key role in the QuIP process and is responsible for managing all aspects of data collection. They will typically be experienced qualitative researchers from the country where the evaluation is taking place, with a track record of conducting high quality fieldwork, and recruiting, training, and managing a research team. A commitment to the goal of enabling the authentic voices of intended beneficiaries to be heard is also critical.

The main responsibilities of the lead researcher are:

- to recruit the research team;
- to train the research team in the QuIP methodology;
- to gain access to the research site and the pre-selected sample of respondents;
- to ensure the research team conduct interviews in a way that is consistent with both the ethical principles set out above and an agreed code of conduct;
- to ensure interview data is of a high standard and is written up in a timely manner;
- to write a brief report detailing experiences and any comments pertinent to the study.

Analyst. The analyst is responsible for:

- coding all the interviews (individual household and focus group discussions);
- entering the data into suitable software to enable analysis;
- analysing the data and pulling out key findings in preparation for de-briefing with the lead evaluator.

The role of analyst and lead evaluator can be combined. But they require very different skills, and hence there is a good case for separating them, so long as they can communicate closely and collaborate effectively. In some situations the commissioner may opt to choose the analyst and this person may even be a member of staff, provided they are not too closely associated with the specific project and are clearly mandated to conduct the study as objectively as possible. By authorizing a suitably trained member of its own staff to do the analysis, the commissioning organization can internalize findings more fully and directly: no summary report can match the more detailed insight that comes from reading through all the primary data. Even if they don't actually do the coding and analysis, encouraging staff to review the primary data can also be useful in building interest in the study, increasing 'buy-in' to its findings, no matter how uncomfortable they may be.

In all cases the quality of the analyst depends critically on appropriate capabilities, reinforced by training. An effective analyst must be able to immerse themselves in the data and identify patterns in it, both *within* the written reports on each interview and focus group, and *between* them. This includes being able to identify and untangle often complicated causal claims and stories of change, both positive and negative. Being able to see broad patterns and pay attention to detail, in order to accurately reflect what respondents said, is a demanding skill; most QuIP analysts to date have training and experience of qualitative analysis at doctoral level. Other important skills are the ability to code comprehensively, inclusively, and systematically, following the more detailed guidelines developed for use with formatted Excel spreadsheets and business analysis software adapted specifically for analysis of QuIP data. The analyst is expected to pull out the main findings from the data,

construct the relevant tables and data visualizations, and present these to the lead evaluator as the foundations for the QuIP report.

Timing of the study

The timing of a QuIP study may be dictated by the timetable of a project's external funders or budget cycles, but also depends on the phasing of the project's implementation, seasonality, and the expected impact trajectory. The data is based on respondent recall over a specified period, although recall bias is less of a worry for narrative data since this does not require respondents to recollect precise figures (e.g. their income) for a specific date. Nevertheless, the ideal is to conduct repeat studies before, during, and after project completion to permit comparisons over time. Scope for using the QuIP will also obviously depend on the availability of funding, and when to collect data is likely to be dependent on an assessment of when feedback is most likely to be useful: in order to influence decisions about whether to close, extend, adjust, replicate or scale-up a project, for example. Other cost considerations include choice of sample size, as well as the nature and extent of project monitoring activities on which the QuIP can build. These issues are discussed below. One advantage of the QuIP is that it is possible to start small with a pilot, and then enlarge or repeat the studies. Confronted with complex projects and contexts there is an even stronger case for this approach, rather than risking all available resources on one large study.

Case selection

There is no universal best practice method for selection of cases for a QuIP study since it depends upon many contextual factors. The most important of these are (a) the main purpose of the study, including its role in assessing an explicit theory of change; (b) availability of relevant data about variation in the characteristics of expected gainers and losers from the project; (c) availability of relevant data about variation in their exposure to project activities; (d) time and resource constraints; and (e) how much data one analyst can manage. This section briefly explores these factors, and then (f) outlines the sequence of sampling decisions and actions needed prior to starting data collection.

Main purpose of the study. Deciding who to interview, how many people to interview, and how best to select them requires clarity about what information is being sought, by whom, and why. Neglecting this not only leads to poor practice, but also to misunderstandings about the quality of a study. For example, sample bias is not an issue for a QuIP study that deliberately sets out to identify drivers of successful outcomes by interviewing positive deviants. Deliberately selective or explicitly biased sampling is, in this instance, fit for purpose.

More generally, differences in sampling strategy arise from whether the priority is to confirm and quantify the overall impact of a completed project on a defined population in relation to a predetermined set of measurable indicators and theory of change, or to explore what is happening in a more open-ended way – to improve implementation of an ongoing project, for example. The QuIP is a relatively flexible and open-ended approach. Its primary purpose is to gather evidence of causal processes at play, not to quantify them.[1] Deciding on the number of interviews and focus groups to conduct depends less on reducing sample bias than on assessing at what point the extra insight into causal processes gained from more data is unlikely to justify the extra cost.[2] As a benchmark, a standard QuIP consists of 24 individual household interviews and four focus group discussions. But this may need adjusting for many reasons, including the time required to locate respondents. For example, it is common to do a 'double QuIP' that doubles the data collection, often in order to draw sub-samples from two contrasting segments of the population.

Contextual variation. Random selection of respondents across the entire population affected by the project is a good starting point for thinking about sampling for a QuIP study, but there are also good reasons for departing from it. For example, if there are good grounds for expecting impact to vary for different sub-groups, and we already have data that enables us to identify those sub-groups, then stratification of the sample would be a useful strategy. A project may cover two areas with marked geographical differences, justifying including a minimum quota of people living in each (e.g. urban and rural areas, irrigated and non-irrigated villages). Stratification of the sample on these grounds is an art that depends on prior thinking about what contextual factors are most likely to be a source of variation in project outcomes. Where baseline and endline monitoring data has already been collected and analysed then there are additional possibilities for QuIP sample selection. For example, quota samples can be selected for 'positive deviant' households that have experienced rapid improvement in key indicators in order to find out more about the drivers of their success. Conversely there is a case for deliberately biasing the sample towards households that have done badly, in order to learn why. A third option is to do both in order to be more confident about picking up the full diversity of causal changes experienced by households. Or a double QuIP might quota sample four groups: richer and improving; richer but declining; poorer but improving; poorer and getting worse. In all cases the number of interviews that it is worth conducting depends not only on minimizing sampling error, but also on the marginal benefit (in terms of extra evidence of key drivers of change) obtained from each extra interview.

Exposure or 'treatment' variation. This refers to variation in how project activities are expected to affect different people, including those who receive different packages of goods and services. In addition, there are those who may only be affected indirectly: because their neighbours are affected and may

share things with them, for example. If data is available on variation in who directly received what and when, and it is expected that these differences will have different causal effects, then there is a case for stratifying the sample to ensure it reflects a range of treatment exposure. This is particularly the case if part of the purpose of the study is to aid decisions about which of a range or combination of project activities to expand or stop. Impact assessment using the QuIP does not require a control group of people completely unaffected by the project. There may nevertheless be an argument for interviewing some people unaffected by the project (but similar to those affected by it) in order to explore whether they volunteer different or additional drivers of change. For example, if monitoring data indicates they achieved desired outcomes ('equifinality') then it may be useful to identify what alternative package of causes contributed to this.

Time and resource constraints. A third reason for departing from pure randomization in sample selection is to cluster respondents geographically in order to reduce the time and cost of data collection. One way to do this is to adopt two stage random sampling, with the first stage based on geographical units (e.g. villages, districts or census areas) listed according to some known criterion that is likely to be an important source of variation in project outcomes (e.g. distance from a main road or market centre; agro-ecological zones). One locality is then selected at random, and additional localities are selected by counting X down the list, where X is the number of localities divided by the desired sample number. For example if there are 40 villages with an equal number of intended beneficiaries in each, and it is agreed to sample four of them, then every 10th village should be selected from a random starting point on the list. In the second stage the procedure is repeated, except starting with a list of all intended beneficiary households in each selected village.

Ultimately, budget constraints (dictated by factors beyond the control of the lead researcher or even the commissioner) may also limit the total number of interviews and focus groups that a QuIP study can cover. The challenge is then to make decisions that maximize potential value, subject to this constraint. This is less precise but no less reasonable than using power calculations to work out the minimum 'required' sample size for estimating the value of a key indicator to an acceptable level of statistical significance.

Absorptive capacity of the analyst. An additional influence on sample size and selection is the limit to the amount of data in which the analyst can immerse themselves, yet still code comprehensively, systematically, and inclusively. Going beyond a double QuIP is likely to stretch all but the most gifted and experienced analyst. If a larger sample is nevertheless justified then parallel QuIPs can be conducted and analysed separately, and the reports can then be subjected to synthesis or meta-analysis.

This constraint may also reinforce the argument for staggering studies – i.e. conducting two smaller studies a few months apart rather than doing a single larger study. This can help to build understanding of project impact

lags, pathways, and cumulative processes, as well as those of other drivers of change. Sampling strategy for repeat studies can also be informed by lessons from earlier studies. Again, the principle here is that credibility of findings builds incrementally with the addition of each extra piece of evidence.

Interviews and focus group discussions. A further design and sampling issue concerns the balance of individual interviews and focus group discussions to conduct. The main reason for including both is to permit triangulation between what people volunteer about their own personal experience, and what they say in the presence of peers, given that group testimony often has more reference to shared experiences and norms. Focus groups are generally undertaken after interviews by inviting interview respondents to attend a follow-up meeting with a friend, or delegate another member of the household to do so. The default pattern for focus groups is to plan four for each set of 24 interviews: one each for older and younger men and older and younger women. However, precise details of how to do this will depend on the precise goals of the project, including the composition of intended beneficiaries, and the geographical dispersion of the initial interviews.

Provision of background information

As the analyst is not blindfolded, they will need information about project activities to enable them to code for attribution more accurately and to identify project activities that are 'missing' (i.e. activities that respondents don't mention). It is the commissioner's responsibility to ensure that the theory of change underpinning the project and relevant project documentation is also made available to the lead evaluator in a timely fashion. It is possible to undertake a QuIP study without a theory of change, or with a loose or tacit one, but the QuIP will be both harder to design and correspondingly more limited in how far it can explicitly confirm or challenge prior expectations and assumptions. Other important documentation includes any contextual analysis or research undertaken prior to the project design, project terms of reference, activity plans, and any monitoring data available. Ensuring the buy-in of local staff implementing a project is often critical to this, as vital information to inform timely questionnaire design and sample selection may have to come from them.

Refining data collection instruments

The QuIP employs two main data collection instruments: semi-structured household level interviews and facilitated focus group interviews. The questionnaires for both are based on a series of livelihood and wellbeing domains, each comprising generative, supplementary, and closed questions. The domains are

Box A.1 Example questions for a domain on food production

Generative question

- Please tell me how your ability as a household to produce your own food has changed in the past two years, if at all.

Supplementary questions

- What do you do more?
- What do you do less?
- In which seasons have changes been most pronounced?
- What are the reasons for these changes?
- Have you taken up any new activities to help you produce more food?
- Is there anything you have stopped doing?
- Are you doing anything differently?
- Why did that happen?

Closed question

- Overall, how has the ability of your household to produce enough food to meet its needs hanged in this time? [Improved, No change, Worse, Not sure]

designed to cover outcomes specified in a project-specific theory of change. For example, a project designed to promote household-farm livelihoods, food security, and nutrition might include domains for food production, food consumption, income, cash spending, intra-household relationships, inter-household relationships, assets, and overall wellbeing. An example question is provided in Box A.1. Generative questions are designed to stimulate discussion in an open way, with lists of supplementary questions available to sustain and deepen conversations about changes observed by the respondent and the reasons behind them. Closed questions follow open-ended discussion of a domain, and are a useful way of drawing discussion of it to an end before moving onto the next domain. There are two reasons why both household level interviews and focus groups start discussion of any topic with broader generative questions before focusing on more specific ones: (a) to maximize the opportunity for respondents to raise unknown and unexpected issues; and (b) because information about reasons for change (including those arising from specific activities) that is provided voluntarily or without prompting is more credible.

Carrying out QuIP fieldwork

Recruiting the lead researcher

The lead evaluator's first task is to recruit a lead researcher to collect the data, generally as part of a team of staff they recruit and employ. The lead researcher should not be an employee of the organization implementing the project, although they could work in the same organization as the lead evaluator, so long as the latter is able to conceal from them the identity

Box A.2 Criteria for selecting the lead researcher and research team

1. Qualifications and experience (particularly with qualitative research methods).
2. Qualifications and experience of named researchers to assist them.
3. Evidence of the quality of similar work they have carried out to a high standard in the past.
4. Knowledge of general context, including relevant languages.
5. An appropriate mix (within the team) of gender and other attributes: e.g. more women if primary respondents are likely to be mostly women.
6. The quality of context specific proposals about how they will conduct the study, including: how long data collection will take; logistics of travel and accommodation; compliance with the commissioner's timetable; proposed modifications to the research guidelines; overall feasibility; and the quality of codes of conduct used to guide staff.
7. Proficiency in English and basic computer skills, plus ability to communicate quickly with the lead evaluator (whether face-to-face, by phone, e-mail, Skype or most often a combination of these).
8. Evidence of their awareness of different forms of potential bias, and how the process of data collection and reporting will affect its credibility. One potential source of such evidence is the quality of comments and queries they provide on the QuIP guidelines, including explanations of how they might adapt them to a particular context.
9. No prior direct involvement with the project, given that the aim is to provide independent evidence. While the QuIP seeks to limit the researchers' prior knowledge of the project, such 'blindfolding' cannot be guaranteed, and so is no substitute for recruiting researchers with a high level of professional integrity as social scientists.
10. Price.

of the commissioner and project implementing agency. Finding the best person to lead data collection is perhaps the single most important determinant of the outcome of a QuIP study, so it is worth investing in a rigorous search and selection process. Open and transparent selection also adds to the credibility of the findings. Criteria for selection of the lead researcher are set out in Box A.2.

Briefing the lead researcher

By the time they are contracted, the lead researcher should already be familiar with these QuIP guidelines, having been invited to offer comments on them as part of their own selection. Initial briefing by the commissioner should cover the following.

- Lists and locations of households from which to select respondents, along with instructions on how to do so, and how to handle replacement in the event of non-response.
- The interview and focus group discussion schedules. It is essential that the researcher pre-tests this in order to identify problems of translation and interpretation, and to gauge likely interview times. These should be followed up by a second meeting to discuss issues raised and to agree any changes to the fieldwork plan and time schedules.

- Details of how the researcher will be introduced to selected households and how focus groups will be organized.
- Details of the format of expected research outputs and how they will be checked.
- Research ethics, including codes of conduct for fieldwork and use of data.

To maintain their distance from the project implementing agency, the research team should make their own logistical arrangements, including avoiding all contact with immediate project staff to locate respondents. Instead, the lead researcher will need to arrange for introductions to official gatekeepers at the appropriate level, and provide necessary supporting documentation. Researchers should each be given an appropriate letter that can be shown to respondents and to any other interested party, introducing them personally and explaining their affiliation and role.

Note that the statement of the purpose of the research should not refer directly to the project itself but to the underlying issue(s) it seeks to address. The main reason for this is to reduce pro-project response bias. While the project may have been the immediate prompt for the study, its ultimate purpose is to contribute to the wider development goals the project addresses. Being less than fully transparent about the purpose of the interview is ethically contentious, but can be defended on the basis that it results in more reliable and therefore more useful information.

The lead researcher's role

The first job of the lead researcher is to recruit the research team and to ensure that they are fully trained in the QuIP methodology. It is important that the researchers have a detailed understanding and familiarity with the questionnaire so that the interview flows smoothly in a conversational way. Particular attention should be given to how key concepts will be translated and explained in the languages to be used during data collection. Training often includes mock interviews with the draft questionnaire and practice in writing up the findings from notes.

Questionnaires should also be piloted with intended beneficiaries of the project, or people as similar as possible to them. The piloting should also include practice in writing up interviews from notes onto the Excel templates. Ideally, these transcripts should be shared with the lead evaluator and analyst to ensure they are meeting study objectives and allow for further adjustments to be made.

Access to respondents can be challenging and/or time limited, so the lead researcher needs to be able to overcome problems as they arise. Interviews (individual household and focus group) should be conducted in a sensitive and courteous manner, showing appropriate respect towards respondents. Consent must be obtained at the start of each interview. During the interviews,

Box A.3 List of outputs required from the lead researcher

1. A brief activity report on the work undertaken: pre-testing of instruments, training, sampling, timeline, plan and departures from the agreed plan, and an account of difficulties encountered.
2. Original schedules of semi-structured interviews with hand-written notes and timeline sheets (one per household).
3. Data from semi-structured household interviews and focus group discussions (in Excel).
4. Digital sound recordings of interviews and focus group discussions.
5. A brief report summarizing the researchers' experiences and their own perceptions of drivers of change in the areas visited.

it is important that the research team keep asking respondents about the reasons behind any reported change until they get to the root driver or drivers behind those outcomes.

The research team must be able to type up the interviews as soon as possible after their completion. The lead researcher is responsible for ensuring that the team adequately translates the data (into the language in which analysis will be conducted) and enters the responses in the required format.

Once the fieldwork is complete, the lead researcher writes a report outlining any difficulties in data collection, as well as observations of the team about the communities visited. It likely that they will also be called upon to clarify any questions that the lead evaluator and analyst have about the interview data. However, the key 'deliverable' of the research team is the primary data, subject to this being of the expected quality. A summary of expected outputs is listed in Box A.3. One advantage of this division of roles from the perspective of lead researchers is that they are insulated to a large degree by the lead evaluator from much of the risk and uncertainty (e.g. of delayed payment by the commissioner) arising from discussion of draft final reports.

Data analysis and use

Data analysis and presentation

As alluded to in the previous section, the lead evaluator plays an important quality assurance role in reviewing the primary data generated by the research team in terms of expected detail, quality of writing, and rigour. Wherever possible the lead evaluator and/or analyst also arranges a face-to-face or virtual debriefing meeting with all members of the research team. This should review both substantive findings and issues arising from the data collection process: what went well, obstacles, difficulties, doubts, and any other thoughts relevant to interpretation of the data. The meeting also provides the research team with an opportunity to share additional material and ideas arising from the fieldwork.

With the transfer of the primary data, responsibility for the quality of the study now passes to the analyst. One criteria for assessing analytical rigour

is whether similar findings would be replicated by two analysts working in parallel and independently of each other. The semi-standardized process of doing QuIP analysis increases the prospects of this happening, particularly for a more confirmatory study that is informed by a detailed theory of change. This section sets out this process, with particular emphasis on forms of analysis and reporting used to identify patterns and provide an overview of findings. At the same time, it is important to emphasize that the QuIP is designed to be used in complex contexts and to assess multi-faceted projects. This means that it is also to be expected, and indeed welcomed, that findings will reflect the *positionality* of the analyst, and/or their personal background, discipline, experience, and insight. This is particularly the case for more exploratory studies that are primarily intended to generate new findings and insights, rather than to confirm whether the project is having the expected impact or not.

Assessing the data can be divided into five steps: (a) familiarization with all the data by reading and rereading it; (b) allocation of segments of the texts to different codes; (c) identification of wider themes, stories or arguments that may combine different coded elements together; (d) back-checking these themes, and the clusters of coded data supporting them, against the original data; and (e) reporting findings to others. However, this process is rarely strictly linear, serving as a particular and important reminder that the analytical process is iterative. At the same time, the QuIP does also involve more tightly structured tasks, thereby distinguishing it from even more fluid ways of doing thematic analysis in social research.

One of the more mechanical steps is to analyse the closed questions about each domain. An overview of these results is illustrated in Figure A.2. This enables both the analyst and users of the study to gain a quick sense of who the interviewees were and what their perception of change was, within a specified period, across all domains. However, even this data can be presented and interpreted in many different ways. For example, patterns can be revealed by ordering the list according to different socio-economic characteristics

	Wealth Group	*Food production*	*Money from livestock*	*Money from other sources*	*Quantity of food*	*Variety of diet*	*Health of children*	*School attendance*	*Amount children working*
DHFC-2	Middle	–	–	–	–	–	=	+	=
DHMC-4	Middle	–	–	–	–	–	=	=	=
DHMC-5	Middle	=	–	+	–	+	+	+	+
DHMC-6	Middle	+	+	+	+	+	=	=	=
DHFC-7	Middle	+	+	–	–	–	+	+	+
DHMC-11	Middle	=	=	+	+	=	+	+	–
DHFC-12	Middle	=	=	–	–	+	+	+	=

Figure A.2 Example of automatic tabulation of the closed question responses

(e.g. by age, gender, location, and/or wealth group) as shown in the left hand columns of Figure A.2. The data can also be triangulated against changes measured using quantitative baseline and endline monitoring data.

This initial analysis provides a useful profile of the sample and experience of change across it, but does not reveal anything about the causal processes behind observed changes. To get at this, the QuIP analysis entails coding segments of the narrative data that make causal claims (e.g. 'X caused Y', or 'Y happened because of X and Z'). The same causal claim can also be coded in one, two or three different ways:

- as a driver of change (i.e. as a specified cause or contributor to an outcome) – based on inductive classification of different reasons behind any change or outcome;
- as an outcome (or a specified consequence or effect of a driver) – based mainly on inductive classification;
- as an attribution claim – based on predetermined codes that provide an initial indication of the nature of the attribution claim (see below).

The drivers of change and the outcome codes are developed uniquely for every study, since they depend entirely on the evidence presented in the narrative statements. These codes can therefore be generated by the analyst without reference to the project theory of change: indeed, an initial round of data coding when the analyst has not yet reviewed material about the project adds to the rigour of the analysis. However, subsequent rounds of refining codes and identifying linking themes can also draw on background information about the project, given that the final goal of most QuIP studies is to interrogate project theory (and commissioners' prior expectations of its impact) against 'reality' as perceived by intended beneficiaries. Drivers of change and outcome codes are mostly explicitly labelled as either positive or negative in relation to the wellbeing of the respondent. Where this normative ascription is not self-evident it can usually be inferred by viewing the causal claim being coded in the wider context of what the respondent said.

Attribution codes are the same for all studies, as shown in Table A1. In most cases, causal claims are again coded as explicitly positive or negative. A distinction is then made between causal links that are: explicitly attributed to project activities (1, 2); implicitly consistent with project theory (3, 4); or incidental to it (7, 8). This part of the coding clearly does require the analyst to be as familiar as possible with project activities as planned and implemented.

Triple coding the data using this system makes it easier to produce tables and diagrams based on frequency counts of different codes to provide an overview of what the study reveals about drivers of change, outcomes, and (most importantly) the relationship between the two. The data can also be used to draw more complex causal chains and maps. These can serve both to summarize how far the evidence confirms or contradicts prior

Table A.1 QuIP attribution coding key

Description	*Positive code*	*Negative code*	*Explanation*
Explicit project link	1	2	Positive or negative change explicitly attributed to the project or to explicitly named project activities or project partners
Implicit project theory of change link	3	4	Change confirming (positive) or refuting (negative) the specific mechanism (or theory of change) by which the project aims to achieve impact, but with no explicit reference to the project or named project activities
Other (incidental) attributed	5	6	Change attributed to other forces (not related to activities included in the project's theory of change)
Other not attributed	7	8	Change not attributed to any specific cause
Neutral	9		Responses that were felt to be of interest, not related to change

expectations of the commissioner and to construct an inductive model of change that incorporates unanticipated drivers and outcomes. At the same time the semi-automated generation of a range of reports from the primary data has its limitations. Frequency counts provide only one indication of the importance of different coded drivers or outcomes – the emphasis respondents place on them also matters, including how often they are repeated in the same interview, as do the logical connections between different links and arguments. For these reasons, the analyst still has an important, active, and reflexive role to play in deciding which outputs are most meaningful and how to complement summary reports with discussion and direct quotations of selected text in full.

Bath Social and Development Research has developed an interactive QuIP dashboard to help the analyst interrogate the data flexibly by switching from summary visualizations to the full text underpinning them. This helps to ensure that rich details in intended beneficiaries' individual voices are not irrevocably lost behind summary numerical data. A few examples of visualisations are reproduced below, but since they often use colour they are best viewed at www.bathsdr.org.

Although coding and analysis of QuIP data can be greatly facilitated by use of bespoke spreadsheet and business analytics software, the work remains primarily manual, as well as dependent on the analyst's active, skilled, and reflective personal engagement with the data. Likewise, while coding and analysis can be facilitated by drawing on project theory (about the causal links from project activities and contextual factors to outcomes) a good analyst will also be open to the unanticipated drivers of change, outcomes, and patterns identified by eyeballing the data, drawing on wider experience, and discussing possible themes emerging from the analysis with others.

Table A.2 Example of relationships between selected drivers and outcomes

Driver	*Outcome*														
	Invested in or started business	*Stopped/ reduced piece work*	*Increased income*	*No longer go hungry*	*Increased WASH knowledge*	*Increased purchasing power*	*Increased food security*	*Increased yield*	*Increased savings/ loans*	*Increased assets*	*Able to pay school costs*	*Improved hygiene practices*	*Increased resilience*	*Better soil conservation*	*Increased livestock numbers*
Social cash transfer (CW)	36	40	24	28		148	78	2	3	36	16		8		25
Village savings group	11	4	1			8	5	1	70	18	2		21		2
Agricultural training and advice	1		8	1			13	28					4	20	
WASH information					26							27			
Started a business		1	10	1		1	1			5					
Business training	7	1	2			1									
Gender training															

Note: colours are used to reinforce frequency: darker indicating more citations; WASH = water, sanitation, and hygiene.

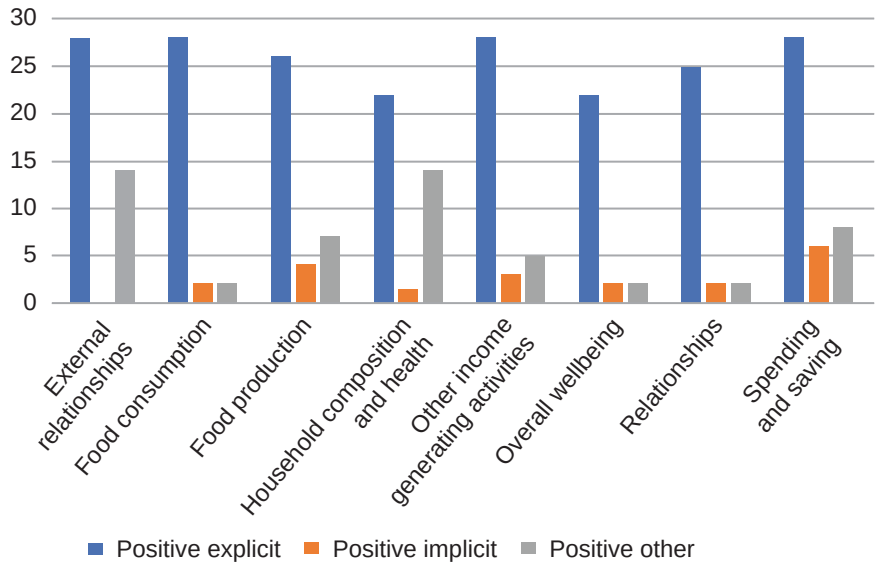


Figure A.3 Example of frequency counts (0–30) of attribution of positive change across pre-selected domains

Table A.3 Example of attribution frequency of positive outcomes by domain with respondent codes

	Explicit attribution of positive change to INGO	*Implicit attribution of positive change to INGO*	*Positive change attributed to other source*
Food production and cash income	DHMC-1 DHFC-2 DHMC-3 DHMC-4 DHMC-6 DHMC-11 DHMP-12 DHFP-9 DHFP-10 DHMG-4 UEMC-6 UEFC-1 UEFC-3 UEMP-5 UEFG-1 UEMG-3 UEFP-2 SEMP-5 SEMP-2	UEFP-4 UEMP-5 DHFP-8 DHFP-9 DHFP-10 DHMP-12 SEMP-5 SEMP-2	DHFC-2 DHMC-5 DHMC-6 DHFC-7 DHFP-8 DHFP-10 DHMP-12 DHFG-2 UEFC-1 UEFC-3 UEMC-6 UEFP-4 UEFP-2 SEMP-5 UEMP-5 UEFG-1 UEMG-3 SEFC-1 SEFC-3 SEFC-4 SEMP-2 SEFP-6
Food consumption	DHMC-1 DHFC-2 DHMC-3 DHMC-4 DHMC-5 DHMC-6 DHFC-7 DHMC-11 DHFP-9 DHFP-10 DHMP-12 DHMG-4 UEFC-1 UEFP-2 UEFG-1 SEFC-1 SEMP-2	DHFP-8 DHMG-4	DHMC-1 DHMC-5 DHMC-6 DHFC-7 DHFG-2 DHMG-4 UEFC-1 UEMC-6 UEFC-3 UEFP-2 UEFP-4 UEMG-3 SEFC-1 SEFC-3 SEMP-2

Note: Using codes for each respondent where each letter represents information about them allows the reader to easily see any patterns across the data set and also to find specific quotes in the coded extracts. This can include categories for location, wealth rank, gender, programme type, etc. Font colours can be varied to indicate different sub-groups: e.g. by village.

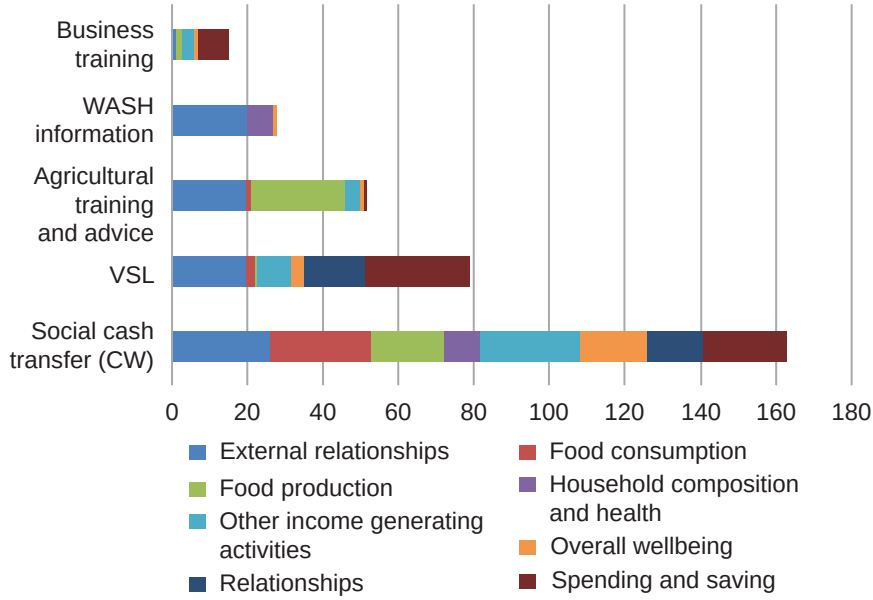


Figure A.4 Example of frequency counts positive drivers of change by domain

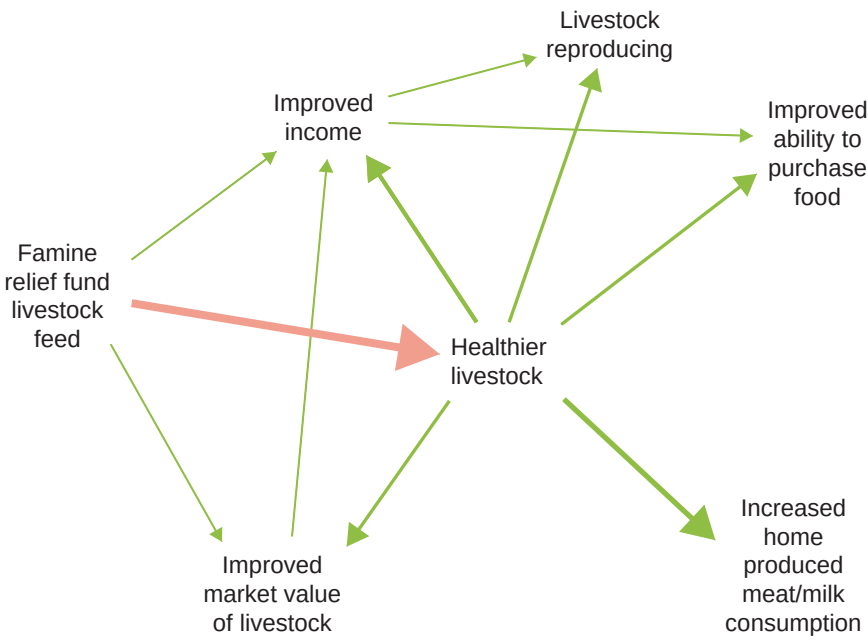


Figure A.5 Example of causal chain
Note: Thickness of arrows denotes strength of relationship between driver and outcome, calculated by the number of times the driver and outcome were cited together.

This is particularly true when it comes to deciding on key themes. Data relevant to each can be charted in a table, with rows for each respondent and columns for topics within the theme. This can be filled out by cutting and pasting material from the original reports into a spreadsheet, or using more advanced software such as NVivo, MAX QDA or Quirkos. The final step is then to produce a written theme-by-theme account of the data, generalizing across households and focus groups to the extent that the data permits this. A simple way of doing this is to write a paragraph or two on each topic that weighs up the quality and frequency of narrative information supporting or opposing particular hypotheses about changes as reported by the respondents, along with the explanations they offered for them. The report may also include discussion of how different concepts are understood by respondents, explore variation in what they said (e.g. according to household type), and identify other patterns in the responses. The evidence can also be systematically compared and contrasted with evidence obtained through monitoring. This can then be used to create case studies, building up a more detailed picture of change in different types of households, and what factors may have influenced outcomes.

More discussion on the use of numbers generated from the analysis of QuIP data can be found in the paper *'QuIP and the Yin/Yang of Quant and Qual: How to navigate QuIP visualizations'* (BSDR, 2017) – available at www.bathsdr.org/resources.

Drawing recommendations from findings

The final stage of a QuIP study is to explore the consistency of the evidence generated with the prior expectations and ideas of the commissioner and other stakeholders. Key interpretive questions include:

- To what extent are findings consistent with both transmission mechanisms and intended outcomes set out in the theory of change?
- What evidence of processes and outcomes is generated that is not consistent with the original theory of change, and how can these be explained?
- What scope is there for generalizing reasonably from findings to the whole project, taking into account characteristics of the whole sample of intended beneficiaries and of the sample of those interviewed?
- What explains differences in intended and observed processes and outcomes of the project and what are the implications for future activities?
- Is the data consistent or at odds with quantitative monitoring data, as well as data collected from other sources (including meetings with project staff)? How can differences and similarities best be interpreted?

The commissioner should be provided with a report which will include core summary tables and other data visualizations picking up on the most interesting patterns in the data, appended by coded extracts which make it easy to find the source data. This ensures that all the data is available rather

than only the quoted extracts selected by the evaluator, and that there is a clear reason for any selected extracts. In addition, if trained in dashboard use commissioners may be supplied with the anonymized data dashboard so that they can further interrogate the data.

Closing the feedback loop

Once analysis has taken place (and if the research team is unlikely to be asked to participate in further blindfolded studies) then a powerful final stage of any QuIP study is to organize one or more fully unblindfolded triangulation or sense-making workshops involving project staff, the research team, respondents, and other relevant stakeholders. This ensures greater transparency, and also allows for deeper discussion and sharing of the findings. The discussions from such workshops can be useful for putting the QuIP findings into a broader context, and starting to draw up internal recommendations for practical action.

Negative or unexpected findings may be a source of internal tension, with some staff or stakeholders preferring to see them buried or dismissed without proper reflection (an issue that can also emerge in discussion of draft reports). Such tensions can be viewed as obstacles to completion of studies, and make unplanned and unwarranted demands on time and resources. But they can also in themselves be powerful learning opportunities.

An alternative follow-up initiative is for commissioners, with or without help from the lead evaluator and research team, to report anonymized findings back to the respondents who were interviewed for the study, through one or more focus groups. This provides an opportunity to thank respondents for their participation, and to explore how they interpret the findings in more detail. Uncertain findings, and specific questions which were not answered in the original interviews, can be explored further, and scope for follow-up project activities discussed.

Glossary of key terms

Analyst: Responsible for coding and analysing the data collected by the research team. They work closely with the lead researcher to produce the final report. This role can also be carried out by the lead evaluator if they have the appropriate skills.

Attribution: Evidence that an action (X) causes change in an outcome (Y), which is the same as saying that action (X) is a necessary condition for change in an outcome (Y) in the presence of a package of other drivers of change (Z). The causal package (X, Z) is sufficient to cause the change in Y, but need not be necessary, because there may be other causal packages that are also sufficient to do so. Some authors define attribution more narrowly as a quantifiable effect of X on Y, but here the term is used more generally and in a way that is synonymous with contribution.

Attribution code: a code that indicates whether a causal claim is having either a positive, negative or neutral effect on a specified outcome variable. Codes may additionally distinguish between causal claims that explicitly link an outcome to a particular organization or project, or implicitly do so by confirming the theory of change underpinning its activities. Causal claims may also link outcomes back to drivers of change that are unrelated or incidental to the actions of a particular organization or project.

Blindfolding: The process of deliberately restricting what interviewers and/or interviewees know about a project (and the organization behind it) in order to reduce the potential bias in favour of emphasizing the importance of this activity or actor relative to other drivers of change.

Causal claim: A proposition that a specified outcome (Y) was a direct consequence of a specified action (X) or (Z).

Causal driver: See driver of change.

Causal mechanism: The process by which one or more drivers has one or more outcomes. This is often hidden – e.g. because it depends on a change in thinking or feeling within a person's head, or on a change in the ideas, norms, and values shared by a group of people.

Citation count: One count per domain per respondent.

Commissioner: The organization contracting a QuIP study, and the primary user of the evidence to be collected. Responsibility rests with them to decide what sort of evidence they want, as well as when, where, how, and why to collect it.

Credibility: How believable a particular finding or conclusion is to a particular person or audience. It acknowledges that their capacity to assess the validity and reliability of findings depends upon their own independent knowledge, experience, and opportunity for cross-checking or triangulation against other sources. This contrasts with the quest to establish universal truths that are valid and reliable independently of the perceiver. In aspiring to produce reasonable or 'good enough' evidence the success of the QuIP ultimately hinges on the credibility of findings.

Credible causation: X credibly causes Y in a particular context if (a) there is strong evidence that X and Y happened; (b) several stakeholders independently assert that X was a cause of Y, with minimal prompting; (c) there is no more credible counter-explanation for why they might have said this; and (d) their account of how X caused Y is consistent with a plausible theory of change.

Domain: A field or category of outcomes agreed in advance with the commissioner and used to structure interviews and focus group discussions. Most studies address a set or group of domains that are consistent with a theory of change. For example, they may refer to different aspects of the wellbeing of individual intended beneficiaries.

Double blindfolding: exchange of information where both researchers and respondents are blindfolded (see blindfolding).

Driver (of change): An action or state (X or Z) behind outcomes (Y). These are generally self-reported by respondents, in answer to questions like 'why did that happen?' or 'what was the reason for that?' This term is synonymous with causal driver. Coding is used to group similar drivers together into groups or clusters.

Evaluation: Systematic enquiry into how a project worked or is working: how far, how cost-effectively and how sustainably it is realizing its intended goals, and the appropriateness of those goals. This can incorporate impact assessment, but goes beyond it.

Intended beneficiary: Those people that a specified organization is aiming to benefit, by achieving outcomes specified in its theory of change. In the case of capacity building projects the intended beneficiaries may be organizations or associations of people.

Impact: Evidence that a specified project credibly caused a specified set of outcomes. In some cases the term impact may refer specifically to final or higher level outcomes.

Lead evaluator: Responsible for liaising with the commissioner, designing and managing the study, commissioning data collection from a research team, and overseeing analysis and reporting.

Lead researcher: Responsible for recruiting and training the research team, and overseeing the collection of data.

Monitoring data: The QuIP works well when used in conjunction with systematic quantitative monitoring of change in selected indicators of project activities, outcomes, and/or wellbeing domains. This aids both sample selection and assessment of how generalizable findings are likely to be across the project's full population of intended beneficiaries.

Outcomes: Changes (positive or negative) reported by respondents, often in answer to the question 'during the last [specified time period] has anything changed in relation to [domain of wellbeing]?' Since outcomes can also become drivers of change, we code first, second and third outcomes if required. For example, X may lead to Y_1 leading to Y_2 leading to Y_3. In this case Y_1 and Y_2 are both drivers of change and outcomes (first and second). These intermediate outcomes may also be referred to by others as outputs or results, but in QuIP studies these terms are generally avoided.

Positive and/or negative deviants. These terms refer to sub-groups of the population of intended beneficiaries of a project revealed by monitoring data to be doing significantly better/worse than is typical or average in relation to specific indicators (e.g. of wellbeing). Their identification can usefully inform purposive sample selection.

Project: A specified set of activities, interventions, or investments over a given period of time aimed at achieving a specified set of intended outcomes for a specified group of intended beneficiaries. This is the object of a specified QuIP study, and it is the responsibility of the commissioner to define it, as well as the theory of change behind it, as precisely as possible. Others may refer to the project as a 'treatment' but in QuIP studies this term is generally avoided.

Qualitative (data): Qualitative data tends to be expressed in words or pictures, and its analysis entails a more subjective process of eliciting meaning – in contrast to quantitative data in the form of numbers that are taken to express facts, and facilitate mathematical manipulation. Qualitative research generally entails delaying when data is coded and how narrowly this is constrained by prior theory and/or measurement conventions. Likewise data recording, analysis and presentation is often more open-ended and synthetic. But all research (quantitative and qualitative) entails framing, coding, decoding, and synthesis processes to some extent.

QuIP: Qualitative Impact Protocol.

Reliability: Replicability or the probability that the same conclusions would result from repeating the study. Poor application of a method by unqualified researchers undermines the reliability of evidence produced even if the design itself is valid. However, given that no two studies can ever be replicated exactly it is often very difficult in practice to distinguish problems of validity and reliability.

Respondents: These are the main source of causal claims, linking drivers of change (including but not limited to project activities) to outcomes, both intended and unintended. Respondents are usually a sample of intended beneficiaries, and data is collected from them through a mix of semi-structured interviews and focus group discussions.

Respondent count: One count per interview (individual household and focus group discussion).

Theory of change: The causal processes by which the commissioner of a QuIP study expects a specified project to achieve intended outcomes and impact. Not all causal drivers originate with the project. Theories of change also identify incidental drivers of change and may also assess the risks associated with their occurrence or non-occurrence.

Validity: The extent to which the research design can be defended from criticisms of bias or false inference. It is common to distinguish between:

- *Construct validity*. Are key concepts understood in the same way by users, researchers, and respondents, or is some of the meaning being distorted or lost?

- *Internal validity*. Are conclusions rigorous in the sense of having been logically derived from the evidence obtained and presented, subject to explicitly stated assumptions?
- *External validity*. On what basis can findings be generalized to other times and places?

Wellbeing: The quality of a person's existence can be defined in many ways, including as an emotional condition (e.g. happiness), as a state of self-knowledge (e.g. satisfaction with a life spent meaningfully), and as achievement of observable thresholds of functioning or capability (e.g. healthy; not poor). It can refer to a unified overall feeling, state or judgement, and also be broken down into components or domains. The QuIP can be used to evaluate changes in wellbeing defined in many different ways, including in response to the prior ideas and values of the commissioner and/or respondents. It is also possible to use the QuIP to explore how different people view the collective wellbeing of a community or the state of an organization (e.g. strong, independent, healthy, resilient).

Notes

1. If the primary purpose is to quantify specific causal effects then there are two options. The first is to use an appropriate experimental or quasi-experimental approach instead. The second is to build a simulation model, using both QuIP data to identify the main causal factors and quantitative monitoring data to calibrate their magnitude. The first is more precise, the second potentially more flexible.
2. To do this formally would not entail estimating statistical sampling errors but a Bayesian process of assigning confidence parameters to prior expectations and assessing how these change with each extra observation.

Reference

Bath Social and Development Research (2017) *QuIP and the Yin/Yang of Quant and Qual: How to Navigate QuIP Visualizations* [pdf], Bath, UK: BSDR <http://bathsdr.org/wp-content/uploads/2017/07/How-to-navigate-QuIP-visualisations.pdf> [accessed 17 October 2018].

About the author

Bath Social and Development Research (BSDR) has a broad remit to provide research, training, advisory, and consultancy services contributing to policies and practices promoting sustainable local, national, and global development, well-being, and social justice. The company was set up at the initiative of staff from the Centre for Development Studies (CDS) at the University Bath, and aims to conduct activities that complement and

enhance its work. Our primary aim at present is to promote and protect the QuIP approach as widely as possible. We are a non-profit organization; all revenue is reinvested in dissemination and development activities which make the QuIP more accessible to as many organizations as possible. BSDR uses the QuIP methodology and name under licence from the University of Bath. Use of the methodology without a licence is limited to non-profit purposes only. Please contact us at info@bathsdr.org if you plan to use the QuIP in your work and we will do what we can to support you.

Index

Page numbers in **bold** refer to tables and in *italics* to figures; page numbers followed by *gl* refer to the glossary.

ATTRIBUTING DEVELOPMENT IMPACT

Substantiating cause and effect is one of the great conundrums for those aiming to have a social impact, be they an NGO, social impact investment fund, or multinational corporation. All face the same quandary: how do you know whether, or how, you contributed to an observed social change? A wide range of impact evaluation methodologies exist to address this need, ranging from informal feedback loops to highly elaborate surveys. But generating useful and credible information in a timely and cost-effective way remains an elusive goal, particularly for organizations working in complex, rapidly evolving and diverse contexts.

Attributing Development Impact brings together responses to this challenge using an innovative impact evaluation approach called the Qualitative Impact Protocol (QuIP). This is a transparent, flexible and relatively simple set of guidelines for collecting, analysing and sharing feedback from intended beneficiaries about significant drivers of change in their lives. Innovative features include the use of 'blindfolded' interviewing to mitigate pro-project bias, and the application of a flexible coding system to make analysis and reporting faster and more transparent.

The QuIP has now been used in many countries, and this book uses case studies from seven countries (Ethiopia, India, Malawi, Mexico, Tanzania, Uganda and UK) assessing a range of activities, including food security, rural livelihoods, factory working conditions, medical training, community empowerment and microcredit for house improvement. It includes comprehensive 'how to' QuIP guidelines and practical insights based on these case studies into how to address the numerous methodological challenges thrown up by impact evaluation.

Essential reading for evaluation specialists within NGOs, governments and donor agencies; social impact investors; community development practitioners; and researchers and students interested in evaluation methodologies.

James Copestake is Professor of International Development at the University of Bath, and has thirty years' experience at the interface between development research, policy and practice. Marlies Morsink worked on the book whilst a Research Officer at the University of Bath, and has since joined Bath Social & Development Research as a QuIP Project Manager. Fiona Remnant helped to develop the QuIP and has spearheaded the creation of Bath Social & Development Research.

'QuIP offers a simple, transparent method to deliver timely, cost-effective and credible causal attributions.' Nancy Cartwright, University of California San Diego and Durham University, UK

PRACTICAL ACTION Publishing

BSDR Bath Social & Development Research Ltd

ISBN 978-1-78853-024-8