EBA

05

2016

PATHWAYS TO CHANGE: EVALUATING DEVELOPMENT INTERVENTIONS WITH QUALITATIVE COMPARATIVE ANALYSIS (QCA)

Barbara Befani

Pathways to Change: Evaluating development interventions with Qualitative Comparative Analysis (QCA)

Barbara Befani
Independent consultant
University of Surrey
University of East Anglia

Rapport 2016:05

till

Expertgruppen för biståndsanalys (EBA)

Barbara Befani is an independent Researcher/Consultant with a European PhD in Socio-Economic and Statistical Studies (thesis in Evaluation Methodology) and MSc in Social Statistics. Former Research Fellow at the Institute of Development Studies, she is currently a Research Fellow at the University of Sussex and Research Associate at the University of East Anglia. Barbara has over ten years of experience in evaluation methodology research and Consultancy. Her fields of expertise are set-theoretic approaches to impact evaluation (QCA and Process Tracing/Bayesian Updating); criteria to select methods in IE designs; and models for causal analysis.

This report can be downloaded free of charge at www.eba.se

ISBN 978-91-88143-16-7

Printed by Elanders Sverige AB
Stockholm 2016

Cover design by Julia Demchenko

Acknowledgements

This report would not have been possible without the leadership of Jan Pettersson and Kim Forss, and their belief that the distance between evaluators and QCA experts could and needed to be reduced. Content-wise, it benefited immensely from the feedback received by an incredibly engaged reference group composed by Laura Camfield, Rick Davies, Lennart Peck and Jos Vaessen. Renowned QCA experts Benoit Rihoux, Adrian Dusa and Alrik Thiem provided important suggestions, while Karel Chambille helped me understand the perspective of an evaluation commissioner. One of the report's strengths is that it builds on several real-life applications, which would not have been possible without the many colleagues who worked with me on those evaluations and with whom I've had precious discussions that have helped me shape many of the arguments made: Carrie Baptist, Gavin Stedman-Bryce, Jeneen Garcia, Jeremy Holland, Florian Schatz, Liz Turner and Katharina Welle. Lucie Moore was a great help in clarifying the differences between QCA and regression analysis.

The DFID Working Paper 38 (2012) was a great opportunity for me to introduce QCA to my eminent colleagues, who helped disseminate different aspects of the method: Elliot Stern, who gave it a lot of space in the report; John Mayne, who discovered INUS and asymmetric causality and started using it to qualify contribution claims; and Rick Davies, who revived his longstanding interest in configurational methods and started refining set-theoretic ideas in a way that made them more accessible to evaluators. They all played an incredibly important role in making QCA more well-known among evaluators.

If Stefano Abruzzini had not dropped a 2-page memo titled "QCA in a nutshell" on my desk back in 2003, when I was at the ULB in Brussels during my PhD, remarking "this has such a great potential for evaluation, I think it's a great scientific investment for the future", I might have discovered QCA a lot later and this report might have been written a lot later, or by someone else. That memo made me feel like I knew what I wanted to work on and led me to Gisele de Meur at ULB, to whom we owe QCA Venn diagrams, and to the leader of the Compasss network, Benoit Rihoux, who opened up the "QCA world" for me and eventually introduced me to Charles Ragin. I am also

deeply indebted to Fritz Sager and Simone Ledermann for the first substantial discussions on how to apply QCA to evaluations, at a time when nobody else was even thinking about it, let alone doing it. Last but not least, Patrick Grasso gave me the opportunity to apply the method to a synthesis of World Bank's evaluations and present the findings in a Brown Bag Lunch during my internship at OED (now IEG) in 2004. That was a one-off experience with no follow up, which made me understand and experience first-hand the complex dynamics of methodological innovation, and cultural innovation more generally.

I hold sole responsibility for any weaknesses, errors or omissions in the report, for which I apologise in advance.

Brighton, May, 2016

Barbara Befani

Table of contents

Preface

What to evaluate, when to evaluate, and how to evaluate are questions of central importance for both those who are commissioning and those who are carrying out evaluations. This EBA report addresses the last of these questions. Proper evaluation demands appropriate evaluation methods. Given the abundance of methods available, knowing when (or when not) to use a method in relation to questions posed in a specific evaluation context is often a difficult task. This is true for evaluators (who also need to know how to apply the method) as well as for purchasers (who also need to have an opinion about the usefulness of the method being proposed by evaluators).

Procurement organisations sometime lack the capacity to assess proposed methods of evaluation against the organisation's evaluation objectives. Thus, the technical advantage of tenderers becomes an informational disadvantage for purchasers. Knowing what to ask for, and when to question proposed suggestions, are important tasks when commissioning evaluations.

In this report, Barbara Befani presents one specific evaluation method, Qualitative Comparative Analysis (QCA). The report intends to provide a non-technical brief description of the method as well as a technical how-to guide. It also provides guidance on when it might be wise to consider the method in a terms of reference and how to assess the relevance of the method as well as how to quality-assure QCA evaluations.

Issuing methodology reports is not the core of EBA:s work, even though the first ever EBA report concerned the role of randomized controlled trials in evaluating aid financed activities. However, it is believed that this report does fill a gap in the literature and that it addresses an interesting combination of qualitative and quantitative analysis that increases our understanding of the sometimes complex workings of development interventions. Increasing this level of understanding is at the heart of EBA:s remit. It is my hope that this report will find its intended audience among both evaluators and commissioners of evaluations.

The author's work has been conducted in dialogue with a reference group chaired by Kim Forss, member of the EBA. However, the author is solely responsible for the content of the report.

Stockholm, June, 2016

Lars Heikensten

Lars Heikensten

Sammanfattning

Inom utvecklingsbiståndet pågår ett ständigt sökande efter nya, tillförlitliga och välanpassade metoder för utvärdering. Ett ändamålsenligt metodval förutsätter förståelse för styrkor och svagheter med olika metoder. Denna rapport bidrar till detta genom att presentera en specifik utvärderingsmetod: kvalitativ komparativ analys, QCA (för det engelska Qualitative Comparative Analysis). Rapporten innehåller en stegvis guide till hur QCA används, baserad på verkliga exempel. En diskussion förs om vilka utvärderingsfrågor som är möjliga att besvara med metoden och under vilka förutsättningar QCA fungerar särskilt bra, men också i vilka situationer den är mindre lämplig. Rapporten riktar sig både till beställare av utvärderingar och till dem som genomför utvärdering, med för varje grupp relevanta frågor om QCA.

QCA särskiljer sig från övriga metoder på ett flertal sätt. Den ***bidrar till att minska avståndet (eller avgrunden) mellan kvalitativa och kvantitativa metoder***. Genom att numeriskt koda kvalitativ data kan efter systematisk analys kausala mönster spåras vilket möjliggör prövning av kausala hypoteser utan en kontrafaktisk situation. På detta sätt kombinerar QCA informationsdjupet i kvalitativa data med kvantitativa metoders stringens och replikerbarhet. Transparens och replikerbarhet bidrar till en högre trovärdighet för analysramen (dvs. hög inre validitet) än vad som normalt kan hävdas i kvalitativa studier. Dessutom kan metoden användas för att analysera såväl ***små dataset*** (från tre fall) som större datamängder.

Kvantitativa upplägg är ofta kostnadskrävande, vilket ibland omöjliggör användande av sådana metoder. QCA kan användas som primär, ex ante bestämd utvärderingsmetod, och därmed styra insamlandet av primärdata. Ofta används dock QCA för att ***göra det bästa av redan existerande data*** och möjliggör prövning av olika konkurrerande förändringsteorier. QCA är därför ofta ***en relativt billig metod.*** Den är även ***väl anpassad för teoriutveckling*** då den ger utvärderaren en snabb indikation på lovande hypoteser och vilka som behöver omformuleras eller avfärdas.

QCA är lämplig för att ***sammanfatta resultat, t.ex. från fallstudier***, och bedöma hur generaliserbara de är. Detta skapar nya möjligheter för meta-utvärderingar, synteser och systematiska översikter, där QCA bidrar med klarhet och stringens.

Den typ av lärande som underlättas av QCA besvarar både frågor som “gjorde vi saker rätt?” och ”gjorde vi rätt saker?”. Därmed kan metoden bidra till en djupare förståelse for ***vad som fungerar bäst för olika grupper, under olika förutsättningar, i olika sammanhang***. Genom att i konstruktionen av det dataset som analyseras bevara fallens olikheter och komplexitet bidrar QCA inte med kunskap om de studerade fallens genomsnittliga egenskaper utan ger snarare en ”kontrollerad förenkling”, en sammanställning av data i vilken en begränsad uppsättning kombinerade bestämningsfaktorer (”causal packages”) kan förklara ett utfall.

QCA är ***väl lämpad för att fånga komplexa orsakssamband***: kausala bestämningsfaktorer som kan utgöra nödvändiga men inte tillräckliga villkor, eller enbart tillräckliga men inte nödvändiga villkor. Vissa faktorer är inte generellt nödvändiga eller tillräckliga villkor för ett utfall, men ändå nödvändiga för att en samling bestämningsfaktorer i en viss situation ska vara tillräckliga – på samma sätt som en insats kan ha stora och tydliga effekter i en kontext (och vara nödvändig för utfallet i detta sammanhang) men inte i andra.

Utifrån tidigare utvärderingar illustreras möjligheterna med QCA och hur metoden stegvis kan tillämpas för att utveckla vår kunskap om hur och under vilka förutsättningar biståndet uppnår resultat. I rapporten uppmärksammas också ***ett flertal fallgropar, utmaningar och begränsningar med QCA***, exempelvis behovet av jämförbara data för de fall som jämförs; behovet av teknisk kompetens i utvärderargruppen; svårigheten att uppskatta antalet iterationer som behövs för att kunna dra meningsfulla slutsatser; och behovet av att göra rimliga, verklighetsförankrade tolkningar av resultaten. Detta kan uppnås på flera sätt, bland andra genom att använda kompletterande utvärderingsansatser, som Contribution Analysis, Realist Evaluation och Process Tracing.

Eftersom möjligheten att generalisera utifrån enskilda fallstudier är ett viktigt mervärde av QCA ägnas ett avsnitt till en diskussion om de typer av generaliseringar som QCA möjliggör. I takt med att efterfrågan på QCA i utvärderingar ökar är det viktigt att förbättra, och kanske även standardisera, system för kvalitetssäkring av metoden. En ***checklista för kvalitetssäkring*** föreslås för att fullt ut utnyttja möjligheterna med QCA och för att undvika felanvändning av metoden.

Summary

In the search for new, more rigorous and more appropriate methods for development evaluation, one key task is to understand the strengths and weaknesses of a broad range of different methods. This report makes a contribution in this sense by focusing on the potential and pitfalls of Qualitative Comparative Analysis (QCA). The report aims at constituting a self-contained 8-step how to-guide to QCA, built on real-world cases. It also discusses issues of relevance for commissioners of evaluations using QCA, in particular on how to quality-assure such evaluations.

QCA stands out as capable of filling a series of important gaps. The method can drastically ***shorten the distance between qualitative and quantitative methods***, sometimes referred to as a divide. By translating qualitative data, including potential causal factors, into a numerical format and systematically analysing it, causal patterns in the data can be found, thus allowing for causal claims to be tested without the need of a counterfactual situation. As such, QCA marries the depth of qualitative information with the rigour of quantitative methods, and allows processes and findings to be replicated. Transparency and replicability gives a higher credibility to the analytical set up (i.e. high internal validity) than what is normally achieved in qualitative studies. In addition, the approach can be used to analyse both ***small sets of data*** (as small as 3 cases) as well as larger sets of cases.

Quantitative "rigorous" research designs are often expensive, and sometimes prohibitively so. While QCA can be used as a first hand evaluation method, and thus letting the evaluation questions guide primary data collection, QCA is often used to ***make the best of existing resources and data***, drawing insight from information which is already available and allowing it to support or refine a number of possible theories of change. In this sense QCA can be ***relatively cheap***, and a ***useful tool for theory development***, allowing the evaluator to quickly recognise which theories are promising and worthy of being taken forward, and which ones need fundamental changes and reformulations.

QCA has the possibility to ***synthesise case-based findings***, and assess the extent to which findings can be generalised. This creates

new possibilities for meta-evaluations, syntheses and systematic reviews, where QCA adds both conceptual clarity and rigour.

The type of learning facilitated by QCA is not simply to answer evaluation questions like "did we do things right?" but also questions of the form "did we do the right things?", allowing an understanding of ***what works best for different groups, under different circumstances, and in different contexts***. By preserving case diversity and complexity, QCA does not capture an average picture of a situation but rather a sort of "controlled simplification", a synthesis of the dataset that identifies a limited number of patterns explaining an outcome.

Finally, ***QCA is ideally suited to capture causal asymmetry***: causal factors that are – although possibly strongly and consistently associated with an outcome – only necessary but not sufficient for it, or only sufficient but not necessary. Some factors are neither, but they can still be important as necessary conditions for a causal package to be sufficient, just like an intervention can make a strong, demonstrable difference in a specific context (and be necessary for the achievement of an outcome there) but not in others (or, in other words, not necessary in general but only under specific circumstances).

Drawing on several real-life applications of QCA to evaluation, the report illustrates the opportunities offered by the method, showing how it can be applied step by step to develop, test and refine theories explaining how and under what circumstances outcomes are achieved. But it also draws attention to ***several pitfalls, challenges and limitations***: for example, the need for consistently available data across comparable cases; the need for technical skills in the evaluation team; the relative unpredictability of the number of iterations needed to achieve meaningful findings; and finally the need for sense-making of the synthesis output, which can be accomplished in many ways, including drawing on other evaluation approaches like Contribution Analysis, Realist Evaluation and Process Tracing.

Since generalisation of case-based findings is such an important added value of QCA, a specific section is devoted to the discussion of the different types of generalisation QCA facilitates. As the demand for inclusion of QCA components in evaluations increases, it is important that QCA quality assurance is improved and perhaps standardised: hence the inclusion of a proposed ***Quality Assurance checklist*** aiming to ensure that the opportunities offered by the method are caught and the pitfalls evaluators can run into are avoided.

How to read this report

This report aims to cover different aspects of QCA and increase the reader‘s knowledge at different levels: from what QCA is useful for and what its place is in the range of useful methods for development evaluation, to offering a detailed step-by-step guide on how to apply it in practice and to exploring new frontiers of significance and generalisability of the findings. Special attention to the accessibility of the report has required many technical terms normally used by QCA researchers to be defined, in the text or in the glossary; with special efforts being made to simplify the language of Chapter 1. Nonetheless, non everyday use terms are still scattered throughout the text and reading the report with the Glossary at hand is strongly recommended. Where allowed by the format, hyperlinks to the glossary, to other sections and to Annexes have been included: when first mentioned, glossary terms are highlighted as ***underlined + italics + bold + green***.

Chapter 1 is divided in two parts: the first locates QCA within a broad range of rigorous and appropriate methods for development evaluation, with a particular focus on impact evaluation. An in-depth understanding of the chapter requires a basic knowledge of a range of different methods; and the specific references to QCA are best understood when reading the second part (in particular, Section 1.3). However, even without this in-depth understanding, the case made that QCA is one of many possible options, each with different strengths and weaknesses in terms of evaluation questions answered, types of validity achieved and types of learning encouraged, should be clear. QCA has specific comparative advantages as well as comparative limitations in relation to other evaluation methods, and the section attempts to locate it rather precisely in a broader “methodological map”. It is worth noting that all the methods addressed, including QCA, can (and sometimes need to) be combined in the same evaluation design, complementing each other’s strength and weaknesses.

Part two of Chapter 1 introduces the potential and relevance of QCA for development evaluation, briefly demonstrating how it can be used to develop and test a programme theory and what kind of recommendations it can inform. The chapter includes a section with practical information for commissioners on what to expect when applying QCA to a real-life evaluation: requirements, limitations and the procurement process.

Chapter 2 is a “how-to”, practical guide to the application of QCA to development evaluation, proposing a step-by-step approach with challenges, opportunities, pitfalls and other “issues at stake” identified for each of the eight steps.

Chapter 3 addresses some of the major critical issues, like bias and generalisation. The chapter also provides a checklist for quality assuring QCA components of evaluations and illustrates how QCA can be combined with explanatory evaluation approaches, like Contribution Analysis, Realist Evaluation and Process Tracing.

The report comes with three Annexes. Annex A, an in-depth discussion of causal frameworks underpinning scientific inference and specifically Impact Evaluation, is important to understand how QCA contributes to the formulation and testing of causal claims, and helps relate QCA to other evaluation methods using different frameworks. Annex B deals with the differences between QCA and regression analysis, which evaluators with a quantitative background are usually interested in discovering. Finally, Annex C includes information and data about the evaluations used as case examples throughout the report.

1 The search for appropriate rigorous methods and the potential of QCA

While evaluation research has always aspired to scientific credibility, efforts to strengthen the rigour and robustness of evaluative evidence have received a strong push in the last 15 years (Savedoff, Levine, & Birdsall, 2006; Duflo & Kremer, 2003; Duflo, Glennerster, & Kremer, 2006; White, 2013; Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011; Treasury, 2011; Leeuw & Vaessen, 2009). Before 2012, when a major study commissioned by DFID on "broadening the range of designs and methods for impact evaluations" was published (Stern, Stame, Mayne, Forss, Davies, & Befani, 2012), the vast majority of these efforts[1] focused on promoting randomised controlled trials (RCTs) and quasi-experiments, the latter considered a second-best option compared to the former. From an epistemological standpoint, the notion of rigour was seen as a prerogative of a) quantitative methods and b) *counterfactual* thinking[2].

Quantitative methods were considered inherently superior to *qualitative methods* and occupied a higher place in the hierarchy of methods ranked by rigour: mixed methods followed quantitative experiments or quasi-experiments, and were followed by qualitative methods. Counterfactual thinking rose to prominence because reconstructing a counterfactual, no-*intervention* situation – and comparing this to the observable, post-intervention one – was considered the only strategy that could lead to a credible demonstration of causality between the intervention and observed *outcomes* of interest.

These ideas are still widespread; however, the seminal DFID Working Paper 38 of 2012 (Stern, Stame, Mayne, Forss, Davies, & Befani, 2012), a synthesis of which has been recently produced for managers and commissioners of evaluations (Stern, 2015), seems to

[1] With the exception of the NONIE 2 guide: (NONIESubgroup2, 2008) and the study "Voices for the Poor" funded by the World Bank in 1999 (Shah & Narayan, 1999).

[2] By "counterfactual thinking" here we not only mean reconstructing a non-observable situation, but also the specific non-observable situation of what would have happened without the intervention, informed by Mill's Method of Difference.

have started a new era where the dominance of quantitative and counterfactual methods is being increasingly eroded by a growing interest in qualitative methods and in alternative, non-counterfactual approaches to the demonstration of ***causal linkages***.

Building on the Stern paper and subsequent work (Befani B. , 2016), this chapter begins with an attempt to take a snapshot of the current methodological situation in development evaluation (with a particular focus on ***impact evaluation***), in terms of what methods are needed and available; and later presents the basic features of QCA, discussing its feasibility and relevance for development evaluation.

1.1 Why we need methods (including QCA) in the first place

The sometimes frantic search for rigorous methods could perhaps benefit from going back to basics and asking why we need development evaluation methods[3] in the first place. Measuring the effectiveness and relevance of policies, improving projects and being accountable to donors and tax payers is a business which can no longer be run using anecdotal evidence or methods that don't maximise learning or ensure empowering experiences for the stakeholders involved: education levels are rising, democratic processes are more and more widespread, and citizens as well as donors are becoming more and more demanding of the organisations they fund, expecting results or at least maximum effort towards achieving outcomes. In this context, scientific research methodology offers:

1. Widely recognised principles, as well as tried and tested ways to answer specific types of questions (including evaluation questions) (see Section 1.1.1)
2. Widely recognised principles, as well as tried and tested methods, to maximise the validity of findings, including evaluation findings and answers to evaluation questions (in particular internal, external and ***construct validity***: see Section 1.1.2)

[3] The discussion in this report is largely limited to "methods" as different from "data collection techniques", like surveys, interviews, or desk reviews. This is because each method can draw on multiple data collection techniques, and the same technique can be useful for multiple methods: in other words they deserve two separate discussions. QCA is a method rather than a data collection technique, so it makes much more sense to compare it to methods rather than techniques.

3. Widely recognised principles, as well as tried and tested ways to demonstrate causality: including assessing the contribution of interventions to given outcomes (see also Annex A and Section 1.1.3).

Other methodologies more specific to evaluation and in particular development evaluation focus on ensuring that stakeholders are involved in evaluation processes and eventually enough empowered to "own" the findings (Burns, 2014).

1.1.1 Answering Evaluation Questions

The first advantage of using scientific research methods is that they enable us to answer (as rigorously as possible) **evaluation questions**, and methodological choice should be influenced, first and foremost, by the type of question we are interested in answering: *different questions require different methods* (see Table 1). This is relevant for us because we will see in subsequent chapters that *QCA is much more appropriate to answer some questions than others*.

For example, commissioners of development impact evaluations are interested in understanding **what difference the intervention under study made, if any**, preferably in relation to a series of outcomes of interest. This broad, overarching question takes different forms when made more specific:

1. How much of a difference did the intervention make, on average?
2. For whom and under what circumstances did the intervention make a difference?
3. How and why did the intervention make a difference?
4. Will the intervention make a difference elsewhere/in the future?
5. What difference is relevant? For whom?

Both the DFID paper (Stern, Stame, Mayne, Forss, Davies, & Befani, 2012) and ongoing work (Befani B. , 2016) link these questions with specific groups of methods that are more or less appropriate for answering each of the questions above. We will see in the rest of the report that *QCA is not appropriate to answer question #1 and needs to be combined with other methods to answer question #3*.

1.1.2 Ensuring Validity of the Findings

Although the benefits of applying scientific research methods are sometimes overestimated (Forss, 2007), these include highly desirable properties of the findings, like *internal validity*, *external validity*, and *construct validity*. In general, it's important that studies can be replicated and the findings considered reliable.

A methodological approach with high **internal validity** protects the research process from selection bias in human resources, informants, sampling, and documentary sources. Researchers with specific skills or knowledge can first influence the range of possible findings that are considered likely (and hence are actually tested); by considering (and testing, implicitly or explicitly) different sets of hypotheses. Secondly, choice of human resources influences the way data is collected, processed and interpreted; so that different groups of researchers might select different samples, different groups of key informants, and different sets of documentary evidence; affecting the reliability and "objectivity" of the findings. An internally valid research process is such that the findings are not likely to differ if other researchers replicate the study following the same protocol. As we will see in the rest of the report, *the process of applying QCA can be fully transparent, with clear standards and parameters; its synthesis procedures are perfectly replicable*.

Aiming for **external validity** ensures that the findings are to some extent generalisable, and the knowledge we acquire is relevant because it concerns a relatively large slice of reality. *QCA operates a form of generalisation called "contingent", "modest", or "limited generalisation", synthesising information in a dataset; however, it can also, in some cases, aspire to wider generalisation* (see Section 3.1).

Findings that are both externally and internally valid, for example robust knowledge about the value of an indicator over a large population, might still present limited utility unless that indicator is a good proxy for the phenomenon we are trying to describe; in other words, unless the findings present high **construct validity**. *QCA is unable to ensure this type of validity by itself and needs intense dialogue with theory and substantive knowledge, including the one embodied by stakeholders*. For this purpose, it might need to be combined with other methods (see Section 3.2).

1.1.3 Increasing our confidence that the intervention has caused the outcome

Methods help us discover or confirm *causal connections* between the intervention and its effects (see Annex A). Contrary to what the proponents of experimental methods often state or imply, reconstructing the counterfactual situation is *not* the only way to demonstrate causal connections. Methods can draw on a plurality of causal inference *models* or frameworks in order to infer causality from the intervention to the outcomes (Befani B. , 2012). The general causal question "is the intervention causally related to the outcome" can be made specific in a few different ways:

- Can we measure the net effect of the intervention?
  - More specifically: can the influence of all plausible causes except one *(e.g. the intervention)* be isolated so that we can attribute the marginal (net) effect to that one cause?
- How often is the cause *(e.g. the intervention)* observed together with the effect (*or outcome*)?
- Does the effect (*or outcome*) decrease or increase as the cause *(e.g. the intervention)* increases or decreases?
- What role does the cause *(e.g. the intervention)* play in producing the effect (*or outcome*)?
- What explains the effect (*or outcome*)? How and why does the cause (*e.g. the intervention*) produce it?
- Finally, is the cause (*e.g. the intervention*) satisfying various notions of *necessity* and *sufficiency* for causality?[4] *QCA enjoys a unique comparative advantage in answering this question*: it will be addressed in the rest of the report, but in a nutshell, the question refers to *whether an intervention, or other factors* like for example specific contextual or historical conditions, *are required to achieve a certain outcome* or if the latter can be achieved without them; and to *whether the intervention is good enough by itself to produce the outcome* or needs other factors and components.

All the above are legitimate impact questions in that they illuminate or reflect on specific aspects of the causal link between the intervention and the outcome.

[4] The notions of necessary, sufficient, and the terms INUS and SUIN causality are clarified in the next chapters and in the Glossary.

1.1.4 Ensuring ownership and empowerment of stakeholders

Other properties of methods which are relevant for development evaluation are not necessarily included under the "scientific research methods" group, but nonetheless are important to ensure that relevant "political goals" are met, like for example empowering stakeholders and ensuring that they influence the evaluation process. These methods aim at taking into account the perspectives of a broad range of groups, in particular the most vulnerable, and ensure that the evaluation contributes to **different types of learning**, in the form of single-loop, double-loop and triple-loop learning (Hummelbrunner, 2015).

The calibration phase in QCA (see Section 2.3) can serve this aim, in that it *encourages evaluation teams and stakeholders to make their value judgements* (and "opinions" about what makes a difference) *explicit* and open to discussion.

1.2 What methods are appropriate?

This section addresses in more detail the abilities of methods to answer questions, ensure (different types of) validity, demonstrate causal relations, and empower stakeholders. The arguments are based on the literature cited and the author's experience and ongoing work (Befani B. , 2016). The aim of the section is to show that methodological choice in development evaluation is complex and QCA can fill some of the gaps identified, but not all.

No single method is a "supermethod": it cannot at the same time answer all types of evaluation questions, be strong on all types of validity, illuminate or reflect on all aspects of causal relations, and empower stakeholders. However, in order to be considered suitable for an evaluation, methods need to satisfy the following requirements:

1. Answering at least one type of evaluation question;
2. Guaranteeing at least one type of validity (the more, the better); and
3. Encouraging at least one of many possible types of learning while "declaring" which perspectives are taken into account

In addition, in order to be considered adequate for an impact evaluation, methods also need to answer causal questions. Annex A presents a detailed discussion of causality and strategies to answer specific causal questions.

1.2.1 A tentative alignment between characteristics of impact evaluations

The five questions in Section 1.1.1 are all related to the overarching impact evaluation question "did the intervention make a difference?". For example, if we are able to answer the subquestion "how did the intervention make a difference?" this automatically implies that it has, indeed, made a difference. If we measure the magnitude of the effect and answer "how much of a difference has the intervention made" or "how large is the net effect of the intervention" we automatically know if the intervention made a difference or not. If we answer "what difference did it make for whom, under what circumstances?" and discover that an intervention has made a difference in a specific context and for a specific group, we automatically provide the more general answer that is has, indeed, made some difference (somewhere).

Knowing that an intervention and/or other factors are *necessary* or *sufficient* for the outcome helps us understand their role in different contexts and can be considered a variant of the "what difference for whom" question (see Table 1). The latter and the how/why question also inform the future-oriented "will the intervention make a difference (in the future, in other contexts)". Finally, knowing what difference is relevant for whom ensures that we are not wasting our time focusing on irrelevant differences.

A number of methods are suggested below that can answer the above questions, with references to the underlying *causal frameworks*, types of validity and types of learning. An attempt is made to locate QCA in the group as precisely as possible.

1.2.1.1 How much of a difference did the intervention make, on average?

This question can also be formulated as "how large was the average net effect of the intervention". It can be answered with methods aimed at measuring the average net effect of the intervention, like ***experiments***, ***quasi-experiments***, and ***statistical modelling*** (Duflo, Glennerster, &

Kremer, 2006; Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011; Leeuw & Vaessen, 2009), which are underpinned by Mill's Methods (see Annex A). In particular, experiments and quasi-experiments are based on Mill's Method of Difference, and the specific causal question underlying this type of investigation is "can the influence of all plausible causes except the intervention be isolated so that we can attribute the marginal (net) effect to the intervention?".

These methods tend to perform well on internal validity – if steps are taken to protect the process from the known threats to internal validity (Campbell & Stanley, 1963); while construct validity and external validity are not guaranteed unless other methods are employed in combination. For example, barely knowing that an intervention has worked somewhere does not guarantee that it will keep working in the future and that it will work elsewhere, unless assumptions are made on why it worked and that the conditions or supporting factors (Cartwright, 2012) that have allowed it to work will be in place elsewhere. These methods also usually use quantitative indicators or ***variables*** which might not faithfully represent the constructs they are meant to. Finally, the type of learning involved is single-loop, aiming to understand if a specific goal has been achieved, with the question answered being "did we do things right? (Hummelbrunner, 2015)?

1.2.1.2 How and why did the intervention make a difference?

This question can be answered by those methods aimed at either direct observation of the mechanism at work or at reconstructing the latter, seeking evidence of its existence. Some examples of these methods, all based on generative frameworks, are: ***Contribution Analysis*** (Mayne, Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly, 2001), ***Realist Evaluation*** (Pawson & Tilley, Realistic Evaluation, 1997), ***Process Tracing*** (Bennett & Checkel, 2014; Beach & Pedersen, 2013; Befani & Stedman-Bryce, 2016) and a wide range of ***Systems-Based evaluation*** approaches (Williams & Hummelbrunner, 2010; Befani, Ramalingam, & Stern, 2015). The specific causal question answered by these methods is "what explains the outcome? How and why does the intervention produce the outcome?"

Here it is usually not possible to isolate the intervention from other contributing factors when describing the usually complex or complicated mechanism responsible for the outcome, and the aim of the method is to describe how the various factors are interrelated within the complex, sometimes "systemic" mechanism. Because of the depth and level of detail with which they describe the mechanism responsible for the outcome, these methods tend to perform well on construct validity; but at the same time the number of cases their findings apply to can be limited, hence they are known mostly as "*within-case methods*". They offer limited external validity unless the total number of possible cases is limited, like in systemic representations of reality or in some applications of ***Realist Synthesis*** (Pawson, 2006); and because of their mostly qualitative nature they are usually weaker than others on internal validity. In terms of learning, these methods lend themselves well to double-loop learning, answering questions of the "did we do the right things" kind; and if multiple perspectives are made transparent or taken into account, even triple-loop learning, answering "who decides what things are right" (Hummelbrunner, 2015).

1.2.1.3 What difference did the intervention make, for whom and under what circumstances?

This question does not focus on the average difference the intervention makes but rather on the diverse performance of the intervention in different settings, trying to understand which combinations of factors, including the intervention or not, worked better under different circumstances or for different groups. The specific causal questions "was the intervention *necessary*, *sufficient*, *INUS* or *SUIN* for the outcome" (terms to be defined in the following and in the glossary) and "what role does the intervention play in producing the outcome" can be answered with methods for synthesis and systematic *cross-case* comparison, like ***Qualitative Comparative Analysis (QCA)*** (Rihoux, Ragin, & (eds), Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques, 2009; Schneider & Wagemann, Set-Theoretic Methods for the Social Sciences, 2012) and ***Realist Synthesis (RS)*** (Pawson, 2006; Pawson & Tilley, 1997), used alone or in combination (Befani, Ledermann, & Sager, 2007). These methods have good potential to be strong on external, internal and construct validity. The external validity advantages are based on the ability to generalise *case-based* findings, adding leverage for "thick" cross-case comparison (Rihoux & Lobe, 2009) (see Section 3.1 for more details)[5]. Construct validity is found in the depth offered by the realist approach; and internal validity in the algorithmic procedures used in QCA to synthesize case-based findings (see Chapter 2).

Some differences between realist synthesis and QCA are illustrated in Table 1: the former is underpinned by generative causation while the latter by multiple-*conjunctural* (or *configurational*) causality (see Annex A); QCA tends to be stronger on internal validity than RS because many of its procedures are perfectly replicable, and has also higher potential in terms of external validity because it can synthesize information covering a high number of cases. By contrast, RS can be stronger than QCA on construct validity thanks to potentially richer descriptions of contexts and mechanisms. These methods contribute to double-loop learning in that they aim at understanding the best way to achieve an outcome under specific circumstances rather than simply checking whether the outcome has been achieved or not (Hummelbrunner, 2015). QCA in particular can also make a small contribution to triple-loop learning, if the calibration phase (see Section 2.3) is seen as an attempt on behalf of the stakeholders and the evaluation team to define "right": for example, defining "success", "failure", and intermediate degrees of achievement of an intervention.

[5] An argument can potentially be made that, since QCA can return different solutions depending on the conditions included in the model, calibration rubrics, cases included in the sample, and other parameters, this makes it weak on external validity. However, the fact that the findings depend on the initial assumptions and can potentially change if data is measured differently and if different variables are included, is true for all methods equally. It can be argued that *large-n* methods can account for random error and their findings are insensitive to the addition of cases and as such perform better on external validity; but this is true of QCA as well when it is applied to a large sample: inclusion thresholds based on frequency are meant to protect against random measurement errors, and make the findings robust with regard to the addition of cases. QCA has been described as an "accordion" because it can handle large-n samples and maintain several properties of variable-based methods, but is also designed to be flexible enough to extract as much "rich information" as possible when comparing cases in smaller samples (*medium-n* or small-n) (Vis, 2012). In any case this transparency and flexibility do not diminish its external validity potential compared to other methods, especially considering that it can take diversity into account as well as frequency.
Another argument against the external validity of QCA can potentially be made, stressing that when stricter consistency thresholds are used, the solution of the Boolean minimisation covers a lower number of cases (see Chapter 2 and Section 2.7). Rather than being an external validity problem, this possibility shows the flexibility of QCA as a tool that – depending on the chosen parameters – can either be externally valid and cover a high number of cases, or accurately represent a smaller subset of complex sufficiency statements, maximising construct validity.

1.2.1.4 What difference is relevant for whom?

Even if we know everything about the difference an intervention has made, where, how much, why and for whom, this difference can still be irrelevant if the perspectives of stakeholders about "what differences matter" are not taken on board. In order not to waste our time, it's important to answer the "what difference is relevant for whom" question. This question should be answered before the causal analysis starts, and produce a selection of outcomes on which the latter will focus on. If it starts after the causal analysis is completed, it can still be helpful to assess the relevance of the findings.

Among the methods answering this question are the *Most Significant Change* (Davies & Dart, 2005), *Outcome Mapping* (Earl, Carden, & Smutylo, 2001) and a variety of *Systemic Approaches* (Williams, 2015; Williams & Hummelbrunner, 2010) like for example Soft Systems Methodology (Checkland & Scholes, 1999). These methods usually ensure that the constructs being measured or assessed are meaningful for stakeholders and can thus be predicted to be relatively strong on construct validity. External validity will depend on the ratio between the number of cases covered and the total number of possible cases (in systemic approaches it might be only one). In comparison with other methods these might not offer strong protection against internal validity-related bias, but in terms of learning they are strongly driven towards both double-loop and triple-loop learning.

Table 1 shows the comparative strengths and weaknesses of QCA with regard to other potentially useful methods. Sections 1.3 and 1.4 will make a stronger case of why QCA is an important addition to the development evaluator's toolkit. In the meantime, it's important to remember that – while enjoying specific comparative advantages compared to other methods – QCA doesn't need to be used on its own: on the contrary, it can easily be combined with methods complementing its weaknesses. For example, ***variable-based*** methods when fine-grained correlations or net effects need to be measured, or explanatory or interpretative approaches to develop its theoretical basis and interpret the findings, improving construct validity (see Section 3.2).

Table 1: A tentative alignment between questions, causal frameworks, methods, validity and types of learning

Overarching Impact Question	Did the intervention make a difference?			Is the difference relevant? For whom?
Specific Impact Question	*How much of a difference (on average)?*	*For whom, under what circumstances?*	*How/why so?*	
Specific Causal Questions	- Can the influence of all plausible causes except the intervention be isolated so that we can attribute the marginal (net) effect to the intervention? - What is the net effect of other factors?	- Was the intervention necessary, sufficient, INUS or SUIN for the outcome? - What role did the intervention play in producing the outcome?	- How and why does the intervention produce the outcome? - What explains the outcome? - What role does the intervention play in producing the outcome?	n.a.
Causal Frameworks	Single-cause frameworks measuring the average net effect (Method of Difference, Method of Concomitant Variation)	Multiple-conjunctural causality frameworks for synthesis and systematic comparison (MCC)	Generative/ Mechanism-based frameworks for in-depth explanation	n.a.
Methods	*Experiments, quasi-experiments, statistical modelling*	*QCA, Realist Synthesis*	*Contribution Analysis, Realist Evaluation, Process Tracing, Systemic Approaches*	*Most Significant Change, Outcome Mapping, Systemic Approaches*
Prevalent Type of Validity	Internal (Es, QEs), External (SM)	External, Construct (RS), Internal (QCA)	Construct	Construct
Prevalent Type of Learning	Single-loop	Double-loop	Double-loop	Triple-loop

Note: Please refer to the Glossary and Annex A for explanations of the technical terms

1.3 QCA in a nutshell

Qualitative Comparative Analysis is a method for systematic cross-case comparison that was first introduced by Charles Ragin in 1987 (Ragin, 1987) to understand which qualitative factors are likely to influence an outcome. It has, since then, undergone several developments (Ragin, 2000; Ragin, 2008; Rihoux, Ragin, & (eds), 2009; Schneider & Wagemann, 2012; Caren & Panofsky, 2005), increasing the interest of social scientists and philosophers in the synthesis of ***Boolean*** datasets (Baumgartner, 2012). Despite its name and despite being a case-based method, QCA is not always considered "qualitative", particularly in the academic traditions of some latin cultures which translate it "Quali-Quantitative Comparative Analysis" (Meur, Rihoux, & Yamasaki, 2002) because of its mathematical basis.

Compared to other case-based methods, QCA's selling point is its ability to compare case-based information systematically, leading to a replicable (rigorous) generalisation of case-specific findings, which is normally considered an advantage of quantitative/variable-based/ statistical methods. Compared to the latter group of methods, however, QCA does not require a large number of cases in order to be applied (although it can handle it); and retains some the "thickness", richness or complexity of case-based in-depth information (Berg-Schlosser, De Meur, Rihoux, & Ragin, 2009; Befani B. , 2013).

Because of these abilities at the crossroad of two methodological cultures (Goertz & Mahoney, 2012), QCA has been said to incorporate the "best of both worlds" (Vis, 2012; Befani B. , 2013). Historically, the method has always been very popular with political scientists and other scholars interested in cross-country generalisation[6].

At its core, QCA requires conceptualising *cases* (for example projects, or groups of projects within countries) as *combinations* or "***packages***" of characteristics that are suspected to causally influence an *outcome*. For example, the availability of spare parts and adequately trained manpower are assumed to influence the chance that broken water points are repaired (Welle, Williams, Pearce, & Befani, 2015). These characteristics of the "case" are called "*conditions*" rather than

[6] A classic, proto-application of the method before it was even "codified" by Ragin in 1987 is included in (Rokkan, 1970). For an updated database of QCA applications, see the website compasss.org

"variables" to emphasise the distinction between QCA and statistics (see Annex B).

Once the characteristics of the cases are known, together with their outcomes, a systematic cross-case comparison is carried out to check which factors are consistently associated with a certain type of outcome (e.g. success of the intervention) and can potentially be considered causally responsible for it[7]. This allows for a potentially quick, simultaneous testing of multiple theories of change.

In the basic version of QCA (called *crisp-set QCA*), both the conditions describing the case and the outcome are defined in terms of "presence" or "absence" of given characteristics across a set of cases: the analysis will reveal which conditions are needed and which ones are most effective for the outcome to occur.

In order to compare the cases systematically, in a way that can be replicated and can be completed quickly – automatically, by software programmes, even over large datasets – presence of conditions is usually denoted with "1" while absence with "0".[8] Both presence and absence need to be defined in order to assign the values of zero and one to the cases. When this process (known as "calibration", see Section 2.3) is completed the result is a matrix of zeros and ones indicating the presence or absence of a series of conditions (represented as columns) over a set of cases (represented as rows), see Table 2.

Table 2: Conditions influencing repairs of broken water points

Project	FUNDSF	RESPCL	SPAREP	REPAIR
Smart Handpumps Kenya	1	1	1	1
M4W Uganda	0	1	0	0
Maji Matone Tanzania	0	0	1	1
Maji Voice Nairobi	1	1	1	1
Next Drop Bangalore	1	1	1	1
Human Sensor Web Zanzibar	0	1	1	0

Table 2 illustrates the characteristics of a series of projects aimed at encouraging the use of ICT to report water point failures, as a way to

[7] The causal inference models used are variants of Mill's Methods: for more details, see Annex A.

[8] This is the case of crisp-set QCA; fuzzy-set QCA allows the assignment of values between 0 and 1, in addition to these.

increase the chances that broken water points are repaired (Welle, Williams, Pearce, & Befani, 2015). The outcome is represented by the last column and denoted as REPAIR: 1 broadly means "presence of repairs" or that the project was a success in that repairs were being made on the basis of ICT reports, while 0 means "absence of repairs" or that the project was not a success in this sense (for a more precise definition of success in this model, see Section 2.3 and Annex C). The table shows that four projects have been successful (Smart Handpumps, Maji Matone, Maji Voice, Next Drop) while two haven't (M4W and Human Sensor Web).

The first column (FUNDSF) shows whether funds are considered sufficient for carrying out the repairs: this is the case in the three projects denoted with 1 or "presence of sufficient funds": Smart Handpumps, Maji Voice and Next Drop. The second column (RESPCL) shows whether Operations & Monitoring responsibilities were clear: which was the case in all projects except one, indicated with 0 or "absence of clear responsibilities" (Maji Matone). Finally, the column SPAREP indicates the availability of spare parts for the repair, which was always registered except in the M4W project.

1.3.1 Evaluation questions answered by QCA

Once the case-specific information has been represented as matrix of zeros or ones (or Boolean matrix, similar to the one illustrated above), the dataset is synthesised through the systematic comparison of the cases on the characteristics embodied by the conditions. The broad, overarching question answered by QCA is "what sets of factors are likely to influence an outcome"? Four different procedures can be used (see details in Chapter 3) to gather empirical support for three related evaluation questions:

1. What causal factors are needed for the outcome to occur?

   - More technically: which causal factors are necessary for the outcome to occur?

2. What causal factors are most effective (alone or combination) for the outcome?

   - More technically: which causal factors are sufficient (alone or in combination) for the outcome to occur?

3. What causal factors make the difference for the outcome, under what circumstances?

The first question asks if there are any factors which are absolutely or normally required (necessary) for the outcome to occur, on the basis of the available data and knowledge. The second if any factors "guarantee" or dramatically increase the chances of the outcome materialising, alone or in combination, even if they are not normally required (they are "sufficient" for the outcome). Finally, the third question identifies "special", stand-out factors that appear to make the difference between a positive and a negative outcome, in specific contexts.

1.3.1.1 The necessity analysis

The typical answer to question one will be a list of conditions, or ***disjunction*** of conditions. [9] For example, in the case presented above, SPAREP is a necessary condition for a positive outcome: it is observed in all four cases where a positive outcome is present. This allows the evaluator to develop the *hypothesis that repairs are not carried out unless spare parts are available*, which could be further tested in additional cases. Note that the necessity analysis in QCA follows a similar logic to Mill's Method of Agreement: grouping the cases presenting the outcome (or the effect) and checking for regularly present factors (see Annex A for more details).

The following table illustrates this necessity analysis in more detail. Note that only one condition is perfectly necessary while the other two are only "75% necessary"(!). This idea of "**imperfect necessity**" can be thought as the presence of the condition being needed more than its absence: a necessity score of 50% indicates that, according to the data analysed, the presence of the condition is needed exactly as much as its absence. If the odds are calculated, we can say that a 75% necessity score means that the presence of the condition is needed three times more than its absence. Extending the concept even further,

[9] A disjunction is logically known as the "union" of a group of conditions, which requires only one of those to be present in order for the union/disjunction to be present. This is in contrast with a combination, conjunction or intersection of conditions, which requires all conditions to be present in order for the combination/conjunction/intersection to be present.

we can say that it's three times "more useful for the outcome" than its absence.

Back to our example illustrated in Table 3, we can make the following statements:

1. Spare parts are necessary to carry out the repairs
2. Clarity in Operations & Monitoring Responsibilities is three times more needed (useful) for repairs than lack of clarity
3. Sufficiency of funds is three times more needed (useful) for repairs than insufficiency of funds

Table 3: Findings from the Necessity Analysis

Condition	# successful cases where it is observed	Odds	Comments
FUNDSF	3/4 (75%)	3 to 1: presence needed 3 times more than absence	One case (MM) is successful with no sufficient funds.
RESPCL	3/4 (75%)	3 to 1: presence needed 3 times more than absence	One case (MM) is successful but responsibilities unclear
SPAREP	4/4 (100%)	-	In all 4 successful cases spare parts are available

When no single condition is perfectly necessary, we can enquire further in two ways. One is calculating the necessity score, estimating how much the presence of the condition is needed compared to its absence. Another is to check whether the condition is part of a group, one condition of which is always present when the outcome is present. In the latter case the condition will be a SUIN cause (Befani B. , 2013) or part of a necessary ***disjunction*** (see Annex A). This means that, even though the condition is not strictly required for success, it's part of a group of "functionally equivalent" conditions which represent "equivalent requirements" for the outcome: no single condition of the group is needed, but at least one of the group must materialise, in order for the outcome to be achieved.

1.3.1.2 The sufficiency analysis

The typical answer to question two (what causal factors are effective/sufficient for the outcome) will be a series of "equivalent" combinations[10]. These pathways are equivalent because they all lead to the outcome even though they are qualitatively different (this is also known as *equifinality*, see Schneider & Wagemann, 2012).

In the example above, the presence of all three conditions, denoted with FUNDSF*RESPCL*SPAREP[11] is effective: it leads to (is "sufficient" for) a positive outcome. Whenever we observe this combination (in all three cases), the outcome is positive. This regularity allows the evaluator to develop the *hypothesis that the simultaneous presence of sufficient funds, clear responsibilities and available spare parts guarantees that repairs are made*. This hypothesis can be further tested on additional cases.

Note that the above combination identified as sufficient is not necessary: the case Maji Matone is successful but does not present it. Repairs are made there despite insufficiency of funds and lack of clarity on O&M responsibilities. The sufficiency analysis returns two effective (successful, sufficient) pathways: the first (FUNDSF* RESPCL*SPAREP) covering three cases and the second (fundsf* respcl*SPAREP) covering just one (see Table 4).

Table 4: (Sufficient) Pathways to success: findings from the sufficiency analysis

Combination	Number of successful cases covered
FUNDSF*RESPCL*SPAREP	3
fundsf*respcl*SPAREP	1

1.3.1.3 The INUS analysis

Finally, the typical answer to question three (what causal factors make the difference for the outcome, under what circumstances) will be the identification of two almost identical cases, one with a positive and the other with a negative outcome, that differ only in one condition. This will allow the evaluator to make *the hypothesis that this different condition is what makes the difference to the outcome in that specific context*.

[10] Technically, a disjunction of combinations.

[11] In QCA notation, the star sign (*) means "logical intersection" or "AND", while "logical union" is indicated with the plus sign (+). Note that presence of the condition is usually denoted with upper case, while absence with lower case.

For example, if we compare any of the three cases covered by the first combination above (FUNDSF*RESPCL*SPAREP) with the project Human Sensor Web, which can be represented with fundsf*RESPCL*SPAREP, we notice that the two combinations only differ in sufficiency of funds (FUNDSF: present in the three cases and absent in Human Sensor Web); the only other difference is in the outcome. This can be interpreted as follows: where funds are sufficient, repairs are made, while where funds aren't sufficient, repairs aren't made; funds are what makes the difference between success and failure.

However, this is not a general regularity: it only happens in the context of RESPCL*SPAREP, or in other words when spare parts are available and responsibilities clear. In Maji Matone, where responsibilities are unclear, funds are not sufficient and yet repairs are made: this case is clearly different from the other three and warrants a different explanation for its success.

Yet, the data allow the evaluator to develop the hypothesis that, when responsibilities are clear and spare parts available, sufficiency of funds makes the difference between repairs being made or not. This special condition (sufficiency of funds) is known as an "INUS cause" (see Annex A) because in itself is neither needed (necessary) nor effective (sufficient) in an absolute sense; and yet it is needed (necessary) for the combination FUNDSF*RESPCL*SPAREP to be effective. If the presence of FUNDSF is removed from the combination and replaced with its absence fundsf, the package loses its effectiveness (sufficiency). In other words, the condition is only needed (necessary) for success in the context of "RESPCL*SPAREP".

The following tables illustrate the findings of the INUS analysis conducted on the above dataset. Note that the INUS analysis is based on Mill's Method of Difference; but instead of eliminating the identical factors as "redundant" it makes them an integral part of the "*causal package*" explaining the outcome (see Annex A for more details).

Table 5: Hypothesis from the INUS analysis: sufficiency of funds makes the difference when spare parts are available and responsibilities clear

Project	FUNDSF	RESPCL	SPAREP	REPAIR
SHP, MV, ND	1	1	1	1
HSW	0	1	1	0

Table 6: Hypothesis from the INUS analysis: sufficiency of funds makes the difference when spare parts are available and responsibilities clear

Combination	OUTCOME
FUNDSF*RESPCL*SPAREP	1, REPAIR (repairs are made)
fundsf*RESPCL*SPAREP	0, repair (repairs are not made)

1.3.2 Developing and testing programme theories with QCA

Although it can potentially be used “theory-free”, QCA is normally used to develop theories from case-based knowledge or to test theories on empirical cases: it can be used for both exploratory and confirmatory purposes. This section reports a stylised example of how a programme theory could be tested, articulated and refined on the basis of an empirical boolean dataset. The aim is to give readers a more articulate idea (compared to the introductory example above) of how QCA can contribute to the development of complex theories of change in development evaluation, and what lessons can be learned from the process (including in terms of recommendations).

Let‘s consider the evaluation of a policy influence programme implemented in eight countries to improve evidence-based policymaking in the health sector, to ultimately improve access to the health system for the poorest segments of the population[12]. The basic idea behind the programme is that improving transparency and accountability in the sector (improving data collection and dissemination and strengthening multi-stakeholder platforms at the national level) will lead to the development of evidence-based policy making in the health system.

The theory for this example is relatively well-developed and QCA is initially used for confirmatory purposes: one of the most influential theories of policy process development, Kingdon’s agenda-setting theory (Kingdon, 2010), is used. Kingdon identifies three process ‘streams’ that influence the setting of policy agendas and the framing of policy options. These are:

1. Problems (PROB): the way socio-economic conditions are framed as undesired or problematic;

[12] This fictitious example is freely inspired by the evaluation of the Medicines Transparency Alliance (MeTA) (Stedman-Bryce, Schatz, Hodgkin, & Balogun, 2016)

2. Policies (SOL): the solutions generated to address problems, constrained by technical feasibility, compatibility with prevailing values, etc.

3. Politics (POL): political factors, e.g. the power/influence of interest groups, other urgencies and burning issues, elections.

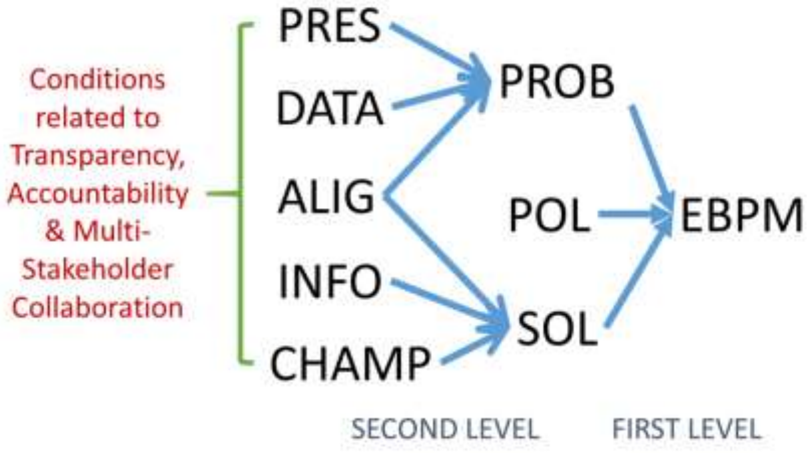
The agenda-setting theory proposes that policy issues are more likely to be addressed by policymakers when at least two of the above streams converge to create a ‘policy window’. For example the removal of technical constraints to a solution coupled with the election of a champion. Part of the programme theory can thus be reformulated as follows (see Figure 1 below):

Evidence-Based policies addressing access to the health system for the poor are put in place when at least two of the following conditions are met: a) the lack of accessibility is unanimously framed as undesired and problematic; b) the solutions proposed in terms of policy options are feasible and acceptable from a socio-cultural viewpoint; and c) the political context is favourable to the adoption of these solutions.

The simple formulation of the theory directly linking transparency and accountability with evidence-based policymaking is thus complicated with the addition of an intermediate step: it is assumed that transparency, accountability and stakeholder collaboration do not directly contribute to Evidence-Based Policy Making (EBPM), but to either the framing of problems (PROB) or the creation of solutions (SOL) or both; and if only one of these is achieved, then a favourable political context (POL) is needed for successful EBPM.

In other words, Kingdon’s three streams are considered intermediate outcomes contributing to the ultimate outcome of evidence-based policymaking (EBPM); and two partially different lists of factors, related to transparency, accountability and multi-stakeholder collaboration, are created as presumably leading to two intermediate outcomes (framing of problems and creation of solutions, see Figure 1).

Figure 1: The programme nested theory of change


Conditions related to Transparency, Accountability & Multi-Stakeholder Collaboration
PRES
DATA
ALIG
INFO
CHAMP
PROB
POL
SOL
EBPM
SECOND LEVEL
FIRST LEVEL

1.3.2.1 Testing the first-level theory: intermediate outcomes to ultimate outcome

Three different theories of changes, the last two nested into the first, are then built and put to the test with QCA. The first is the the application of Kingdon's own theory to the health sector in the eight countries. The QCA model reads as follows:

- ATHSP (access to the health system for the poor) is unanimously identified as a problem by stakeholders (PROB) +
- Feasible and acceptable solutions to ATHSP are identified in the course of an active multi-stakeholder dialogue (SOL) +
- the political context is favourable to addressing ATHSP issues (POL) =
- evidence-based policies around ATHSP (EBPM) are put in place.

Presence and absence of the above conditions are defined as follows:

Table 7: Definitions of presence and absence of the three condition-EBPM model

Condition	Presence (1)	Absence (0)
PROB	ATHSP has been defined as a “problem“ by multiple,	ATHSP is seen as a problem only by a handful of stakeholders or only by the weakest stakeholders

	powerful stakeholders (PROB)	(prob)
SOL	The ATHSP community discusses specific solutions which are deemed feasible and acceptable (SOL)	The ATHSP community either does not discuss specific solutions, or discusses solutions which are unfeasible or unacceptable (sol)
POL	There is clear and sustained political support for ATHSP (POL)	There is either unclear or discontinuous support for ATHSP (pol)
EBPM	Evidence-based ATHSP policies are either in place or about to be (EBPM)	Evidence-based ATHSP policies are not in place and are not predicted to be any time soon (ebpm)

The conditions are scored with 1s and 0s across the eight countries, and the following dataset is created[13]:

Table 8: Dataset for the EBPM model

	PROB	SOL	POL	EBPM
Vietnam	1	0	0	0
Kenya	1	0	0	0
Zimbabwe	1	1	0	1
Bolivia	1	1	1	1
Indonesia	1	1	0	1
Ethiopia	1	1	1	1
Laos	1	1	0	1
Tajikstan	1	1	1	1

Note that some countries have identical combinations of the three conditions and the same outcome. It is thus possible to construct another table, a "Truth Table" (see Section 2.6), to compare the different configurations more clearly:

Table 9: Truth Table for the EBPM model

PROB	SOL	POL	EBPM	Countries
1	0	0	0	Kenya, Vietnam
1	1	0	1	Zimbabwe, Indonesia, Laos
1	1	1	1	Bolivia, Ethiopia, Tajikstan

All countries analysed except Kenya and Vietnam are successful in that they either have evidence-based policies on ATHSP or are about

[13] The complete dataset can be found in Annex C.

to adopt them. All 6 successful countries have defined ATHSP as a problem and proposed feasible and acceptable solutions. However, in only 3 of these countries the political situation is favourable to the adoption of such policies.

The analysis of this data allows the evaluator to develop the hypothesis that a politically favourable context is not necessary/needed for success, while defining ATHSP as a problem and proposing feasible and acceptable solutions are. More specifically, all cases that have adopted ATHSP policies on the basis of evidence have defined ATHSP as a problem and proposed feasible and acceptable solutions; but only some have a favourable political situation. It seems the latter is irrelevant.

Another way to show the irrelevance of the political context is to merge the two successful combinations PROB*SOL*pol and PROB*SOL*POL into a simpler, two condition combination: PROB*SOL. When the first two conditions are present, they are sufficient for success by themselves and the third is irrelevant. This is the logic employed by one of the two QCA procedures to analyse sufficiency illustrated in detail in Chapter 3: the boolean minimisation.

The synthetic QCA findings – technically known as “*solutions*“ – now read as follows:

PROB*sol*pol => ebpm
PROB*SOL => EBPM

The pathway leading to lack of success (no evidence based-policymaking) shows that even when multiple, powerful stakeholders frame ATHSP as a problem, this is not enough for success: the inability to turn this framing into feasible and acceptable solutions prevents its achievement.

1.3.2.2 Lesson learned from synthesising the dataset

What we learn from these findings can be summarised as below:

1. The Kingdon-inspired programme theory that at least two of the three factors need to be present for success is confirmed. PROB and SOL are both necessary for success.
2. The Kingdon-inspired programme theory is refined as follows:

- Framing ATHSP as problematic and desirable AND devising feasible solutions to this problem is more important than having a favourable political context. For what concerns ATHSP policies in these eight countries, these two particular conditions are sufficient for success by themselves, no matter the political context.
- Merely seeing ATHSP as a problem is not enough to translate this concern into evidence-based policies: the cases of Vietnam and Kenya show that the inability to find viable solutions neutralises the benefits of framing the problem appropriately.

3. The political context cannot be used as an excuse to "justify" the situation for those countries lagging behind in terms of evidence-based policy on ATHSP. The *QCA solution* shows that when a multistakeholder community is able to agree that ATHSP is insufficient and identify viable policy solutions, policymakers will follow up accordingly, whether political conditions are favourable or not.

1.3.2.3 Testing the second-level theories: explaining the intermediate outcomes

The second theory put to the test aims at identifying conditions that need to be in place in order for stakeholders to reach a consensus in defining ATHSP as a problem. The QCA model reads as follows:

- Focused events are organised and other forms of public pressure on ATHSP are put in place (PRES) +
- Interest groups are generally aligned on policy priorities (ALIG) +
- Groups are able to access credible data on ATHSP (DATA) =
- ATHSP is unanimously identified as a problem by stakeholders (PROB)

Table 10: Truth Table for the PROB model

PRES	ALIG	DATA	PROB	Countries
1	0	0	1	Laos
0	0	1	1	Bolivia, Zimbabwe, Indonesia
1	1	1	1	Vietnam
1	0	1	1	Kenya
0	1	1	1	Ethiopia, Tajikstan

The dataset shows that a consensus is reached in all countries. The solution DATA + PRES*alig*data => PROB show that access to credible data is the most important condition that needs to be in place in order to reach an agreement. In all cases except Laos, public pressure and alignment of policy priorities vary substantially while access to credible data stays the same (this is the logic used in the Method of Agreement, see Annex A). However, DATA is not necessary for success: the second successful pathway "PRES*alig* data", covering Laos, shows that when credible data is not accessible, success is still possible.

The third theory explaining the identification of viable policy solutions is "translated" in the following QCA model:

- Interest groups are generally aligned on policy priorities (ALIG) +
- Information-sharing agreements or protocols exist in the multi-stakeholder ATHSP community (INFO) +
- Skilled policy entrepreneurs or "champions" are active in the ATHSP sector (CHAMP) =
- Feasible and acceptable solutions to ATHSP are identified in the course of an active multi-stakeholder dialogue (SOL)

Table 11: Truth Table for the SOL model

ALIG	INFO	CHAMP	SOL	Cases
0	1	1	1	Indonesia, Laos
0	0	1	1	Bolivia, Zimbabwe
1	1	1	1	Ethiopia
1	1	0	1	Tajikstan
1	0	0	0	Vietnam
0	0	0	0	Kenya

The solutions confirm the low importance of actors agreeing on general policy priorities (ALIG), which happens for both successful and unsuccessful cases. The two most important factors for the identification of viable policy solutions appear to be the presence of skilled policy entrepreneurs (or "champions", CHAMP) and the existence of information sharing agreements/protocols in the multi-stakeholder ATHSP community (INFO). The latter is not necessary but it becomes sufficient when combined with either "champions" (Indonesia, Laos, Ethiopia) or "alignment on political priorities"

(Ethiopia/Tajikstan). The absence of both “champions“ and “information sharing protocols“ seems to guarantee lack of success.

ALIG*INFO (Ethiopia/Tajikstan) + alig*info*CHAMP (Bolivia, Zimbabwe) + INFO*CHAMP (Indonesia, Laos, Ethiopia) => SOL (viable solutions identified)

info*champ (Vietnam, Kenya) => sol (no viable solutions identified)

1.3.2.4 Making recommendations

What conclusions can we draw from testing this relatively complex, nested theory of change with QCA? First of all, from the first test we learned that we should focus on the factors leading to problem identification and viable policy solutions, in particular the latter, rather than worry about the political context. Secondly, among the factors associated with viable policy solutions, the existence of champions and information sharing protocols seems to be particularly critical (more than general alignment on policy priorities). In sum, we have reason to believe that focusing efforts on supporting policy entrepreneurs and encouraging the drafting of information sharing protocols can lead to great success in identifying viable policy solutions. Which in turn seems to lead to evidence-based policymaking in ATHSP, in particular if there is agreement on problem definition.

Agreement on problem definition does not appear as critical as the conditions mentioned above (being present also in Vietnam and Kenya which have not been successful on ATHSP evidence-based policy making). For this reason, if interventions focus exclusively on those conditions which are key to facilitate agreement on problem definition – like improving access to credible data and focused events – immediate results on evidence-based policy making should not be expected. On the other hand, if interventions facilitate the identification of viable solutions by supporting champions and encouraging the use of information sharing agreements/protocols, especially in addition to facilitating agreement on problem definition, more immediate results are likely.

In sum, QCA has helped the evaluator make recommendations on which types of interventions to prioritise in the future, associating different mixes of interventions with different expectations of success. One simple way of illustrating this recommendation is: “if you can support one intervention, focus on improving access to credible data; if you can support two, add support to champions". It has accomplished this by building on a theory of change expressed in simple and relatively broad terms; and drawing on social science theory and empirical data to develop a more articulated, nested theory of change connecting different types of interventions, including but not limited to those supported by evaluation commissioners.

Table 12: Evaluation of the chances of success of different mixes of policy options

Policy Mix			Chances of Success in ATHSP evidence-based policy making
Improving access to credible data	Supporting Champions	Encouraging information sharing	
X	-	-	LOW
X	-	X	MEDIUM
X	X	-	HIGH
X	X	X	VERY HIGH

Notice that, while QCA can lead to the confirmation of a theory, it is fully open to its rejection: for example the first test clearly shows that one element of the initial programme theory (political context) is not relevant for success; and both the other two strongly downplay the role of general alignment on policy priorities. Even though the evaluator might have an unconscious preference for some conditions, in QCA models all conditions have initially the same weight: the synthesis procedures return findings where some (combinations of) conditions are shown to be more important than others through automatic, algorithm-based procedures which are blind to evaluator biases (see Chapters 3 and 4). This is why QCA solutions are sometimes surprising for evaluators, who might struggle to make sense of the configurations.

1.4 The Relevance of QCA for development evaluations

In introducing the landscape of methodological innovation in impact evaluation, Chapter 1 has highlighted the shortcomings of well-known quantitative methods: for example the presence of construct validity concerns, or the complexity of settings or of the service being provided not allowing the reconstruction of a counterfactual situation;

or the number of available cases not being large enough to obtain statistically significant findings.

As a theory-informed method handling qualitative constructs, which is suitable for (though not solely confined to) small-n analysis, and which is compatible with at least three different causality frameworks[14], QCA is uniquely positioned to overcome these shortcomings. At the same time, QCA offers opportunities which are normally provided by quantitative methods, like internal validity and the ability to generalise. The method offers advantages of both quali and quanti methods, combining the richness and diversity of case-based studies with the rigour, ***replicability*** and generalisation potential of variable-based methods.

Another advantage offered by QCA is that, in addition to testing theories within a confirmatory perspective, it can also be used within an exploratory perspective, to select and adjudicate among multiple theories those which appear more promising on the basis of available empirical data.

1.4.1 The role of the intervention and Contingent Causality

A problem in development evaluation is that many development goals, like institutional reform, empowerment of disadvantaged groups, social change, etc. are not achieved by simple "injections" of activities as the idea of "treatment", popular in experimental approaches, would imply. External interventions might be required but might not be sufficient; it is only in combination with other factors, perhaps other interventions, and favourable contextual conditions, that interventions are likely to achieve their goals (Stern, Stame, Mayne, Forss, Davies, & Befani, 2012; Mayne, 2012). In technical terms, the causal power of many interventions is conjunctural: or contingent on a number of other factors. Interventions do not make an impact by themselves[15]. This is what makes QCA's unique ability to answer the question

[14] Mill's Methods of Difference, Mill's Method of Agreement, Mechanism-Based and Generative frameworks (see Annex A)

[15] Suppose that, in the example of water points repair (see Annex C), the intervention provided spare parts. The dataset shows that the availability of spare parts alone – although necessary – is not sufficient for success (there is one case where spare parts are available but repairs are not carried out). It is spare parts in combination with sufficient funds and clear responsibilities that ensures that repairs are made (i.e. that the intervention is successful). The intervention by itself is not always sufficient.

"what combinations of factors are sufficient for the outcome" relevant.

1.4.2 Equifinality: which interventions are needed?

Donor-funded development interventions are usually not the only way to achieve the goals that they set for themselves. Local governments, the private sector, country demographics and the geopolitical situation might all conspire to make empowerment of women harder, for example; but might also work in synergy to accelerate it. Interventions create entry points into complex systems (Garcia & Zazueta, 2015) and interact with the rest of the system to achieve positive change and create opportunities. Since systems are complex, failing to implement a specific intervention does not automatically make outcomes worse. This is the intuition behind the need for impact evaluation: interventions might be redundant, or even harmful. Their performance must not be taken for granted.

It is then reasonable to assume that most development goals are achieved through a number of different pathways, and only some of these pathways necessarily need to include a specific intervention in order for outcomes to be achieved.[16] In other words, some interventions might not be needed (necessary), or they might be needed (necessary) only under particular circumstances. This is what makes QCA's unique ability to answer the question "what factors are needed (necessary) for the outcome" relevant.

Thinking in terms of necessity and sufficiency, and viewing contributing factors in this sense, constitutes a paradigm shift compared to statistical or econometric models where effects are most often thought of as incremental, contributing to the outcome proportionally, as if "topping it up", or gradually eroding it. In the example of young farmers in Annex B, incrementally adding remedies to face the demand crisis does not increase, on average, the chances of success. A method, like regression analysis, that identifies the incremental contribution of each factor to the outcome is not

[16] In the example of water points repair, if the intervention provides spare parts, it means the intervention provides a necessary ingredient for repairs to be made; however, the intervention in principle is not the only way to make spare parts available. In some water points the local authority might find ways to make spare parts available without that particular intervention, and perhaps without any external/donor intervention at all.

designed to capture the complexity of successful (or unsuccessful) pathways to the outcome. It is through the systematic comparison of cases seen as "wholes", allowed by QCA, that qualitatively different typologies of cases can be built, and that success (or lack thereof) can be explained by the combination[17] (constellation or *configuration*), rather than the sum, of the different remedies that farmers can potentially take. For the differences between QCA and Regression Analysis see Annex B and Thiem, Baumgartner, & Bol (2015).

1.4.3 Limited Diversity: understanding what can be synthesised/generalised and what can't

Many times the interest of donors and commissioners does not lie in discovering the average effect of the intervention over a sample of projects or countries, but in understanding what made the difference or didn't for different groups or under different circumstances. Most of the times, cases present considerable variety; but at the same they are not completely different and can be grouped into typologies describing contextual dynamics, situations, backgrounds or histories. Rather than considering cases totally unique or similar enough for the average to be informative, QCA allows the conceptualisation and analysis of a "middle range" of differences and similarities; and allows the construction of typologies of cases that are likely to experience partially similar and partially different pathways to the outcome; where "partially" can be defined in many ways and degrees. The question answered by QCA "what makes the difference for whom and under what circumstances" recognises that the intervention is likely to play different roles in different contexts, and demonstrates interest in learning more about these diversified dynamics; in understanding what they have in common and in what they differ. It acknowledges that the intervention has the potential to indeed make a difference, but that its impact is likely to depend on contextual and/or contingent circumstances.

[17] This difference is akin to the difference between set-theory (with its concepts of "logical intersection" and "logical union") and mathematical operations (like "addition" and "multiplication"). However, the same notation of "+" and "*" signs can be used to denote concepts from both categories.

1.4.4 Robustness and rigour: creating credible evidence of effectiveness

Methodological gaps in impact evaluation are often thought of as inability to establish causality or to generalise. QCA complements existing methods well[18] in testing causal hypotheses (see Annex A) and allows for modest or contingent generalisation, even aspiring to broader, "outside-the-dataset" generalisation depending on the characteristics of the dataset (see Section 3.1).

1.5 Deciding to use QCA in a development evaluation

While initially popular with political scientists, QCA has now been tried and tested in many fields like business (Romme, 1995; Kask & Linton, 2013), education (Stevenson, 2013), environmental science (Basurto, 2013), and health research (Blackman, 2013).[19] Since the development field is just beginning to open its doors to QCA (Raab & Stuppert, 2014; Stedman-Bryce, Schatz, Hodgkin, & Balogun, 2016; Baptist, Edouard, & Batran, 2015; Holvoet & Inberg, 2013; Welle, Williams, Pearce, & Befani, 2015; Baptist & Befani, 2015), a commissioner or an evaluator approaching QCA for the first time would probably like to know what QCA can do, but also have an idea of its practical challenges, requirements and limitations. While previous sections should have "whetted the reader's appetite" for QCA, this section attempts at managing expectations by providing a brief discussion of the process of commissioning and applying QCA to a development evaluation.

1.5.1 Requirements

Using QCA requires at least 3 conditions to be in place: appropriate theories of change, at least a small or medium set of comparable cases providing suitable data in relation to the theory of change, and an adequate mix of technical skills and sectoral expertise in the evaluation team.

[18] As (Schneider & Wagemann, 2012) put it, "QCA is inherently a multi-method approach".

[19] For an updated overview of QCA applications, see www.compasss.org

1.5.1.1 Appropriate Theories of Change

Theories of Change, including factors that are assumed to causally explain/contribute to the outcome, are required for QCA. Note that "contribute to the outcome" is not to be intended in an incremental sense, as in "more of this factor produces more of the outcome", but in a "chemical" sense, as if thinking of ingredients for a recipe (see Annex A and Section 2.1.2).

This will usually require going beyond the logframe-like, popular linear sequences of activities, outputs, outcomes and impacts, to explore additional intermediary outputs and outcomes, and most importantly conditions pertaining to the historical, socio-economic, cultural, institutional and organisational context. The ideal conditions for QCA models make a substantial difference to the outcome. Moreover, appropriate theories here do not require a high (30+) number of conditions. The best QCA models include *a small number of critical conditions* (see Section 2.1.2).

Note that "appropriate" here is not to be intended as solid, proven theories, but more like theories including a specific type and number of elements. QCA is often most useful when adjudicating between competing theories of change, so the theories don't necessarily need to be well-established from the start: they can be tentative, speculative, based on anecdotal evidence or even hunches: what matters is that they are expressed and formulated in a certain way.

1.5.1.2 A (small, medium or large) set of comparable cases

The conditions identified in QCA models need to be assigned numerical scores: usually binary scales, but sometimes 4-point or 6-point scales if using *fuzzy set QCA*. This has the following implications:

- Data need to be available for all conditions across all cases: if this doesn't happen either the case or the condition needs to be dropped as QCA does not tolerate missing data. In some cases, it will make sense to undertake targeted primary data collection to fill in some data gaps or to rely on expert judgement; but this might not always be feasible or it might undermine the robustness/ internal validity of the findings

- Data for the same condition need to be comparable across all cases, in the sense that a standard *rubric* for the condition (assigning the same numerical values to similar cases) needs to be able to capture the whole range of cross-case variation for that condition[20]. If cases are too diverse for rubrics to capture the empirical variation of a condition, that condition cannot be used for QCA; if the same happens for most conditions, QCA cannot be used at all.

Finally, QCA is a case based approach but it doesn't work with a single case. There is usually some flexibility in defining what "a case" is for a particular evaluation: for example a country, region or city; but also an intervention type or implementing partner; or some combination of these (see also Ragin, Becker, & (eds), 1992). However, it's fundamental to have at least a small or medium number of cases to work on. It's difficult to imagine QCA fulfilling its potential with 3 or fewer cases.

Notice that while QCA is suitable for *small-n* analysis, it is not limited to it, and can handle *medium* and even *large* number of cases.

1.5.1.3 An adequate mix of skills in the evaluation team

Evaluation teams usually include experts of the sector the intervention is operating in; while QCA technical skills or experience in using the method are harder to find. Unfortunately, teams not only need to include such skills, but also ensure that a constant dialogue takes place between the sectoral experts with the case-based knowledge and the QCA experts, due to the iterative nature of QCA. As far as possible, and in order to exploit the full range of QCA's possibilities, the technical experts should be familiar with multiple software platforms, at least with fsQCA, R and Tosmana which have different strengths and weaknesses, and mentor the experts holding the case-based knowledge on at least the basic characteristics of the method.

[20] Note that the complexity of a given case can be captured by collecting information on several conditions and analyse their interactions and the way they combine with each other; it doesn't need to be *entirely* captured by the variation of *one* condition. See also Section 2.3.

1.5.2 Limitations

As Chapter 1 attempted to clarify from the very start, QCA is not suitable for use in all situations and has been subject to critique (see Rihoux B., 2015 and De Meur, Rihoux, & Yamasaki, 2009 for comprehensive reviews of critiques to QCA). QCA is not put forward here as the best possible method; but rather as a useful tool in the evaluation toolbox, with clear comparative advantages under specific evaluation circumstances (particularly when combined with other theory-based, comparative or case-based approaches – see also Chapter 3). We have seen in Chapter 1 that different methods have different comparative advantages in answering specific questions and have different strengths and weaknesses in different types of validity.

In relation to evaluation, the three main critiques addressed to QCA (see Chapter 3 for more details) are:

1. its inability to stand by itself without appropriate theories and adequate conceptual development (which often means adequate substantive/thematic, case-based expertise in the sector the intervention is operating in);
2. its inability to generalise the findings under specific conditions; and
3. the risk of over-simplification lying in synthesising rich qualitative information into numerical scores, which removes detail and nuance from qualitative data.

All have a bright side or can be minimised, if one considers that:

1. QCA's dependence on other methods for full theory development allows it to incorporate insights from and set up a "dialogue" with, several different methods;
2. quantitative methods are subject to the same challenges and constraints (model specification, sample composition and size) when it comes to generalisation; and
3. the risk of over-simplification is even greater for quantitative methods, and it is the price we must pay, sometimes, for internal validity, precision and generalisation. The good news with QCA is that the qualitative nuances and details can be recovered at a second stage, when QCA solutions are interpreted and made sense of theoretically (see Chapter 3).

QCA's main protocols are "tried & tested" in academic research; but the method is still getting its feet wet in evaluation, so it's important that in this early phase – in order to avoid future path dependency stemming from initial bad practices – quality standards are met and rigorous application protocols respected. This is why this report includes a section on Quality Assurance (Chapter 3) and strongly encourages use of the checklist reported there.

1.5.3 The QCA process: Step by step

This section – drawing largely on Baptist & Befani (2015) – illustrates four broad, basic steps of the process of applying QCA to a real-life development evaluation: assessing whether QCA is appropriate and feasible; model specification and case selection; data collection and analysis; interpretation of the findings and iteration. Chapter 2 will expand on the description of this process significantly, adding details in terms of intermediate steps and issues at stake for each step.

1.5.3.1 Assessing whether QCA is appropriate and feasible

This is the phase where evaluators and/or commissioners decide to use QCA. The first issue to consider are evaluation questions. What are the relevant questions for the stakeholders? Is QCA suitable to answer the questions of interest (see Sections 1.1.1 and 1.3.1)? QCA might not be the ideal approach, particularly if used alone, to answer all evaluation questions that might be of interest to donors and commissioners.

The second issue is the possibility to identify at least a small number of distinct cases: QCA is not a method for *within-case analysis* and requires at least 3-5 cases to express its potential. The third issue is the availability of appropriate theories of change including a small or medium number of explanatory factors. QCA will be hard to apply to models including 30 or even 20 conditions, so theories where such a high number of conditions are considered equally plausible and cannot be prioritised are problematic. Statistical methods might be more appropriate in these cases.

Fourthly, in order to create a dataset of zeros and ones (as those presented in Section 1.3), the evaluation needs to accommodate the collection of data across all cases for all conditions included in the model. If this is not possible because data is unavailable or uncollectable, either the case or the condition need to be removed from the analysis. If data is available on too small a number of cases or only on conditions that do not make sense as explanatory of the outcome (due to missing data on the conditions that were thought to be explanatory), QCA might be a waste of time. It can still be theoretically applied, but the risk of it adding zero value to the evaluator's knowledge is higher.

Finally, while the basic logic of QCA can be mastered by relatively unspecialised staff, a certain level of technical skills is required in order to use the full range of QCA procedures, for which different software platforms are available. A high quality QCA application might require the ability to use at least three software platforms: fsQCA, Tosmana and R. R is the least user friendly overall, but the only package allowing a complete necessity analysis and the most user-friendly *subset* analysis; Tosmana is the best package to visualise Venn diagrams; and fsQCA is the most user friendly for the creation of the Truth Table and the Boolean minimisation (see Chapter 2 for more details).

1.5.3.2 Model specification and case selection

In order to decide whether to use QCA or not, at least a broad idea of what cases can be studied and what data can be collected is usually needed. Since the combination of numbers of cases and conditions influences the robustness of the findings of some QCA procedures (see Chapter 3), there will need to be a balance between number of conditions and number of cases. Once the evaluation starts, a more detailed list of conditions to be included in QCA models and a complete list of cases need to be developed.

As for the conditions, usually experts with high-level, thematic or substantive knowledge of the field, who are able to develop appropriate Theories of Change in collaboration with the QCA expert, are needed. Other evaluation approaches might be combined with QCA at this stage, like Contribution Analysis or Realist Evaluation (see Chapter 3). If more than one outcome needs to be explained, different lists of explanatory conditions will need to be devised for each outcome (see Chapter 2).

As for case selection, the first thing to consider is the balance between cases with a positive and with a negative outcome. Some causal questions cannot be answered unless the two groups of cases are compared (see Chapter 3). This might be difficult when the outcome is a measure of the intervention success, and commissioners or other stakeholders are unwilling to disclose information about less successful cases where the outcome would be negative.

1.5.3.3 Data collection and analysis

Once the conditions and cases have been identified, the skeleton of a QCA dataset can be created. At this point the matrix needs to be filled with data: all cells need to be filled, as QCA does not tolerate missing data. If data is missing for one condition in one case, either the case or condition needs to be removed from the analysis. Data can be collected from a variety of sources and the evaluator has plenty of freedom to define presence and absence of a condition (as well as intermediate degrees, for fuzzy-set QCA). However, once rubrics for each condition are created (or in other words, once conditions are calibrated), the same standard needs to be scrupulously applied across all cases, lest comparability (and thus internal validity) be compromised.

The Boolean matrix is analysed with multiple automatic procedures in order to understand which conditions are needed (necessary) for the outcome to occur and which combinations are effective (sufficient) (see Chapter 2 for more details). To some extent, when datasets are created from a very small number of cases or conditions, relationships can also be spotted by eye, although this is usually risky and prone to error. The software platforms provide rigorous ways of analysing patterns the eye might miss, and usually misses for more than 3-4 cases or conditions. In any case, the final products of these analyses will be configurations representing logical relations of "***union***" and "***intersection***" among conditions presenting the same outcome. Analysing both positive and negative outcomes is strongly recommended: some causal questions can only be answered by comparing the findings from the two groups (see also DeMorgan's Law in Section 2.5.3.3).

1.5.3.4 Interpretation of findings and iteration

At this point the consultant or researcher with substantive expertise of the sector (who presumably helped develop the QCA model to be tested and is often the principal investigator (P.I.)) returns at centre stage to make sense of the QCA solutions and interpret the findings. Do the QCA configurations synthesising the dataset support the original theory of change? What theory refinements do the QCA findings suggest? Below is a list of typical scenarios encountered at this stage:

- New data about new conditions or new cases that the P.I. thinks are relevant have become available: a new dataset is built and synthesis procedures are run again, with findings from both datasets compared (fairly common).
- QCA findings confirm the initial hypotheses (fairly rare)
- QCA findings draw the P.I.'s attention to a limited number of conditions which trigger new theoretical hypotheses and hence new models to test (fairly common)
- QCA findings are too complex to be easily interpreted and various strategies are used to remove conditions from the analysis and run the synthesis procedures on the more parsimonious models (fairly common)

After a number of iterations, a "good solution" is found which is a) clear and easily interpretable; b) covering most cases; and c) reliable enough to be considered either a necessity or sufficiency statement on the basis of the available data.

Note that, since QCA is an iterative method where it is difficult to know in advance how many iterations are needed, predicting its cost precisely is not easy, as each iteration will have additional costs.

2 QCA Step-by-Step: opportunities and pitfalls

This chapter provides detailed information on the application of QCA to development evaluations, and requires either a basic knowledge of QCA or having read Section 1.3. It illustrates a sequence of activities (or steps) that should be taken in a high-quality application of QCA (see also Section 3.3.3 for Quality Assurance in QCA). Every step is described in terms of either:

1. ***opportunity*** (what it is useful for, what specific questions it helps answer);
2. ***pitfall*** to look out for, requiring solutions and protection against;
3. or more generically ***issue at stake***, when the issue presents both opportunities and challenges.

The title of each step is followed by the main question it helps answer.

The chapter draws on material from several completed or ongoing evaluations[21], focusing on specific elements that are particularly illustrative of QCA's potential or challenges. In order not to disrupt the flow of argumentation, more information about these evaluations is reported in Annex C.

Note that, while every step carries different bias risks and most are reported in each step, a summary of biases can be found in Section 3.3.

The first 3 steps (model specification, ensuring data availability, and calibration) are consecutive and focus on the organisation of empirical data; the main differences between crisp-set QCA and fuzzy-set QCA lie in these phases, particularly in calibration (Step 3). The last 5 steps (the Venn diagram, the SuperSubset Analysis, the Truth Table, the Boolean minimisation, and the INUS analysis) illustrate ways to analyse and synthesise the data organised in the previous steps and are largely identical for crisp-set and fuzzy-set

[21] Please note that the findings/interpretations of the configurations included in this section are purely indicative: QCA was just one of the methods used in these evaluations, and the final findings included in the cited final reports were triangulated against other methods and findings. The examples reported here are only intended to demonstrate the use of the method.

QCA (particularly the last three). Note that – even though QCA "proper" is typically known as the consecutive completion of Steps 6 (building the Truth Table) and 7 (the Boolean minimisation) – the last five steps can be considered optional – or at least less strictly required than the first three. In particular, Steps 4, 5 and 6 can be carried out independently of each other, which means that either all three, or any two, or any one single step can be carried out to synthesise the dataset. Steps 7 and 8, however, are dependent on Step 6 and cannot start before Step 6 is completed. Finally Steps 7 and 8 are also independent of each other and both don't necessarily need to be conducted (see Figure 2).

The definition of "QCA analysis" can vary, but we argue that a QCA analysis is never possible unless the first three steps are completed. Similarly, neither Step 7 (the Boolean minimisation) nor Step 8 (the INUS analysis) are possible unless Step 6 (building the Truth Table) is completed. For the rest, none of the other steps are strictly required but – as information gathered during the different steps is usually useful and helps improve knowledge about the cases, it is advisable to make the most of the opportunities offered by the approach and complete all the steps; while being keenly aware of the pitfalls hidden along the way, taking decisive action to avoid them.

It is difficult to assess, in general and before knowing the specific details of the single case, the value of a QCA application that only takes some of the optional steps or overlooks some of the pitfalls under the steps taken. Our general recommendation is to be as comprehensive as possible, and justify on a case by case basis why some steps have been omitted or why specific pitfalls have not been paid attention to.

A brief, comparative summary of the procedures reported in Steps 5 to 8 can be found in Box 1.

Note that completing the steps does not necessarily mean concluding the analysis; it might be just the end of the first round. In many cases, the findings are not satisfactory at first; either because the solutions are not consistent or reliable or they don't cover enough cases; or because they are too complex and difficult to interpret. In most cases we need to test more than one model before we find a solution that is satisfactory: essentially, easy to interpret, meaningful, covering enough cases and having sufficient consistency.

Box 1: comparative summary of the different QCA procedures

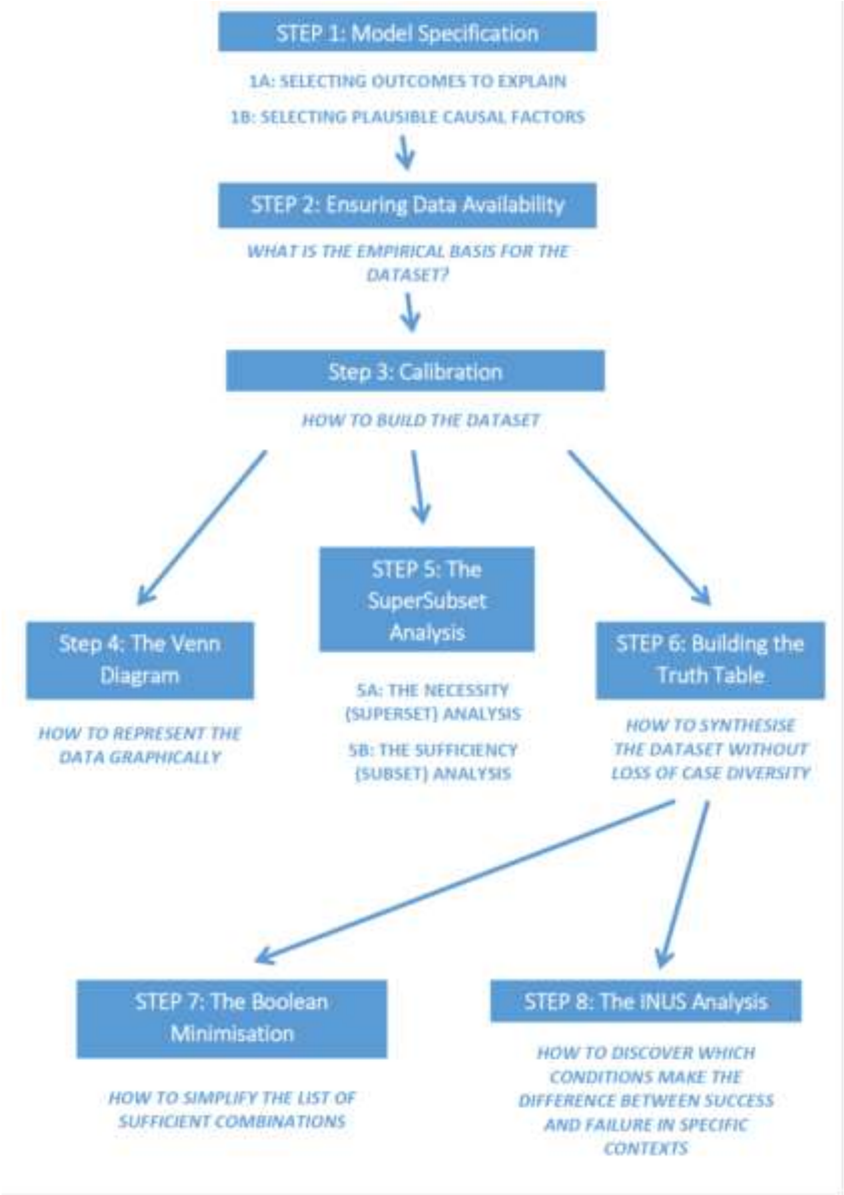
- the necessity analysis groups the cases with the same outcome and calculates frequencies;
- the subset analysis groups cases with the same conditions and calculates frequencies;
- the Truth Table procedure groups identical cases (on all conditions and the outcome) and merges them into one single combination;
- the Boolean minimisation merges Truth Table combinations with the same outcome and at most one different condition;
- the INUS analysis compares/isolates Truth Table combinations with a different outcome and at most one different condition.

Not rarely we acquire new information after one or more QCA iterations have been completed; and might be able to fill data gaps about a new case or a new condition, or change values that are already included in the dataset. We might also realise that the calibration of some conditions needs to be improved; or that two conditions are conceptually similar and can be grouped. All these potential changes affect the dataset and require that the analysis starts again from the Step 2 (ensuring data avilability) included.

Perhaps the biggest challenge in QCA evaluations is integrating technical expertise in evaluation teams. If QCA expertise is not held within the team, and in particular with the P.I., the analysis will require close collaboration with an external consultant. The closeness and duration of this collaboration will have a strong impact on the robustness and relevance of the QCA findings, particularly in datasets with many cases and conditions where many different models can be tested and many sets of findings need to be interpreted.

If the above conditions cannot be met, QCA will usually be used to test a pre-defined set of hypotheses and mostly either confirm or disconfirm them, without much room for refinement (and re-testing). This will still expand the evaluator's knowledge, but will fail to make the most of the full potential of QCA.

Figure 2: Map of the mandatory and optional steps of a QCA analysis


STEP 1: Model Specification
1A: SELECTING OUTCOMES TO EXPLAIN
1B: SELECTING PLAUSIBLE CAUSAL FACTORS
STEP 2: Ensuring Data Availability
WHAT IS THE EMPIRICAL BASIS FOR THE DATASET?
Step 3: Calibration
HOW TO BUILD THE DATASET
Step 4: The Venn Diagram
HOW TO REPRESENT THE DATA GRAPHICALLY
STEP 5: The SuperSubset Analysis
5A: THE NECESSITY (SUPERSET) ANALYSIS
5B: THE SUFFICIENCY (SUBSET) ANALYSIS
STEP 6: Building the Truth Table
HOW TO SYNTHESISE THE DATASET WITHOUT LOSS OF CASE DIVERSITY
STEP 7: The Boolean Minimisation
HOW TO SIMPLIFY THE LIST OF SUFFICIENT COMBINATIONS
STEP 8: The INUS Analysis
HOW TO DISCOVER WHICH CONDITIONS MAKE THE DIFFERENCE BETWEEN SUCCESS AND FAILURE IN SPECIFIC CONTEXTS

2.1 Step One: Model Specification

The first step in QCA involves selecting outcomes to explain and plausible explanatory conditions for each outcome. The basis for this selection is one or more theories of change linked to case-based knowledge, which can be created in many different ways: drawing on social science theory, as in the Evidence-Based Policy evaluation exemplified in the previous chapter or in (Stedman-Bryce, Schatz, Hodgkin, & Balogun, 2016); asking stakeholders more or less directly which are the most relevant factors for a given change, as in the QUIP protocol (Copestake, 2014) or in Most Significant Change approaches (Davies & Dart, 2005); conducting (at least some steps of) a contribution analysis (Mayne, 2008; Mayne, 2001; Baptist, Edouard, & Batran, 2015); converting realist CMO configurations into conditions (Befani, Ledermann, & Sager, 2007); using a mechanism that has been shown to exist in one case, for example with with Process Tracing (see Chapter 3); and more generally drawing on one or more programme theories for the intervention. The previous chapter has provided an example of how QCA can begin from a social science theory explaining how the policy agenda is set. Chapter 3 reports on the integration between QCA and explanatory/generative evaluation approaches, like Contribution Analysis, Realist Evaluation and Process Tracing (see also Amenta & Poulsen, 1994), which can be used as entry points for QCA.

No matter the sources of case-based knowledge or theories of change used, in order to create a dataset that can be analysed with QCA, data always need to be collected on a series of characteristics of the case, not too differently from what is done when using statistical variables (see also Annex B); except that these can be qualitative as well as quantitative and are hence known as „conditions". Most importantly, conditions need to be used to explain an outcome, so each QCA model is made of **one outcome** and **a list of conditions** that are assumed to affect that outcome:

Condition A + Condition B + Condition C + Condition D + ... + Condition P = Outcome.

2.1.1 Step 1A – Selecting Outcomes: "what are the main outcomes to explain?"

In selecting outcomes we answer the question "what are the main outcomes to explain" in this evaluation? This section will address the opportunities offered by QCA to test nested Theories of Change, synthesise evaluations, create typologies of cases, and take contextual influences into account; but also the constraints imposed by having to repeat most steps of the analysis to test additional outcomes.

2.1.1.1 OPPORTUNITY: prioritising outcomes for analysis when testing nested Theories of Change

In many evaluations, the logframe or the Theory of Change will include a series of intermediate outcomes that are assumed to be achieved before the ultimate, most highly desired outcome is. Normally, each outcome needs to be tested and explained separately; however, some of these outcomes can be redundant and fully explained by others, so the whole series of intermediate outcomes can also be tested as a QCA model in an attempt to explain the ultimate outcome. The such-developed "nested" Theory of Change can produce a nested QCA model, where n lists of conditions are associated with n intermediate outcomes; and an overarching model where n intermediate outcomes are associated with the ultimate outcome.

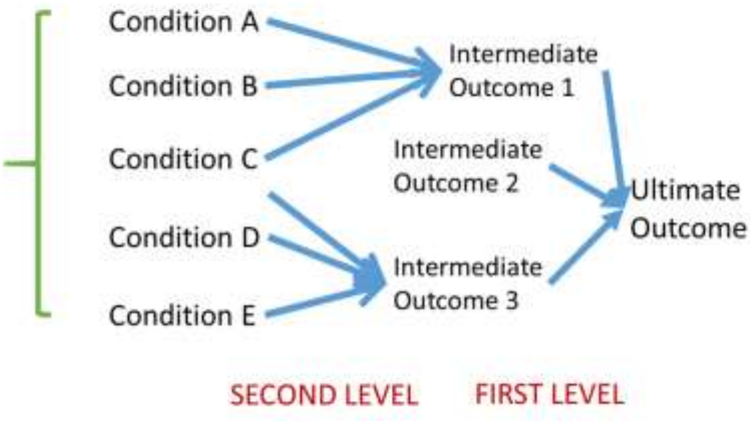
In Figure 3, a generic nested Theory of Change is represented, where an ultimate outcome is associated with three intermediate outcomes. Two of these intermediate outcomes are explained with two models, while a separate model can be envisaged for the ultimate outcome:

FIRST LEVEL: IO1 + IO2 + IO3 = O (Ultimate Outcome)

SECOND LEVEL: A + B + C = IO1 (Intermediate Outcome 1)

SECOND LEVEL: C + D + E = IO3 (Intermediate Outcome 3)

Figure 3: Representation of a nested Theory of Change



In the MAVC study (Welle, Williams, Pearce, & Befani, 2015) outcomes one (use of ICT) and two (processing of ICT-based reports) are intermediate outcomes assumed to be required for (ultimate) outcome 3 (repairs to water points being carried out): the ToC assumes that in order for rural water points to be repaired (based on ICT reports and data analysis), local government authorities need to process and follow up on ICT-based reports, and users need to use ICT in the way specified by the initiative. A quick analysis of the outcomes dataset for this project (see Annex C) reveals that, in the majority of cases (4/6), the theory is confirmed, with 75% of successful cases (3/4) showing that repairs are made when data processing and use of ICT are observed; plus one case where all outcomes are negative. However two unexpected pathways emerge:

- Repairs are carried out but there is no use of ICT for reporting and no data processing either.
- ICT are used and data is processed but no repairs are carried out.

We learn that the two intermediate outcomes are not necessarily needed for repairs and the relations are complex; which encourages us to keep our focus on all three outcomes separately.

In the Evidence-Based Policy Evaluation in Chapter 2, the three outcomes “Problems”, “Policies” and “Politics” are considered intermediate and included in a QCA model in an attempt to ultimately explain “evidence-based policymaking in access to the health system for the poor". Two are discovered to be necessary for the ultimate outcome while the third isn't. This shifts the attention of the evaluator to the analysis and explanation of the two necessary outcomes.

2.1.1.2 PITFALL: only one outcome at a time

QCA can explain the presence (or absence) of only one outcome at a time. If the explanation of more than one outcome is sought, each outcome will need a different QCA test: in other words, most of the steps in Chapter 3 will need to be repeated to explain another outcome. In terms of time and resources needed, adding one more outcome can have serious implications: the additional workload is proportional to the additional conditions needed to explain the new outcome. It's important to have a clear idea of which outcomes are to be explained as early as possible in the analysis, in order to estimate the workload reliably.

2.1.1.3 ISSUE AT STAKE: development outcomes vs. judgments of intervention success (synthesis of evaluations)

QCA has been proposed as a viable method to conduct systematic reviews or syntheses of evaluations (Befani, Ledermann, & Sager, 2007; Sager & Andereggen, 2012; Vaessen, Garcia, & Uitto, 2014). In these cases the evaluations might have already proceeded to assess, in each specific case, the causal power of the intervention to produce a certain outcome using within-case methods. As a consequence, the outcome condition used during the synthesis might not be a single indicator the intervention aims to have an impact on, like empowerment or resilience, but rather a more general indication of whether the intervention has been successful or not, as emerging from the single-case evaluations. In addition, the intervention will not be a condition, because all cases analysed have seen the implementation of the intervention. The analysis becomes a synthesis of existing cases where a group of interventions has been implemented, with the aim of discovering which factors have enabled or triggered success.

In such cases the conditions revealed as necessary or sufficient by QCA will not explain a socio-economic or policy regularity (like for example whether certain policy decisions have been taken or not), but rather the success of a group of similar interventions, on the basis of a series of contextual conditions or project characteristics that are assumed to influence the functioning of the intervention. For example, in a pilot round of analysis under a macro evaluation of DFID's Strategic Vision for Girls and Women[22], the conditions analysed for projects aimed at reducing violence against women reflected the characteristics of activities implemented in a given project: media campaigns, school-based interventions, activities targeting men and boys, policy frameworks, justice systems, referral and support services, legal support for victims, safe spaces, etc. and the aim was to understand which project types and contextual features were most consistently associated with success.

When QCA is used to explain the success of similar interventions on the basis of judgements of success made by different evaluators in different single-case evaluations, its findings will need to be interpreted in terms of which conditions have contributed to the success of the intervention, or have been obstacles to it; and care needs to be taken that the judgments of success are comparable across the cases considered.

2.1.1.4 ISSUE AT STAKE: observed vs. anticipated outcomes (creating typologies of interventions)

Sometimes reviews are conducted before the outcomes have had the time to materialise, or before evaluations of single cases are conducted to assess the outcomes empirically, ex-post. In such cases a QCA analysis cannot help evaluating impact; however it can be used to synthesise the characteristics of the interventions and create intervention typologies. This is the case of an ongoing (at the time of writing) review of policy instruments implemented in different developing countries to improve the livelihoods of vulnerable households and preserve their consumption capacity in times of crisis. The review is targeting policies that have adopted an integrated social protection/climate change approach and creating typologies of climate change + development integration on the basis of funding sources, financial instruments, policy instruments, risks covered, type of implementing organisation, target group, etc. The review is aimed at

[22] http://www.itad.com/knowledge-and-resources/dfids-macro-evaluations/

informing a future impact evaluation of these policies, by helping shape hypotheses on which type of "environment + development" integration work better, for whom and under what circumstances.

This is an opportunity because it shows that QCA can also be used when data about outcomes is not yet available; however it also has implications on the interpretation of the findings, which merely help conceptual development without having any significance in terms of causal inference.

2.1.2 Step 1B – Selecting Plausible Causal Factors: "how to identify the main contributing factors?"

Selecting conditions to explain each outcome is an opportunity to understand how the combinations of these factors influence each other's ability to contribute to the outcome. QCA can simultaneously test all combinations of conditions, supporting multiple theories of change, which is useful when theory is poorly developed and the evaluation objective is to adjudicate between multiple Theories of Change. However, this opportunity is constrained by the limited ability of QCA to handle a high number of conditions at the same time, which requires a good knowledge of a small number of "stand-out" factors.

2.1.2.1 OPPORTUNITY: identifying contextual and other conditions that affect the intervention's performance

QCA offers the opportunity to assess the effect of packages of factors, or combinations of causes, on an outcome. This is different from identifying the single best-fitting mathematical model explaining the average relation between a group of independent variables and a dependent variable, as regression analysis does (see Annex B and (Thiem, Baumgartner, & Bol, 2015). The conditions that work best for a QCA model do not make small incremental differences to the outcome, especially not directly: they rather affect substantially the causal power of the intervention or other conditions to contribute to it. For example, in the QCA model explaining the use of food storage facilities reported in Chapter 4 (the Food Trade Evaluation), the requirement to share the facility with other farmers determined, together with the existence of agreed quality standards, whether the facility was used or not. Agreed quality standards were needed for success if the requirement was present, but not needed if it were not. The issue is not how much quality standards affect use, on average, but when they are needed or not.

2.1.2.2 PITFALL: working with a small number of conditions

For reasons related to its qualitative and set-theoretic nature, QCA struggles to handle large (30+) or sometimes even medium (20+) numbers of conditions. This might be a problem in the following situations:

1. when the evaluator is interested in the incremental, average contribution of each factor and tends to include all factors somehow thought to make a minimal contribution;
2. when knowledge about the relevant factors is very poor and the evaluator is unable to prioritise a small number of factors for inclusion in the model.

Since QCA can only handle effectively a small number of causal factors for the explanation of each outcome, aiming for full coverage of all the possibly relevant factors is usually fruitless. QCA helps with testing how causal factors influence each other's ability to contribute to the outcome, rather than measuring the extent of their dominance or importance. This is why QCA is not the optimal choice when the impact question of interest is "how large is the effect of the intervention or of other causal factors on the outcome" (see Chapter 1), which requires isolating the influence of each single factor and obtaining the "net effect" by subtraction. At the same time, it enjoys a unique comparative advantage in spotting "winning recipes" or complex combinations that are successful, while their single components might not be so.

Fortunately, using another method is not the only solution to dealing with a high number of conditions. More constructive strategies are suggested below and in Sections 2.7.1.8 and 2.7.1.9.

2.1.2.3 PITFALL: associating all factors to all outcomes

In a typical situation falling under point 2 (when knowledge about the relevant factors is very poor and the evaluator is unable to prioritise a small number of factors for inclusion in the model), the evaluator might identify a list of outcomes and a list of factors, without making assumptions as to which factors presumably explain which outcome. The implicit assumption is that all factors might potentially contribute to each outcome. In these cases it very difficult to reduce the number of conditions. It is highly recommended that different lists of factors are associated with different outcomes, since it's usually likely that some conditions won't make sense for at least some outcomes. In other words, lists of factors assumed to explain different outcomes can overlap but do not need to be identical, even when the outcomes are related. Table 13 illustrates how partially overlapping lists of conditions were associated to three different outcomes in the evaluation of the African Regional Empowerment and Accountability Programme (AREAP) (Baptist, Edouard, & Batran, 2015).

In the AREAP evaluation, all three of the outcomes tested (strengthened relationships with policy makers, civil society engaging in regularised accountability spaces, and focus on policy change in key sectors) had in common 5 conditions (highlighted in green: transparency, political stability, engagement with a broad range of partners, research capacity, competing work). Outcome 1 and outcome 2 had in common 3 additional conditions (government sensitivity to criticism, use of evidence in policy making, and champions, highlighted in orange); outcome 1 and 3, 4 additional conditions, highlighted in blue. Finally, six conditions were used to explain only one outcome (highlighted in yellow).

2.1.2.4 OPPORTUNITY: testing multiple Theories of Change simultaneously

Table 13 also shows us how multiple theories of change can be tested simultaneously with QCA, which is useful when the theory of change is not well-developed or multiple competing theories seem all well supported and adjudicating among them is difficult.

Table 13: Conditions selected for inclusion in different models (explaining different outcomes)

	OUTCOMES		
CONDITIONS	Outcome 1: The extent to which civil society coalitions have developed strengthened relationships and a shared understanding with policymakers on key issues	Outcome 2: The extent to which civil society coalitions engage in regularised accountability mechanisms and spaces for civil society participation in policymaking	Outcome 3: The extent to which strengthened civil society coalitions supported by the Trust or SOTU focused on policy change on key national and regional topics targeted by AREAP implementing partners
Government opens up space for engagement with non-state actors (GovEng)	X		X
National partner coalition engagement in spaces that enable citizens to engage with policy makers/government (CitizenEng)	X		
Policymakers are transparent about decision making processes (Transp)	X	X	X
Level of national/ regional political stability (PolStab)	X	X	X
Enabling environment for non-partisan and empowered CSOs at national level (EnEnv)	X		X
Level of government sensitivity to criticism (GovSens)	X	X	
Strengthened technical skills within civil society coalitions for policy engagement (StrengthTech)	X		
Strength of use of evidence in policy making (EvidencePm)	X	X	
NSAs have adequate knowledge of stakeholders and their needs, agendas (KnowStake)	X		
In-house capacity of IP for effective packaging of information around civil society views (IPinHouseCap)	X		
Strength of national partner reputation and credibility on relevant issues (StrengthPlat)	X		
Constant engagement by national	X	X	X

partner coalitions with diverse audiences to develop ownership of ideas and map the national political economy (EngDiv)			
Existence of champions for non-state actor's cause and activities (ChampNSA, here termed IPChampGov)	X	X	
Internal research capacity of national partner coalitions (IntResCap)	X	X	X
National partner coalitions have other competing work (NatParComp)	X	X	X
Internal general capacity of national partner coalitions (IntGenCap)	X		X
Visibility and perceived credibility of national partner (NatParCred)	X		X
Political pluralism (PolPlur)			X

2.1.2.5 PITFALL: some (ultimate) outcomes cannot be analysed

In many programmes, evaluations are commissioned at a programme stage where the ultimate effects still haven't had time to materialise. This will make it difficult to collect evidence on whether the intervention contributed to the ultimate outcomes, as data on the latter might be sparse or low quality while it might be easier to observe changes in proximate outcomes. In practice, this would likely make tests possible only on the intermediate outcomes.

The above situation can be problematic with all methods and in all situations, but it's particularly challenging in a small-n environment. In some cases, automatic procedures and theory, as well as additional data collection, can help estimate the missing outcome values; but when n is small, low-quality data on even a small number of outcomes can hinder the robustness of findings substantially.

This risk materialised in the course of a pilot round of analysis under a macro evaluation of DFID's Empowerment & Accountability initiatives[23], where improvements were registered in local-level service delivery but not in higher-level (national) service delivery. This discrepancy was attributed to the assumption that changes caused by

[23] http://www.itad.com/knowledge-and-resources/dfids-macro-evaluations/

the intervention are quicker to materialise at the local level than at the national level. Likewise, improvements were visible across most cases in accountability and response of government to Violence Against Women (VAWG), but no changes were observed in reports of Gender-Based Violence (GBV) and social norms: the latter outcomes were assumed to take longer to change than the former. The analyses of national service delivery and change in reports of GBV and social norms were attempted; but eventually not considered reliable.

2.2 Step Two: Ensuring Data Availability: "what is the empirical basis for the dataset?"

QCA is a case-based approach that offers indications on how to compare case-study data; it is not a data collection technique providing suggestions on how to collect data. Data that is suitable for QCA can be collected with a wide range of techniques, including desk reviews, interviews and questionnaires. As long as the data is relevant to improve the available information on the various conditions, and it fits the standards applying to the various techniques, it can be considered adequate for QCA use. Chapter 4 reports on the typical biases involved in data collection, which QCA is subject to like all other research methods.

Perhaps the biggest difference between QCA and other approaches in terms of data requirements is that missing data implies a relatively high information loss, which in turn can affect the evaluation design. The strict data requirements can constrain the choice of conditions and outcomes that is possible to analyse.

On the bright side, the same methodological feature that makes handling missing data difficult is responsible for returning synthetic configurations that represent all cases equally and preserve the diversity and richness of every single case, even when their frequency is low.

2.2.1.1 PITFALL: the high cost of missing data

QCA does not handle missing data well. If data is missing on one of the factors included in a model the evaluator wants to test, the default option is removing the case(s) for which data is missing. However, this might be very costly if data is missing on a high number of cases. Another option to consider is removing the condition for the model under test. But this might also be costly, either because that condition is considered important to include, or because it is well covered on all cases except one or two. Removing the condition here would mean losing all the information across the other cases, for that condition.

If either the case(s) or the condition(s) for which data is missing must be removed, the best choice will depend on how much information is lost with either action. One good strategy is to compare how many complete cells are lost removing the case vs. removing the condition. Usually a model includes more cases than conditions, so if data is missing on only one cell, removing the condition will be more expensive (it will mean removing more cells) than removing the case. But if the condition has created problems on many cases, the "good cells" lost by removing the condition might be fewer than those lost removing the case.

A more forgiving option to handle missing data is to estimate it through expert input, building on the little available information; which might, however, be subject to internal validity problems. One typical solution is to conduct additional primary data collection.

2.2.1.2 OPPORTUNITY: drawing on a wide range of data sources and data collection techniques

While the choice between removal of the condition and removal of the case may seem hard, it can be compensated by QCA's neutrality in terms of sources: the missing data can be collected with a wide variety of techniques, drawing on a broad range of sources. In practice, it is usually not too difficult to complete the dataset by conducting additional interviews or reviewing additional documentation. Data that is suitable for QCA can be collected with a wide range of techniques, including desk reviews, interviews and questionnaires. This opportunity was taken during the MAVC study. As long as the data is relevant to improve the available information on the various conditions, and it fits the standards applying to the various techniques, it can be considered adequate for QCA use.

2.2.1.3 PITFALL: some conditions cannot be included in the model[24]

The list of causal factors to include in the analysis might be severely influenced by data availability. A QCA short note for commissioners (Baptist & Befani, 2015) reports that reducing the factors that can realistically be included in a QCA analysis after data collection is a typical step to undertake. We have seen above that, if data is not available for a condition in a specific case, the evaluator needs to either remove the case or the condition. If the case is removed, the findings will apply to a smaller dataset than originally planned. If the condition is removed, the analysis might miss discovering the role of an important factor. If this condition is an outcome, then that outcome cannot be explained/analysed. All possible efforts should be made to obtain a complete dataset.

2.2.1.4 OPPORTUNITY: the synthetic findings represent the full spectrum of diversity

Statistical methods offer information on the average state of a set of cases, neglecting outliers and "deviant" cases. For this reason, statistical procedures can easily estimate the value of missing data while preserving the robustness of findings. By contrast, QCA findings are derived from a systematic comparison of cases that all count equally: one single case can profoundly change the shape of a parsimonious/synthetic configuration obtained from a dataset, when it is added to the dataset. The same methodological feature that makes handling missing data difficult provides the benefit of returning parsimonious configurations that faithfully represent the full spectrum of diversity embodied across the cases. In QCA, differences are respected and treated as equally worthwhile alternatives, instead of "noise" to be removed. Notice that this does not prevent QCA from restricting the analysis to the most frequent combinations, if the evaluator chooses to do so; nor from estimating the statistical significance of its findings, as will be shown in Section 3.1.1.

[24] See also next step

2.3 Step Three: Calibration: how to build the dataset

Calibration is the process of assigning numerical values to conditions across the cases. In crisp-set QCA, values can only be 0 or 1, while in fuzzy-set QCA there is room for more fine-grained scales, for example 4-point scales like 0-0.33-0.67-1 or 6-point scales like 0-0.2-0.4-0.6-0.8-1. Fuzzy-set QCA does not accept middle values like 0.5, which can be used in *multi-variate* (Cronqvist & Berg-Schlosser, Multi-Value QCA (mvQCA), 2009) QCA (mvQCA). The latter, however, uses an algorithm which makes synthesis (minimisation) increasingly difficult as the number of categories increase (Cronqvist & Berg-Schlosser, Multi-Value QCA (mvQCA), 2009).

In crisp-set QCA, 0 and 1 denote (respectively) absence and presence of a certain construct that defines the condition or the outcome. In fuzzy set QCA, values describe the degree of presence of a factor, which makes "zero" indicate its complete absence and "one" indicate its full presence. Hence zero will mean "zero degrees of presence" while other values will denote intermediate degrees of presence of that factor (or that outcome).

Calibration is an extremely important moment in QCA: all findings from subsequent steps are based on the numerical values identified here and might be highly dependent on those. If we calibrate the data differently, we might obtain very different findings, although this does not always happen. In other words, QCA findings might be highly sensitive to calibration; hence changing the calibration method is part of the *sensitivity* analysis (see Section 3.3).

Since QCA is a qualitative approach, the available data will be textual or mixed (text + numbers/pattern data). Transforming this data into numerical values requires creating rubrics, or qualitative descriptors of each numerical value.

For example, in the AREAP evaluation, the condition "Political Pluralism" was calibrated as follows:

Table 14: Example of a condition calibrated using a fuzzy (non crisp) approach

1	Multiple strong political parties (4 or more)
0.67	Multiparty, with 3 strong parties.
0.33	State with two dominant political parties
0	State with one dominant political party

In the MAVC study, the three outcomes were calibrated as in Table 15.

Table 15: Definitions of achievement and non-achievement for each outcome

Achievement of outcome 1 <u>Successful ICT reporting:</u> Users or their representatives, including government staff, directly or indirectly, use ICTs in the way specified by the initiative to report water supply functionality to the local government authority or relevant stakeholder; this could be either through ad hoc crowdsourcing or through government- or service provider-led, regular updating mechanisms.	Non-achievement of outcome 1 <u>Unsuccessful ICT reporting</u>: Users, or their representatives fail to use ICTs to report water supply functionality, or bypass the ICT channel using other forms of communication with the local government authority or relevant stakeholder.
Achievement of outcome 2 <u>Successful processing of ICT reports:</u> Local government authority (national sector government, if relevant) or service provider process ICT reports.	Non-achievement of outcome 2 <u>Unsuccessful processing of ICT reports</u>: Local government authority (national sector government, if relevant) or service provider do not process ICT reports.
Achievement of outcome 3 <u>Successful service improvement:</u> Water points are repaired based on ICT reports and processing.	Non-achievement of outcome 3 <u>Lack of service improvement:</u> Water points are not repaired based on ICT reports and processing.

Source: Welle, Williams, Pearce, & Befani (2015), page 15

In evaluation, calibration can be a strongly value-driven enterprise, and requires evaluation teams, usually in consultation with project staff and relevant stakeholders, to define not just success, but also lack of it and perhaps degrees of it. It makes teams think of causal factors as non-linear, step-like functions with thresholds that make a difference to the outcome. The mere calibration phase can be useful in itself, even without creating the dataset, let alone synthesizing the information in it.

The exercise can also be excessively challenging, particularly when several degrees of presence need to be defined and assigned numerical values when trying to explain a multiple point-scale outcome. But it can also be difficult with simpler scales where the boundary between different outcomes and different degrees of one single outcome is unclear.

These opportunities and challenges are discussed in more detail below, including pros and cons of fuzzy vs. crisp sets. Multi-variate QCA is not discussed in detail because this variant of QCA has attracted special controversy (Vink & Van Vliet, 2009; Thiem, 2013;

Vink & van Vliet, 2013). It might be wise for evaluators to wait until scholars reach a broader agreement on the usefulness of this variant, which is starting to offer some of the advantages shared by the other, more established ones (Thiem, 2015). Section 2.3.2 illustrates two strategies that can be used for calibration.

2.3.1.1 OPPORTUNITY: empowering teams to define success (and other constructs) on their own terms

Since explaining and causally attributing the success of an intervention is of high interest in evaluation, in many cases the value "1" for an outcome condition will be associated with a definition of success, and the "0" with a definition for lack of success. In fuzzy set QCA, values will denote the degree of success of an intervention, with "zero" indicating the lowest grade. Furthermore, causal factors won't be thought of having linear dynamics but more as step-like trends that make a substantial difference to an outcome if a certain benchmark is reached.

Even before data collection starts, establishing thresholds and building rubrics can be a useful, empowering exercise for evaluation teams and other stakeholders, if interested parties are involved through a democratic process and their views are incorporated. Stakeholders can be offered a chance to sharpen constructs and define success on their own terms – particularly in those evaluations that start with unclear, vague or excessively broad concepts. At the same time calibration is ridden with pitfalls and risks.

2.3.1.2 PITFALL: when several cases fall close to the middle point[25]

In a typical Boolean calibration, 1 is assigned to the presence of a given concept or construct in the case, for example "child-labour free zone", while 0 is assigned to its absence ("zone not free from child-labour"). Since the mathematical basis for QCA is *set theory*, "absence" is to be intended as the negation of presence, or as "any other situation

[25] When several cases fall close to the middle point, one option to consider is multi-variate QCA (mvQCA), which accepts the value of 0.5. However at the current state of knowledge this choice can be problematic for the reasons outlined in the introduction to Step Three.

outside presence"; so that *the logical union* of presence and absence covers the whole set of cases. In other words, when using crisp-set QCA, presence and absence should be thought of as the only two possibilities: all cases should be coded as either one or the other. They should also be fully separate i.e. with zero overlap: fuzzy boundaries between the two categories increase the risk of mis-labelling/mis-categorisation, which puts internal validity in danger.

When it's difficult to identify clear qualitative differences among cases, and in particular when it's difficult to divide cases in two separate groups, the use of fuzzy sets should be considered. However, if several cases populate the area close to the middle point (e.g. would be scored 0.4 or 0.6), even the use of fuzzy-set QCA is risky because the scoring process might not be able to take measurement errors into account, thus compromising the internal validity of the findings. The benchmark chosen should be "safe" enough to make the evaluator confident that, for example, a case that scores as successful is indeed successful.

In an evaluation of child-labour free zones (Millard, Basu, Forss, Kandyomunda, McEvoy, & Woldeyohannes, 2015), a successful outcome (a child labour free zone) had been initially defined as a zone meeting at least 4 of 7 criteria: this rubric produced 24 cases recognised as successful. However meeting each of these criteria had a different meaning in each case which was not always easy to grasp precisely with the available data. The team felt they weren't always too certain that a certain case was meeting a specific criterion for the right reasons, which made the state of the boundary cases (those meeting 4 criteria) problematic. As a consequence, the team decided to raise the bar from 4/7 to 5/7 criteria met, in order to increase their confidence that cases labelled "one" were actually successful.

This change reduced the number of cases considered successful to 15, which was still high enough to allow the use of QCA synthesis procedures for this evaluation[26]. But in other evaluations with more uncertainty or with a smaller set of cases, transitioning to a stricter definition of success might reduce the number of successful cases substantially, for example to 2 or 3, making use of QCA software for synthesis less useful.

[26] A subsequent revision of the rubric removed one of the seven criteria and defined successful those cases meeting at least 4 out of 6, which produced a group of 22 successful cases.

2.3.1.3 PITFALL: providing qualitative descriptors of values between 0 and 1

Being based on *dichotomous* data is considered one of the biggest weaknesses of crisp-set QCA; however, since QCA is a qualitative method and numerical values need to be associated to qualitative descriptors and concepts, using fuzzy sets sometimes overcomplicates the procedure. The values between 0 and 1, just like the extremes, need to be precisely defined and hold a specific qualitative meaning (Schneider & Wagemann, 2012); in addition, they need to be ordered, with "1" representing an "ideal type", "0" the corresponding worst case, 0.2 a less desirable situation that 0.4, 0.6 than 0.8, and so on. In practice it can be extremely challenging to order qualitative characteristics of a case as required above. One typical problem is that we are unable to determine if a situation is less desirable than another, and deserves a lower value of the same condition, or it is simply qualitatively different, in which case it deserves a different condition.

For example, when calibrating a condition related to normative change about child labour, there was an ambiguity between two states being two different conditions or two values of the same: "neighbouring communities change their norms" and "institutions are sensitised to the need to reduce the barriers to communities changing norms (around the idea that no child should work)" were initially considered two different conditions; a later review however established that the latter can be considered an initial step toward the former, as describing a process where communities change their norms after institutions are sensitised to contribute in this direction; and was considered part (a less desirable state) of the same condition (Millard, Basu, Forss, Kandyomunda, McEvoy, & Woldeyohannes, 2015).

Another problem is that, because of how the fuzzy-set procedures work (see Section 2.6), small differences in the numerical values assigned during calibration can end up determining inclusion (or exclusion) of the combination in the Truth Table, affecting the membership scores of cases to a "crisp" combination; or declaring a certain configuration as sufficient instead of necessary (or viceversa) for the outcome (for example if the outcome is 0.8 and the condition can be 0.7 but also 0.9). Given these risks, it's important to be able to justify inclusion vs. exclusion or sufficiency vs. necessity on a theoretical basis.

If the above risks do not materialise, the application of fuzzy sets is easier; however, it might also mean that the conditions (and related data) are more crisp than fuzzy. When this happens, the findings from the fuzzy analysis might be very close, if not identical, to the findings from the crisp set analysis, because the rows considered eligible for inclusion in the Truth Table might be (largely) the same. This is what happened in the AREAP evaluation.

2.3.2 Calibration Strategies: "how to assign numerical values to conditions across the cases?"

Ideally, rubrics should be consistent with available, established theory and at the same time measure the diversity observed in the empirical data set. However, fulfilling both criteria might not always be possible; either because the available cases do not represent the full spectrum of possibilities (and using theory-consistent rubrics cases would be labelled either mostly "ones" or mostly "zeros"), or because available theory is not developed enough to capture the empirical diversity observed in the dataset. These two principles (consistency with established theory and coverage of empirical information) underpin two different calibration strategies: the Theory-Consistency and the Empirical-Coverage strategies. The two strategies are not mutually exclusive and, when possible, calibration should both reflect theory well and fit empirical data adequately.

2.3.2.1 The Theory-Consistency Calibration Strategy

The first strategy consists of the following steps:

- Define "1" on the basis of a commonly accepted/well-known/widely understood definition of the condition of interest in literature or practice
- Define the other scores in relation to the above, taking the definition of "1" as a starting point.

This strategy makes data easily comparable across studies; however, the "0" category might include very different situations that the evaluator might not be happy associating with the same value. In other words the first strategy might not fit the empirical data as well as the evaluator might like, particularly if s/he wishes to explain unsuccessful cases using the same dataset.

2.3.2.2 The Empirical Coverage Calibration Strategy

The second calibration strategy consists of the following broad steps:

- Define “1” and “0” on the basis of the “best” and “worst” case(s) included in the dataset (the “empirical extremes”)
- Define the other scores (if you are not using crisp-set QCA) on the basis of the distance of the other cases from the two empirical ends.

In contrast to the first one, the second calibration strategy tends to work well when theory is poorly developed or does not reflect the empirical cases well; for example, when we are trying to compare the characteristics of the users of a service, while the intervention has specified the target group poorly or when the real users are different from the intended users (though the latter might be well-specified). In this case the conditions describing the characteristics of the user would be better calibrated not on the basis of existing theory but following a more exploratory approach, using the information empirically available about the actual users. The downside is that the findings obtained from a dataset calibrated in this way might not test a well-established theory and as such might be more difficult to compare to other studies.

2.3.2.3 The Raw Data Matrix

The Empirical Coverage Calibration strategy could benefit from creating a so-called “raw data matrix”, which consists in copy-pasting quotes, notes, figures, and other relevant narrative/textual information into the cells of a large matrix, where each cell reports the data available on the condition on its column, for the case on its row. The structure of this matrix is the same as the Boolean dataset (same cases, same conditions) except that the cases are not calibrated/scored yet, and the cells include narrative information instead of Boolean data. At first, the evaluator can proceed on a row-by-row basis, completing one case (row) at a time.

When selecting textual or pattern information to include in the cells, care should be taken to avoid pre-conceived specific notions or definitions of a given condition and include in the cell all the relevant available information for that case: this will allow the evaluator to be open to different definitions of the condition. Instead, if a definition/ rubric is used, it should be as broad as possible at this stage, making room for different meanings and interpretation of the condition.

When all cells are complete, the evaluator compares the information relating to the same condition across all cases, along columns: running through cases/rows for a given column will provide the evaluator with an idea of the kind of cross-case information that is available for a specific condition; which will be the basis for the assignment of “0”, “1” and possibly other values.

At this stage the cells can potentially include a large amount of information, so the process might be cumbersome and time-consuming, possibly not feasible in situations of limited resources. However, it’s important to gradually reduce the cross-case information available as the comparison process selects what information will be included in the definition of the condition.

2.4 Step Four: The Venn Diagram: “how to represent the data graphically?”

From the author’s perspective, the Venn diagram (Meur, Rihoux, & Yamasaki, 2002) (Cronqvist, 2011) is possibly the most important tool in QCA. It displays, in an intuitive and visual form, all available information about the set of cases, as reported in a boolean dataset. A trained eye can spot at a glance relations of necessity, sufficiency, which assumptions can help strengthen such relations, and which cases are covered by which configurations.

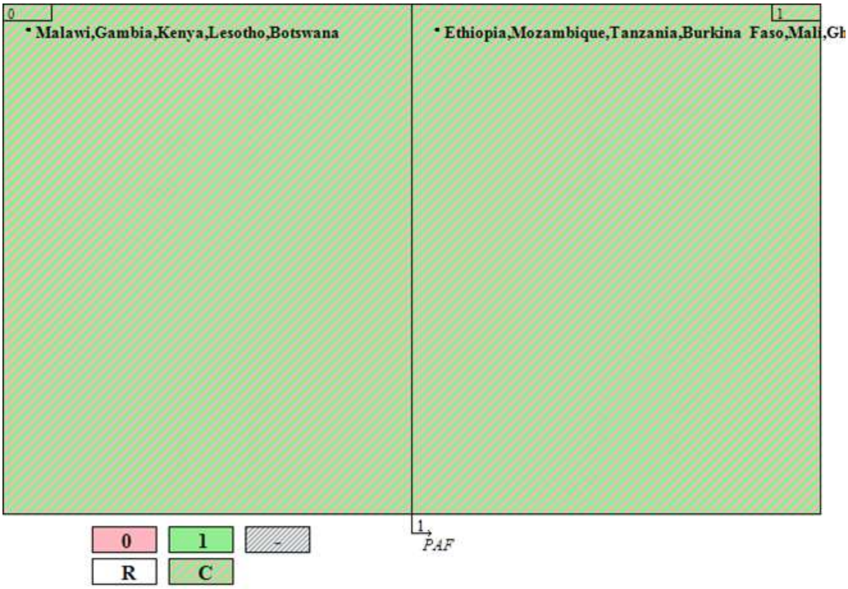
This section will take the reader through the construction of the Venn diagram step by step, starting from one condition, then two, three and finally four. It will then discuss its main limitations: inability to handle fuzzy datasets, inability to show consistency scores, and inability to handle more than 5 conditions (on the Tosmana software).

2.4.1.1 A Venn diagram with one condition

The Venn diagram looks different, depending on how many conditions it represents. Let's start from the simple (almost trivial) case of the Venn diagram representing one condition, in the Budget Support Evaluation (see Annex C) (Holvoet & Inberg, 2013). In Figure 4, the vertical line in the middle divides the bi-dimensional space into two special areas: the right side including countries where the budget support programming document includes gender-sensitive indicators (presence of PAF – Performance Assessment Framework), and the left side countries where the same document does not include such indicators (absence of PAF).

Notice that all cases are either on the left or on the right of the space; in other words, PAF is either present (which would put them on the right) or absent (which would put them on the left). The two areas are both striped because they both include successful and unsuccessful cases (i.e. both countries where primary school enrolment of girls has increased and countries where it hasn't; technically these are called "*contradictory cases*" hence the "C" in the key).

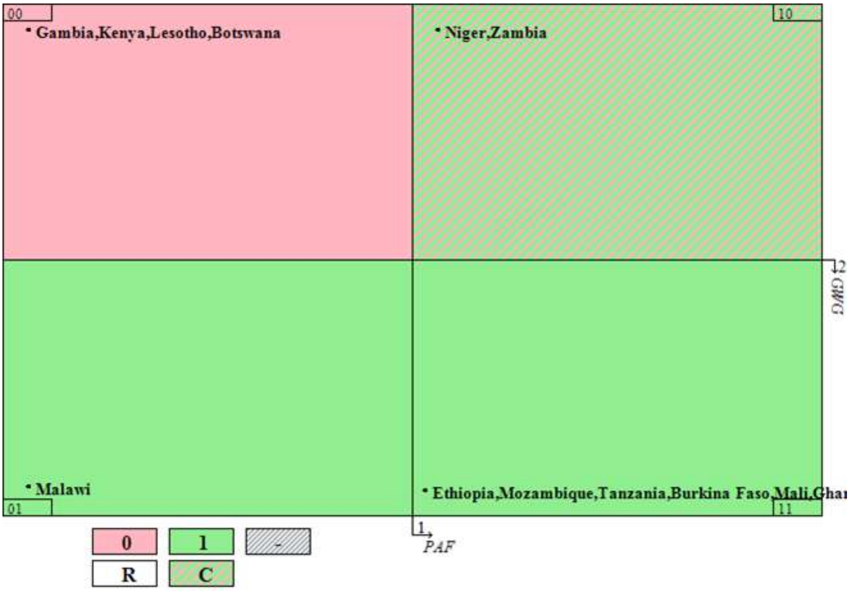
Figure 4: A Venn diagram with one condition


0
• Malawi,Gambia,Kenya,Lesotho,Botswana
1
• Ethiopia,Mozambique,Tanzania,Burkina Faso,Mali,Gh
1
PAF
0
1
-
R
C

2.4.1.2 A Venn diagram with two conditions

Let’s now add a second condition: gender-sensitive working groups (Figure 5). The second condition is a horizontal line in the middle which divides the space into two areas: top and bottom. The cases where these working groups are present are located at the bottom; while countries where these groups are absent are located at the top. Countries are either at the top or at the bottom. Note that this way of dividing the space (with a horizontal line) creates areas which intersect with the other way of dividing the space (the vertical line). In other words, the two special areas of the first condition, “right and left”, intersect with the two special areas of the second condition, “top and bottom”, to create a space with four quadrants: top-left, top-right, bottom-left and bottom-right. Each quadrant includes countries displaying a specific combination of presence or absence of the conditions.

Figure 5: A Venn diagram with two conditions


00
• Gambia,Kenya,Lesotho,Botswana
10
• Niger,Zambia
2
GWG
• Malawi
01
• Ethiopia,Mozambique,Tanzania,Burkina Faso,Mali,Ghai
11
1
PAF
0
1
-
R
C

For example, in the top-left quadrant, Gambia, Kenya, Lesotho and Botswana display absence of both conditions, which is also denoted in the “00” in the corner; in the top-right, Niger and Zambia have indicators in the main planning document (PAF is positive) but do not have gender-sensitive working groups (the second condition

GWG is negative); which is denoted by the “10” in the corner. Most countries display the presence of both conditions (“11” in the corner) and are thus located in the bottom-right quadrant.

Notice that only one quadrant is now striped, while the other three are either pink or green. The green areas represent combinations which are consistently successful (all cases included present a positive outcome, see key with “1”) while all the pink quadrant’s cases display a negative outcome (see key with “0”). The top-right quadrant is striped (key says “C” for “contradictory”) because one case with that combination is positive (Niger) and the other negative (Zambia).

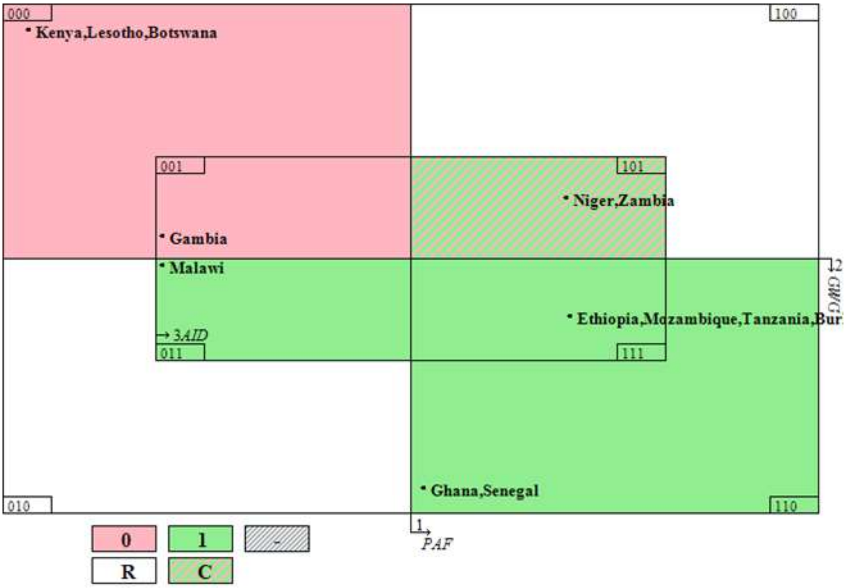
2.4.1.3 A Venn diagram with three conditions

If we want to analyse combinations of three conditions we can add a third special area, this time represented by a central rectangle in Figure 6. The cases where the third condition is present will go inside the rectangle, while the cases where it’s absent will go outside. The third condition represent levels of aid for primary education, so Kenya, Lesotho, Botswana, Ghana and Senegal, which present low levels of aid, all sit outside the central rectangle, while all the other cases with comparatively higher aid are inside.

Notice that drawing this rectangle in the space brings the total number of areas from 4 to 8: each quadrant is now divided into two areas (inside or outside the central rectangle) offering the possibility to locate cases more precisely. For example. The top-left quadrant including four cases is now divided in two areas: one with Gambia (see the “001” in the corner) which has high levels of aid and sits closer to the centre; and the other with Kenya, Lesotho and Botswana which have low levels of aid. Notice that the old “00” in the corner now has an additional, third condition and reads “000”.

The majority of cases still belong in the bottom-right quadrant, but now the highly populated “111” area has been distinguished from the “110”, the latter describing the situation of Ghana and Senegal. Note that, while we still have both pink, for consistently unsuccessful; green, for consistently successful; and striped for cases with an inconsistent outcome, we also have two white or “blank” areas: “010” and “100”. White means that these combinations are not empirically supported in the dataset and are merely theoretical (“R” in the key stands for “*remainder*”: see Section 2.6.2 for more details).

Figure 6: A Venn diagram with three conditions

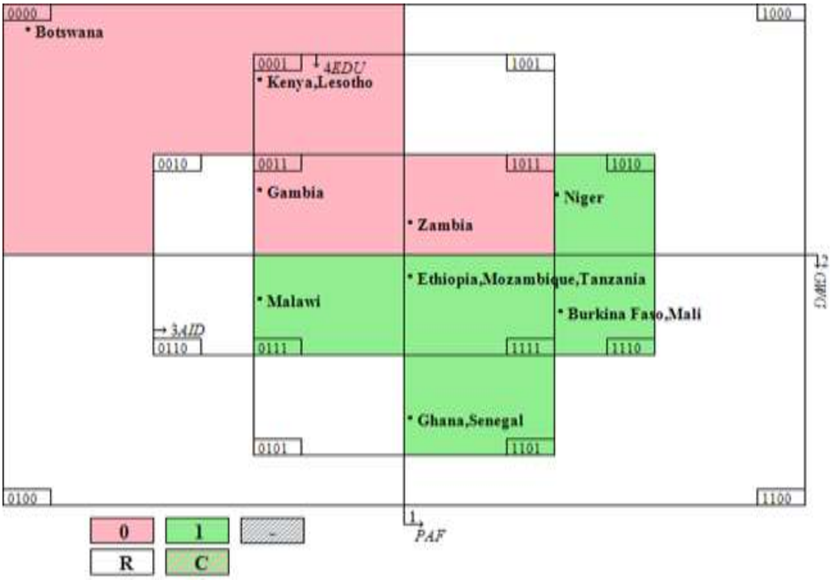


2.4.1.4 A Venn diagram with four conditions

Adding a fourth condition (e.g. free education, or EDU) in the form of another central, but this time vertical instead of horizontal, rectangle, doubles the number of areas to 16; the same as all the logically possible combinations of ones and zeros in a four-term sequence (Figure 7). In this particular case it also removes the striped areas: no combination of four conditions is inconsistent. It is either consistently successful, consistently unsuccessful, or not supported by the dataset.

In summary, the Venn diagram divides the bi-dimensional space into a number of "special areas", each representing presence or absence of a single condition included in the model. Every special area can intersect with every other special area (top with right, inside-rectangle with left); just like single conditions can form combinations with any other condition. The intersections between two or more special areas represent the conjunctions/combinations between two or more conditions. We will see later in the chapter how to combine the information on the outcome (the colours) with the information on the conditions (location of the cases).

Figure 7: Venn diagram with four conditions



While the benefits of the Venn diagram will be clearer in the next sections, it is already possible to see its limitations: the inability to display fuzzy scores, the maximum limit of 5 conditions (in the Tosmana software), and the failure to indicate the level of consistency of those combinations with inconsistent outcomes.

2.4.1.5 PITFALL: can only handle crisp datasets

Technically, the Venn diagram is only available for crisp set QCA. The software returns the diagram after analysing a Boolean dataset of zeros and ones. If we want to use the diagram when working with fuzzy sets, we need to "crisp" the dataset first, which is what was done in the AREAP evaluation (Baptist, Edouard, & Batran, 2015). Another option is to use the diagram to display, not the dataset, but the Truth Table (see Section 2.6), which is always crisp even in fuzzy-set QCA. So the diagram cannot be used to represent a fuzzy dataset graphically, but it can be used to represent the Truth Table obtained towards the end of a fuzzy-set QCA analysis.

2.4.1.6 PITFALL: works with a maximum of 5 conditions

Perhaps the most glaring constraint of the diagram is that it gets increasingly complicated the more conditions are added, particularly with 5 or more conditions; the Tosmana software does not even handle more than 5 conditions. This limitation is mitigated by the fact that findings from models with more than 5 conditions tend to be more difficult to interpret for the evaluator, and every analysis includes at least some relatively simple model of 3 to 5 conditions (which can then be visualised with the diagram).

At the time of writing, a new type of Venn diagram using oval shapes which handles more than five conditions has just been released by Adrian Dusa and is available in the R software (Dusa, 2007) under the "venn[27]" package.

2.4.1.7 PITFALL: does not show levels of consistency

While pink or green areas are fully informative on the consistency of the associated combinations (either all successful or all unsuccessful), "striped", which stands for simple "inconsistency" can hide very different situations. As an illustration, let's consider two different combinations, each supported by six different cases. In the first situation, five cases present a positive outcome and one case a negative one; that is, the combination is mostly sufficient for a positive outcome, although not perfectly (83% consistency). In the second situation, 5 cases present a negative outcome and one case a positive one; that is, the combination is mostly sufficient for a negative outcome (same consistency as the first). The Venn diagram will be insensitive to these details and will simply paint both areas with the same, pink-green striped texture.

In the next sections we will see how the Venn diagram can help us interpret the findings of all the QCA procedures: the necessity analysis, the two sufficiency analyses, and the INUS analysis. It also helps with the sensitivity analysis; in particular, it provides an immediate and intuitive understanding of the consequences of adding specific cases to the dataset/combinations to the Truth Table.

[27] Note that "venn" is lower case – the package "VennDiagram" is not as closely integrated with QCA.

2.5 Step Five: the SuperSubset Analysis: "what conditions are necessary or sufficient for the outcome?"

The SuperSubset Analysis is a group of two types of analysis, the *superset* or necessity analysis and the subset sufficiency analysis. These focus on measuring the consistency of association between conditions and outcomes. Even when the relations are perfectly consistent as measured on the dataset, the sufficiency and necessity statements are not to be intended in an absolute sense, but as synthetic descriptors of associations that are observed across the dataset. Sometimes the dataset represents the entire population of existing cases, so no statistical inference or "extrapolation" is needed. When the sample is a subset of the total population, the issue of generalising to a larger population emerges. In Chapter 3 we will see that, under specific conditions, necessity and sufficiency statements can be generalised outside the dataset.

This section addresses in detail only the case of dichotomous data (crisp-set QCA), for simplicity of illustration and also because of the reasons illustrated in Section 2.3.1.3 (in relation to calibration) and 2.6 (in relation to the creation of the Truth Table). When the dataset is fuzzy (includes values that are different from 0 and 1), the consistency and coverage scores are computed differently. We cannot group cases with exactly the same value because cases will likely have many different fuzzy scores. We need to use different formulas (Ragin, 2000; Schneider & Wagemann, 2012). What is important to remember here is that, both in crisp and fuzzy sets, a consistency score of 1 means perfect consistency and a coverage score of 1 means perfect coverage.

Although the two types of supersubset analyses are different and their findings have a different meaning, particularly in relation to evaluation questions and recommendations, sometimes we can easily deduct the findings of one type of analysis (e.g. necessity) from the findings of the other (sufficiency), and viceversa. This is due to DeMorgan's Law, which is covered in Section 2.5.3.3, together with practical suggestions on how to use the software to conduct the analyses in practice.

2.5.1 Step 5A: The Necessity (Superset) Analysis: “what conditions are necessary for the outcome”

The necessity analysis aims at discovering if there are any conditions that necessarily need to be in place for the outcome to materialise. If a causal factor is discovered to be necessary, it can be considered a requirement, or a “pre-condition”: it means that, on the basis of the empirical data available, it’s impossible to achieve the outcome unless the factor is in place.

The section illustrates the basic logic of the necessity analysis, drawing on several evaluation examples. It later discusses available options when no single perfectly necessary conditions is found, for examples disjunctions and consistency/coverage scores.

In crisp-set QCA, in order to identify necessary conditions, the superset analysis groups all successful cases[28] (see Table 16) and searches for common conditions they might share. If the number of cases is relatively low, this can be done by just looking at the dataset, without the help of a software platform. However, using software is recommended: in particular, the R software (Dusa & Thiem, 2014; Thiem & Dusa, 2012; Dusa, 2007) is currently the best option for the necessity analysis, at least among the freely available platforms (see http://www.compasss.org/software.htm).

The logic of the necessity analysis can be illustrated in a 2x2 table as follows (Schneider & Wagemann, 2012): the key indicator of necessity is that no cases are observed where the outcome is present and the configuration is absent.

Table 16: Logic of the necessity analysis

A condition or configuration is (perfectly) necessary if		Outcome is	
		Present	Absent
Configuration is	Present	Some cases	Not relevant
	Absent	No cases	Not relevant

This section will illustrate the opportunities offered by the necessity analysis in three evaluations, including using the Venn diagram, and will suggest courses of action when no single condition is

[28] or all unsuccessful cases: the procedure is exactly the same when the evaluator wants to discover the necessary conditions for lack of success.

necessary (considering consistency scores and analysing the necessity of disjunctions/unions of conditions). Note that when more than one condition is necessary, then the combination of all necessary conditions is also necessary. The section also warns against the "triviality" pitfall.

2.5.1.1 OPPORTUNITY: Necessity relations as displayed in a Venn diagram

The Venn diagram shows the findings of the necessity analysis at a glance. When looking at the diagram, the "necessity question" is "where are the green areas located?" In particular, "are the green areas located within one or more special areas (corresponding to single conditions)?" "What overall shape do the green areas take?" In the last chart above, the green areas are all included in the union (disjunction) between the bottom area and the right area; which means they are either at the bottom or on the right. More information about how the Venn diagram helped locate necessary conditions in the three evaluations mentioned in this section is provided below.

More generally, if all the green areas are located in the right side of the space, it means that the presence of the first condition is necessary for success, because no successful case is observed in the area corresponding to absence of the first condition. It they are all located at the bottom, it means that it is the second condition that is necessary, because no successful case is observed in the area corresponding to absence of the second condition. And so on. If the green areas are located in two or more special areas corresponding to the presence of one or more conditions, then the union (disjunction) of these conditions will be necessary; or in other words in order to be successful/green it is necessary for a case to be located in one of these areas.

2.5.1.2 Evaluations were the Necessity Analysis provided important findings

In the MAVC study[29], all four conditions included in the model explaining repairs (outcome 3) were initially assumed to be necessary for repairs to be made. However, the necessity analysis revealed that only three of these were necessary: the presence of accountability mechanisms to ensure that repairs are made (ACCMEC, see also Table 18), the availability of spare parts (SPAREP), and the alignment of the ICT initiative with existing responsibilities in the sector (EXRESP). Unlike these, the other condition included in the model was not consistently observed across all four successful cases: in other words, the fact that the intervention provided sufficient funds for the repair to be made was not necessary, although present in 3 of the 4 successful cases (see Table 17 and Figure 8).

In Tanzania repairs were made but – although there were many schemes across a district, and some committees had sufficient funds while others didn't – overall, the judgement was that funds tended not to be sufficient for repairs. This was explained with the assumption that the district water engineers, who got a copy of the text message, then followed up in different specific ways to make sure that repairs were carried out, which did not happen in other cases with insufficient funds. This could have included encouraging the committee to collect sufficient funds or even helping out with budget from the district office. This assumption however was not tested in-depth as the two follow up studies focused on other countries.

[29] Notice that, unlike in the Access to the Health System for the Poor example in section 1.3.2, the theory of change in the MAVC is less well developed initially, and amounts essentially to a list of conditions identified following reasonable and logical assumptions, rather than a well-known, extensive body of literature.

Table 17: Dataset for the QCA model explaining Outcome 3 in the MAVC study (repairs being made) showing necessary conditions

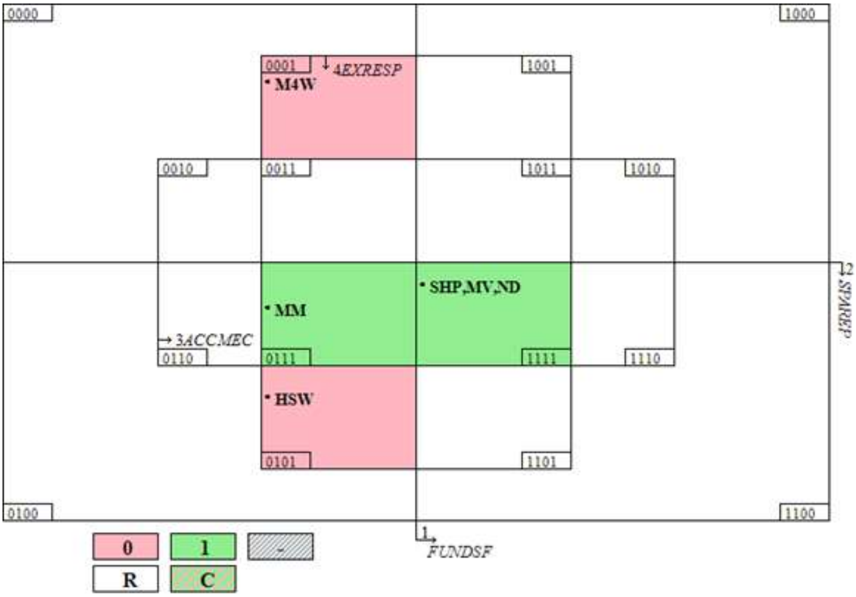
Project	FUNDSF	SPAREP	ACCMEC	EXRESP	REPAIR
SHP	1	1	1	1	1
M4W	0	0	0	1	0
MM	0	1	1	1	1
MV	1	1	1	1	1
ND	1	1	1	1	1
HSW	0	1	0	1	0

Notice that if three conditions are necessary, their combination is also necessary: indeed, in the Venn diagram (Figure 8) all green areas are located in the intersection of the bottom area (spare parts) and the two central rectangles (sector responsibilities and accountability mechanisms).

Table 18: Necessity Table for the ACCMEC condition

		REPAIR is	
		Present	Absent
ACCMEC	Present	4	Not relevant
	Absent	No cases	Not relevant

Figure 8: Venn diagram for the model FUNDSF + SPAREP + ACCMEC + EXRESP = REPAIR



Similarly, in GEF IEO, (2015) (the GEF/UNDP Biodiversity evaluation, see Annex C) the necessity analysis provided important findings for the outcome “decrease in trends of illegal activities within the protected area”. In all the 27 successful protected areas, out of a total 30, activities were in place to provide information to communities: out of the many conditions included in the model, this was the only necessary one (see Tables 19 and 20).

Table 19 Excerpt from the dataset of the GEF-UNDP Biodiversity Evaluation showing necessary conditions

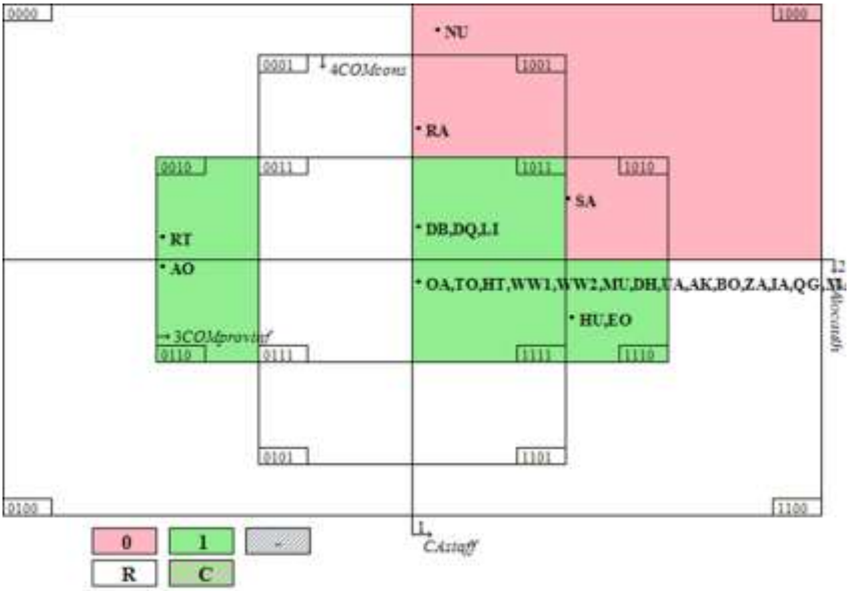
Protected Area	CAstaff	CAlocauth	COMprovinf	COMcons	BIOtrend
OA	1	1	1	1	1
RA	1	0	0	1	0
TO	1	1	1	1	1
DB	1	0	1	1	1
HT	1	1	1	1	1
WW1	1	1	1	1	1
WW2	1	1	1	1	1
MU	1	1	1	1	1
DH	1	1	1	1	1
DQ	1	0	1	1	1
LI	1	0	1	1	1
RT	0	0	1	0	1
UA	1	1	1	1	1
NU	1	0	0	0	0
AK	1	1	1	1	1
BO	1	1	1	1	1
SA	1	0	1	0	0
ZA	1	1	1	1	1
HU	1	1	1	0	1
AO	0	1	1	0	1
IA	1	1	1	1	1
EO	1	1	1	0	1
QG	1	1	1	1	1
MA	1	1	1	1	1
KA2	1	1	1	1	1
AI	1	1	1	1	1
AU	1	1	1	1	1
RO	1	1	1	1	1
NA	1	1	1	1	1
PZ	1	1	1	1	1

Table 20: Necessity Table for the COMprovinf condition

		BIOtrend is	
		Present	Absent
COMprovinf	Present	27	Not relevant
	Absent	No cases	Not relevant

The Venn diagram in Figure 9 shows that none of the 27 successful cases lie outside the central horizontal rectangle.

Figure 9 illustration of the 4-condition model explaining BIOtrend



2.5.1.3 OPPORTUNITY: Necessity of Disjunctions (and SUIN causes)

When no single condition is necessary, one way to extract meaningful necessity-related information from the dataset is to pay attention to the necessity of “disjunctions”, or logical unions of conditions. As we add more conditions to a disjunction, the chances of that disjunction being necessary increase. In practical terms this means that, when no single condition is invariably observed in successful cases, there is a chance that at least one out of two always is; and even a higher chance that one out of three always is, and so on. Disjunctions represent "equivalent requirements" for success: on the basis of the empirical data, at least one term of the disjunction needs to be present in order to achieve success.

In the Budget Support Evaluation[30], neither of the two policy instruments analysed are necessary for success (see Table 22): but their logical union (PAF + GWG) is. This means that either gender indicators in the PAF or gender working groups (GWG) are required in order for primary school enrolment of girls to increase in the countries analysed (see Table 21).

Table 21 Dataset from the Budget Support Evaluation showing necessary conditions

Country	PAF	GWG	AID	EDU	OUT
Ethiopia	1	1	1	1	1
Mozambique	1	1	1	1	1
Tanzania	1	1	1	1	1
Burkina Faso	1	1	1	0	1
Mali	1	1	1	0	1
Ghana	1	1	0	1	1
Senegal	1	1	0	1	1
Malawi	0	1	1	1	1
Niger	1	0	1	0	1
Zambia	1	0	1	1	0
Gambia	0	0	1	1	0
Kenya	0	0	0	1	0
Lesotho	0	0	0	1	0
Botswana	0	0	0	0	0

In the Venn diagram, the necessity of the disjunction PAF + GWG is reflected in the shape of the "green patch": spanning both the bottom (GWG) and right (PAF) areas of the diagram. In order to be successful/green, it is sufficient that a case is located in either the bottom OR the right (Figure 10).

[30] Notice that – unlike in the MAVC study and the GEF/UNDP evaluation – the theory of change was quite well-developed for this evaluation, and an extensive literature review was carried out before selecting the conditions to include in the analysis.

Figure 10 Graphic illustration of the dataset for the Budget Support Evaluation

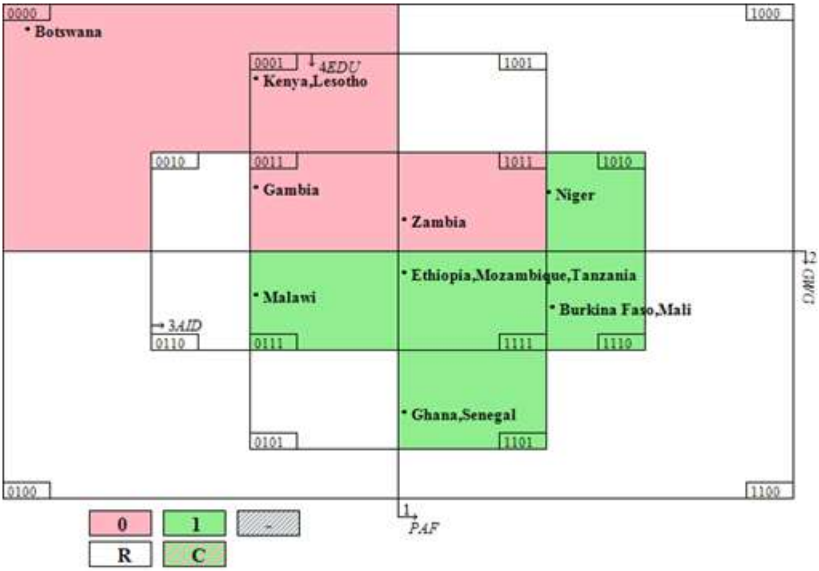


Table 22: Necessity Table for the GWG condition (it would be the same table for PAF)

		OUT is	
		Present	Absent
GWG	Present	8	Not relevant
	Absent	1	Not relevant

The opportunity to identify necessary disjunctions or SUIN causes for an outcome is important in evaluation because the intervention might not be necessary to achieve that outcome, but we might be able to show that it is a SUIN cause, or a sufficient part of a necessary (but insufficient) disjunction. Saying that an intervention is one of two factors, as in this case, or one of three factors, one of which is at least required for success, is more informative that simply saying it is not required for success by itself.

2.5.1.4 OPPORTUNITY: Parameters of Fit (in crisp-set QCA)

When no single condition is necessary, another way to extract meaningful necessity-related information from the dataset is to pay attention to the so-called *parameters of fit*: **consistency** and **coverage**. These have a different meaning depending on whether they are being used for the superset (necessity) or subset (sufficiency) analysis.

For the necessity analysis, the consistency of a condition (a.k.a. necessity-consistency) is the number of successful cases that condition is observed in, divided by the total number of successful cases. This indicator takes a minimum value of zero (when the condition is never observed in successful cases) and a maximum value of one (when the conditions is observed in all successful cases and is thus perfectly necessary). Intermediate values of necessity-consistency like, for example, 0.9 mean that the condition is present in 90% of successful cases.

In the GEF/UNDP Biodiversity evaluation, the presence of adequate staff is not necessary for success, because it is observed in only 25 of the 27 protected areas with a positive outcome. However, its necessity consistency score is high: 0.926 (or 25/27) which means that 25 out of a total 27 successful cases present the (almost necessary) conditions.

The notion of consistency in the necessity analysis can again be illustrated by a 2x2 table, as in Table 23 (Schneider & Wagemann, 2012): note that unlike in the case of perfect necessity, we can observe some cases where the outcome is present which do not present the configuration.

Table 23: The notion of consistency in the necessity analysis

A condition or configuration is necessary to some degree if	Outcome is	
	Present	Absent
Configuration is Present	Some cases (many, e.g. 90%)	Not relevant
Absent	Some cases (few, e.g. 10%)	Not relevant

Following this logic it might appear than any condition or configuration is necessary to some degree, which might seem to devoid the notion of necessity of meaning. However, consistency scores are only introduced to make sense of those situations which are very close to perfect necessity, with scores over 85% or 90%, if not higher. Saying that a condition is 50% necessary means that its presence is not more required than its absence for the outcome; but saying that it is 95% necessary means that – although the condition is not absolutely required – it's difficult to imagine a positive outcome without it. Which is more informative that saying that condition is "simply not necessary".

The opposite of (1 minus) the necessity-consistency score can be interpreted as the chance of the outcome being observed without observing that condition as well. If the consistency is 95%, it means that there is only a 5% chance of observing the outcome without observing that condition together with it. It's not impossible, as would be with perfect, 100% necessity; but still unlikely.

In the example from the GEF/UNDP Biodiversity evaluation above, having adequate staff is not absolutely necessary for success: however, success without adequate staff is very unlikely. Only 7% of the successful cases present inadequate staff (see Table 24).

Table 24: Necessity Table for the CAstaff condition

		BIOtrend is	
		Present	Absent
CAstaff	Present	25	Not relevant
	Absent	2	Not relevant

Put differently, using consistency scores for necessity only makes sense if the number of cases in the bottom left quadrant in Table 24 is low compared to the number of cases in the top-left quadrant: if these numbers are similar, it means that successful cases have similar chances of presenting and not presenting the condition.

2.5.1.5 PITFALL: Triviality of conditions and necessity-coverage

When a condition is observed to be perfectly necessary for a positive outcome, this discovery might not necessarily be very informative or insightful. When the same condition is also necessary for the negative outcome, it means that the condition is required in all cases independently of the outcome: in other words, it does not help discriminate between successful and less successful cases. It doesn't allow us to fully understand what makes the difference between a positive and negative outcome and how we can improve the programme to increase our chances of being successful: it is required for success as much as for its absence. This is why conditions which are required in both positive and negative cases are called "trivial". They might provide important information, but obviously we want to know more.

In order to evaluate how insightful a necessary condition is, it is advisable to measure a second parameter of fit, called "coverage". Coverage (a number between 0 and 1) is also measured in the sufficiency analysis, hence in this section we address the "necessity-coverage".

The necessity-coverage is obtained by computing the % of successful cases within the group of cases that present the necessary condition: it measures the "exclusivity" with which a condition is necessary for the positive outcome, or the extent to which it is "non-trivial". Table 25 illustrates the notion of necessity-coverage: when the cases presenting the configuration are only or mostly successful (only or mostly present a positive outcome) and the top-right quadrant includes either no or a low number of cases (compared to the top-left quadrant), necessity-coverage is high and the configuration is non-trivial.

Table 25: Coverage in the necessity analysis

A condition or configuration is perfectly necessary and has perfect (good) coverage if		Outcome is	
		Present	Absent
Configuration is	Present	Some cases	No (or few) cases
	Absent	No cases	Not relevant

For example, in the MAVC study, the condition "the ICT initiative supports sector responsibilities" (EXRESP) was included in the model and discovered to be necessary (present in all four cases with a positive outcome): however it was also always present in the two cases with a negative outcome (see Tables 17 and 26). In other words, it is always present, in all six cases analysed, independently of the value of the outcome. Its necessity-consistency is obviously 1.00, but its necessity-coverage is much lower (0.67). It means that only two thirds of the cases presenting that condition are successful; the other third aren't.

Table 26: Necessity Table for the EXRESP condition

		REPAIR is	
		Present	Absent
EXRESP	Present	4	2
	Absent	No cases	Not relevant

When necessity-coverage is low, the condition can be trivial and perhaps equally necessary for success as for its absence. When it is 0.5, it means that – out of all the cases presenting that condition – half are successful and half aren't. By contrast, when necessity-coverage is high, we are reassured that the condition is necessary for success, while not being necessary for lack of success. In the extreme case of the parameter of fit having the value of one (and the top-right quadrant presenting zero cases), the condition is not only necessary but also sufficient for success – whenever it is observed, the case is always successful. When both consistency and coverage equal 1.00, the condition is both perfectly necessary and perfectly sufficient for success (this applies also to the subset sufficiency analysis). This was the case of "accountability mechanisms" (ACCMEC) in the MAVC study presented above.

In practice, the assessment of triviality will not only depend on the measurement of consistency and coverage, but also on what else we know about the condition. If we have reason to believe that it can't have a strong causal link with failure, its consistent presence with both a positive and a negative outcome might mean that in the latter cases it still hasn't been able to realise its potential. This can happen when outcomes are measured prematurely as in the case mentioned in Section 2.1.1.4.

2.5.2 Step 5B: The Sufficiency (Subset) Analysis: "What conditions are sufficient for the outcome?"

The goal of the subset analysis is to assess the sufficiency of (combinations of) causal factors (conditions) for the outcome. It is one of two types of sufficiency analyses, both aimed at measuring how consistently combinations lead to a certain outcome. While the Boolean minimisation (the best-known QCA procedure, see Section 2.7) is conservative in terms of reducing the number of conditions representing cases, the subset analysis measures the sufficiency-consistency (and coverage) of any possible combination of causal factors, including single conditions and simple combinations of conditions. In other words, while the Boolean minimisation tends to consider cases as "wholes" and it is relatively conservative in simplifying the way cases are described, the subset analysis covers the sufficiency of a variety of groups of conditions equally, even single conditions and combinations of two or three. In the latter the values of all the other conditions are ignored and do not have any bearing on the findings.

When the subset analysis declares a condition (or a combination) sufficient for success, it means that whenever that condition (or combination) is observed in the dataset, the outcome is always positive. This is a key insight for policy making: while necessary conditions inform about the required, but not sufficient, ingredients, the sufficient conditions might not be required but, when met, guarantee, on the basis of available information, a successful outcome. A sufficient condition (or combination) doesn't have to be (and in most cases won't be) necessary: in most cases the outcome is achieved with different pathways and "recipes", each sufficient but none strictly required.

This section covers some of the opportunities offered by the subset analysis, both in theory, like Venn diagram visualisation, what to do when no single condition is subset-sufficient (e.g. measuring consistency/coverage and examining combinations of conditions); and in practice, with examples from 3 evaluations. It also invites the evaluator to check how much a condition or combination is representative of the dataset, before rejoicing about its perfect sufficiency. The "within-dataset" representativeness of a combination or condition is not to be confused with "outside-the-dataset generalisation", which is covered in Section 3.1.

In crisp-set QCA, in order to identify sufficient conditions, the subset analysis groups all cases where a given (combination of) condition(s) is observed and measures the frequency with which that group of cases presents a positive outcome (or a negative outcome if we are looking for conditions that guarantee lack of success). If the number of cases is relatively low and we are interested in the sufficiency of a small number of single conditions or combinations of two, this can be done by just looking at the dataset in excel, without the help of a specific software platform. However, using software is always strongly recommended, and necessary, if we are interested in testing the sufficiency of several conditions or combinations, particularly across a large dataset.

Table 27: Logic of the subset sufficiency analysis

A condition or configuration is (perfectly) (subset) sufficient if		Outcome is	
		Present	Absent
Configuration is	Present	Some cases	No cases
	Absent	Not relevant	Not relevant

The notion of subset sufficiency can be represented by a 2x2 table as in Table 27 (Schneider & Wagemann, 2012): the key indicator of perfect sufficiency is that no cases are observed where the configuration is present and the outcome is absent.

2.5.2.1 OPPORTUNITY: Subset Sufficiency in Venn diagrams

Many of the Tosmana Venn diagrams illustrated above show conditions which are sufficient in a subset sense. The evaluation of general budget support on gender equality shows that setting up gender working groups appears sufficient for an increase in female primary school enrolment: all countries where these groups are set up, located in the bottom area of the Venn diagram, are painted green (Figure 11). This condition is not necessary, as the green rectangle representing Niger and lying on the upper side shows; but it is subset sufficient as no "pink" areas can be spotted in the bottom side of the diagram.

Once the evaluator becomes familiar with the special areas of the Venn diagram, the sufficient conditions become visible at a glance, where by special areas we mean right/left, bottom/top, inside/outside horizontal rectangle, and inside/outside vertical rectangle. In order for a condition/special area to be subset-sufficient, the only requirements are that at least one green case and no pink cases lie within it. For example, setting up gender working groups (GWG) is the second condition in the model, and cases where it is present lie in the bottom area. While some green cases lie there, no pink/unsuccessful cases do, which means the condition is subset-sufficient.

The questions we ask here is "what colour are the special areas painted with"? We can start from the larger special areas, for example the right side of the space, or the bottom space, or the central horizontal rectangle. "Is any pink observed in this particular special area?" If not, the condition corresponding to that special area is sufficient in a subset sense.

In other words, in order to be sufficient in a subset sense, white/blank areas are tolerated: unlike for sufficient areas in a "Boolean minimisation sense" (which we will see in Section 2.7.1.5), where the entire area corresponding to the condition needs to be green, no blank/white areas tolerated.

When a special area only includes green and white/blank cases, it is subset-sufficient, even if other "green" cases are observed outside of it (as in the case of budget support). It means that the area is sufficient but not necessary: if it were necessary it would have included ALL the green cases within itself.

After having checked the special areas related to single conditions, we can check their intersections. The quadrant on the bottom-right represents the combination of the presence of the two first conditions; the one on the top-left the combination of their absence, and so on.

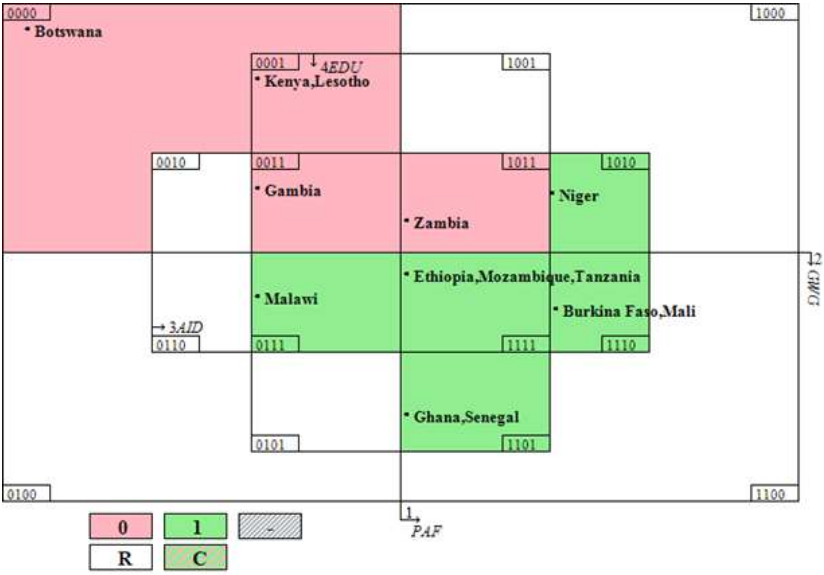
2.5.2.2 Evaluations where the subset analysis provided important findings

In the Budget Support Evaluation, the presence of gender working groups (GWG) in the process leading to the formulation of the national budget support plan was observed to be sufficient in itself for an increase in school enrolment of girls. All 8 cases where gender working groups were set up ended up being successful (Table 28).

Table 28: Dataset from the Budget Support Evaluation showing subset sufficient conditions

Country	PAF	GWG	AID	EDU	OUT
Ethiopia	1	1	1	1	1
Mozambique	1	1	1	1	1
Tanzania	1	1	1	1	1
Burkina Faso	1	1	1	0	1
Mali	1	1	1	0	1
Ghana	1	1	0	1	1
Senegal	1	1	0	1	1
Malawi	0	1	1	1	1
Niger	1	0	1	0	1
Zambia	1	0	1	1	0
Gambia	0	0	1	1	0
Kenya	0	0	0	1	0
Lesotho	0	0	0	1	0
Botswana	0	0	0	0	0

Figure 11: Graphic illustration of the Budget Support Evaluation data



If we look at the bottom area of the Venn diagram (Figure 11), representing the presence of GWG, we can see that no pink/unsuccessful case is located there[31].

Finally, the sufficiency table for condition GWG (Table 29) shows that 8 cases present both the condition and the outcome but no cases present the configuration without presenting the outcome at the same time.

Table 29: Sufficiency table for the GWG condition

		OUT is	
		Present	Absent
GWG	Present	8	No cases
	Absent	Not relevant	Not relevant

In the MAVC study, two single conditions – the intervention providing sufficient funds (FUNDSF) and the presence of accountability mechanisms (ACCMEC) were found to be subset-sufficient by themselves (and hence also in combination) for repairs to be made to broken water points (Table 30).

Table 30: Excerpt from the MAVC study dataset explaining outcome 3 showing subset sufficient conditions

Project	FUNDSF	SPAREP	ACCMEC	EXRESP	REPAIR
SHP	1	1	1	1	1
M4W	0	0	0	1	0
MM	0	1	1	1	1
MV	1	1	1	1	1
ND	1	1	1	1	1
HSW	0	1	0	1	0

Table 31 shows that whenever the combination is present, the outcome is never absent.

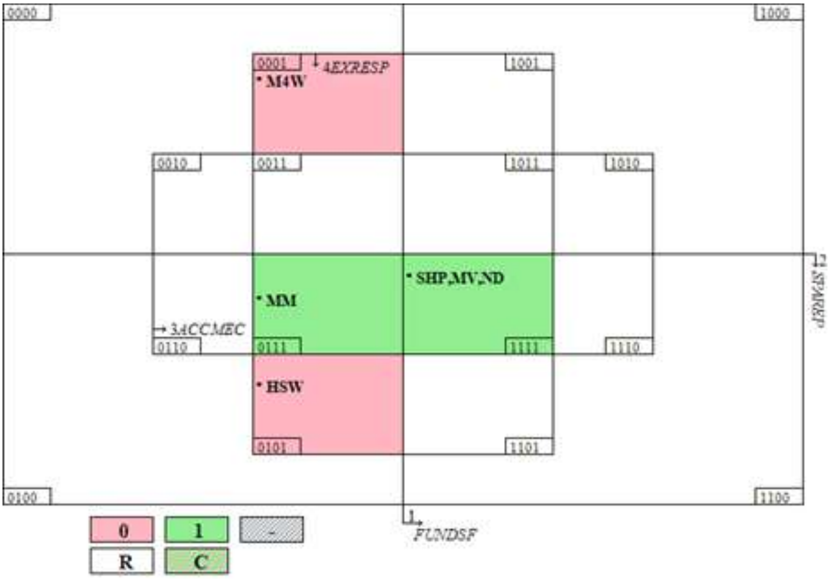
[31] We also notice that working groups are not necessary: Niger is successful without them. The data show that the combination PAF*AID*edu (gender indicators included in the programming document, high aid levels, primary education not free) – representing Niger, Burkina Faso and Mali – is also subset-sufficient for success: all the three cases covered by it are successful. This also emerges from the Venn diagram, where the intersection between the central horizontal rectangle (AID), the right area (PAF) and the outside of the central vertical rectangle (edu) is fully green.

Table 31: Sufficiency table for the FUNDSF*ACCMEC combination

		OUT is	
		Present	Absent
FUNDSF*ACCMEC	Present	3	No cases
	Absent	Not relevant	Not relevant

Finally, the Venn diagram (Figure 12) shows that neither the right hand side (FUNDSF) nor the central horizontal rectangle (ACCMEC) include any pink areas (lack of outcome achievement).

Figure 12: Graphic illustration of the QCA dataset explaining outcome 3 in the MAVC study



2.5.2.3 OPPORTUNITY: Sufficiency of Conjunctions (Combinations)

If more than one condition, let's say n conditions, are found to be perfectly sufficient (ie. constantly leading to success in 100% of cases where they are observed), both the disjunction and the combination of these n conditions will also be perfectly sufficient. However, in many cases, no single condition is perfectly sufficient for a positive outcome. This happens because, intuitively, most outcomes we are interested in are complex and demanding, and will be achieved only when a combination of conditions align at the same time. In such cases it is advisable to focus on the sufficiency of "conjunctions" or "causal packages": a.k.a. the ***logical intersection*** of a number of conditions. As we add more conditions to a conjunction, the chances of that conjunction being perfectly sufficient increase (Befani B. , 2013) (see also Chapter 4). In practical terms this means that, although no single condition might invariably lead to success, there are higher chances that a combination of two will, and even higher that a combination of three will, and so on.

In the GEF/UNDP Biodiversity Evaluation, one line of inquiry addressed the factors responsible for a functional Protected Area (PA) System at the national level[32]. Thirteen factors were identified which could potentially be responsible for achieving a functional PA system; out of which 5 were found to be necessary for success:

- Transparency of financial flows and management (TRANSPFIN)
- Adequate legal framework for conservation (ADQLEG)
- Transparency of decision-making procedures (TRANSPDEC)
- Unified and clear mandates among institutions (e.g. no overlaps) (CLRMAND)
- CSO-Corporate sector-Government collaboration within government framework (COLLAB)

None of these factors was sufficient for success by itself: for example, out of the six cases presenting CLRMAND, only 5 were successful (Table 32).

[32] A group of consultants conducting fieldwork gathered during a two-day workshop defined a functional PA system as one meeting the following 3 criteria: (i) Sufficient human resources, including staff with specific skills and expertise, to carry out management functions and objectives (i.e. timely planning); (ii) Availability of an operational management information system that generates knowledge used for adaptive management ; and (iii) Ability to be resilient against catastrophes and shocks (e.g. market forces, climate change).

Table 32: Sufficiency table for the CLRMAND condition

		OUT is	
		Present	Absent
CLRMAND	Present	5	1
	Absent	Not relevant	Not relevant

The subset analysis conducted on these 5 factors, however, returned three sufficient combinations of two conditions each[33]. It was interesting to note that the clarity of mandate and the lack of overlap among institutional mandates (CLRMAND) was particularly key, needing only one of three other conditions (either TRANSPDEC, ADQLEG, or TRANSPFIN) in order to fully account for a functional PA system.

Table 33: Sufficiency table for the TRANSPDEC*CLRMAND combination

		OUT is	
		Present	Absent
TRANSPDEC*CLRMAND	Present	5	No cases
	Absent	Not relevant	Not relevant

The sufficiency table for the combination of TRANSPDEC* CLRMAND, for example, shows that whenever this configuration is present, the outcome is also present: there are no cases presenting the combination without presenting the outcome at the same time (Table 33).

2.5.2.4 OPPORTUNITY: Parameters of Fit for Sufficiency in crisp-set QCA

When no condition is perfectly sufficient, another way to extract sufficiency-related information from the database is to pay attention to the parameters of fit: consistency and coverage, which we have already seen for necessity. In the subset analysis, parameters of fit have a different meaning than in necessity analysis and are called

[33]
```
incl  cov.r
1 TRANSPDEC*CLRMAND 1.000 1.000
2 ADQLEG*CLRMAND    1.000 1.000
3 TRANSPFIN*CLRMAND 1.000 1.000
```

sufficiency-consistency and sufficiency-coverage. Sufficiency-consistency of a condition is the number of cases presenting that condition which are also successful, divided by the total number of cases presenting that condition. This indicator has a minimum value of zero (when no case presenting that condition is successful) and a maximum value of one for perfect consistency (when all cases presenting the condition are successful). If the sufficiency-consistency of a condition is 0.9, it means that condition leads to success not all the time, but in 90% of cases where it is observed.

The notion of consistency in the subset sufficiency analysis can again be illustrated by a 2x2 table as in Table 34 (Schneider & Wagemann, 2012): note that unlike in the case of perfect sufficiency, we can observe some cases presenting the configuration but not presenting the outcome at the same time.

Table 34: The notion of consistency in the subset sufficiency analysis

A condition or configuration is subset sufficient to some degree if		Outcome is	
		Present	Absent
Configuration is	Present	Some cases (many, e.g. 90%)	Some cases (few, e.g. 10%)
	Absent	Not relevant	Not relevant

The notion of sufficiency is more meaningful for high consistency scores. Saying that a condition is 50% sufficient is not very informative: it means that the condition is as likely to lead to a positive as to a negative outcome. On the other hand, saying that it is 95% sufficient means that – although not a "guarantee" – the chances of success are very high, or – put differently – the condition is associated with a risk of failure of only 5%. This is more informative than simply claiming that the condition is not perfectly sufficient.

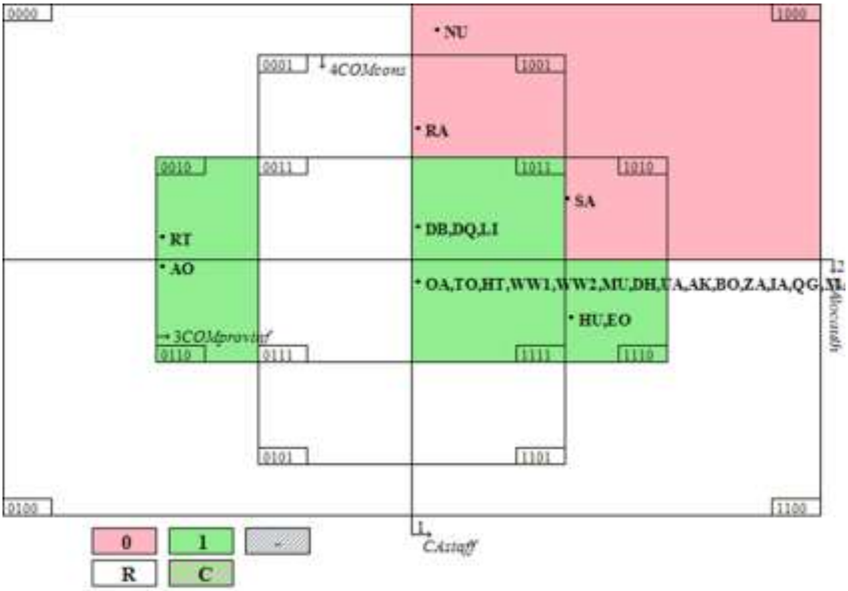
In the GEF/UNDP Biodiversity Evaluation, "provision of information" to the communities in the Protected Area (COMprovinf) was discovered to be a necessary, but not sufficient condition for "decreased trends of illegal activities" (BIOtrend); and no other single condition was subset-sufficient. However, in the 28 PAs where information was provided, a decrease in illegal activities was almost always observed; there was only one exception (the case labelled "SA"). This would amount to a 27/28 = 96% sufficiency-consistency level for the condition (see Table 35).

Table 35: Sufficiency table for the COMprovinf condition

		BIOtrend is Present	Absent
COMprovinf	Present	27	1
	Absent	Not relevant	Not relevant

The Venn diagram in Figure 13 shows that the central horizontal rectangle is not completely free of "pink areas" as perfect subset-sufficiency would require: one case ("SA") inside of it (meaning associated with presence of that condition) is located in a pink area.

Figure 13: Graphic illustration of a 4-condition model explaining BIOtrend in the GEF/UNDP evaluation



This puzzling finding seemed to be explained by the fact that information was indeed provided to the community, but to only one community, which might have been insufficient to influence other communities that were conducting the illegal activities. There was no further follow-up, but one possible action could have involved checking how widespread provision of information was in other protected areas, and if no other successful case presented such a low diffusion of this aspect of the intervention, a recalibration of the condition (raising the threshold for "presence" of COMprovinf) would have resulted in a consistent finding.

2.5.2.5 PITFALL: Non-representative conditions and sufficiency-coverage

When a condition is observed to be sufficient for a positive outcome, this might not necessarily be very relevant; it might not tell us the whole story or even the most important part of the story: for example when that condition only covers a very limited amount of successful cases. In the Budget Support Evaluation this was not the case, and the subset sufficient GWG covered 8/9 successful cases (89%). In the MAVC repairs dataset, likewise, the subset sufficient conditions FUNDSF and ACCMEC covered, respectively, 75% and 100% of successful cases.

Table 36: Dataset of 8-condition model explaining outcome 1 in the MAVC study, showing a perfectly subset-sufficient conditions with low coverage

Project	RECEPT	DEVCHG	ACCDEV	DATCOLL	HUMAUT	WHORPS	PREFRP	COSTNO	USEICT
SHP	1	1	1	0	1	1	1	1	1
M4W	0	1	0	1	0	0	0	0	1
MM	0	0	0	1	0	0	0	1	0
MV	1	1	1	1	0	0	1	1	1
SIBS	1	1	0	0	0	1	1	1	1
RiR	1	1	1	0	0	1	1	1	1
ND	1	1	1	1	0	0	1	1	1
HSW	1	1	0	1	0	0	0	1	0

However, it is not uncommon to observe that one sufficient condition is present in a very low portion of successful cases, at the extreme only one or two. In the MAVC study, the model attempting to explain use of ICT in reporting faults of water points included a condition (HUMANAUT) addressing whether the report of the fault was automatic or required human interaction. The subset analysis software returned a perfect sufficiency-consistency score for "automatic reporting"; except that only one case out of the successful 6 presented this reporting mechanism, while all the other 5 required human interaction (see Table 36 and 37). In other words, "automatic reporting" had perfect consistency but covered only 17% of the cases:

its sufficiency-coverage was the lowest possible, given the number of successful cases (6).

Table 37: Sufficiency table for the HUMANAUT condition

		USEICT is	
		Present	Absent
HUMANAUT	Present	1	No cases
	Absent	5	Not relevant

The sufficiency-coverage is a measure of how much one single explanation, however robust, is telling the whole story; if it's just one of many ways to achieve the outcome or the only or most prevalent way. It is obtained by dividing the number of cases presenting the sufficient condition by the total number of successful cases. It is equivalent to the percentage of successful cases presenting the sufficient condition. It measures the relative importance of that condition with respect to others, in terms of the frequency with which it is present across the successful cases. The higher the percentage of successful cases where the condition is observed, the higher its coverage.

Table 38: coverage in the subset sufficiency analysis

A condition or configuration is perfectly sufficient and has perfect (good) coverage if		Outcome is	
		Present	Absent
Configuration is	Present	Some cases	No cases
	Absent	No (or few) cases	Not relevant

For example, when sufficiency-coverage is 0.5, it means that the sufficient condition is observed in half of the successful cases: the other half do not present it. This would imply that the two cells in the left column of Table 38 have exactly the same number of cases. In the extreme situation where the parameter of fit has the value of one (and the bottom-left cell includes zero cases), the condition is not only sufficient but also necessary for success – as it would be present in all (100%) of successful cases. In other words, when both consistency and coverage equal 1.00, the condition is both perfectly sufficient and perfectly necessary for success (see also similarities with the necessity section).

A high level of sufficiency-consistency is more meaningful when it refers to a condition or combination covering a high proportion of cases, when it is associated with high levels of sufficiency-coverage; if the consistent association is observed in only one or two cases, it might be due to chance. In addition to the sufficiency-coverage it is useful to calculate the probability that the consistency of the association is not due to chance, but to an underlying mechanism that links the condition with success. Section 3.1.1.1 reports the levels of confidence we can associate with perfect supersubset relations, on the basis of how many cases they are observed in. If the association is due to random factors, as more cases are added the probability that the association is still consistent decreases; so if the empirical association is still consistent over a higher number of cases, our confidence on the robustness of the association increases.

2.5.3 The SuperSubset Analysis in Practice

This section covers more practical aspects of the supersubset analysis, like recommended software and the opportunity to deduct sufficiency relations from necessity relations and viceversa, offered by DeMorgan's Law.

2.5.3.1 OPPORTUNITY: software for the SuperSubset Analysis

The most complete software platform for the SuperSubset Analysis is R (Dusa & Thiem, 2014; Thiem & Dusa, 2012; Dusa, 2007). R will read the dataset and list the findings of the procedure superSubset() starting from the consistency and coverage scores of single conditions; then groups of two conditions, groups of three conditions, and so on. Since all possible groups of conditions are considered, the findings can potentially be returned in the form of very long lists, even with relatively modest numbers of conditions. An advantage of R is that this list can be reduced to its most relevant part, by specifying minimum benchmarks for the parameters of fit, under which the findings are not relevant for the evaluator. The consistency cut-off point is denoted with "incl.cut" and the coverage cut-off point with "cov.cut" (Thiem & Dusa, 2012). For example, the instructions "incl.cut = 0.9" and "cov.cut = 0.5" in a superset/necessity analysis tells the software to only list those configurations that have a necessity consistency higher than 0.9 (are present in 90% of successful cases) and coverage higher than 0.5, that is a proportion of successful cases over unsuccessful cases higher than 0.5. In a subset analysis, the same commands and parameters tell the software to return (groups of) conditions that have a sufficiency-consistency higher than 0.9 (lead to success 90% of the times they're observed) and cover at least 50% of all successful cases (a sufficiency-coverage of 0.5).

The subset analysis can also be easily performed by the freely downloadable software fsQCA. However, fsQCA reports the values of the parameters of fit for all possible combinations, which can result in extremely long lists even with modest numbers of conditions. More specifically, fsQCA lists the consistency and coverage values of all single conditions plus all possible combinations of the conditions included in the specified models. For 4 conditions the software lists 15 rows, for 5 31, for 6 63, for 7 127 rows; for 8 conditions 255 rows; for 9 511 and for 10 conditions, 1023 combinations… R is able to do the same but it can also trim down the long list on the basis of benchmarks, or cuts: minimum values of consistency and coverage.

For these reasons, R is recommended for models of five or more conditions, while fsQCA works well with four or less. Both software platforms will start from analysing single conditions, then groups of two, then groups of three, and so on.

2.5.3.2 PITFALL: only a small number of conditions (at a time)

In the author's experience, the superSubset procedure in R doesn't work when the number of conditions analysed at the same time is higher than 12. In such situations, it is advisable to cluster the conditions into groups. For example, in the GEF/UNDP Biodiversity Evaluation the dataset included a total of 28 conditions, which were grouped in three typologies: capacity, community and context. The supersubset analysis was conducted separately on each of these groups (see Section 2.7.1.9).

2.5.3.3 OPPORTUNITY: DeMorgan's Law

DeMorgan's Law illustrates a logical equivalence: the statement "P implies Q" is identical to the statement "non Q implies non P". It is relevant here because it can help the evaluator convert the findings from a superset analysis on a positive outcome to findings from a subset analysis on a negative outcome and viceversa: that is, deriving the former from the latter. More specifically:

1. If a condition is necessary for a positive outcome (C <= O), then its negation is sufficient for a negative outcome (non C => non O)
2. If a condition is sufficient for a positive outcome (C => O), then its negation is necessary for a negative outcome (non C <= non O)
3. If a condition is necessary for a negative outcome (C <= non O), then its negation is sufficient for a positive outcome (non C => O)
4. If a condition is sufficient for a negative outcome (C => non O), then its negation is necessary for a positive outcome (non C <= O).

In the Budget Support Evaluation, we can convert the finding that either gender working groups (GWG) or the inclusion of gender-sensitive indicators in the main planning document (PAF) are necessary for success (which can be written as GWG + PAF <= O) into the finding that absence of both policy instruments guarantees (is sufficient for) lack of success (which can be written as non GWG*non PAF => non O or gwg*paf => o depending on notation). This is because the negation of a disjunction or logical union of conditions is the conjunction/combination/intersection of the negations of the same conditions. This can also be seen in the Venn diagram: saying that all green areas are located in the right area plus the bottom-left quadrant means that no green area is located in the remaining top-left quadrant; therefore any case in that quadrant is guaranteed to be unsuccessful.

Notice that DeMorgan's Law works well only when necessity and sufficiency are analysed in a supersubset sense: when these statements are made while interpreting the findings of a boolean minimisation (see Step 7) the logical symmetry often doesn't hold because of limited diversity: for this procedure, the use of DeMorgan's Law is more controversial (Schneider & Wagemann, 2012).

2.6 Step Six: building the Truth Table[34]: "How can the dataset be synthesised without loss of case diversity?"

This section explains what a Truth Table is and why it is useful to synthesise the dataset without loss of case diversity, drawing on a real-life evaluation example. It further elaborates on the pitfalls connected with building it: namely that similarity of cases depends on model selection, and that several discretionary decisions can be made when building it (and as a consequence the same dataset can produce many different Truth Tables). The section concludes with a presentation of the so-called "complete" Truth Table, which includes those logical combinations of the conditions included in the model that are not empirically supported in the dataset.

We have seen in the previous section that, once the dataset is built, the subset-sufficiency analysis is conducted simultaneously on single conditions, combinations of two, three, etc. and does not necessarily take into account the entire set of conditions we have information on across the cases (for example when it addresses single conditions). In the rest of the section we address a procedure that, unlike the above, considers all the conditions included in the dataset at the same time and simplifies the dataset without loss of information on the diversity of cases.

The procedure returns a table, called the "Truth Table", describing a series of combinations of conditions which are considered sufficient for the outcome[35], and are relatively complex compared to the findings of the subset analysis: no case information is lost compared to the dataset, and yet the Truth Table is simpler than the dataset. While the subset analysis isolated conditions and measured consistency and

[34] In fuzzy-set QCA, the Boolean minimisation follows the same rules as in crisp-set QCA. The big difference between csQCA and fsQCA lies in the way the Truth Table is built. In fsQCA, consistency scores of combinations are calculated differently than illustrated above; and frequency benchmarks are replaced with a measure of average distance or closeness of the cases in the dataset to a given combination. However, once the combinations are selected and the Truth Table is built, the Boolean minimisation proceeds in the same way as for crisp sets.

[35] See section on Generalisation for more details.

coverage across all cases (cutting, so to speak “up and down” across the dataset), building the Truth Table requires merging similar cases (and thus cutting, so to speak “right and left”).

Broadly speaking, a Truth Table represents the list of sufficiency statements supported by the empirical dataset, which are also necessary for the outcome, when they are considered as a logical union/disjunction. This means that in order to observe the outcome, at least one Truth Table combination is required, it doesn’t matter which one; and each Truth Table combination is sufficient for the outcome.

The Truth Table building process is different depending on whether we are working with fuzzy datasets, very large datasets, or datasets with relatively uncertain sufficiency relations. Under specific circumstances, the associations can be argued to be due to chance; however, there are ways to measure our confidence that the combinations represent real sufficiency relations (see Chapter 4 for more details).

Finally, the notion of “case similarity” is totally dependent on which conditions the evaluator decides to consider; or on “model specification”. Cases which are identical on conditions A, B and C might differ with regard to D, E and F.

In crisp-set QCA, building the Truth Table means merging all identical cases with the same outcome into one single combination of conditions. Identity is meant as "perfect identity" and it is established following a “zero-difference” rule whereby two cases cannot be merged if they differ in one or more condition. The Truth Table displays only different combinations (if two combinations were identical they would be merged).

For example, the Budget Support Evaluation dataset including 14 cases (Table 39) becomes a Truth Table with 9 different combinations (Table 40), only five of which cover one single case, while 3 cover 2 cases each and one covers three cases. Note that, unlike in the dataset in Table 39, all rows in Table 40 are different.

Table 39: Dataset of the Budget Support Evaluation showing identical rows that can be merged

Country	PAF	GWG	AID	EDU	OUT
Ethiopia	1	1	1	1	1
Mozambique	1	1	1	1	1
Tanzania	1	1	1	1	1
Burkina Faso	1	1	1	0	1
Mali	1	1	1	0	1
Ghana	1	1	0	1	1
Senegal	1	1	0	1	1
Malawi	0	1	1	1	1
Niger	1	0	1	0	1
Zambia	1	0	1	1	0
Gambia	0	0	1	1	0
Kenya	0	0	0	1	0
Lesotho	0	0	0	1	0
Botswana	0	0	0	0	0

Table 40: Truth Table of the Budget Support Evaluation (for the 4-conditions model)

Country	PAF	GWG	AID	EDU	OUT
Ethiopia, Mozambique, Tanzania (3)	1	1	1	1	1
Burkina Faso, Mali (2)	1	1	1	0	1
Ghana, Senegal (2)	1	1	0	1	1
Malawi (1)	0	1	1	1	1
Niger (1)	1	0	1	0	1
Zambia (1)	1	0	1	1	0
Gambia (1)	0	0	1	1	0
Kenya, Lesotho (2)	0	0	0	1	0
Botswana (1)	0	0	0	0	0

The advantage of building a Truth Table compared to working on the dataset is that the Truth Table is simpler without being less informative or less diverse: *every different case counts*. What's kept in the Truth Table are the different sufficient pathways to the outcome, some covering more cases than others.

2.6.1.1 PITFALL: similarity is dependent on model selection

The construction of the Truth Table is fully dependent on the notion of case similarity. In crisp datasets, adherence to a strict "zero-difference" rule is required, which means that only cases that do not differ in any condition at all, amongst those included in the model, can be merged. However, depending on the particular conditions considered, different cases can be merged: cases sharing identical values on a group of conditions might not do so on another group. For example, if we remove the condition "PAF" from the budget support model above, and consider for comparison purposes only the conditions "GWG", "AID" and "EDU" (in addition to the outcome), we obtain the Truth Table below (Table 41):

Table 41: Alternative Truth Table of the Budget Support Evaluation (for the 3-condition model)

Country	GWG	AID	EDU	OUT
Ethiopia. Mozambique, Tanzania, Malawi (4)	1	1	1	1
Burkina Faso, Mali (2)	1	1	0	1
Ghana, Senegal (2)	1	0	1	1
Niger (1)	0	1	0	1
Zambia, Gambia (2)	0	1	1	0
Kenya, Lesotho (2)	0	0	1	0
Botswana (1)	0	0	0	0

It's easy to see the differences between Table 40 and Table 41, obtained from the same dataset but where different conditions have been used to establish identity. Here only 2 combinations cover only one case each, while 4 combinations cover 2 cases each and one combination covers 4.

2.6.1.2 ISSUE AT STAKE: simplifying and strengthening the Truth Table requires discretionary decisions

In some cases, particularly with crisp datasets, creating the Truth Table is straightforward: it's easy to select the conditions, the combinations are consistently sufficient for the outcome, and all different combinations are included no matter their frequency. In other cases a series of discretionary decisions need to be made which might affect the final result. For quality purposes it is thus essential that the evaluator discloses the Truth Table building process in detail, and perhaps illustrates alternative Truth Tables obtained with slightly different decisions.

First of all, for large or very large datasets, the "every different case counts" rule might lead to overly complex Truth Tables: fortunately evaluator can set "frequency benchmarks" and include in the Truth Table only those combinations that are represented in a minimum number of cases, for example 2, 3, 4, or 5. For example, if the frequency threshold is higher than 1, a Truth Table row might no longer simply represent a combination present in the dataset, but also a combination with a minimum coverage. In this case the Truth Table would not represent the full spectrum of diversity, but only the main pathways to the outcome (which could still cover a significant amount of diversity).

The evaluator can choose the level of diversity which best fits the evaluation questions and the data: however, different frequency thresholds might produce different Truth Tables: the higher the benchmark (the stricter the inclusion rules), the lower the chance that any given combination is included and thus the smallest the Truth Table (and the lower the diversity). The sensitivity analysis will reveal the extent to which these choices affect the findings (see Section 3.3)[36].

The Truth Table in Table 42 has been built from the Budget Support Evaluation dataset, operating an inclusion threshold of two cases: its difference compared to the "original" Truth Table in Table 40 should be obvious: the five combinations covering only one case have been removed, while the others have remained the same.

[36] Some might say that increasing the threshold for inclusion decreases external validity: it is actually the opposite, external validity is strengthened because the Truth Table would only represent frequently occurring cases and the sufficiency statements will be more robust (i.e. applicable outside the dataset). On the other hand, internal validity might be affected because changing the threshold for inclusion might result in considerably different Truth Tables and thus very different solutions (which is something for the sensitivity analysis to establish).

Table 42: Alternative Truth Table for the 4-condition model, with inclusion threshold of 2 cases

Country	PAF	GWG	AID	EDU	OUT
Ethiopia, Mozambique, Tanzania (3)	1	1	1	1	1
Burkina Faso, Mali (2)	1	1	1	0	1
Ghana, Senegal (2)	1	1	0	1	1
Kenya, Lesotho (2)	0	0	0	1	0

The second issue is that the combinations associated with/considered sufficient for an outcome, are not necessarily consistently sufficient for it. The inclusion decision might have been based on consistency thresholds, a.k.a. "cut-off points", representing minimum values of the sufficiency-consistency scores of each combination (see Section 2.5.2). In brief, if too few combinations are perfectly sufficient, the evaluator might decide to include in the Truth Table all combinations with a consistency score higher than a benchmark, e.g. 0.9 or 0.8, which means that all combinations leading to the outcome in at least 80% or 90% of the cases are considered sufficient, for the purpose of Truth Table construction. The evaluator needs to specify which cut-off points have been used, because different cut-off points lead to different Truth Tables. Namely, the lower the cut-off point, the higher the chance that combinations are included (and of having larger Truth Tables), while the consistency of the sufficiency statements decreases. It's essential that evaluators specify the cut-off points used to build the Truth Table, and ideally use different cut-off points to build and compare different Truth Tables.

Finally, in fuzzy datasets, the Truth Table is still crisp, and represents those combinations that empirical cases are most similar, or "closest" to, in the multi-dimensional space (Ragin, 2000; Schneider & Wagemann, 2012). Inclusion criteria take into account both the "crispness"/"fuzziness" of the case (how close it is to a "crisp", Boolean combination, or "ideal type") and whether its membership score to the closest crisp combination is higher or lower than its membership score (or closeness) to the outcome. In addition, the number of cases that are closest to given crisp combinations (a variant of the frequency threshold) can also affect inclusion. Changing any of these parameters can produce a different Truth Table. It is good practice to disclose these choices and to check how small changes in the parameters affect the Truth Table.

2.6.2 The “complete” Truth Table

All the combinations reported in the Truth Table are different; but usually not all possible different combinations are reported in the Truth Table, because of the “limited diversity” that is observable in empirical cases. It is possible to combine presence and absence of n conditions in 2^n different ways (the permutations with repetitions of 2 values in n-tuples): which means that the Truth Table can have a maximum of 4 rows for 2 conditions, 8 rows for 3 conditions, 16 for 4 conditions, 32 for 5, etc. Usually not all the rows are supported empirically, so the Truth Table never reaches these limits. However it is possible to construct one particular type of Truth Table on the basis of the logically possible combinations and simply assign “0” to the column indicating the number of cases if no empirical case supports a specific combination. The outcome will be indicated as a question mark because it is not observed for that combination. The theoretically possible combinations not observed in the dataset are called “remainders” or “*logical cases*”[37]. A Truth Table constructed as such is called the “complete” Truth Table.

2.6.2.1 ISSUE AT STAKE: assessing the proportion of theoretical diversity that is covered empirically

Listing all the combinations that are logically possible from a combinatorial perspective can be useful because it allows the evaluator to assess the proportion of theoretically possible diversity that is covered empirically in the dataset, by comparing the number of empirically supported combinations to the maximum number of possible combinations. In the Budget Support Evaluation, the complete Truth Table (Table 43) shows that only slightly more than half of all the possible combinations (9/16, or 56%) are covered empirically, while the others are “remainders” or “logical cases”.

[37] In the fsQCA software platform they are called “counterfactuals”, to stress their nature of “unobserved cases”.

Table 43: Complete Truth Table of the Budget Support Evaluation, 4-condition model

Country	PAF	GWG	AID	EDU	OUT
Ethiopia. Mozambique, Tanzania (3)	1	1	1	1	1
Burkina Faso, Mali (2)	1	1	1	0	1
Ghana, Senegal (2)	1	1	0	1	1
Malawi (1)	0	1	1	1	1
Niger (1)	1	0	1	0	1
Zambia (1)	1	0	1	1	0
Gambia (1)	0	0	1	1	0
Kenya, Lesotho (2)	0	0	0	1	0
Botswana (1)	0	0	0	0	0
Logical case/Remainder (0)	0	1	1	0	?
Logical case/Remainder (0)	1	1	0	0	?
Logical case/Remainder (0)	1	0	0	1	?
Logical case/Remainder (0)	0	0	1	0	?
Logical case/Remainder (0)	0	1	0	0	?
Logical case/Remainder (0)	1	0	0	0	?
Logical case/Remainder (0)	0	1	0	1	?

The situation is very different for the other Truth Table constructed with a three-condition model: this one covers empirically almost all theoretically possible cases (7/8, or 87%), with only one exception (Table 44).

Table 44: Complete Truth Table of the Budget Support Evaluation, 3-condition model

Country	GWG	AID	EDU	OUT
Ethiopia. Mozambique, Tanzania, Malawi (4)	1	1	1	1
Burkina Faso, Mali (2)	1	1	0	1
Ghana, Senegal (2)	1	0	1	1
Niger (1)	0	1	0	1
Zambia, Gambia (2)	0	1	1	0
Kenya, Lesotho (2)	0	0	1	0
Botswana (1)	0	0	0	0
Logical case/Remainder (0)	1	0	0	?

Creating the complete Truth Table can also have negative consequences, particularly if the number of conditions is higher than 5 or 6.

2.6.2.2 PITFALL: a high number of conditions produces dysfunctional complete Truth Tables

The negative consequence of considering all logically possible combinations is that their number rises exponentially as new conditions are added to models: for 6 conditions 64 rows are required, for 8 conditions 256 and for 10 conditions 1024 rows, etc. If the number of cases does not keep up with the growth of the complete Truth Table as conditions are added, and in most cases it doesn't, adding conditions will result in an extremely high amount of logical cases compared to the empirically covered combinations. Put differently, when the number of conditions is high, the proportion of theoretical diversity covered by the empirical cases, which in the examples above decreased from 87% to 56% when going from 4 to 5 conditions, decreases substantially. Consider a situation where 9 conditions are included and a good 35 combinations are covered empirically. This will only amount to 35/512, or 7% of theoretically possible diversity covered in the dataset.

The balance between the number of conditions and number of cases that need to be respected in QCA is further discussed in Section 3.1.

2.7 Step Seven: the Boolean minimisation: How can the list of sufficient pathways be simplified?

In the previous step the dataset was simplified into a Truth Table, with fewer rows, without losing any information on the richness and diversity of cases. This section illustrates a way to simplify the dataset even further, without losing relevant information. The procedure reduces the Truth Table into a shorter list of simpler combinations, without losing information on causally sufficient pathways. It is known as the Boolean minimisation or the Quine-McCluskey algorithm: it pairs combinations in the Truth Table on the basis of their similarity and replaces two similar combinations with a simpler one sharing the conditions they have in common. The algorithm operates by merging two combinations of a Truth Table sharing the outcome and all conditions except one (the "one-difference rule"), into a simpler combination presenting all the identical conditions (and the same outcome) but not the different condition (its logic is explained in any QCA textbook).

This procedure is based on the deduction that, if two almost identical combinations with only one difference lead to the same outcome, this one difference is irrelevant for the outcome and the condition can be removed. It is an application of a variant of Mill's Method of Agreement (see Annex A), where the consistently present "cause" is actually a combination rather than a single cause, and it is consistently present together with the same outcome while the other condition varies.

The Quine-McCluskey algorithm of Boolean minimisation is not the only possible way of synthesising the information included in a Truth Table. Rick Davies has shown how Boolean datasets can be synthesised using decision tree modelling[38]. In terms of social science academic developments, a new algorithm, called "coincidence analysis", has been pioneered in (Baumgartner, Detecting Causal Chains in Small-n Data, 2012); applied to an empirical case in (Baumgartner & Epple, 2013; Thiem, 2015); and recently developed into a new function in the R package (Baumgartner & Thiem, 2015). Given the early stages of this new development, coincidence analysis will not be addressed in this report. However, the author agrees with (Thiem, 2015) that this procedure will likely not replace the "traditional" Boolean minimisation and that using both types of minimisation in applied research will provide a useful triangulation of the findings in the future and likely become a good practice in handling Boolean datasets.

When the software platforms fsQCA or R are used, the findings of a Boolean minimisation, also known as the "solution", indicate the consistency and coverage scores of each combination (see Step 5). Coverage is measured both in "raw" terms (representing the % of cases logically covered by the combination) and in unique terms (representing the % of cases that are uniquely covered by that combination, that is they are not covered by any other combination of the solution). Unique coverage is a measure of how much that specific combination is needed in the solution: if it's high, removing that combination from the solution will mean obtaining a solution covering a much smaller number of cases. This is not the case for raw coverage: if combinations with high raw coverage are removed, the

[38] http://mande.co.uk/2012/uncategorized/where-there-is-no-single-theory-of-change-the-uses-of-decision-tree-models/

cases covered by those can still be potentially covered by other combinations.

This section will illustrate the advantages of the Boolean minimisation with a practical example from a real-life evaluation, and then discuss a way of simplifying the data even further using the logical cases included in the complete Truth Table introduced in the previous step. After illustrating the difference between minimisation-sufficiency and subset-sufficiency, both conceptually and visually with the help of the Venn diagram, arbitrary choices that can be made in the process will be discussed, together with their implications and tradeoffs. Finally, the Boolean minimisation's need of a small number of conditions will be addressed, and two different strategies proposed to reduce the number of relevant conditions when there is theoretical or conceptual uncertainty as to which conditions should be included.

2.7.1.1 OPPORTUNITY: identifying a small number of moderately complex sufficient pathways

The Budget Support Evaluation illustrates the advantages of the Boolean minimisation well. The 4-condition Truth Table reported in Table 40 sees its 9 combinations of 4 conditions each reduced to 5 combinations of 3 conditions each; in particular, the 5 combinations considered sufficient for a positive outcome are reduced to 3 (Table 45), and the 4 combinations associated with the negative outcome are reduced to 2 (Table 46).

Table 45: Complex Solution of the Boolean Minimisation for the BSE (4-condition model, positive outcome)

COMPLEX SOLUTION: PAF, GWG, AID, EDU (positive outcome)

Combination	Raw Coverage	Unique Coverage	Consistency
PAF * AID * edu (1-10)	0.333333	0.333333	1
PAF * GWG * EDU (11-1)	0.555556	0.222222	1
GWG * AID * EDU (-111)	0.444444	0.111111	1

Solution Coverage: 1.000000
Solution Consistency: 1.000000

The three successful pathways are all perfectly consistent (1.00) and cover, altogether, all 14 cases (solution coverage is 1.00): in particular, the first combination PAF*AID*edu covers one third of the cases uniquely, while the second covers more than half of the cases (22% uniquely). The third combination covers 44% of the cases (11% uniquely, see Table 45).

The solution above means that an increase in the primary school enrolment of girls seems to be achievable, on the basis of the data analysed, through three different pathways (Table 45):

1. in those countries with a relatively high aid budget for education, the policy instrument that works best depends on whether there is free education or not.
   – Gender working groups work well when there is free education, while
   – if this is not the case including gender-sensitive indicators in the planning document is preferable.
2. Finally, independently of aid levels, combining the two policy instruments guarantees success, provided there is free education.

Table 46: Complex Solution of the Boolean Minimisation for the BSE (4-condition model, negative outcome)

COMPLEX SOLUTION: PAF, GWG, AID, EDU (negative outcome)

Combination	Raw Coverage	Unique Coverage	Consistency
paf * gwg * aid (000-)	0.6	0.6	1
gwg * AID * EDU (-011)	0.4	0.4	1

Solution Coverage: 1.000000
Solution Consistency: 1.000000

The solution for the reduction of negative outcomes shows that lack of success is explained by two pathways, both perfectly sufficient (Table 46):

1. in contexts with a low aid budget, whether education is free or not, failing to implement both policy instruments guarantees that no progress will be made on primary school enrolment of girls. But
2. where aid levels are high and education is free, even just failing to implement gender working groups, independently of the other policy instrument, seems very costly.

2.7.1.2 OPPORTUNITY: incorporating remainders/logical cases

If the solution returned by the algorithm is too complex to understand and make sense of, the Boolean minimisation offers a further simplification opportunity through the assignment of an outcome value to logical cases and including these combinations in the Truth Table as if they had been empirically observed.

The current standard Boolean minimisation procedure in fsQCA offers three types of analyses, depending on whether logical cases are included or not, and if so which type. The COMPLEX solution is the one obtained without recurring to logical cases at all; the PARSIMONIOUS solution is the one obtained by including all useful remainders, while the INTERMEDIATE solution only includes a particular type of remainders, called "easy *counterfactuals*". Easy counterfactuals represent hypotheses that are easy to support on the basis of the empirical data, while the rest of logical cases (that aren't "easy counterfactuals") are called difficult counterfactuals.

All logical cases are either easy or difficult counterfactuals, depending on the directional expectations of each condition. If the presence of a condition, say "C", is expected to be associated with the outcome, and we know from the dataset that the combination of, say, A*B*c is sufficient, then we can safely assume that the combination A*B*C will also be sufficient. This means that A*B*C is an easy counterfactuals and can be safely incorporated in the minimisation. If, on the contrary, A*B*C were empirically shown to be sufficient, and A*B*c a remainder, the latter would be a difficult counterfactual.

In the Budget Support Evaluation, including easy counterfactuals in our minimisation of the negative outcome returns a simpler solution, where one of the two combinations (gwg*AID*EDU or -011) is reduced from three to two conditions: gwg*EDU (Table 47). It means that whenever there is free education, failing to implement gender working groups in the budget planning process is costly in terms of the outcome, no matter how high the aid levels are. This simplification has been obtained by assuming that the combination gwg*aid*EDU (-001) is sufficient for the (negative) outcome. This was an easy assumption to make because gwg*AID*EDU (-011) is empirically shown to lead to a negative outcome in the Truth Table; and it can be argued that a lower aid budget would not improve the situation in that context. Put differently, the directional expectation is that "AID" is associated with a positive and "aid" with a negative outcome.

Table 47: Intermediate Solution of the Boolean Minimisation for the BSE (4-condition model, negative outcome)

INTERMEDIATE SOLUTION: PAF, GWG, AID, EDU (negative outcome)

Combination	Raw Coverage	Unique Coverage	Consistency
paf * gwg * aid (000-)	0.6	0.2	1
gwg * EDU (-0-1)	0.8	0.4	1

Solution Coverage: 1.000000
Solution Consistency: 1.000000

If all useful logical cases are used to simplify the solution, the Truth Table is further simplified, to the point of having combinations of maximum two conditions (see next section on the Venn diagram).

2.7.1.3 OPPORTUNITY: Identifying simplifying cases on the Venn diagram

Remainders to include in the minimisation can also be chosen by hand, one by one, by locating them on the Venn diagram and assessing their simplification potential. This allows the evaluator to have more control on which remainders are included and which aren't, and to justify their inclusion on a combination-by-combination basis. In the Venn diagram, remainders are associated with blank/white areas: assigning outcome values to the logical cases in order to include them in the minimisation (in the hope of simplifying the solution) is equivalent to adding the missing pieces to a "green puzzle" (or pink puzzle, depending on the type of outcome) to complete a certain "shape"; for example, painting the white areas of the bottom-right quadrant in Figure 14 green, so that the combination of the first two conditions in the model (corresponding to the intersection of the right side and the bottom area) can be considered sufficient (in a minimisation sense).

In the Budget Support Evaluation, the intermediate solution of the negative cases was obtained by adding the easy counterfactual "1001", so that EDU*gwg can become a sufficient combination for lack of success. As for the positive outcome, there are no easy counterfactuals so the intermediate solution is identical to the complex one. However, the computation of the parsimonious solution provides opportunities for further simplifying the solution, listing three simple combinations of max 2 conditions each (Table 48). The first (GWG) is obtained by adding 4 remainders (all the blank spaces in the bottom of the diagram); the second (PAF*aid) by adding 3 remainders (the blank spaces on the right side which are external to the central horizontal rectangle) and the third (AID*edu) which is obtained by adding two logical cases (the blank areas on the left of the central horizontal rectangle, outside the central vertical rectangle EDU).

Table 48: Parsimonious Solution of the Boolean Minimisation for the BSE (4-condition model, positive outcome)

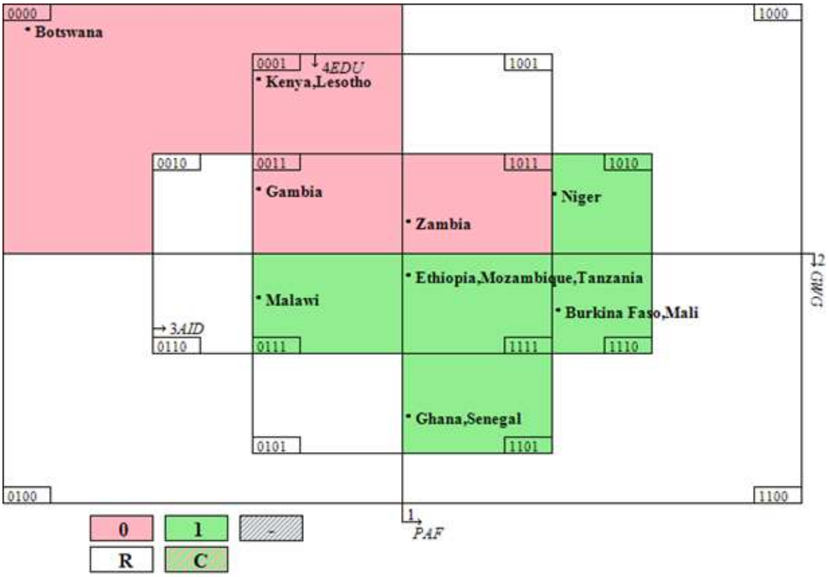
PARSIMONIOUS SOLUTION: PAF, GWG, AID, EDU (positive outcome)

Combination	Raw Coverage	Unique Coverage	Consistency
GWG (-1--)	0.888889	0.444444	1
PAF * aid (1-0-)	0.222222	0	1
AID * edu (--10)	0.333333	0.111111	1

Solution Coverage: 1.000000

Solution Consistency: 1.000000

Figure 14: Graphic Illustration of the Budget Support Evaluation dataset (4-condition model)



The advantage of using the Venn diagram to add logical cases is that a higher number of simple solutions can be discovered. For example, adding paf*GWG*aid*EDU (0101) returns the two combination-solution GWG*EDU (-1-1) and PAF*AID*edu (1-10). As a consequence, selecting the best simplified solutions on the basis of the assumptions that make most sense theoretically becomes easier.

2.7.1.4 PITFALL: credibility of assumptions on logical cases

Using logical cases can be very tempting in evaluation situations, usually characterised by complex models with many conditions. It will produce simpler solutions, sometimes much simpler. The downside is that, unless inclusion of logical cases is theoretically justified, the synthesis embodied by the findings loses validity, because it is founded upon assumptions which are not necessarily credible.

Schneider and Wagemann (Schneider & Wagemann, 2012) list several potential problems that can make the inclusion of remainders difficult to justify. In evaluation situations, time is limited and demands for transparency high, particularly in evaluations that are considered sensitive and can be attacked for political reasons. In the author's opinion, the available time is better spent testing the infinitely high number of models that can be tested, until some strong solution is found that can be justified mostly or exclusively on the basis of empirical findings. If there is time to justify the inclusion of selected logical cases, fine; but if not, the procedures that automatically include logical cases should be avoided.

In particular, the parsimonious solution, which automatically includes all logical cases as long as they simplify the solution, is almost always uninformative when models include more than 5 or 6 conditions. This is because the number of logical cases rises exponentially with the number of conditions, while the empirical diversity doesn't (as mentioned in 2.6.2.2). In such cases the parsimonious solution will most often end up being overly simple and usually list a series of single conditions or combinations of two conditions at most. So not only the parsimonious solution will often risk being not credible (because of the high amount of logical cases used) but it will also be extremely and unrealistically simple.

The complexity of the intermediate solution is, indeed, intermediate between the complex and the parsimonious solutions;

however, the limitation to "easy counterfactuals" for the cases that can be included will not always protect from the risks connected with including remainders.

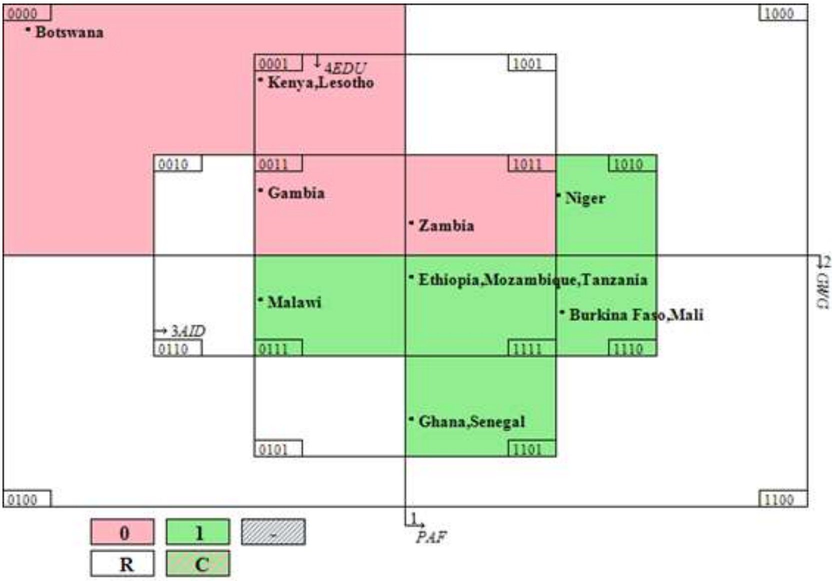
In practical situations, the author suggests that the evaluator tries the complex solution first, and if the latter is too complex that they resort to the wisdom of the Venn diagram, manually identifying the combinations that would simplify the solution, focusing efforts on justifying a limited amount of simplifying hypotheses.

2.7.1.5 ISSUE AT STAKE: the difference between subset-sufficiency and minimisation-sufficiency

The combinations that constitute the "solution" of a Boolean minimisation are obtained by simplifying more complex statements of sufficiency (as identified in the Truth Table) and taking care that the information removed is not relevant in causal sufficiency terms: this is why they are statements of sufficiency in themselves and why the Boolean minimisation is a form of sufficiency analysis. However, this type of sufficiency (which we can call "minimisation-sufficiency") is different from subset sufficiency. If we look at the Venn diagram illustrating the model above, we notice that the condition GWG (represented by the bottom area) is subset sufficient because the area is "pink-free": no cases with a negative outcome are located in the bottom of the diagram. However, this combination is not reported as sufficient in the Boolean minimisation, unless several logical cases are included, including difficult counterfactuals. In order for GWG to be considered sufficient in a minimisation sense, the entire bottom area needs to be covered by green rectangles (or combinations leading to a positive outcome). Not displaying pink areas/unsuccessful cases is not enough: the blank/white spaces need to be "coloured" green.

This is an important difference between subset sufficiency and minimisation sufficiency: in the results of a subset analysis, in order for a combination to be considered sufficient for success, it is enough that no unsuccessful cases are covered by the combination (no pink cases appear in the area). For the minimisation-sufficiency, on the other hand, several other similar combinations must be successful: all specific combinations included in the area must present a positive outcome (and be painted green).

Figure 15: Graphic Illustration of the Budget Support Evaluation dataset (4-condition model)



The Boolean minimisation is more conservative than subset sufficiency: obtaining simple sufficient combinations here is not just a matter of isolating conditions and calculating frequency of success and consistency scores across cases sharing those conditions, as in the subset analysis. Simplification is harder to achieve because it is obtained through merging cases which are almost identical across all conditions except one (while having the same outcome). Complexity is reduced cautiously and gradually starting from combinations with many conditions, while the subset analysis can deliberately focus on any number of conditions and ignore all the others. The Boolean minimisation embraces a more holistic approach than the subset analysis and takes into account all conditions included in the model at the same time. *Minimisation-sufficiency logically implies subset-sufficiency, but not viceversa*: combinations that are sufficient in a Boolean minimisation sense are also subset sufficient (but not viceversa).

2.7.1.6 ISSUE AT STAKE: choosing the right degree of complexity, balancing Consistency and Coverage

The solutions of the Boolean minimisation can present a range of degrees of complexity, and a range of values of consistency and coverage, depending on which combinations the evaluator includes in the Truth Table. Choices affecting which combinations are included are made at various stages (see Box 2).

Box 2: phases of the analysis requiring decisions affecting which cases/combinations are included in the Truth Table

1. When selecting which conditions to include in the model
   a. more conditions usually mean more complexity, more consistency and less coverage for single combinations
2. When selecting a frequency threshold for inclusion
   a. higher threshold means inclusion of combinations with higher coverage and hence single combinations of the solution will have higher coverage
3. When selecting a cut-off point for sufficiency-consistency
   a. lower cut-off point means inclusion of more cases but less consistent cases, resulting in the solution having higher coverage and lower consistency
4. When selecting a minimum degree of membership of the fuzzy case to an ideal, crisp combination
   a. higher threshold means a lower number of cases to be included, and hence lower coverage

The range of possible solutions spans from simple solutions with high coverage, low consistency, possibly requiring logical cases; to more complex, more consistent solutions with lower coverage, obtained without adding logical cases at all.

It is widespread practice to seek reliably consistent combinations first, to ensure that the combinations can be legitimately considered sufficiency statements; and then try to maximise coverage second, which will usually result in the addition of more combinations/ pathways to the solution, rather than simply going for simpler combinations with fewer conditions. Ideally, the solution will include a limited number of perfectly sufficient pathways that, taken together, fully cover the dataset (and as such have perfect consistency and coverage).

It is good practice to make small changes to the parameters identified in the various phases of the bullet point list above and see what impact they have on the solution in terms of complexity, consistency and coverage (see also Section 3.3). For each solution presented, the above parameters should always be specified: list of conditions, and thresholds for consistency, frequency and degrees of membership.

On one hand the flexibility of the Truth Table building process is an opportunity for the evaluator because it can adjust to many different types of dataset; but on the other hand it requires caution when interpreting the findings, as they might be highly dependent on these arbitrary choices.

2.7.1.7 PITFALL: works only with a small number of conditions (at a time)

Like the SuperSubset analysis, the Boolean minimisation becomes dysfunctional when a high number of conditions are analysed at the same time. The problem is exacerbated in the Boolean minimisation, given the more conservative approach of this procedure towards making relatively simple sufficiency statements compared to the SuperSubset analysis. This adds to previous arguments made against working with a high number of conditions, particularly with a relatively small number of cases: working with a high number of conditions creates problems in the SuperSubset analysis, in the creation of the Truth Table (for the limited diversity problem), and for assessing the reliability of Truth Table rows as sufficiency statements (see Section 3.1.1.2).

The combinatorial and set-theoretic nature of QCA – while offering opportunities that other methods don’t – is a serious obstacle to analysing a high number of conditions at the same time. Although there is no established “right” number of conditions in QCA, selecting a relatively small number of conditions for all QCA procedures is crucial.

When theory is poorly developed, discriminating in advance between essential and redundant factors, and selecting a small number of likely important conditions to include in a QCA analysis might be challenging. Below we propose possible solutions: one strategy that has been long known in the literature and a series of pragmatic criteria the author has found useful.

2.7.1.8 OPPORTUNITY: reducing the number of conditions with the two-step procedure

One commonly used procedure to reduce the number of conditions, initially suggested in (Schneider & Wagemann, 2006), makes the most of the fact that, in evaluations, causal factors are often grouped in categories. The idea is to perform the QCA analyses separately within different categories of factors; and successively test a unified model, created with the most important conditions emerged from each group. This strategy was adopted for the PA level analysis in the GEF/UNDP Biodiversity Evaluation, in which the initial model included 28 conditions grouped into 3 categories. The "capacity" category included 12 conditions, among which 4 emerged as the most important (CAstaff CAlocauth, CAotherextupp CAleader); the "community" group included 7 conditions, among which COMprovinf and COMcons were shown to be the most relevant; finally, the analysis of the conditions included in the "context" category (9 conditions) showed CXTpolconf to be the most important one, followed by CXTTHREAT and the combination of CXTTOURCUL and CXTACCPA (see Annex C for details).

After further examining the list of factors, the evaluation team decided to remove CXTtourcul while merging CXTthreat and CXTecval, which became CXTthreatecval. Two other conditions that did not emerge as particularly relevant from the first phase of this analysis but were deemed so on the basis of other strands of the study (which was a multi-method study with several research questions[39]) were kept in the final, 10-condition model: COMconcrben and CXTdevpres. In brief the final model tested included the 10 most important conditions from all the three categories (see Annex C).

When realist evaluation is combined with QCA (Befani, Ledermann, & Sager, 2007), causal factors are "naturally" divided into contexts and mechanisms: two groups on which the analyses could be

[39] https://www.thegef.org/gef/Impact%20Evaluation%3A%20GEF%20-%20UNDP%20Support%20to%20Protected%20Areas%20and%20Protected%20Area%20Systems

conducted separately, if the complete list of conditions is too long (see also Section 3.2.2).

2.7.1.9 OPPORTUNITY: some useful criteria to reduce the number of conditions

One potantial problem with the two-step procedure is that, even after grouping conditions and proceeding to separate analyses, the findings might still be too complex and difficult to interpret. The models might still include too many conditions. One strategy that the author has found very useful in all situations where models are complex and difficult to interpret is to eliminate conditions on the basis of the following criteria:

1. Consistency scores of solution terms (of combinations included in the solution)
2. Coverage scores of solution terms (of combinations included in the solution)
3. Number of times the same condition appears across solution terms
4. Necessity scores

Since the author uses the complex solution in the vast majority of cases to avoid having to justify assumptions on remainders, the consistency scores are almost always 1.00. When working on the solution to the 10-condition multi-category model of the GEF/UNDP Biodiversity Evaluation illustrated above, the author noticed that, among the 12, perfectly consistent combinations in the solution, only 5 covered more than one case: the last 7 combinations, all with 4% coverage scores, described only one case each. The author then decided to focus on the first five combinations, and in particular the first four which had much higher coverage values than the rest of solution terms. Following this strategy, the author noticed that only six conditions were consistently present in all four solution terms: CAleader, CAstaff, CAotherextsupp, COMprovinf, COMcons and CXTecvalthreat. This group of conditions was tested in a new model, which returned the solution illustrated in Table 49, with perfect consistency and coverage.

Table 49: Complex Solution of the 6-condition model explaining BIOtrend in the GEF/UNDP Biodiversity Evaluation

COMPLEX SOLUTION: CALEADER, CASTAFF, CAOTHEREXTSUPP, COMPROVINF, COMCONS, CXTTHREATECVAL

Combination	Raw Coverage	Unique Coverage	Consistency
CASTAFF*COMPROVINF*COMCONS*CXTTHREATECVAL*CALEADER	0.777778	0.074074	1
CASTAFF*COMPROVINF*COMCONS*CXTTHREATECVAL*CAOTHEREXTSUPP	0.777778	0.074074	1
CASTAFF*CALEADER*CAOTHEREXTSUPP*COMPROVINF*comcons	0.074074	0.074074	1
castaff*caleader*caotherextsupp*COMPROVINF*comcons*cxtthreatecval	0.037037	0.037037	1
castaff*CALEADER*caotherextsupp*COMPROVINF*comcons*CXTTHREATECVAL	0.037037	0.037037	1

Solution Coverage: 1.000000
Solution Consistency: 1.000000

The analysis of the above solution revealed that only two of five combinations covered more than two cases; namely 78% of cases each. These two combinations had 4 conditions in common, plus a fifth that could be either CAleader or CAotherextsupp.[40]. In other words, the solution offered an interesting 4-condition model to test, which produced very clear and strong findings: a "magic" combination of 4 conditions covering 85% of the cases, including uniquely (Table 50)[41]. The solution did not have perfect coverage (due to one contradictory combination that was excluded from this analysis) but the combination of four conditions can be considered a very successful recipe, because it consistently leads to success and covers, or explains, 23 out of a total 27 successful cases.

[40] This could be interpreted as meaning that, once the four essential conditions were met, a good leadership or external, non-governmental support were equivalent in their capacity to affect the outcome. However, QCA was just one of the methods used in the evaluation: the final findings in the report were triangulated against other methods and findings. The examples used in the guide are merely intended to demonstrate the use of the method.

[41] Two combinations covering 2 cases each (one each uniquely) were also included in the solution.

Table 50: Complex Solution of the 4-condition model explaining BIOtrend in the GEF/UNDP Biodiversity Evaluation

```
COMPLEX SOLUTION: CASTAFF, COMPROVINF, COMCONS, CXTTHREATECVAL
                                         Raw       Unique
Combination                              Coverage  Coverage  Consistency
CASTAFF*COMPROVINF*COMCONS*CXTTHREATECV
AL                                       0.851852  0.851852            1
COMPROVINF*comcons*cxtthreatecval        0.074074  0.037037            1
COMPROVINF*comcons*castaff               0.074074  0.037037            1
Solution Coverage: 0.962963
Solution Consistency: 1.000000
```

Finally, another useful criterion to reduce the number of conditions is to remove the necessary conditions from the model, including “trivial” conditions (see Section 2.5.1.5). If a condition is found to be necessary, its inclusion in the Boolean minimisation will not add any relevant information, as the latter is a synthesis of cases presenting the same outcome and any combination in the solution will thus include the necessary condition.

In sum, in order to reduce the number of conditions in the model, the evaluator can:

1. Remove the perfectly necessary conditions identified during the superset analysis
2. Focus on the terms from a complex solution with the highest consistency and coverage scores, and run the minimisation again using the conditions observed in these terms. This can be repeated until a small enough set of conditions is identified.
3. If the above strategy does not reduce the number of conditions to a sufficient extent, give priority to those conditions appearing in most solution terms

Following the above strategies will in most cases lead to manageable models and solutions with relatively high consistency and coverage scores.

2.8 Step Eight: The INUS analysis: Which conditions make the difference between success and failure in specific contexts?

The procedures illustrated above provide empirical support to statements of causal necessity or sufficiency. In particular, the necessity of single conditions (the superset analysis) and the sufficiency of single conditions or combinations of conditions (the subset analysis and the Boolean minimisation). When no single condition is necessary in itself, the necessity of disjunctions can be analysed, which might help shed light on the role of the intervention as one of two or three functionally equivalent factors, none of which is necessary in itself but at least one of which is required to produce the outcome[42].

Sometimes one single condition can be necessary, but not in the sense that the outcome can never materialise without it; rather in the sense that the combination it is part of cannot be sufficient without it. This special kind of necessity is called “INUS necessity”. INUS[43] causes are not necessary in an absolute sense for the outcome, but only for a combination to be sufficient. They were theorised by John Mackie (Mackie, 1974) and represent a sort of “local necessity” or “conjunctural necessity”: conditions that are necessary not in general, but within specific contexts and circumstances.

These conditions are only required under particular circumstances for the outcome, not in general, and can be argued to represent the reality of development and public policy well: it might be unrealistic to think that development interventions (or other factors) are always required under any circumstance to achieve goals in education, empowerment, health. Particularly not single interventions: they will always happen together with policies of national and local governments, other interventions, decisions of market operators, migration flows, historical and socio-cultural changes, etc. The notion that interventions likely play different roles under different circumstances is very popular in evaluation (Pawson & Tilley, 1997; Westhorp, 2014) and INUS necessity helps make sense of the role an intervention could play in a specific context. In particular, it can help show that, although the intervention is not always necessary to

[42] This is also known as “SUIN causality”.

[43] **I**nsufficient but **N**ecessary part of an **U**nnecessary but **S**ufficient combination.

achieve the outcome, it has been necessary under particular circumstances.

The packages INUS causes are included in are consistently sufficient. These causes are interesting because – unlike perhaps other factors in the same packages – when they are removed the combination is no longer sufficient: INUS causes make a difference to the outcome, without being necessary in absolute sense, or sufficient by themselves.

In the Boolean minimisation, similar cases differing in only one condition are merged, if they share the same outcome. But similar cases present a different outcome, they cannot be merged and an argument can be made that one condition makes the difference for the outcome, within that combination/context; that it is necessary to that particular statement of sufficiency, because when it is removed the combination is no longer sufficient for a positive outcome. That's why INUS causes can be spotted by following the "one difference rule", as in the Boolean minimisation; comparing cases one by one, and "coupling" those differing only in one condition. The difference with the Boolean minimisation is that for the INUS analysis such cases must differ *also* in the outcome.

INUS conditions are extremely relevant in impact evaluation because they answer the question "did it make a difference, for whom and under what circumstances" in a way that directly and automatically emerges from the data. They are based on the same causal logic (Mill's Method of Difference, see Annex A) underpinning counterfactual analysis; but at the same time a) can be based on "factual" as well as counterfactual data, and b) ask a context specific question, instead of seeking an average net effect.

The key strategy for spotting INUS causes is to look for similar cases that have a different outcome. The Venn diagram can be used for this purpose, for example looking for ***contiguous*** areas that have a different colour. When the evaluator is specifically interested in the role of the intervention, the latter should be the first condition included in the model so the dividing line becomes the central vertical one: cases on the right side have received the intervention and cases on the left side haven't. Ideally the evaluator will find symmetrical areas of different colour around the vertical axis: these would represent the contexts and conditions under which the intervention made a difference. If symmetrical areas of the same colour are observed, it means that the intervention, under those circumstances, didn't make any difference.

The rest of the section shows how the INUS analysis has been applied in real-life evaluations, in the first case isolating the context-dependent contribution of a single factor unrelated to the intervention; in the second case isolating the contribution of the intervention proper; and in the third case comparing the contribution of two different types of intervention.

2.8.1.1 OPPORTUNITY: answering the question "what factors made the difference under what circumstances"

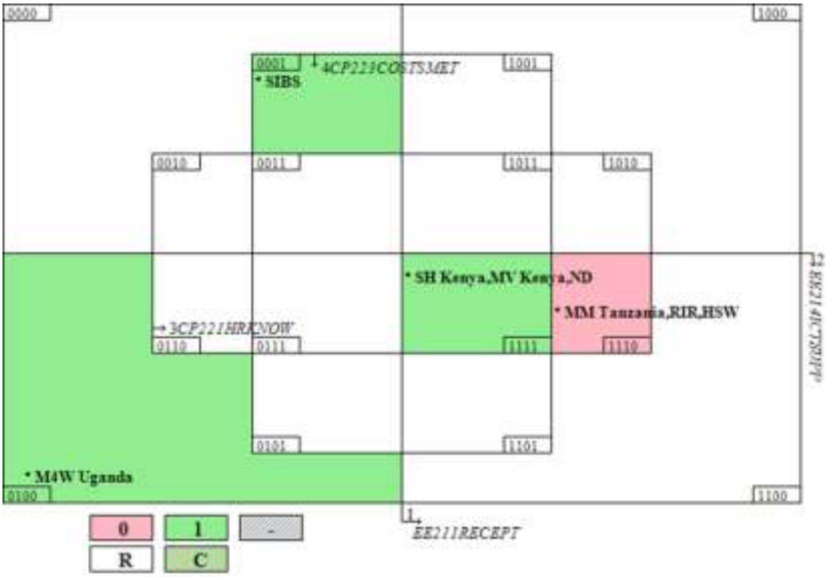
The condition that is discovered to make the difference can be the intervention or any other causal factor. In the MAVC study, one causal model was tested in an attempt to explain whether data about water points failure collected with ICT was processed and used to plan repairs. Four conditions were included: whether reception for mobile phone was of sufficient quality, whether ICT support was available, whether human resources had sufficient skills to process the data, and whether data processing costs were met by the service provider. None of the four conditions was found to be necessary in an absolute sense.

The analysis of "Costs met" tells an interesting story: the factor was subset-sufficient (over 4 cases), but not necessary: the case labelled M4WUganda was successful[44] even though costs were not met by the service provider. However, "costs met" made the difference in the 6 cases where the first three conditions were positive: reception, ICT support and HR knowledge. Under these favourable conditions, the fact that the service provider was meeting costs or not appeared to determine success or failure: the three cases were this happened (SH Kenya, MV Kenya and ND) were all successful, and three cases where

[44] The data processing was judged to be successful because the local government was involved in it: the SMS went straight to the local government system and the hand pump mechanic had to close the ticket when the repair was made. This was then, in theory, double-checked by the district water officer, and a way of increasing the Hand pump mechanic's accountability to the district water officer. However, it was later found in the qualitative case study that the system did not function as well as reported and the positive outcome was less warranted than initially thought. The implication was that the condition "costs met" became more credible as a necessary condition in an absolute sense (not just under the above reported specific circumstances): in all cases were the outcome was reliably judged positive costs were met by the service provider.

this didn't happen (MM Tanzania, RIR and HSW) were all unsuccessful.

Figure 16: Graphic Illustration of the 4-condition model explaining outcome 2 in the MAVC study



The Venn diagram (Figure 16) shows the influence of meeting costs on behalf of the service provider, represented by the central vertical rectangle, in a specific area/context. The two contiguous/ *adjacent* areas [1111] and [1110] are identical except for the last condition, which means they represent the same favourable context; but they are of a different colour/they present a different outcome. On the basis of this data, we can apply Mill's Method of Difference (see Annex A) and argue that – in this favourable context – the difference in the outcome can be attributed to the difference in meeting costs, as no other condition included in the model could be responsible for it.

The local or contingent influence of meeting costs in this favourable context can be represented as follows:

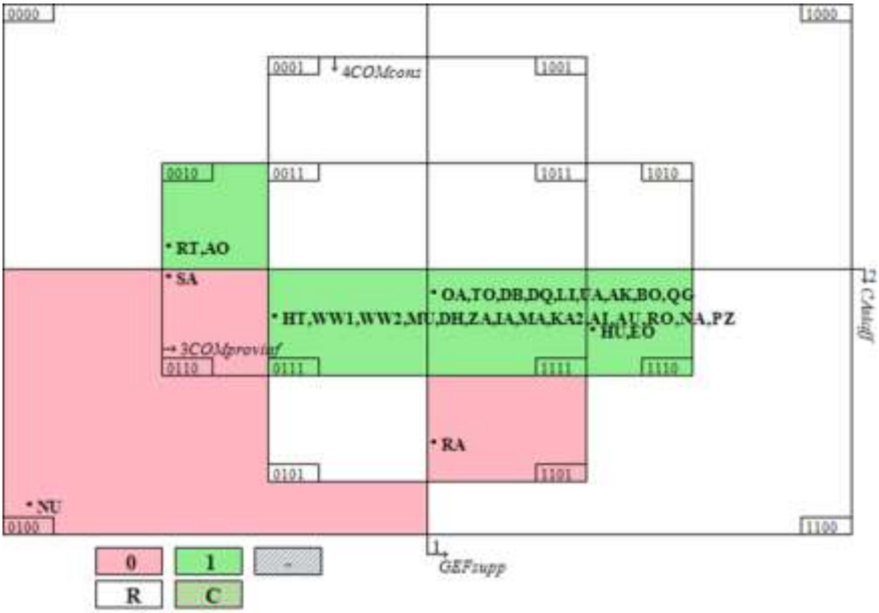
- RECEPT*ICTSUPP*HRKNOW*COSTSMET => DATAPROCESSING
- RECEPT*ICTSUPP*HRKNOW*costsmet => dataprocessing

In the GEF/UNDP Biodiversity Evaluation, using the INUS analysis allowed the evaluation team to have a clearer understanding of the role of GEF and other factors, in relation to the main outcome "decrease of illegal incidents in the PAs". Adding the condition "presence of GEF support" to the 4-condition model illustrated above allowed the evaluation team to make the following assumptions (see Figure 17)[45]:

- GEF support by itself is neither necessary nor sufficient for success in decreasing illegal incidents in the PAs
- When GEF support is provided and two other conditions are favourable (staff, community consultation), it is provision of information that makes the difference (which is what is missing in the "RA" case compared to OA, TO, DB, etc.).
  - GEFSUPP*STAFF*COMCONS*COMPROVINF => SUCCESS
  - GEFSUPP*STAFF*COMCONS*comprovinf => success
- When GEF support is not provided and two other conditions are favourable (staff, provision of information), it is the possibility to organise community consultations that makes the difference (which is what is missing in the "SA" case compared to HT, WW1, WW2, etc.).
  - gefsupp*STAFF*COMPROVINF*COMCONS => SUCCESS
  - gefsupp*STAFF*COMPROVINF*comcons => success

[45] Note that these were simplified, to some extent provoking, interpretations supported by this part of the analysis. The final evaluation findings emerged from a synthesis of multiple research strands, guided by different methodologies: the aim of this section is to merely illustrate the logic of the methodology.

Figure 17: Graphic illustration of a 4-condition model explaining BIOtrend in the GEF/UNDP Biodiversity Evaluation



2.8.1.2 OPPORTUNITY: answering the question "did the intervention make a difference, for whom and under what circumstances"

Learning that GEF support was neither necessary nor sufficient in an absolute sense was not terribly informative: more specific assumptions about the distinctive role of GEF support[46] have been identified by comparing Protected Areas benefiting with those not benefiting from it. Out of the 30 PAs studies, only 12 received GEF support. This comparison allowed the team to make the following hypotheses (see Figure 17):

- The intervention makes the difference between success and failure when two favourable conditions are met (staff, provision of information) and the opportunity of community consultation is not available (context -110). In the three cases where these conditions are present, the intervention is the only difference

[46] See disclaimer in the above footnote.

between the two successful cases (HU, EO) and the one unsuccessful cases (SA). This INUS association can be represented as:
  - STAFF*PROVINF*cons*INTERVENTION => SUCCESS
  - STAFF*PROVINF*cons*intervention => success
- When three favourable conditions are met (staff, provision of information, and availability of community consultation) the intervention does not make any difference. All these cases are successful, no matter if they receive support or not. Of the 23 cases where these conditions are observed, all successful, 9 have received GEF support and 14 haven't. This can be represented as:
  - STAFF*PROVINF*CONS*INTERVENTION (9) => SUCCESS
  - STAFF*PROVINF*CONS*intervention (14) => SUCCESS

These findings, based on a snapshot of the situation that doesn't take into account the dynamics of these phenomena, support the idea that GEF support might be more effective when opportunities for community consultations are rarer; while when the latter are more established its added value is less apparent[47].

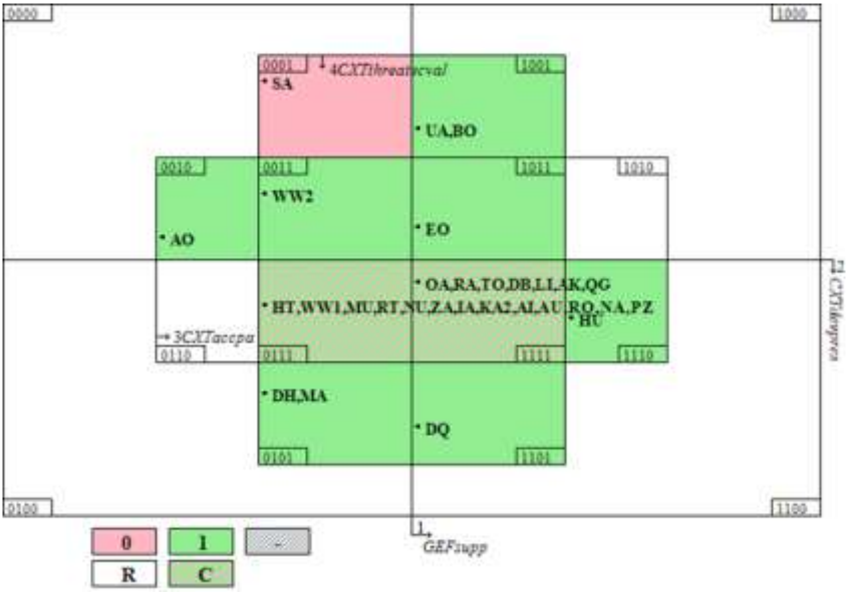
The INUS analysis can be conducted on different models to shed light on other combinations of conditions under which GEF support makes the difference. The analysis of the context conditions allowed the team to make the following assumptions[48] (see Figure 18):

- The intervention makes the difference when the PA has high value (in terms of threatened species or economically); is difficult to access, and is not subject to "development pressures". Among the cases presenting these conditions (the rectangle on the very top), the intervention is the only difference between the two successful cases UA and BO and the one unsuccessful case of SA. This can be represented as:
  - accpa*THREATECVAL*devpres*INTERVENTION => SUCCESS
  - accpa*THREATECVAL*devpres*intervention => success

[47] Note that the INUS analysis is helpful in developing hypotheses which might need to be confirmed at a later stage, often by in-depth case studies, aimed at uncovering mechanisms which explain why those associations are observed. In this specific evaluation these assumptions emerged at a late stage and were eventually not followed up to/verified.
[48] See previous footnote.

- When the PA has high value, is difficult to access, and it is – unlike in the above case – subject to development pressures, the intervention does not make any difference. The group of three cases where these conditions are met (the rectangle at the very bottom) are all successful independently of whether the intervention has been implemented or not. One has received GEF support (DQ) and two haven't (DH, MA). This can be represented as:
  - accpa*THREATECVAL*DEVPRES*INTERVENTION => SUCCESS
  - accpa*THREATECVAL*DEVPRES*intervention => SUCCESS

Figure 18: Graphic Illustration of a 4-condition model showing the role of the intervention



The fact that the absence of development pressures enables GEF support to make a difference under the context "accpa*THREATECVAL" could be explained in a number of ways. However, this type of analysis was experimented with towards the end of the evaluation and in-depth, case-based work to explain the association was not feasible at the time, within the scheduled timeframe.

2.8.1.3 OPPORTUNITY: answering the question “which type of intervention was the most effective in a specific context?”

Another way the INUS analysis can be useful is in comparing the effectiveness of different types of intervention implemented in the same context. In the Budget Support Evaluation, where the increase in female enrolment in primary school was explained with a four condition model including gender working groups, gender indicators in the PAF, free education and aid levels for education, no single condition was found to be necessary for success, but one INUS cause (and one “almost INUS” cause) can be identified.

The INUS cause is observed in the context of [–11], where education is free and aid levels for education are high (the square-ish rectangle in the middle). In such contexts setting up gender working groups makes the difference between success and lack of it: among the 6 cases presenting free education and high levels of education aid (all located in the central rectangle) the two unsuccessful ones (Gambia and Zambia) have not set up gender working groups, while the four successful ones (Malawi, Ethiopia, Mozambique, Tanzania) have. At the same time, having gender indicators in the PAF does not make as much of a difference: in the same context/square-ish rectangle, there are both successful and unsuccessful cases on the left (where the condition is negative) and on the right (where the condition is positive). However, it does make a small difference because the proportion of successful cases improves when gender indicators are included in the PAF (from ½ to ¾).

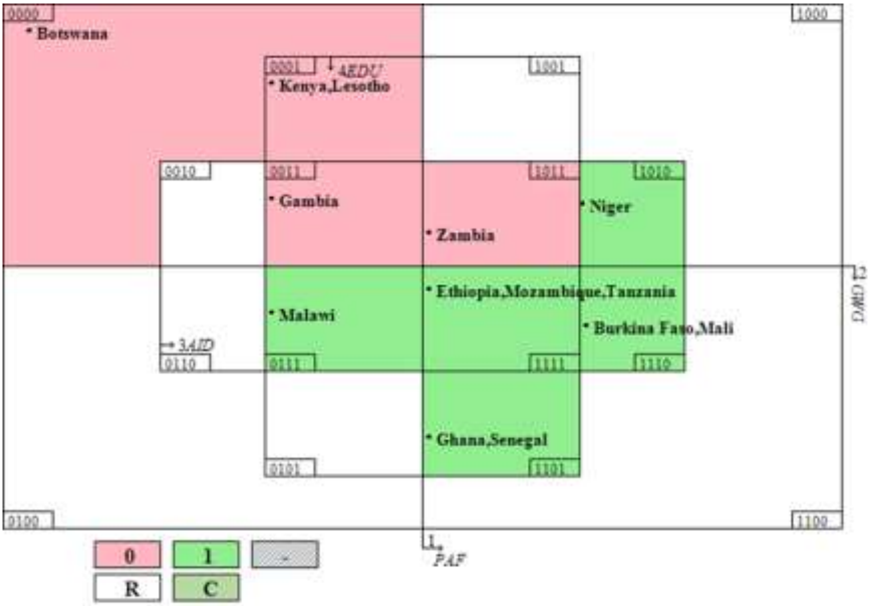
Here, the INUS analysis allows us to answer the question “what makes the difference in favourable contexts with free education and high levels of aid”? And the answer is that setting up gender working groups makes a bigger difference than including gender-sensitive indicators in the programming document.

This INUS associations for GWG and PAF can be represented as follows:

- EDU*AID*GWG => ENROL
- EDU*AID*gwg => enrol
- EDU*AID*PAF => 75% chance of ENROL
- EDU*AID*paf => 50% chance of ENROL

Including GWG brings the chances of success from 0% to 100%, while including PAF only raises those chances from 50% to 75%.

Figure 19: Graphic Illustration of the BSE dataset, showing the different role of the two policy instruments GWG and PAF



2.8.2 Using the INUS logic to identify tipping points in complex dynamic systems

Some authors argue that QCA can't be used to analyse time series or temporal dynamics (Caren & Panofsky, 2005; Schneider & Wagemann, 2012). They refer to QCA's Boolean minimisation algorithm and suggest ways to overcome this limitation. Here we argue that Boolean or fuzzy datasets normally used in QCA are, in themselves, well suited to analyse trends like qualitative time series describing system states; but if they are to serve this function they must not be synthesised with the Boolean minimisation. In particular, we argue that the INUS analysis is useful to understand the conditions under which dynamical systems "tip" from one state to another.

For example, if we think of conflict and peace as two outcomes but also as two "strange attractors" (Ramalingam, Jones, Reba, & Young, 2008) and imagine a series of factors influencing the position of the system at any given time (closer to attractor one, peace, or attractor two, war), we can describe and compare system states as displayed in Table 51.

Table 51 displays fictitious data about the presence and absence of six conditions assumed to contribute to the achievement of peace (vs. conflict). The cases are "system states" recorded either in the same country or in different countries[49].

Table 51: Factors contributing to peace for 13 system states.

	SEC	DRIVER	POWER	GOV	ECO	COHES	PEACE
State A	1	1	1	1	1	1	1
State B	1	0	1	1	1	0	1
State C	1	1	0	0	1	1	1
State D	0	0	1	1	1	1	1
State E	0	0	1	1	1	0	1
State F	1	0	1	1	0	0	1
State G	0	0	1	1	0	1	1
State H	0	1	0	0	1	1	0
State I	1	1	0	0	0	1	0
State J	0	0	0	1	1	0	0
State K	0	0	1	1	0	0	0
State L	1	0	1	0	0	0	0
State M	0	0	0	0	0	0	0

Note: Factors in table are: Physical Security and sense of security (SEC); Acknowledgement of key conflict drivers and commitment to address them (DRIVER); Durable political arrangement for handling power (POWER); Good enough governance - resilient relationship between government and society (GOV); Economic fairness and opportunity (ECO); Social Cohesion (COHES).

Some extreme states with all zeros (like M) and all ones (like A) can be observed, and they have a predictable outcome. However, there is also a middle ground where the outcome is more uncertain.

At the top, after State A with all 1's, we see that states with 2 inconsistent and four consistent conditions (with the outcome), present the expected outcome (States B C and D, highlighted in green). Then there is a middle area (highlighted in gray) where states with three positive and three negative causal conditions exhibit

[49] What matters for this exercise is that cases are comparable on the six conditions plus the outcome: they don't need to describe dynamics observed in a single country, although they could. The author would like to thank Diana Chigas and Peter Woodrow for their help, inspiration and insight in creating and discussing this example.

uncertainty as to the outcome (States E F G H and I; the first three having a positive outcome and the last two a negative one).

Within this subset, we can use Mill's Method of Agreement (see Annex A) and compare the three successful states (E, F, and G): we realise that the only consistently present conditions are POWER, GOV and "driver". This might mean that in uncertain situations it is particularly important to ensure the presence of a durable arrangement to handle power distribution and a resilient relationship between the government and society, even in the absence of commitment to acknowledge key conflict drivers. Similarly, in the unsuccessful states (H and I) the only consistent causal factors are power, gov and DRIVER, which is consistent with the insight gained above.

Most importantly, these "gray" states are those on the "edge of chaos", where a small change can make a big difference.

For example, if we compare the two similar states with a different outcome E & J, we realise that in the context of sec*driver*POWER*GOV*ECO*cohes (case E), losing a durable arrangement for power distribution and transitioning to sec*driver*power*GOV*ECO*cohes (case J) can make the difference and bring chaos/conflict to the country. Similarly, by comparing F and L we see that in the context of SEC*driver*POWER*gov*eco*cohes (case L), improving the relationship between the government and society, and transitioning to SEC*driver*POWER*GOV*eco*cohes (case F), can bring peace.

More generally, the Boolean dataset and in particular the INUS analysis allow us to see what small changes tip the system to the other attractor (to conflict from peace, or to peace from conflict) under different circumstances[50].

[50] A similar metaphor can be drawn from evolutionary theory and the notion of "fitness landscapes" (many thanks to Rick Davies for drawing my attention to this metaphor). High points on a fitness landscape are points of higher fitness (e.g. a configuration with high consistency). Incremental changes to a configuration move it across the landscape, sometimes to higher fitness levels, sometimes to lower fitness levels. Some landcsapes are very rugged: small changes in a configuration can lead to substantial changes to fitness. Other landscapes can be very smooth, with any incremental change in the configurtaion leading to only modest changes in its fitness.

3 GENERALISATION, BIAS AND THE DIALOGUE WITH THEORY

This final chapter includes cross-cutting reflections that are relevant at multiple stages of a QCA analysis and that it was difficult to include in any detail in the discussion of specific steps in the previous chapter. The first is the dialogue with theory: both model specification and interpretation of the findings are possibly the most critical moments in QCA, denoting the iterative nature of the method and often determining the number of iterations in practice. Theory intervenes at multiple stages, before and after every synthesis procedure (supersubset analysis, Truth Table, Boolean minimisation, INUS analysis).

A second major issue in QCA is generalisation, because it's the key advantage of QCA compared to within-case methods, and offers multiple opportunities in terms of both outside-the-dataset and within-the-dataset generalisation (also known as synthesis); yet it needs theory and a minimum number of cases. The goal of all QCA procedures is to support theoretical statements using the information empirically available in the dataset, and hence strengthen the empirical basis of those statements.

Finally, every research method is subject to biases of multiple kinds and QCA is no exception. As a guide aiming to encourage high quality applications of QCA to evaluations, we could not avoid a section on bias and on proposed solutions to mitigate its consequences. Biases intervene potentially in all steps, but their consequences are can be particularly serious in model specification, case selection and data collection (the latter two not technically QCA-specific activities, but still affect QCA findings).

The chapter ends with a Quality Assurance checklist, aimed at helping commissioners and evaluators quality assure the QCA components of their evaluations. We start with generalisation because the end of the section links well with the section on the dialogue with theory.

3.1 Generalisation

One typical concern shown towards QCA by evaluators with a quantitative background is the robustness and generalisability of its findings. When the empirical findings from a QCA analysis support a necessity or sufficiency statement, what exacty does that statement apply to?

The standard answer is that QCA findings apply to the dataset only and QCA is only able to perform a “modest generalisation” (Berg-Schlosser, De Meur, Rihoux, & Ragin, 2009) or “limited generalisation” (Ragin, 1987); the notion of “contingent generalisation” has also been suggested (Blatter & Blume, 2008) to argue that QCA adopts a qualitative and “case-oriented” approach to generalisation.

“Contingent generalisation” refers to QCA procedures that aim to find a simple and parsimonious way to synthesise the information included in the dataset. These procedures can also incorporate information coming from known cases that are not represented in the dataset, but essentially this is a “within-the-dataset” kind of generalisation (because those cases are eventually added to the dataset). This is the type of generalisation we have addressed in the above chapter, the type operated by the mentioned procedures (supersubset analysis, Truth Table, Boolean minimisation, INUS analysis).

In the supersubset analysis we can focus on specific sections of the dataset and find a parsimonious way to represent a necessity or sufficiency relation between one or some conditions and the outcome: the findings will apply to the entire set of cases. Similarly, when building the Truth Table we obtain a limited number of different pathways that are more or less strongly associated with the outcome: the set of pathways as a whole represents the entire set of cases. Finally, the goal of the boolean minimisation is to synthesise the Truth Table even further, without losing either sufficiency-consistency or case coverage. These procedures are covered in detail in Chapter 3.

When the dataset represents the entire population, the relations of necessity and sufficiency identified by the QCA procedures can go beyond “contingent generalisation” and be considered “universally valid": in this sense, under these circumstances, they do not different substantially from descriptive statistics.

In other cases, the dataset represents a sample of a larger set of cases which are all potentially available for analysis. In development evaluation, choice of cases is often limited by resources, time constraints, and quality of relations with local partners. As far as the author knows, no QCA application has ever relied on random extraction for case selection. In the most common development evaluation scenario, the dataset represents the population of cases the evaluation team was realistically able to collect high quality data on, with some attention for the diversity of cases. Care is usually taken so that the sample represents at least some different contexts, with some kind of stratification taking place. In some evaluations, the cases can be argued to have been selected independently of each other: but randomisation is rare.

In spite of this, we will see that in some cases QCA findings can be generalised outside of the dataset, as in "statistical generalisation". We will also see that QCA can aid "conceptual generalisation".

We thus approach the issue of generalisation in two different ways. In addition to "contingent generalisation" which has been addressed in Chapter 3, "statistical generalisation" relates to the way statisticians think about generalisation: asking what the findings would look like if data were random. The objective here is to generalise to a broader population including, but not limited to, the dataset: the idea is that if the associations identified are shown to be significantly different from what would be expected from a random dataset of the same size, they indicate a "real" regularity and can be generalised outside the dataset. We can call this "outside-the-dataset" generalisation.

The second type is generalisation of concepts: or the process leading from specific and detailed concepts applying to a small number of cases to broader and more abstract concepts, applying to a higher number of cases. This kind of generalisation, which is possibly more familiar for evaluators with a qualitative background, is also "within-the-dataset" but uses information about cases which is not necessarily explicit in the dataset and usually can't be obtained through the above mentioned contingent generalisation.

3.1.1 Statistical Generalisation: how many cases are needed to generalise to the whole population?

Commissioners of evaluations analysing multiple cases are often interested in knowing if the findings are generalisable to a broader population. One question they might ask is "**how many cases are needed for robust generalisation to the entire population**"? For example, this might happen at the end of pilot studies conducted on a small set of cases, with the commissioner wanting to enlarge the sample to increase their confidence in the findings. The evaluator using QCA must therefore be prepared to answer such questions.

One approach to generalisation adopted by statisticians is the creation of tests of hypotheses aiming to reject an assumption of interest. Tests are judged robust when they have high levels of sensitivity and/or specificity, and low levels of Type I and/or Type II error (see Table 52). In statistical inference Type I error is denoted with "alpha" and usually set at 5%, which means that findings obtained from a test with an alpha higher than 5% are not considered significant. We can use a similar logic to assess the robustness of QCA statements. We do it in two different ways depending on whether we are testing the robustness of findings from the supersubset analysis, or the robustness of Truth Table rows as sufficiency statements (and hence the robustness of the Boolean minimisation).

3.1.1.1 Generalising Findings from the SuperSubset Analysis

The SuperSubset analysis produces statements like "the combination A*B is sufficient for the outcome" or "has 90% sufficiency-consistency"; and "the condition C is necessary for the outcome" or "has 80% necessity-consistency", and so on. These statements are made on the basis of the frequency with which the condition and the outcome are observed together in the same case, across the set of cases, compared to a maximum frequency (the total number of cases with a positive outcome for the necessity analysis, and the total number of cases presenting that condition/combination for the subset analysis).

Let's take the example of necessity. If the condition "presence of a champion" is observed in all successful cases, the necessity-consistency is perfect. However, if the dataset does not represent the entire population and is extracted from a larger set of cases, the necessity statement might have a different meaning depending on how many cases are included in the dataset – e.g. 15 vs. 3 successful cases. Intuitively, we understand that, if the values of the conditions are random or due to chance, rather than being indicative of an underlying regularity, the probability of all 3 successful cases consistently presenting champions is higher than the probability of 15 successful cases consistently doing the same. Put differently, our data might be misleading, showing a perfectly consistent superset relation, while there is no underlying regularity and the observed consistency is due to chance.

The same risk applies to the subset analysis: with a small number of cases, the observed high levels of consistency might be misleading, indicating a relationship that doesn't exist (is not as consistent in reality). It helps to compare these findings to those we would obtain under random conditions.

We can set up a test of hypothesis where the null hypothesis states that the underlying relation between the conditions analysed is regular in reality, and the alternative hypothesis states that the same relation is random, like a coin being flipped as many times as cases are observed. Table 52 illustrates the logic of the test of hypothesis, showing how sensitivity, specificity, Type I error, Type II error and predictive values are calculated from the numbers of true positives and false positives.

Table 54 shows these values for various sample sizes. If the underlying relation is regular, the test will always be consistent so its sensitivity is perfect (column one). Column two shows the Type I error, or the probability of the same test showing the same consistent values if the underlying relation is random. This value is very important in statistics: it is usually denoted with "alpha" and set at a minimum of 0.05. The values in the column show that the test starts becoming significant when the sample size is 6 or greater: that's when the Type I error becomes lower than 0.05. This means that that **perfect consistency in supersubset analysis is statistically significant when it is measured over 6 or more cases**[51].

Notice that the Type II error of this test is always zero (if the relation is regular it's impossible to observe imperfect consistency), as shown in column three; and that the specificity of this test is greater

[51] Provided they are extracted randomly or assumed to be independent.

than 95% when the number of cases is 6 or higher. Similarly, the negative predictive value is always 1 because observing imperfect consistency automatically proves that the underlying relation is irregular.

Perhaps the most insightful column in Table 54 is #10: the "Positive Predictive Value". This estimates the probability of the underlying relation being regular given that perfect consistency has been empirically observed over x cases. For six cases, the probability of the relation being regular in reality is greater than 97%; and it becomes greater than 99% for 8 cases and over. Put differently: **a perfectly consistent supersubset relation observed over 8 or more cases[52] means that the probability of the relation being regular in reality is higher than 99%.**

Table 52: **Logic of the test of hypothesis "the relation is regular in reality"**

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Reality within the broader population (ontological reality)</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>The relation between the configuration and the outcome is regular</td><td>The relation between the configuration and the outcome is irregular (random)</td></tr>
<tr><td rowspan="2">Evidence from the dataset (observable reality)</td><td>Perfect supersubset relation</td><td>True Positive (A)</td><td>False Positive (B)</td><td>Positive Predictive Value = A / (A + B)</td><td>False Discovery Rate = B / A + B)</td></tr>
<tr><td>Imperfect supersubset relation</td><td>False Negative (C)</td><td>True Negative (D)</td><td>False omission rate = C / (C + D)</td><td>Negative Predictive Value = D / (C + D)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>True positives rate = Sensitivity = 1 – Type II error = A / (A + C)</td><td>False positives rate = 1 – Specificity = Type I error = B / (B + D)</td><td>Likelihood ratio = TPR / FPR = Sensitivity / Type I error</td><td rowspan="2"></td></tr>
<tr><td>False negatives rate = Type II error = 1 – Sensitivity = C / (A + C)</td><td>True negatives rate = Specificity = 1 – Type I error = D / (B + D)</td><td></td></tr>
</table>

[52] See above footnote.

We can take a step further and use the binomial probability distribution to check to what extent consistency scores of subset relations denote a significant sufficient or necessary relation between a configuration and an outcome. Let's say for example that we observe a consistency score of 75%. Is that a significant indication of a supersubset relation? Or is it just due to chance?

In order to answer this question, we can check the values of the binomial probability distribution for p = 0.5, which means that presence and absence of the configuration are equally related to the outcome: the situation furthest away from the underlying "real" relation being regular, where looking for evidence would resemble the flipping of a perfectly symmetrical coin.

Table 53[53] reports a series of relevant probabilities for this test by number of trials, which represent cases extracted randomly (or at least assumed to be independent – column two). Note that the number of successes is always exactly 75% the number of cases or trials (columns two, three and four).

Table 53[54]: Relevant binomial probabilities for p = 0.5 and an observed consistency score of 0.75

Probability of success on a single trial	Number of trials (cases)	Number of successes (x)	Observed Consistency Score	Cumulative Probability: P(X < x)	Cumulative Probability: P(X >= x)
0.5	4	3	0.75	0.6875	0.3125
0.5	8	6	0.75	0.855	0.145
0.5	16	12	0.75	0.962	0.038
0.5	32	24	0.75	0.997	0.004
0.5	64	48	0.75	1	0

The most insightful columns for us are number five and six. In particular, column 6 shows the probability of observing a consistency score of 75% or higher for the different numbers of trials/cases. When the number of trials is 16, this probability is already lower than 0.05 (namely 0.038) so we can argue that our consistency score is likely not to have been generated by the random relation embodied by p=0.5. Put differently, under random circumstances (p=0.5) the probability

[53] The same table can be calculated for different values of the binomial parameter p and different values of observed consistency in the supersubset relation using this calculator http://stattrek.com/online-calculator/binomial.aspx

[54] See footnote above

of observing a consistency value lower than 0.75 is already very high (0.96) for 16 cases (column five). **This means that if we observe a consistency score of 0.75 or higher over 16 cases, the underlying relation is unlikely to be random**. As the number of cases/trials increases, the same consistency score of 75% will be increasingly indicative of a "skewed", rather than symmetrical, supersubset relation.

3.1.1.2 Assessing the robustness of Truth Table rows as sufficiency statements

When building the Truth Table, the intention of the researcher is to identify which (complex) sufficiency statements are empirically supported in the dataset. After identical cases are merged and consistency scores calculated, perfect sufficiency-consistency is usually considered evidence of sufficiency. However, it can be argued that the high values of sufficiency-consistency scores of Truth Table rows are due to chance.

From the previous section we learned that, when condition values are random, consistency of association with the outcome is more likely to be observed with a low number of cases (see Table 54), and this also holds for Truth Table rows. However, the number of conditions included in the model also influences the probability of sufficiency-consistency in Truth Tables. When we add conditions to the model, if the number of cases does not change, the number of possible combinations of conditions rises exponentially, and Truth Tables will likely have a higher number of rows (complete Truth Tables certainly do). Since the number of cases doesn't change, each row will cover, on average, a lower number of cases. This increases the probability of each row being perfectly sufficiency-consistent.

By contrast, if conditions stay the same and we add cases, the complete Truth Table has the same number of rows. Some of these cases might present different combinations not included in the previous standard (empirically supported) Truth Table but on average, as the number of cases increase, the number of cases covered by each Truth Table row also increases; hence the chance of having mixed-outcome cases for each Truth Table increases in turn, and each row will be less likely to be perfectly sufficiency-consistent.

Table 54: Significance parameters for the supersubset analysis

Number of cases	Probability of observing a perfect supersubset relation if the underlying relation is regular in reality (Sensitivity = 1 - Type II error, true positives rate) A / (A + C)	Probability of observing a perfect supersubset relation if the underlying relation is irregular (random) (False positives rate = 1 – Specificity = Type I error = B / (B + D))	Probability of observing an imperfect supersubset relation if the relation is regular in reality (Type II error, false negatives rate) C / (A + C)	Probability of observing an imperfect supersubset relation if the relation is irregular (random)	True Positives (A)	False Positives (B)	False Negatives (C)	True Negatives (D)	Positive Predictive Value = A / (A + B)	Negative Predictive Value = D / (C + D)	Likelihood ratio = TPR / FPR = Sensitivity / Type I error
				True negatives rate = Specificity = 1 – Type I error = D / (B + D)							
1	1	1	0	0	2	2	0	0	0.5	-	1
2	1	0.5	0	0.5	4	2	0	2	0.6667	1	2
3	1	0.25	0	0.75	8	2	0	6	0.8	1	4
4	1	0.125	0	0.875	16	2	0	14	0.8889	1	8
5	1	0.0625	0	0.9375	32	2	0	30	0.9412	1	16
6	1	0.03125	0	0.96875	64	2	0	62	0.9697	1	32
7	1	0.015625	0	0.984375	128	2	0	126	0.9846	1	64
8	1	0.007813	0	0.992188	256	2	0	254	0.9922	1	128
9	1	0.003906	0	0.996094	512	2	0	510	0.9961	1	256
10	1	0.001953	0	0.998047	1024	2	0	1022	0.9981	1	512

When data are randomly generated, the likelihood of observing sufficiency-consistency of Truth Table rows is directly proportional to the number of conditions and indirectly proportional to the number of cases. Under random assumptions, there is a higher chance of observing sufficiency-consistent Truth Table rows when the number of cases is low and the number of conditions is high. This makes sufficiency-consistent Truth Tables rows more believable, and a more reliable indication of an underlying regularity rather than a random phenomenon, when the number of conditions is small and the number of cases is relatively large.

Marx & Dusa (2011) have calculated the probability that at least one Truth Table row is consistent because of chance, by simulating millions of random Truth Tables with fixed numbers of cases and conditions; and calculating the % of these Truth Tables that present at least one not perfectly sufficient (contradictory) combination. The simulation was carried out for many combinations of numbers of cases and conditions: as the former increased and the latter decreased, the probability of Truth Tables presenting at least one contradictory combination increased.

When the probability of obtaining contradictory combinations out of random data is high, it can be argued that the consistent (non-contradictory) sufficiency statements obtained empirically are credible; it is likely that they are not due to chance. Table 55, from Marx & Dusa (2011), reports the frequency of randomly-generated Truth Tables reporting at least one contradictory combination, for various combinations of numbers of conditions and cases. This frequency, or probability, is considered a measure of robustness, or significance, of Truth Table rows as sufficiency statements.

For models with two conditions (plus the outcome), obtaining contradictory combinations out of random data becomes a practical certainty from 12 cases onwards. For models of 3 conditions plus the outcome, 15 cases are needed, and so on. The authors suggest that at least a 0.9 probability benchmark be used in applications, which would make 10 the minimum number of cases that need to be used for models of 3 conditions; 13 for models of 4 conditions; 18 for models of 5 conditions, and so on.

Table 55: Levels of Confidence in Truth Table rows representing sufficiency statements, by number of cases and number of conditions

		# conditions (excluding outcome)								
		2	3	4	5	6	7	8	9	10
# cases	2	0.12	0.06	0.03	0.01	0.01	0.00	0.01	0.00	0.00
	3	0.34	0.18	0.09	0.04	0.03	0.01	0.00	0.00	0.00
	4	0.56	0.31	0.19	0.09	0.05	0.03	0.01	0.00	0.00
	5	0.70	0.47	0.26	0.15	0.09	0.04	0.02	0.01	0.01
	6	0.83	0.62	0.37	0.22	0.11	0.05	0.03	0.01	0.01
	7	0.92	0.72	0.50	0.30	0.16	0.07	0.04	0.02	0.01
	8	0.96	0.84	0.60	0.35	0.19	0.10	0.05	0.03	0.01
	9	0.97	0.89	0.68	0.43	0.24	0.12	0.07	0.05	0.02
	10	0.98	0.94	0.76	0.51	0.30	0.17	0.08	0.03	0.02
	11	0.99	0.97	0.83	0.58	0.36	0.17	0.09	0.05	0.02
	12	1.00	0.98	0.87	0.65	0.41	0.23	0.13	0.06	0.04
	13	1.00	0.98	0.92	0.70	0.48	0.28	0.14	0.08	0.04
	14	1.00	0.99	0.94	0.76	0.49	0.29	0.17	0.10	0.04
	15	1.00	1.00	0.96	0.81	0.58	0.35	0.19	0.10	0.05
	16	1.00	1.00	0.96	0.84	0.61	0.39	0.20	0.12	0.06
	17	1.00	1.00	0.99	0.88	0.65	0.40	0.23	0.13	0.06
	18	1.00	1.00	0.99	0.92	0.71	0.47	0.26	0.14	0.07
	19	1.00	1.00	0.99	0.93	0.72	0.51	0.28	0.18	0.07
	20	1.00	1.00	1.00	0.95	0.78	0.52	0.30	0.17	0.08
	21	1.00	1.00	1.00	0.95	0.82	0.57	0.35	0.18	0.09
	22	1.00	1.00	1.00	0.97	0.84	0.58	0.38	0.20	0.11
	23	1.00	1.00	1.00	0.98	0.87	0.64	0.40	0.20	0.12
	24	1.00	1.00	1.00	0.98	0.89	0.67	0.43	0.23	0.14
	25	1.00	1.00	1.00	0.99	0.91	0.69	0.43	0.26	0.14
	26	1.00	1.00	1.00	1.00	0.92	0.73	0.47	0.26	0.16
	27	1.00	1.00	1.00	1.00	0.94	0.73	0.50	0.30	0.17
	28	1.00	1.00	1.00	1.00	0.94	0.78	0.53	0.30	0.18
	29	1.00	1.00	1.00	1.00	0.97	0.79	0.55	0.32	0.18
	30	1.00	1.00	1.00	1.00	0.97	0.80	0.58	0.36	0.20
	31	1.00	1.00	1.00	1.00	0.97	0.83	0.59	0.35	0.22
	32	1.00	1.00	1.00	1.00	0.98	0.87	0.64	0.36	0.20
	33	1.00	1.00	1.00	1.00	0.98	0.87	0.65	0.40	0.24
	34	1.00	1.00	1.00	1.00	0.99	0.90	0.67	0.42	0.24
	35	1.00	1.00	1.00	1.00	0.99	0.90	0.68	0.45	0.26
	36	1.00	1.00	1.00	1.00	1.00	0.90	0.70	0.46	0.26
	37	1.00	1.00	1.00	1.00	0.99	0.93	0.72	0.50	0.29
	38	1.00	1.00	1.00	1.00	1.00	0.92	0.75	0.48	0.31

39	1.00	1.00	1.00	1.00	1.00	0.95	0.76	0.51	0.29
40	1.00	1.00	1.00	1.00	1.00	0.95	0.78	0.54	0.33
41	1.00	1.00	1.00	1.00	1.00	0.95	0.79	0.59	0.32
42	1.00	1.00	1.00	1.00	1.00	0.97	0.83	0.54	0.35
43	1.00	1.00	1.00	1.00	1.00	0.97	0.83	0.58	0.35
44	1.00	1.00	1.00	1.00	1.00	0.97	0.84	0.59	0.39
45	1.00	1.00	1.00	1.00	1.00	0.98	0.86	0.63	0.37
46	1.00	1.00	1.00	1.00	1.00	0.98	0.87	0.63	0.39
47	1.00	1.00	1.00	1.00	1.00	0.99	0.89	0.66	0.41
48	1.00	1.00	1.00	1.00	1.00	0.98	0.89	0.66	0.42
49	1.00	1.00	1.00	1.00	1.00	0.99	0.90	0.71	0.44
50	1.00	1.00	1.00	1.00	1.00	0.99	0.90	0.70	0.43

Source: Table 4 in Marx & Dusa (2011)

3.1.2 Conceptual Generalisation: merging conditions into more abstract constructs

The types of generalisation covered so far are to some extent automatic, or algorithmic, and are all aided by software tools. They all come with generalisation or synthesis rules:

- the necessity analysis groups the cases with the same outcome and calculates frequencies;
- the subset analysis groups cases with the same conditions and calculates frequencies;
- the Truth Table procedure groups identical cases (on all conditions and the outcome) and merges them into one single combination;
- the Boolean minimisation merges Truth Table combinations with the same outcome and at most one different condition;
- the INUS analysis compares/isolates Truth Table combinations with a different outcome and at most one different condition.

In QCA there is no automatic procedure to synthesise cases differing in more than one condition, even with the same outcome. Nonetheless, it is possible to spot similarities between such cases and merge them, if we are able to find abstract concepts encompassing these cases' differences. Realist syntheses (Pawson, 2006) use a similar logic of grouping cases by increasing the degree of abstraction of the constructs that are used to describe them. In addition to allowing within-the-dataset and outside-the-dataset generalisation, QCA also facilitates such "conceptual syntheses".

Let's take the example of an impact evaluation of a nutrition programme in Bangladesh. The hypothesis under test is that better nutritional knowledge acquired by mothers in poor households leads to better nutritional outcomes for children (Cartwright, 2012). The example considers three factors which are assumed to affect nutritional outcomes: who is the person acquiring the knowledge transferred through the intervention within the household; who controls food distribution within the household; and whether food supplements are distributed by the intervention along with knowledge transfers about nutrition best practices.

Table 56 lists six sufficient combinations of the three factors, two leading to an improvement of nutritional outcomes and four leading to a lack of improvement. Table 57 is a simplification of Table 56 where conditions are indicated by a letter; and the first two combinations of Table 56 are split into two combinations each to indicate who controls the food (which is irrelevant for success because food supplements are not distributed). Finally, Table 58 represents the "Booleanisation" of Table 57, which can be analysed with QCA.

Table 56: Combinations (not) leading to improvement in nutritional outcomes

Case #	Combination	Successful?
1-2	Knowledge acquired by mother (M) * food supplements not distributed (N)	NO
3-4	Knowledge acquired by grandmother (G) * food supplements not distributed (N)	NO
5	Knowledge acquired by mother (M) * food supplements distributed (Y) * food controlled by grandmother (G)	NO
6	Knowledge acquired by grandmother (G) * food supplements distributed (Y) * food controlled by grandmother (G)	YES
7	Knowledge acquired by mother (M) * food supplements distributed (Y) * food controlled by mother (M)	YES
8	Knowledge acquired by grandmother (G) * food supplements distributed (Y) * food controlled by mother (M)	NO

Table 57: Systematic Comparison of Combinations (not) leading to improvement in nutritional outcomes

Case ID	Who acquires the knowledge (KNOW)	Food Supplements Distributed? (FOOD)	Who controls the food? (CTRL)	Success (O)
1	M	N	M	N
2	M	N	G	N
3	G	N	M	N
4	G	N	G	N
5	G	Y	G	Y
6	M	Y	M	Y
7	G	Y	M	N
8	M	Y	G	N

Table 58: Conversion of Table 57 into a Boolean dataset

Case ID	Who acquires the knowledge? M=1, G=0	Food Supplements Distributed? Y=1, N=0	Who controls the food? M=1, G=0	Success? Y=1, N=0
1	1	0	1	0
2	1	0	0	0
3	0	0	1	0
4	0	0	0	0
5	0	1	0	1
6	1	1	1	1
7	0	1	1	0
8	1	1	0	0

The QCA findings are the following:

- FOOD <= O (or food => o)
- food + know*FOOD*CTRL + KNOW*FOOD*ctrl => o
- know*FOOD*ctrl + KNOW*FOOD*CTRL => O

The necessity analysis reveals that distributing food supplements is necessary (but not sufficient) for success (FOOD <= O). In both successful combinations this condition is present, though it doesn't always lead to success (as in cases #7 and #8). To have any chance of success it is necessary that food supplements are distributed; if they aren't, failure is guaranteed (food => o). There are two more pathways to failure: GYM (011), when food is distributed, knowledge is acquired by the grandmothers and food distribution is controlled by mothers (know*FOOD*CTRL), and MYG (110), when food is distributed, knowledge is acquired by mothers and food distribution is controlled by grandmothers (KNOW*FOOD*ctrl). There are two pathways to success: GYG (010), when food is distributed, and grandmothers both acquire the knowledge and control the food (know*FOOD*ctrl); and MYM (111), where food is distributed, and mothers both acquire the knowledge and control the food (KNOW*FOOD*CTRL).

From the comparison of these 4 combinations, we deduce that, provided that food supplements are distributed, the equality of the first and third condition (when the person who receives the training is the same person controlling the food), is sufficient for success; while difference between the first and the third condition (the person who receives the training is different from the one who controls the food) is sufficient for lack of success.

Put differently, we realise that the relevant relation here is the identity or difference between the person controlling the food and the one acquiring the knowledge: so we can replace the two conditions in the previous table with a more abstract one, denoting identity or difference between the two roles.

Table 59: Table 58 modified, with two conditions merged into one

Case ID	Is the same person controlling the food and acquiring the knowledge? Y=1, N=0 (SAME)	Food Supplements Distributed? Y=1, N=0	Success? Y=1, N=0
1	1	0	0
2	0	0	0
3	0	0	0
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	0
8	0	1	0

The QCA findings from Table 59 are the following:

- FOOD <= O (or food => o)
- food + same*FOOD => o
- SAME*FOOD => O

The necessity relations do not change; what changes is that we have identified a more parsimonious model that is able to merge combinations 5 and 6 into "SAME*FOOD" and combinations 7 and 8 into "same*FOOD", providing a clearer solution.

This kind of generalisation didn't result from following the rules of any of the automatic procedures illustrated in Chapter 2, but from our intuition that two conditions can be connected and merged into one and that this will explain the regularities of the dataset better. Similarly to the young farmers case in Annex B, where measures taken by farmers were grouped into high-risk and low-to-medium-risk strategies, similarities were found between combinations that were not being merged automatically by the available, algorithm-based procedures. The work needed was conceptual and focused on the constructs describing the conditions. The computer does not see "conceptual similarities", like combinations of conditions describing a process or a type of strategy: this requires sectoral or substantive knowledge, like imagining the process of the same person acquiring the knowledge and then controlling the food, being able to implement what she has learned.

This type of generalisation requires good substantive knowledge of the field or sector related to the evaluation; and in particular of the theoretical assumptions the evaluator is trying to test or develop with QCA (what characterises successful strategies of farm management? What are the characteristics of successful maternal nutrition projects?, etc.). The dialogue with theory is an essential part of QCA, as the next section will try to show.

3.2 Combining QCA with explanatory/generative approaches

The beginning of a QCA analysis requires the specification of one or more "models": a list of conditions that, according to existing knowledge, play a role in affecting the outcome. Creating this list is a crucial task, because it is only possible to demonstrate the role of those conditions that are put to the test: conditions that are ignored at this stage will not emerge as relevant from the findings. Setting up a process to conduct this selection has thus important implications for the internal validity of the findings (see Section 3.3 on possible biases emerging at this stage).

In previous sections we have mentioned the importance of having appropriate case based knowledge or a Theory of Change to guide the initial selection of conditions and outcomes for model specification. Relevant information can come from a variety of sources: theoretical explanations in the academic literature, context-specific arguments of regional/functional experts, implicit theories of actors involved in the intervention in various roles, understandings participants have of what they are doing and why, etc. Explanations can also be either structural or agent-based (see Table 60).

One strategy that can be used is to draw on other, theory-based evaluation approaches, like realist evaluation, contribution analysis, or Process Tracing, to aid the initial model specification. In this sense it should be clear that *QCA does not replace Theory-Based evaluation, but it is in fact a form of theory-based evaluation aimed at generalising case-based information to a small or medium set of cases*.

Table 60: Types and sources of theoretical knowledge for model specification

		Major theoretical categories of social explanation	
		Agent-based	Structural
Major sources of social explanation	- Literature - Regional Specialists and functional experts - Participants and Journalists	rational calculations, material interest, cognitive biases and other beliefs, emotional drives, normative concerns, inclinations, preferences	material power, institutional constraints and opportunities, social norms and legitimacy

Note: Adapted from Bennett & Checkel (2014)

The end of a QCA analysis generates the various types of configurations mentioned in the previous chapters; but from a theory-developing or theory-testing point of view it is not satisfactory to stop at that point, without explaining why the conjunction of a given group of conditions, or the combination of a given group of "ingredients", is a good "recipe" for the outcome; or why it seems impossible to achieve the outcome unless one or more conditions are met.

As an approach to causal inference, QCA will produce "automatic" statements of causal necessity and sufficiency once the dataset is ready and fed to the software; but these statements are usually not self-explanatory and need to be interpreted conceptually, for example through the development of theoretical mechanisms. Consider as an analogy the way the statistical link between smoke and lung cancer has been explained in terms of a biological mechanism, describing how chemicals contained in cigarette smoke reduce the healthy functioning of lung tissue cells.

Interpreting QCA solutions might require „going back to the cases“ and organising primary data collection on a selection of case studies. If one configuration covers many cases, it might be hard to select a small group of one or two case studies in an attempt to explain why that configuration is important.

In his activity of QCA evaluation quality assurance, Rick Davies suggested using the "Hamming Distance" to calculate the distance among cases covered by the same parsimonious configuration. This distance could be used to identify the most "central" case, the one most similar to most others; and/or to identify the "extreme", most different cases. This method is being used in the macro evaluation of DFID's Empowerment & Accountability initiatives[55]. The section on Process Tracing (3.2.3) discusses other criteria for selecting cases to study in-depth.

The evaluation approaches used to "interpret" the configurations obtained at the end of the QCA analysis might be the same used at the initial, model specification, stage. Combining QCA with these approaches is recommended and usually strengthens construct validity. The rest of this section outlines examples of how explanatory approaches (underpinned by generative causality, see Annex A) like Contribution Analysis, Realist Evaluation and Process Tracing, can be combined with QCA.

3.2.1 Contribution Analysis

As an evaluation approach, contribution analysis has been introduced over 15 years ago (Mayne, 1999). It is essentially a narrative approach that can be supported by various types of evidence, where the evaluator formulates and then tests a contribution story that explains how the intervention has supposedly achieved (or is supposed to achieve) its impact. The contribution story is usually visualised as a causal chain of intermediate steps or outcomes, with assumptions and risks that make each step more or less likely to materialise. In what

[55] http://www.itad.com/knowledge-and-resources/dfids-macro-evaluations/

follows we demonstrate how Contribution Analysis can be integrated with QCA in two different sectors[56]: empowerment and accountability and food trade.

3.2.1.1 Improving accountability in Sub-Saharan Africa

The first example is the evaluation of an intervention aimed at, amongst other things, improving accountability mechanisms and opening up spaces for civil society participation in policymaking in Sub-Saharan Africa (see Annex C). The contribution story explained how project-supported national partners, working with external partners and a diverse groups of stakeholders, were supposedly more likely to contribute to the institutionalisation/formalisation of accountability mechanisms (ACCcsoPART) if they found one or more government officials who took the accountability cause to heart and "championed" their demands within the political establishment. Project funding allowed the partners to strengthen their own organisational capacity, to increase the numbers of stakeholders they were able to work with, and improve the quality of their own outreach and advocacy. However, the presence of champions within the government was included in the partners' theory of change, and it was thought to be necessary to achieve success (Figure 20).

In terms of QCA analysis, this theory resulted in the selection of four conditions, only two of which survived initial tests of relevance for the Boolean minimisation (see Section 2.7.1.9): presence of champions (CHAMP) and demonstrated ability of the national partner to engage with a diverse group of stakeholders (ENGDIV). Some contextual conditions were then added to account for other factors potentially influencing the outcome, based on the intervention theory of change, like relatively unstable, pluralistic states (polstab) vs. one-party, more stable states (POLSTAB); and transparency of decision-making processes (TRANSP), see Table 61. **The goal was to check whether the theory held uniformly across different contexts or not**, and if not, how it needed to be refined to account for local differences.

[56] The examples have been extracted from more complex evaluations and "simplified" to illustrate the interaction between QCA and Contribution Analysis

The Boolean minimisation showed that the combination of champions (CHAMP) and ability to engage with diverse stakeholders (ENGDIV) did not lead to success everywhere: only half the cases where these two conditions were present (bottom right quadrant in Figure 22) showed a positive outcome. In addition, the INUS analysis identified the following two pathways:

1. CHAMP*ENGDIV*TRANSP*polstab =>
   accountability mechanisms improved (KenyaSOTU)
2. champ*ENGDIV*TRANSP*polstab =>
   accountability mechanisms not improved (KenyaTRUST)

Table 61: Truth Table for the accountability mechanisms model

Country	ChampNSA	EngDIV	Transp	PolSTAB	ACCcsoPART
NigeriaSOTU	0	0	1	0	1
RwandaSOTU	1	1	0	1	0
KenyaSOTU	1	1	1	0	1
KenyaTRUST	0	1	1	0	0
ZimbabweTRUST	1	1	0	0	1
MozambiqueSOTU	1	1	0	0	0
MozambiqueTRUST	0	0	0	0	0
BotswanaTRUST	0	0	0	1	0

The comparison of the first and second pathway showed that "champions" were an INUS cause: they were not necessary in general, but when national partners were able to engage with a diverse range of stakeholders and decision-making was transparent (both cases in Kenya), champions made the difference. If these two conditions were not met, for example where decision making was not transparent (RwandaSOTU), champions seemed to become irrelevant for the outcome.

The contribution story was thus reformulated: an additional assumption was added to the main explanation (see Figure 20), describing how project-supported national partners, working with external partners and a diverse groups of stakeholders, were more likely to contribute to the institutionalisation/formalisation of accountability mechanisms if they found one or more government officials who took the accountability cause at heart and "championed" their demands. More emphasis was added to the fact that champions required pluralistic states with transparent decision-making in order to be effective. If these assumptions were not met the link would not hold and the "causal chain" would break (see Figure 21).

Figure 20: Representation of pre-QCA contribution story

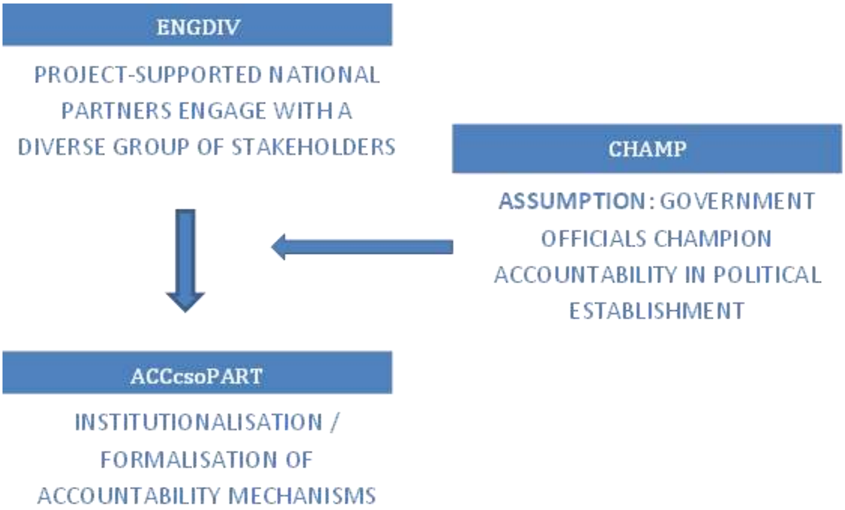

ENGDIV
PROJECT-SUPPORTED NATIONAL PARTNERS ENGAGE WITH A DIVERSE GROUP OF STAKEHOLDERS
CHAMP
ASSUMPTION: GOVERNMENT OFFICIALS CHAMPION ACCOUNTABILITY IN POLITICAL ESTABLISHMENT
ACCcsoPART
INSTITUTIONALISATION / FORMALISATION OF ACCOUNTABILITY MECHANISMS

Figure 21: Representation of post-QCA contribution story

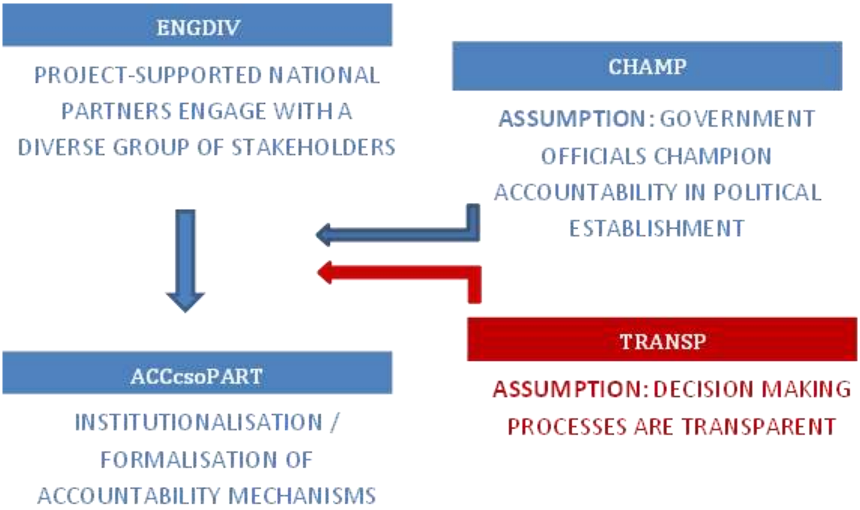
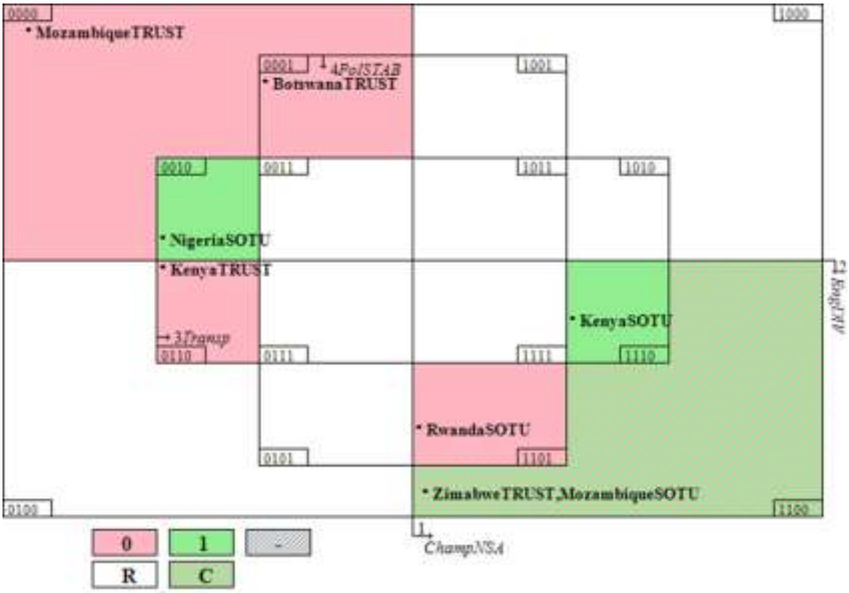

ENGDIV
PROJECT-SUPPORTED NATIONAL PARTNERS ENGAGE WITH A DIVERSE GROUP OF STAKEHOLDERS
CHAMP
ASSUMPTION: GOVERNMENT OFFICIALS CHAMPION ACCOUNTABILITY IN POLITICAL ESTABLISHMENT
TRANSP
ASSUMPTION: DECISION MAKING PROCESSES ARE TRANSPARENT
ACCcsoPART
INSTITUTIONALISATION / FORMALISATION OF ACCOUNTABILITY MECHANISMS

Figure 22: Venn diagram for the model CHAMP + ENGDIV + TRANSP + polstab = ACCcsoPART



3.2.1.2 Improving Food Trade in East Africa

The following example[57] is taken from the evaluation of a group of projects aiming at the improvement of food trade in East Africa. A preliminary step in contribution analysis is the statement of the problem at hand; in this case that farmers do not have enough physical space to store surplus crops. This is a problem because if surplus crops can't be preserved, the farmer will be forced to waste them or, alternatively, to sell them at a low price. The ability to store surplus crops and sell them when the price is high would not only bring additional revenue to the farmer and stabilise the market, but also provide an incentive for the farmer to increase production. In general, availability of storage space to preserve crops can be seen as a means of

[57] This example has illustration purposes only and has been extracted from a more complex evaluation. It has been developed with Liz Turner (Itad) as part of the design of the evaluation of FoodTrade East and Southern Africa and later "stylised" to illustrate the interaction between QCA, Contribution Analysis and Realist Evaluation.

improving the matching between supply and demand in the food market.

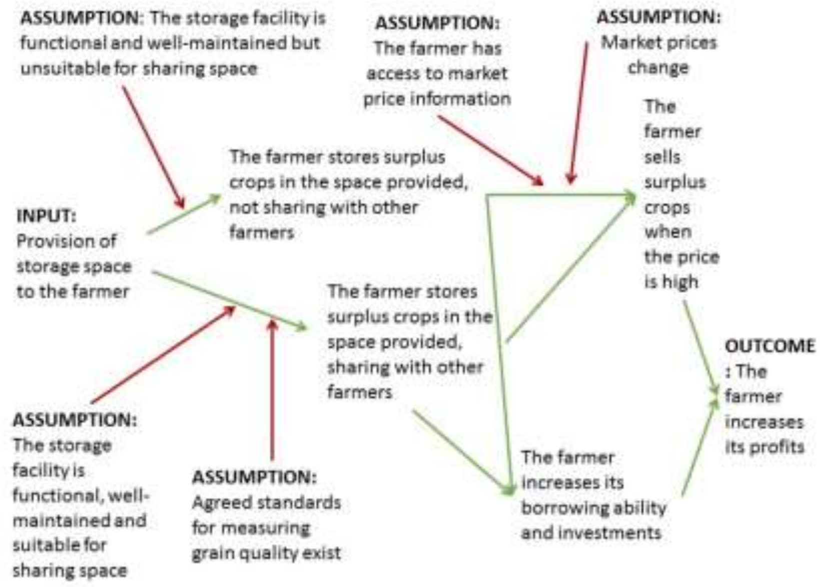
The programme aims at providing physical space for the farmer to store crops. This would allow the farmer to store surplus crops and sell them when the price is high, provided the farmer has information on the market price (and the market price changes). The storage facility would also need to be functional and well-maintained (no damage, no water leaks), and allow for the distinction/separation of crops of different quality, particularly if the farmer has crops of different quality or is sharing with the other farmers.

The farmer could maximise the utility of the storage space s/he is provided with by sharing it with other farmers, reducing storage costs per unit stored and ultimately increasing profits. This would be easier if the access rights to the facility did not foster conflicts between farmers, which might depend on trust but also on the existence of agreed standards for the measurement of grain quality.

In the last part of the contribution story, all the above would increase the farmer's capital and thus their borrowing ability because the banks would be protected in case of insolvency. Higher borrowing ability potentially increases investment and thus ultimately profits (see Figure 23).

Before transforming the above contribution story into QCA-ready conditions, an intermediate step is useful: applying realist evaluation and extracting CMO configurations from the CA causal chain. This allows us to show how QCA can be combined not only with Contribution Analysis, but also with Realist Evaluation.

Figure 23: Representation of the contribution story for the Food Trade evaluation


ASSUMPTION: The storage facility is functional and well-maintained but unsuitable for sharing space
ASSUMPTION: The farmer has access to market price information
ASSUMPTION: Market prices change
The farmer sells surplus crops when the price is high
The farmer stores surplus crops in the space provided, not sharing with other farmers
INPUT: Provision of storage space to the farmer
The farmer stores surplus crops in the space provided, sharing with other farmers
OUTCOME: The farmer increases its profits
ASSUMPTION: The storage facility is functional, well-maintained and suitable for sharing space
ASSUMPTION: Agreed standards for measuring grain quality exist
The farmer increases its borrowing ability and investments

3.2.2 Realist Evaluation

Realist evaluation is an application of scientific realism to evaluation. Scientific realism (Bhaskar, 2009) is an ontology framing reality as a stratified object made of nested layers, sometimes represented as an onion, where action is entirely embedded and as such dependent on the context. As an approach for evaluation research, it was introduced in a seminal book (Pawson & Tilley, 1997) and has been widely applied ever since (Westhorp, 2014).

The basic message of realist evaluation is that evaluation research needs to focus on understanding what works better for whom, under what circumstances; and in particular what it is within a programme that makes it work. In order to do so it needs to unravel the “inner mechanisms” at work in different contexts, because interventions do not work in the same way everywhere and are opportunities that individuals might or might not take.

Technically, Realist Evaluation entails identifying one or more Context-Mechanism-Outcome (CMO) configurations, where contexts are made of resources, opportunities and constraints available to the beneficiaries; mechanisms are choices, reasoning or decisions that beneficiaries take on the basis of the resources available in their context; and outcomes are the product of individual behaviour and choices.

A number of CMO configurations can be extracted from the contribution story of the Food Trade evaluation illustrated above. These explain how contextual factors influence the farmers' behaviour and ultimately determine whether the facility is used or not. Table 62 illustrates some of these mechanisms: note that information on market prices, change in market prices and functionality of the storage space are not included because they are assumed to have the same influence across all configurations.

Table 62: CMO configurations explaining why the storage facility is used or not

CONTEXT	MECHANISM	OUTCOME
There are agreed quality standards that farmers use to measure the quality of their crops	Farmers trust the quality measurement and believe that their crops are of the same quality when the measurement says so	The facility is USED
Farmers can use the facility on condition that they share it with other farmers		
There are no agreed quality standards that farmers use to measure the quality of their crops	Farmers don't trust quality measurements but don't need to in order to use the facility	
Farmers can use the facility all for themselves (no need to share it)		
There are no agreed quality standards that farmers use to measure the quality of their crops	Farmers don't trust quality measurements and are afraid that different quality crops will be mixed up in shared storage (in particular that final quality will be lower and the low price they would be able to sell it will not justify storing their crop)	The facility NOT USED
Farmers can use the facility on condition that they share it with other farmers		

The realist configurations explain how different contextual factors combine in different ways to elicit three different reactions in the farmers; only two of which ultimately lead to use. These allow us to have clear explanations of why the facility is used or not; but we still don't know how consistently those mechanisms lead to those outcomes; nor how consistently those mechanisms are triggered under those contextual conditions.

Jon Elster makes a distinction between Type A and Type B mechanisms (Elster, 1998). Type B mechanisms can be triggered simultaneously, but might work in opposite directions, with the outcome being uncertain and depending on which force prevails in a given circumstance. By contrast, Type A mechanisms are mutually incompatible: only one can be triggered and there is no uncertainty on the outcome. For example, the storage facility may be functional and present separate spaces which makes the farmer open to the idea of space sharing, but the absence of agreed measurements for grain quality might work in the opposite direction, decreasing trust and neutralising the effect of the facility: these are type B mechanisms. On the other hand, it can be assumed that if the farmer does not have information on market prices, they will have no incentive storing surplus crop and the facility won't be used: this is a type A mechanism.

Table 63: Dataset obtained out of the CMO configurations

Case	STAND (presence or absence of agreed quality standards)	SHARECOND (necessity to share the facility with other farmers in order to use it)	FUNCTIONAL (storage facility is functional and well-maintained)	INFO (information on market prices is available to farmers)	CHANGE (market prices change)	OUTCOME (USE of facility)
A	1	1	1	1	1	1
A1	1	1	1	0	0	0
A2	1	1	0	1	0	0
A3	1	1	0	0	1	0
A4	1	1	1	1	0	0
A5	1	1	1	0	1	0
A6	1	1	0	1	1	0
A7	1	1	0	0	0	0
B	0	0	1	1	1	1
B1	0	0	1	0	0	0
B2	0	0	0	1	0	0
B3	0	0	0	0	1	0
B4	0	0	1	1	0	0
B5	0	0	1	0	1	0
B6	0	0	0	1	1	0
B7	0	0	0	0	0	0
C	0	1	1	1	1	0
C1	0	1	1	0	0	0
C2	0	1	0	1	0	0

C3	0	1	0	0	1	0
C4	0	1	1	1	0	0
C5	0	1	1	0	1	0
C6	0	1	0	1	1	0
C7	0	1	0	0	0	0
D	1	0	1	1	1	1
D1	1	0	1	0	0	0
D2	1	0	0	1	0	0
D3	1	0	0	0	1	0
D4	1	0	1	1	0	0
D5	1	0	1	0	1	0
D6	1	0	0	1	1	0
D7	1	0	0	0	0	0

For type B mechanisms, the uncertainty lies on which mechanism will prevail in influencing the outcome, out of a range of mechanisms which have all been triggered; while for type A mechanisms the uncertainty lies on which single mechanism is triggered, out of a range of possibilities.

In this context, a QCA analysis is useful to understand the nature and type of these mechanisms more in-depth, checking if they are triggered alone or together with others, and which outcomes they seem to produce. It can also be used to see if the contextual conditions in Table 63 are consistently associated with the hypothesised outcome over a medium or large number of cases, and if the conditions omitted because assumed to influence the outcome equally across the configurations reported in the table actually do so.

3.2.2.1 Transforming context characteristics into binary conditions

The model used to test the CMOs thus incorporates the contextual information formally included in the CMOs (SHARECOND and STAND) and three other conditions from the contribution analysis (FUNCTIONAL, INFO and CHANGE) which are assumed to influence use of the storage space independently of which mechanism will be triggered from the ones reported in Table 62.

Carrying out QCA necessity and sufficiency analyses on the model STAND + SHARECOND + FUNCTIONAL + INFO + CHANGE shows the following:

- the combination FUNCTIONAL*INFO*CHANGE is necessary for success, and the three single conditions "change", "info" and "functional" are each sufficient for failure, compatibly with DeMorgan's Law. This means that the facility is never used unless it is functional, and unless the farmer has information on changing market prices.
- No matter what happens with the three conditions above, "stand*SHARECOND" is sufficient for failure: which means that whenever the opportunity to use the facility is conditional on sharing and there are no agreed quality standards, the facility is never used.

In short, while the disjunction "change + info + functional + stand*SHARECOND" is sufficient for failure, the combination "FUNCTIONAL*INFO*CHANGE" is necessary but not sufficient for success, because in the case of "stand*SHARECOND" the farmer still doesn't use the facility, even if all the other three conditions are positive (case c). In addition to these three conditions, either STAND or sharecond are needed to cover the successful cases (are necessary for success), which means that when the facility is used, there are either agreed quality standards or the sharing conditionality does not apply.

In symbols, the successful cases can be synthesised with the Boolean minimisation and appear as the following two pathways:

1. FUNCTIONAL * INFO * CHANGE * STAND => SUCCESS
2. FUNCTIONAL * INFO * CHANGE * sharecond => SUCCESS

while the negative cases appear as:

- change + info + functional + stand * SHARECOND => success.

3.2.2.2 Generalising CMO configurations to 32 cases

QCA confirms the initial CMO configurations, "translating" them into the necessity of the disjunction STAND + sharecond: if there are agreed quality standards, conditionality doesn't matter; and if there is no conditionality and farmers are not forced to share the space, it doesn't matter whether there are agreed quality standards or not. In other words, the CMO configurations associated with success “hide” two SUIN causes.

It is also confirmed that the three conditions FUNCTIONAL, INFO and CHANGE influence the CMOs in Table 62 equally, the absence of each preventing other contextual characteristics (SHARECOND and STAND) from making a difference to the outcome. If the storage space is not functional, farmers won’t want to use it anyway, no matter the values of SHARECOND or STAND; the same goes for changes in market price (if the price doesn’t change the best strategy is to sell everything as soon as possible), and availability of information on market prices, which both neutralise the need for storage space. Table 64 reports the refined, post-QCA CMO configurations, which we now know synthesise 32 cases.

Table 64: Refined, post-QCA CMO configurations, synthesising 32 cases

CONTEXT		MECHANISM	OUTCOME
Market prices change; information about market prices is available to farmers; the facility is functional and allows the preservation of quality crops.	There are agreed quality standards that farmers use to measure the quality of their crops; there might or might not be conditionality in terms of sharing the facility	Farmers know when it’s most convenient to sell crops; they value the opportunity to use a functional storage space; including sharing it if necessary, as there are quality measurements they trust.	The facility is USED
	Farmers can use the facility all for themselves (no need to share it); there might or might not be agreed quality standards that farmers use to measure the quality of their crops.	Farmers know when it’s most convenient to sell crops; they value the opportunity to use a functional storage space; they might not trust quality measurements but don't need to in order to use the facility	
Market prices can change or not; information about market prices might be available or not; the facility might be functional or not.	There are no agreed quality standards that farmers use to measure the quality of their crops	Farmers might not: know when it’s most convenient to sell crops; value the opportunity to use a dysfunctional storage space; trust quality measurements and be afraid that different quality crops will be mixed up in shared storage (in particular that final quality will be lower and the low price they would be able to sell it will not justify storing their crop)	The facility NOT USED
	Farmers can use the facility on condition that they share it with other farmers		

3.2.2.3 Understanding the type of mechanism

As for the types of mechanisms, the findings support the idea of quality standards triggering trust being a Type B mechanism which can lead to success or not depending on how it interacts with other mechanisms. As for functionality of the storage space, its presence also triggers a type B mechanism which leads to success or not depending on other conditions, but its absence triggers a type A mechanism which will reliably prevent use of the space. Type A mechanisms are also triggered by absence of market price information and absence of change in market prices.

More generally, QCA is well positioned to improve our understanding of both Type A and Type B mechanisms. As for Type B, the INUS analysis in QCA allows us to learn under which conditions a given mechanism of the group (the INUS cause) will prevail over others triggered at the same time, and make a difference for the outcome. As for Type A, the idea of sufficient but unnecessary configurations represents confidence about the outcome a given mechanism is producing, and at the same time uncertainty about which one (which configuration) will be triggered in a given case.

3.2.3 Process Tracing

Process Tracing is increasingly gaining attention as a method and an approach to data collection in impact evaluation (Befani & Stedman-Bryce, 2016). Like Contribution Analysis, it is essentially a within-case method analysing a sequence of events or intermediate outcomes that are assumed to follow a triggering event (Befani & Mayne, 2014). This chain of events is usually thought of as a hypothesis explaining an outcome (Beach & Pedersen, 2011; Beach & Pedersen, 2013; Bennett & Checkel, 2014). Unlike Contribution Analysis, Process Tracing focuses on data collection and is sometimes called a causal inference "technique" (Bennett & Checkel, 2014) aiming at minimising inferential error.

Process Tracing operates a clear distinction between a) ontological reality (e.g. an event that is assumed to have happened); b) the hypothesis made by the researcher that that event has actually happened; and c) the observable evidence increasing or decreasing the researcher's confidence in its hypothesis. Variants of Process Tracing are classified differently by different authors but they all make a distinction between Inductive and Deductive Process Tracing. Inductive Process Tracing starts from an outcome and proceeds backward, trying to reconstruct a chain of events that goes back in time until the triggering event (for example the beginning of the intervention); much like “a detective piecing together the last few hours and days in the life of a victim” (Bennett & Checkel, 2014). Deductive Process Tracing starts when the theory is well enough developed to be tested empirically.

Deductive Process Tracing comes with a small arsenal of concepts and tools that are able to rigorously measure the strength of given pieces of evidence to prove or disprove a specific theory/mechanism. These tools allow the evaluator to operate a clear distinction between “absence of evidence” and “evidence of absence” (Bennett & Checkel, 2014). After the theory is ready to be tested, a key step in deductive Process Tracing is the development of case-specific, observable implications of the theory: in terms of what specific pieces of evidence the researcher expects to observe if the theory holds, and of what specific pieces of evidence are not expected under the theory, but would confirm it if observed (these are also known as, respectively, the Hoop Test and the Smoking Gun, see Evera, 1997; Bennett, 2010; Collier, 2011).

Many of these expectations concern the behaviour of specific key informants or actors, either in the data collection or desk review phase. The desk review is not simply a literature review but involves collecting evidence of media behaviour, meeting minutes, and other specific information which might or might not be public at the time of the investigation (this is because its probative value[58] for some interesting theories would likely be very high).

3.2.3.1 Using QCA to generalise Process Tracing mechanisms

Process Tracing can be combined with QCA in essentially two ways (see also (Schneider & Rohlfing, 2013). Before a QCA analysis, inductive Process Tracing can be one way of developing a theory to test with QCA, together with Contribution Analysis, Realist

[58] Probative value is a legal term related to the notion of relevance of a given item of evidence to prove or disprove one of the legal elements of the case. Probative is a term used to signify "tending to prove” https://en.wikipedia.org/wiki/Relevance_(law)

Evaluation, or other sources like major social science theories, interviews with regional specialists and other stakeholders, journalists, etc. If this theory has already been tested with deductive Process Tracing and has been shown to hold in a specific case, QCA provides a chance to test and generalise that theory over multiple cases.

Let's take the example of Oxfam Great Britain's evaluation of a Health Advocacy Campaign in Ghana (Stedman-Bryce, 2013). The evaluation demonstrated the existence of a mechanism explaining the impact of a campaign report on the decision of the Government of Ghana to change the formula it used to calculate the percentage of the population who are registered with the national health insurance scheme. The mechanism included five key ingredients that were assumed to have led to the reform, only some of which related to the campaign's report: ELECT + MATHS + PREVCOM + REACT + HIST = SUGGCH (Table 65).

1. Upcoming elections (ELECT)
2. The mathematical nature of the suggested change/the objectivity of its value (MATHS)
3. The appreciation of the change on behalf of the Government, paradoxically signalled at first by a violent reaction to the campaign's report with no substantial criticism to it, and later by public admission (REACT)
4. (The above while the Government had no previous history of violent reactions to similar reports) (HIST)
5. A possible previous commitment to the change/the Government was already working on it, which was not however strongly supported by the evidence in the specific case (PREVCOM)
6. The outcome being extreme similarity between the measure taken by the Government and the measure suggested in the campaign's report (SUGGCH).

We can hypothetically use this mechanism as (part of) a policy influence theory to test over multiple cases with QCA. Table 65 presents a fictional Truth Table showing what the data could potentially look like across multiple cases where similar campaigns have been conducted. No single condition is necessary for success, although the disjunction "PREVCOM + ELECT" is: it means that apparently successful influence cases either had the government planning those changes before the campaign, or happened under upcoming elections (or both). PREVCOM (already existing government plans) is also subset sufficient: whenever the government is planning the change, it will eventually happen; but not necessary, as one case where the government was not planning the change is still successful.

The Boolean minimisation conducted in Table 65 shows us three pathways to success:

1. PREVCOM * react * hist
2. elect * maths * PREVCOM * react
3. ELECT * MATHS * prevcom * REACT * hist

Table 65: Hypothetical Truth Table of the Ghana Health Care Campaign mechanism

Case	ELECT	MATHS	PREVCOM	REACT	HIST	SUGGCH
A	1	1	0	1	0	1
B	1	1	1	0	0	1
b1	1	0	1	0	0	1
b2	0	1	1	0	0	1
b3	0	0	1	0	0	1
C	0	0	0	1	1	0
c1	0	1	0	1	1	0
c2	1	0	0	1	1	0
c3	1	1	0	1	1	0
d	1	0	0	1	1	0
d1	0	0	0	1	1	0
d2	1	0	0	1	0	0
d3	0	0	0	1	0	0
d4	1	0	0	0	1	0
d5	0	0	0	0	1	0
d6	1	0	0	0	0	0
d7	0	0	0	0	0	0
e	0	0	1	0	1	1

Combinations 1 and 2 show that, in successful cases, when the Government already had ongoing plans to make the specific change at the time of the campaign, no violent reaction was observed, whether the Government had a history of similar reactions or not. The second pathway shows that the change was implemented even if there were no upcoming elections and the nature of the change could not be 'objectively' or 'mathematically' assessed. Combination 3 shows that when the government was not planning to make that specific change, exceptional, violent reactions take place and the government in the end recognises the ‘objectively’ superior value of the suggested change, thinking it is particularly convenient to do so under elections.

In brief, QCA shows that the process tracing mechanism argued to demonstrate the impact of the campaign is only one of the pathways at work across the cases that can explain the outcome (implementation on behalf of the government of a policy change which is identical to the change suggested during an advocacy campaign). In particular, the first two combinations explain how this change can happen when the government was considering it prior to the campaign.

QCA also identifies pathways to “failure”: that is leading to the government not implementing a change that was suggested during the campaign; all unsuccessful cases are observed when the government had no previous plans to implement such change:

1. maths*prevcom
2. prevcom*REACT*HIST

Combination 1 shows that under such conditions, if the value of the change cannot be mathematically /objective assessed, this is ‘bad enough’ for the change not to be implemented; while combination 2 shows that, under the same conditions, the government reacting violently to the campaign while having a history of similar reactions is also not a good sign that influence has taken place.

3.2.3.2 Using Process Tracing to explain QCA combinations in-depth

We have previously mentioned that QCA findings can be obscure unless an expert of a substantive field, for example the same expert who contributed to model specification, makes sense of them. Some guiding questions here are “how can a condition be necessary for an outcome? What is it within the outcome that requires it?” and “why is that combination sufficient for the outcome? What reactions do the “ingredients” produce when they are mixed together, and how does that reaction relate to the outcome”?

Sometimes QCA findings, while providing part of the answers, raise even more questions. Even when an expert of the field develops a mechanism that is supported by QCA findings, that might not be convincing enough in “generative” terms (i.e. in understanding in-depth how the outcome is produced). Let's go back to the previously mentioned example (Section 2.1.1.4) of the review of policies combining social protection and climate change elements to protect the livelihoods of vulnerable households and preserve their consumption capacity in times of crisis. Let's assume that we analyse the following dimensions with QCA:

- Funding sources (FUND)
  - Mostly public vs. mostly private
- Financial instrument (INSTR)
  - Mainly insurance or conditional cash transfer vs. mainly unconditional cash transfers
- Implementing organisation (IMPL)
  - Mainly the government vs. mainly NGOs or the private sector

Let's also assume that QCA identifies three main typologies of policies:

- FUND*INSTR*IMPL
- fund*INSTR*impl
- fund*instr*impl

A sectoral expert could link the first pathway with traditional social protection schemes run by the government, like pensions or other forms of social insurance; the second with privately-funded commercial insurance, implemented by companies in the private sector; and the third with livelihoods programmes based on unconditional transfers, implemented by NGOs and funded by private donors.

Suppose that these three pathways are associated with success: how can the latter be explained? One hypothesis could be that traditional social protection schemes with high coverage integrating protection for climate change risks (associated with the first typology/combination) are effective because they scale up rapidly in response to drought. Another hypothesis could be that commercial insurance (second typology/combination) is effective, because it links payments ensuring livelihoods directly to climate change risks like droughts, and so on.

For each hypothesis, the principles, tests and best practices (Bennett & Checkel, 2014) of Process Tracing could be implemented and the hypotheses tested in single cases.

3.2.3.3 Selecting cases for in-depth study

At the beginning of Section 3.2 we have highlighted the need of conducting a selection of in-depth case studies to interpret the configurations obtained with QCA, and suggested the Hamming distance as a method of selecting cases covered by the same configuration. Process Tracing/Bayesian principles recommend that case selection is informed by the relevance of cases for the demonstration of a certain theory, and that cases with the highest probative value for that theory are selected.

For example, if the aim is to confirm a theory, a case where the theory is unlikely to hold should be selected; and if the aim is to disconfirm it, a case likely to confirm it should be picked. This is because if the theory is confirmed in a case we were expecting it to be confirmed in, such case shall have a lower probative value than a case where the theory is confirmed but we were not expecting the case to confirm it (Table 66). Similarly, a case disconfirming the theory that we were expecting to be confirmatory, shall have a higher probative value (against the theory) than a case disconfirming the theory that was expected to do so.

Table 66: Probative value of case X for theory T depending on prior expectations

	Theory T confirmed in case X	Theory T disconfirmed in case X
Case X expected to confirm Theory T	LOW	HIGH
Case expected to disconfirm theory T	HIGH	LOW

More guidance on Process Tracing can be found in Bennett & Checkel (2014) and Beach & Pedersen (2013); but hopefully this section has offered broad-brush, proof-of-concept argument that their integration is feasible.

3.3 Bias Control and Quality Assurance

As we get to the end of QCA journey it's useful to discuss the issue of bias more broadly, before focusing on quality assurance.

Scientific research and in particular social science research are exposed to several types of biases; impact evaluation is not less so. Camfield et al. (Camfield, Duvendack, & Palmer-Jones, 2014) propose four broad categories of bias that qualitative researchers should be aware of: empirical, researcher, methodological and contextual biases. At least three of these (empirical[59], researcher and contextual) can potentially influence the selection of cases and/or conditions to include in a QCA analysis. We now discuss the potential manifestations of these biases and suggest solutions for anti-bias protection.

3.3.1 Biases affecting the selection of conditions

Researcher biases are the evaluator's attachment to a particular theory, explanation or discipline, a.k.a. confirmation bias, which usually stems from political or disciplinary background; but also the conservative bias, which is the slow revision of beliefs in light of new evidence (in Bayesian terms, a disproportionate weight given to pre-observation vs. post-observation confidence, see Kahneman, 2012). Contextual biases, on the other hand, are related to social or political pressures to demonstrate a specific result, coming from sponsors, friends, of friendly relations developed between the evaluator and project staff, a.k.a. in-group bias (when the evaluator and project staff start seeing themselves as part of the same group). Both researcher and contextual biases can affect the selection of conditions, for example including the presence of the intervention in all models despite repeated analysis that show its irrelevance.

[59] Empirical biases include biases related to the distorted observation of empirical phenomena: for example the tendency to see patterns where there aren't (e.g. the gambler's fallacy); over-interpreting observed effects; availability biases (related to memorable or vivid occurrences, for example recent occurrences); and attribution biases (causally attributing effects to one-off events or specific actors while the real causes might have been a combination of several factors operating slowly over time; and self-serving or self-importance biases.

3.3.1.1 Solutions to biases affecting the selection of conditions

In addition to the above, there might be an availability, attribution or a self-serving bias in selecting conditions to include in the model, which can be remedied by investing more on theory-building outside of QCA. Other solutions to condition selection bias include following the suggestions indicated in Chapter 2 (2.7.1.8 and 2.7.1.9) when needing to reduce the number of conditions analysed.

QCA can help overcome the tendency to see patterns that do not exist because of its systematic way of comparing cases and spotting similarities and differences. Not seldom QCA returns counterintuitive or unexpected results, for example the absence of a condition where the evaluator was expecting its presence, or viceversa: the boolean nature of QCA makes negation of the initial hypotheses as clear as it can possibly be (a simple NO rather than YES). In this sense QCA makes it easy for the evaluator to have its initial hypotheses refuted or contradicted.

In other words, QCA comes with an inbuilt ability to maximise specificity or minimise false positives: when a relation is confirmed, particularly if care has been taken to calibrate the conditions and maximise internal validity, and the number of cases is adequate to the standards set in the previous section (minimum 6 cases for the supersubset analysis, the binomial distribution for imperfect consistency, and a minimum number of cases for each number of conditions), the evaluator can be cautiously confident that the relation holds, for the dataset and possibly beyond.

If the relation passes the sensitivity tests on the modified dataset (adding/removing conditions, inclusion of logical cases representing specific hypotheses, changing consistency cutoff points, changing calibration thresholds – e.g. adopting different rubrics), the evaluator can be quite confident about its robustness and focus on its interpretation.

3.3.2 Biases affecting case selection and data collection

Contextual biases like social and political pressures can affect the choice of cases to be analysed, which in development evaluation is usually an enterprise highly dependent on the quality of already established relations with local partners, and thus also subject to the friendship bias. Camfield et al. (Camfield, Duvendack, & Palmer-Jones, 2014) identify a series of impact evaluation biases that affect the quality of data collection processes and ultimately of data. Some are related to the relationship between the informant and the data collector, like similar person bias, charismatic bias (the halo effect), exposure bias, courtesy bias, diplomatic bias; others are more related to data processing, like interpretation, transcription, and translation; yet others to the specific moment when data collection takes place (embodied knowledge, note-taking).

3.3.2.1 Proposed solutions to biases affecting case selection and data collection

Technically, QCA synthesis procedures do not protect against data collection biases, and they take the dataset as given. However, if there is reason to believe that primary data is heavily biased or is unreliable, a possible solution is to recalibrate the condition using more detailed and specific definitions, in order to minimise the risks of misinterpretation and make it easier to assess data quality; or to include in the rubrics defining the conditions only data which are known to be high quality, even if that might affect construct validity.

As for minimising case selection bias, it is good practice to try to have as much diversity in the sample as possible, so for example avoid sets of cases presenting largely the same values in one condition or in the outcome. The sensitivity tests specifically protecting against case selection bias are a) addition or removal of cases and b) changing the frequency cutoffs for including combinations in the Truth Table.

3.3.3 Quality-Assuring QCA evaluations: A checklist

This report has argued that QCA analyses hold many opportunities to increase the quality of evaluations, while hiding – sometimes in plain sight – pitfalls and challenges. An evaluation taking full advantage of the opportunities offered by QCA and implementing all the procedures suggested in this guide will produce triangulated and much richer findings than an evaluation taking only some of these opportunities. At the same time, the principles of rigour, robustness and reliability require that the evaluator demonstrates awareness of a territory densely mined with pitfalls and traps, and outlines the solutions devised to protect the evaluation quality and integrity from those pitfalls. The quality assurance checklist we present below builds directly on the opportunities and pitfalls identified in Chapter 2.

QCA analyses are usually components of multi-method evaluations: this section does not provide criteria to assess the quality of evaluations including QCA in general, but only the quality of QCA components embedded in those evaluations.

General criteria for good practice in QCA can be found in Schneider & Wagemann (2010). Our criteria build on and overlap with those, but also add specific concerns (for example in relation to generalisation) and are tailored to the evaluation community, where most members are QCA newbies and presumably need more guidance.

Our checklist of criteria to assess the quality of QCA components in evaluations is listed below. We realise that in specific cases it might not be possible for the evaluator to equally pay attention to all the points below. If so, we suggest that the evaluator provides convincing reasons for failing to do so.

In order to be considered a good practice, the QCA evaluation component must include a clear specification of:

1. The rationale for adding QCA to the design, including expectations of what QCA will contribute to the analysis
2. The evaluation questions that QCA will presumably contribute to answer
3. The various sources of tacit knowledge, established social science theory, understandings of programme functioning and stakeholder behaviour on behalf of officials or beneficiaries, data availability constraints, and all other factors that have contributed to the initial selection of conditions and outcomes, with particular attention to the sources of condition selection bias indicated in Section 3.3.1
4. The models (conditions + outcome) that have been tested
5. The rationale used to reduce the number of conditions and test simpler models in successive iterations, if such reduction has taken place (Sections 2.7.1.8 and 2.7.1.9)

6. The opportunities and constraints that have influenced case selection, with particular attention to the sources of case selection bias illustrated in Section 3.3.2
7. The strategies used to handle missing data
8. The empirical and theoretical resources used for calibration
9. The rubrics defining presence and absence (and intermediate degrees of presence if using fuzzysets) of conditions
10. The rationales used for setting calibration thresholds, or for changing them if they have been changed during the course of the analysis
11. An assessment of the calibration robustness, including risks that the same conditions in the same cases are calibrated differently by different researchers (and the solutions implemented to protect against such risks)
12. Whether calibration has considered theory-consistency and/or coverage and given priority to either
13. A representation of all the models discussed (up to 5 conditions) with Venn diagrams, and ideally with the [venn] function in R for 6 and 7-condition models
14. The necessity analysis, paying particular attention to/attempting to explain:
    a. Triviality of conditions
    b. Parameters of fit
    c. Disjunctions of conditions
    d. Robustness of necessity statements (Section 3.1.1.1)
15. The subset analysis, paying particular attention to/attempting to explain:
    a. Parameters of fit
    b. Conjunctions (combinations) of conditions
    c. DeMorgan's Law
    d. Robustness of subset-sufficiency statements (Section 3.1.1.1)
16. Building Truth Tables for the models under test, explaining or justifying:

a. Frequency thresholds and consistency thresholds

b. The rationale used to reduce conditions to a manageable number

c. Whether the combination of numbers of cases and conditions is significant according to the Marx & Dusa standards (Section 3.1.1.2)

17. The Boolean minimisation on the above Truth Tables, for both positive and negative outcomes, explaining or justifying:

a. The selection of logical cases for inclusion, if any (why they were included, on what basis inclusion is justified)

b. Comparison with the Truth Table (how much was it simplified? How many fewer combinations? How simpler are the combinations?) and with the findings of the subset analysis (are any combinations subset-sufficient but not minimisation-sufficient?)

c. The balance of consistency and coverage of the different solutions: is there a tradeoff?

18. The INUS analysis on the above Truth Tables, explaining or making sense of:

a. The role of the intervention, if cases with no intervention are available

b. The role of contextual factors

c. The role of mechanisms or different intervention types

19. Provide narrative and mechanism-based explanations of QCA findings, indicating which explanatory/interpretative methods are used.

a. If additional case study work is needed to interpret specific QCA configurations, explain how cases covered by the same configuration were selected.

Bibliography

Amenta, E., & Poulsen, J. (1994). Where to Begin: A Survey of Five Approaches to Selecting Independent Variables for Qualitative Comparative Analysis. *Sociological Methods & Research, 23*(1), 22-53.

Baptist, C., & Befani, B. (2015). Qualitative Comparative Analysis: A Rigorous Qualitative Method for Assessing Impact. *Coffey "How To" Note*.

Baptist, C., Edouard, E., & Batran, M. (2015). *Mid-Term Evaluation Report: Independent Evaluation of the Africa Regional* Empowerment *and Accountability Programme (AREAP).* London: Coffey International Development.

Basurto, X. (2013). Linking multi-level governance to local common-pool resource theory using fuzzy-set qualitative comparative analysis: Insights from twenty years of biodiversity conservation in Costa Rica. *Global Environmental Change, 23*(3), 573-587.

Baumgartner, M. (2012). Detecting Causal Chains in Small-n Data. *Field Methods, 25*(1), 3-24.

Baumgartner, M., & Epple, R. (2013). A Coincidence Analysis of a Causal Chain: The Swiss Minaret Vote. *Sociological Methods & Research, 43*(2), 280-312.

Baumgartner, M., & Thiem, A. (2015). Identifying Complex Causal Dependencies in Configurational Data with Coincidence Analysis. *The R Journal, 7*(1), 176-184.

Beach, D., & Pedersen, R. (2011). What is Process-Tracing Actually Tracing? The Three Variants of Process Tracing Methods and Their Uses and Limitations. *APSA 2011 Annual Meeting* Paper.

Beach, D., & Pedersen, R. (2013). *Process-Tracing Methods:* Foundations *and Guidelines.* University of Michigan Press.

Befani, B. (2012). Models of Causality and Causal Inference - a review prepared as part of DFID Working Paper 38. UK Department for International Development.

Befani, B. (2013). Between complexity and generalization: Addressing evaluation challenges with QCA. *Evaluation, 19*(3), 269-283.

Befani, B. (2013). Multiple Pathways to Policy Impact: Testing an Uptake Theory with QCA. *CDI Practice Paper 5*. Institute of Development Studies.

Befani, B. (2016). Appropriate Methods for Impact Evaluation: criteria and tools for selection. BOND UK (forthcoming).

Befani, B., & Mayne, J. (2014). Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation. (B. Befani, C. Barnett, & E. Stern, eds.) *IDS Bulletin, 45*(6), 17-36.

Befani, B., & Stedman-Bryce, G. (2016). Process Tracing and Bayesian Updating for Impact Evaluation. *Evaluation (forthcoming)*.

Befani, B., Ledermann, S., & Sager, F. (2007). Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application. *Evaluation, 13*(2), 171-192.

Befani, B., Ramalingam, B., & Stern, E. (2015). Introduction – Towards Systemic Approaches to Evaluation and Impact. (B. Befani, B. Ramalingam, & E. Stern, eds.) *IDS Bulletin, 46*(1), 1-6.

Bennett, A. (2010). Process Tracing and Causal Inference. i H. Brady, & D. Collier, *Rethinking Social Inquiry.* Rowman and Littlefield.

Bennett, A., & Checkel, J. (2014). Introduction: Process tracing: from philosophical roots to best practices. in A. Bennett, & J. Checkel, *Process Tracing: From Metaphor to Analytic Tool.* Cambridge University Press.

Bennett, A., Checkel, J., & (eds). (2014). *Process Tracing: From* Metaphor *to Analytic Tool.* Cambridge University Press.

Berg-Schlosser, D., De Meur, G., Rihoux, B., & Ragin, C. (2009). Qualitative Comparative Analysis (QCA) As An Approach. i B. Rihoux & C. Ragin, (eds), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques.* Sage.

Bhaskar, R. (2009). *Scientific Realism and Human Emancipation.* Routledge.

Blackman, T. (2013). Exploring Explanations for Local Reductions in Teenage Pregnancy Rates in England: An Approach Using Qualitative Comparative Analysis. *Social Policy and Society, 12*(1), 61-72.

Blatter, J., & Blume, T. (2008). In Search of Co-variance, Causal Mechanisms or Congruence? Towards a Plural Understanding of Case Studies. *Swiss Political Science Review, 14*(2), 315-356.

Brady, H. (2002). Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory. *Annual Meeting of the* Political *Methodology Group, University of Washington*. Seattle, Washington.

Burns, D. (2014). Assessing Impact in Dynamic and Complex Environments: Systemic Action Research and Participatory Systemic Inquiry. *CDI Practice Paper 8*. Institute of Development Studies.

Camfield, L., Duvendack, M., & Palmer-Jones, R. (2014). Things you Wanted to Know about Bias in Evaluations but Never Dared to Think. (B. Befani, C. Barnett, & E. Stern, Red.) *IDS Bulletin, 45*(6), 49-64.

Campbell, D. (1969). Reforms as experiments. *American Psychologist, 24*, 409-429.

Campbell, D., & Stanley, J. (1963). Experimental and Quasi-Experimental Designs for Research and Teaching. in N. Gage (Ed.), Handbook *of Research on Teaching* (ss. 171-246). Chicago: Rand-McNally.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research on teaching. in N. Gage (Ed.), Handbook *of Research on Teaching* (ss. 171-246). Chicago: Rand McNally.

Caren, N., & Panofsky, A. (2005). TQCA: A Technique for Adding Temporality to Qualitative Comparative Analysis. *Sociological Methods Research, 34*(2), 147-172.

Cartwright, N. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better.* Oxford University Press.

Checkland, P., & Scholes, J. (1999). *Soft Systems Methodology in* Action. Wiley.

Collier, D. (2011). Understanding Process Tracing. *Political Science and Politics, 44*(4), 823-830.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin Company.

Copestake, J. (2014). Full guidelines for the Qualitative Impact Protocol (QUIP). *Working Paper*. Bath: Centre for Development Studies, University of Bath.

Cronqvist, L. (2011). Tosmana: Tool for Small-N Analysis [Computer Programme], Version 1.3.2.0. Trier: University of Trier.

Cronqvist, L., & Berg-Schlosser, D. (2009). Multi-Value QCA (mvQCA). in B. Rihoux, & C. Ragin, *Configurational Comparative Methods: Qualitative Comparative Analysis (*QCA*) and Related Techniques* (ss. 69-87). Thousand Oaks, CA: Sage.

Davies, R., & Dart, J. (2005). *The 'Most Significant Change' (MSC)* Technique*: A Guide to Its Use.*

De Meur, G., Rihoux, B., & Yamasaki, S. (2009). Addressing the Critiques of QCA. in B. Rihoux & C. Ragin (eds), Configurational *Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques.* Sage.

Deaton, A. (2009). Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. *Working Paper 14690*. Cambridge, USA: National Bureau Of Economic Research.

Derwisch, S., & Löwe, P. (2015). Systems Dynamics Modelling in Industrial Development Evaluation. *IDS Bulletin, 46*(1).

Duflo, E., & Kremer, M. (2003). Use of Randomization in the Evaluation of Development Effectiveness. *World Bank Operations Evaluation Department (OED) Conference on* Evaluation *and Development Effectiveness, 15-16 July 2003.* Washington, D.C.

Duflo, E., Glennerster, R., & Kremer, M. (2006). *Using Randomization in Development Economics Research: A Toolkit.* Washington, D.C.: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Dusa, A. (2007). User Manual for the QCA(GUI) package in R. *Journal of Business Research, 60*(5), 576-586.

Dusa, A., & Thiem, A. (2014). Qualitative Comparative Analysis. R Package Version 1.1-4. URL: http://cran.r-project.org/package=QCA.

Earl, S., Carden, F., & Smutylo, T. (2001). *Outcome Mapping: Building Learning and Reflection into Development Programs.* International Development Research Centre (IDRC).

Elster, J. (1998). A plea for mechanisms. in P. Hedstrøm & R. Swedberg (eds), *Social Mechanisms: An Analytical Approach to Social Theory* (ss. 45-73). Cambridge: Cambridge University Press.

Evera, S. V. (1997). *Guide to Methods for Students of Political Science.* Cornell University Press.

Forss, K. (2007). *Utvärdering som hantverk: bortom mallar och manualer.* Studentlitteratur.

Garcia, J., & Zazueta, A. (2015). Going Beyond Mixed Methods to Mixed Approaches: A Systems Perspective for Asking the Right Questions. *IDS Bulletin, 46*(1), 30-43.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., & Vermeerch, C. (2011). *Impact Evaluation in Practice.* Washington, D.C.: The World Bank.

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

Holvoet, N., & Inberg, L. (2013). Multiple Pathways to Gender-Sensitive Budget Support in the Education Sector. *WIDER* Working *Paper No. 105*. UNU-WIDER.

Hummelbrunner, R. (2015). Learning, Systems Concepts and Values in Evaluation: Proposal for an Exploratory Framework to Improve Coherence. (B. Befani, B. Ramalingam, & E. Stern, Red.) *IDS Bulletin, 46*(1), 17-29.

Kahneman, D. (2012). *Thinking, Fast and Slow.* Penguin.

Kask, J., & Linton, G. (2013). Business mating: when start-ups get it right. *Journal of Small Business & Entrepreneurship, 26*(5), 511-536.

Kingdon, J. (2010). *Agendas, Alternatives, and Public Policies.* Pearson.

Leeuw, F., & Vaessen, J. (2009). *NONIE Guidance on Impact* Evaluation. Washington, D.C.: Impact Evaluation Group (IEG).

Mackie, J. (1974). *The Cement of the Universe: a Study of Causation.* Oxford: Clarendon Press.

Marx, A., & Dusa, A. (2011). Crisp-Set Qualitative Comparative Analysis (csQCA), Contradictions and Consistency Benchmarks for Model Specification. *Methodological Innovations Online, 6*(2), 103-148.

Mayne, J. (1999). Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly. *Discussion Paper*. Office of the Auditor General of Canada.

Mayne, J. (2001). Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly. *The* Canadian *Journal of Program Evaluation, 16*(1), 1-24.

Mayne, J. (2008). Contribution Analysis: an approach to exploring cause and effect. *ILAC Brief 16*. Institutional Learning and Change (ILAC) Initiative (CGIAR).

Mayne, J. (2012). Making Causal Claims. *ILAC Brief 26*. Institutional Learning and Change Initiative (ILAC, CGIAR).

Meur, G. D., Rihoux, B., & Yamasaki, S. (2002). *L'analyse quali-quantitative comparée (AQQC-QCA): approche, techniques et* applications *en sciences humaines.* Louvain-la-Neuve: Academia-Bruylant.

Millard, A., Basu, A., Forss, K., Kandyomunda, B., McEvoy, C., & Woldeyohannes, A. (2015). *Is the end of child labour in sight? A critical review of a vision and journey.*

Mumford, S., & Anjum, R. (2013). *Causation: A Very Short* Introduction. Oxford University Press.

NONIESubgroup2. (2008). NONIE Impact Evaluation Guidance.

Pawson, R. (2006). *Evidence-Based Policy: a Realist Perspective.* Sage.

Pawson, R., & Tilley, N. (1997). *Realistic Evaluation.* Sage.

Raab, M., & Stuppert, W. (2014). *Review of evaluation approaches and methods for interventions related to violence against women and girls (VAWG).* London: Department for International Development (DFID).

Ragin, C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* University of California Press.

Ragin, C. (2000). *Fuzzy-Set Social Science.* University Of Chicago Press.

Ragin, C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond.* University Of Chicago Press.

Ragin, C. & Becker, H. (eds). (1992). *What Is a Case? Exploring the* Foundations *of Social Inquiry.* Cambridge University Press.

Ramalingam, B., Jones, H., Reba, T., & Young, J. (2008). *Exploring the science of complexity: Ideas and implications for development and* humanitarian *efforts.* London: Overseas Development Institute.

Rihoux, B. (2015). On trench warfare and hand grenades. Essay from the battlefield of critiques against QCA and set-theoretic/configurational comparative methods. *International conference on 'Qualitative Comparative Analysis - Social Science* Applications *and Methodological Challenges', Tilburg, 15-16 Jan 2015.*

Rihoux, B., & Lobe, B. (2009). The case for qualitative comparative analysis (QCA): Adding leverage for thick cross-case comparison. i D. Byrne, & C. Ragin, *The SAGE Handbook of Case-Based Methods* (ss. 222-242). Sage.

Rihoux, B. & Ragin, C. (eds). (2009). *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related* Techniques. Sage.

Rokkan, S. (1970). *Citizens Elections Parties. Approaches to the* Comparative *Study of the Processes of Development.* Oslo: Universitetsforlaget.

Romme, G. (1995). Self-organizing processes in top management teams: a Boolean comparative approach. *Journal of Business Research, 34*, 11-34.

Sager, F., & Andereggen, C. (2012). Dealing With Complex Causality in Realist Synthesis: The Promise of Qualitative Comparative Analysis. *American Journal of Evaluation, 33*(1), 60-78.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). *When will we ever learn? Improving Lives Through Impact Evaluation, Report of the* Evaluation *Gap Working Group.* Washington, D.C.: Center for Global Development (CGD).

Schneider, C., & Rohlfing, I. (2013). Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research. *Sociological Methods and Research, 42*(4), 559-597.

Schneider, C., & Wagemann, C. (2006). Reducing complexity in Qualitative Comparative Analysis (QCA): Remote and proximate factors and the consolidation of democracy. *European* Journal *of Political Research, 45*(5), 751-786.

Schneider, C., & Wagemann, C. (2010). Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy-Sets. *Comparative Sociology, 9*, 1-22.

Schneider, C., & Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences.* Cambridge University Press.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research. *Journal of* MultiDisciplinary *Evaluation, 5*(9).

Scriven, M. (2009). Demythologizing Causation and Evidence. i S. Donaldson, C. Christie & M. Mark (eds), *What counts as credible evidence in applied research and evaluation practice?* (pp.134-153). Thousand Oaks, CA: Sage.

Shah, M. K., & Narayan, D. (1999). *Consultations with the poor: Methodology guide for the 20 country study for the world development* report *2000/2001.* Washington, DC: The World Bank.

Stedman-Bryce, G. (2013). *Health For All: Towards Universal Health Care in Ghana. End of Campaign Evaluation Report.* Oxford, UK: Oxfam Great Britain.

Stedman-Bryce, G., Schatz, F., Hodgkin, C., & Balogun, P. (2016). *Medicines* Transparency *Alliance (MeTA) Evaluation.* ePact.

Stern, E. (2015). Impact Evaluation: a Guide for Commissioners and Managers. BOND UK.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the Range of Designs and Methods for Impact Evaluations. *DFID Working Paper 38*. UK Department for International Development.

Stevenson, I. (2013). Does technology have an impact on learning? A Fuzzy Set Analysis of historical data on the role of digital repertoires in shaping the outcomes of classroom pedagogy. *Computers & Education, 69*, 148-158.

Stinchcombe, A. (1998). Monopolistic competition as a mechanism: Corporations, universities, and nation-states in competitive fields. i P. Hedstrom, R. Swedberg, & (eds), *Social Mechanisms: An* Analytical *Approach to Social Theory* (ss. 267-305). Cambridge: Cambridge University Press.

The Global Environment Facility Independent Evaluation Office (GEF IEO). (2015). *Impact Evaluation of GEF/UNDP Support to Protected Areas and Protected Area Systems.* Washington, DC: The GEF IEO.

Thiem, A. (2013). Clearly Crisp, and Not Fuzzy: A Reassessment of the (Putative) Pitfalls of Multi-value QCA. *Field Methods, 25*(2), 197-207.

Thiem, A. (2015). Parameters of Fit and Intermediate Solutions in Multi-Value Qualitative Comparative Analysis. *Quality & Quantity: International Journal of Methodology, 49*(2), 657-674.

Thiem, A. (2015). Using Qualitative Comparative Analysis for Identifying Causal Chains in Configurational Data: A Methodological Commentary on Baumgartner and Epple (2014). Sociological *Methods & Research, 44*(4), 723-736.

Thiem, A., & Dusa, A. (2012). *Qualitative Comparative Analysis with R: a User's* Guide. Springer.

Thiem, A., Baumgartner, M., & Bol, D. (2015). Still Lost in Translation! A Correction of Three Misunderstandings Between Configurational Comparativists and Regressional Analysts. *Comparative Political Studies*, 1-33.

Treasury, H. (2011). *The Magenta Book: Guidance for Evaluation.* London: UK Government.

Vaessen, J., Garcia, O., & Uitto, J. (2014). Making M&E More 'Impact-oriented': Illustrations from the UN. *IDS Bulletin, 45*(6), 65-76.

Welle, K., Williams, J., Pearce, J., & Befani, B. (2015). *Testing the Waters: A Qualitative Comparative Analysis of the Factors Affecting Success in Rendering Water Services Sustainable Based on ICT*

Reporting. Brighton: Institute of Development Studies and WaterAid.

Westhorp, G. (2014). *Realist impact evaluation: an introduction.* London: Overseas Development Institute (Methods Lab).

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness, 5*(1), 30-49.

Williams, B. (2015). Prosaic or Profound? The Adoption of Systems Ideas by Impact Evaluation. (B. Befani, B. Ramalingam, & E. Stern, eds.) IDS *Bulletin, 46*(1), 7-16.

Williams, B., & Hummelbrunner, R. (2010). *Systems Concepts in Action: a practicioner's toolkit.* Stanford University Press.

Vink, M., & Van Vliet, O. (2009). Assessing the Potential and Pitfalls of Multi-Value QCA. *FIeld Methods, 21*(3), 265-289.

Vink, M., & van Vliet, O. (2013). Potential and Pitfalls of Multi-value QCA: Response to Thiem. *Field Methods, 25*(2), 208-213.

Vis, B. (2012). The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses. *Sociological* Methods *Research, 41*(1), 168-198.

GLOSSARY

Adjacent: see "contiguous"

Boolean: taking the values of either 0 or 1. In Boolean algebra the values of the variables are the truth values true and false, usually denoted 1 and 0 respectively. Unlike elementary algebra where the main operations are addition and multiplication, the main operations of Boolean algebra are conjunction, disjunction and negation. It is a formalisation of logical relations, like ordinary algebra is for numeric relations.

Calibration: the process of setting rules or creating rubrics to convert case-based information on given conditions into numeric or logical values (e.g. 0 or 1). See Section 2.3.

Case: a case is a unit of analysis for the testing of a theory. Cases used in evaluation can be projects, programmes, policies, countries, regions and other geographical units; individuals, households, target groups, etc.

Case-based analysis/methods: used to indicate approaches and methods drawing on relatively complex, "thick" cases, where a lot of information is available and useful for each single case and a small or medium number of cases are analysed at the same time. It is used in contrast with Variable-based analysis/methods (see below).

Causal connection (see "causal link")

Causal framework: the logic used to infer causality. Options range from Mill's Methods to configurational and generative, mechanism-based approaches (see Annex A).

Causal package (or causal pathway): a combination or conjunction of multiple causal factors, which is associated with the outcome as a whole. It is usually not possible to isolate the causal role of single components of a causal package (see Annex A)

Causal link (also causal connection): a relation between one or more causes and an effect (see Annex A). For example, the "wind made the door slam": the wind is the cause and the door slamming the effect (note that there are also other necessary causes: the door had to be open, movable/not fixed, etc.)

Combination (see "conjunction")

Condition: characteristic or property of the cases under investigation that is being selected for analysis. For example, if the cases are countries, the type of intervention that is being implemented in each country, or the quality of relations between donors and the government, or whether the country is politically stable, politically plural, etc. Since the characteristic is mostly described in qualitative terms and can take a limited amount of values (usually 0/1, presence/absence, high/low, or more generally a meaningful contrast between two options), this term is preferred to "variable", usually associated with quantities, real numbers and intervals.

Configuration: constellation or assembly of conditions associated with an outcome. A configuration can be either a conjunction of conditions, for example "high income AND low education" or a disjunction (e.g. "passport OR driving license"). The elements of a conjunction can be disjunctions (see SUIN causes) and the elements of a disjunction can be combinations, as in a typical solution to the Boolean minimisation (see Section 2.7)

Conjunction: the logical intersection of two or more sets, which is smaller than any of the sets taken individually. Two sets are united in a conjunction by the logical operator AND, which means that any element is a member of the conjunction if it's a member of all sets constituting the conjunction at the same time. For example, the conjunction of blonde people and tall people is the set of people who are both tall and blonde.

Construct Validity: the ability of an indicator or a score to faithfully represent the concept or notion it is meant to represent or measure. Another way to define construct validity is to focus on the extent to which theoretical insight can be gained from indicators or scores or other ways the evidence has been operationalised. For example, is annual income a valid measure of well being? If we want to assess general health, is the result of a specific blood test a valid indicator? And so on. Particularly relevant for this report are measures of success of an intervention.

Contiguous: areas sharing one or more boundaries.

Contradictory case: a case presenting a combination of conditions which does not consistently lead to the same outcome across the entire dataset.

Counterfactual (in relation to impact evaluation): in impact evaluation, the "counterfactual" is normally used to refer to the hypothetical, unobservable situation where the intervention has not been implemented (while in reality it has). It is the answer to the question "what would have happened without the intervention" and by definition it cannot be observed (it is "counter to fact" and distinct from "factual", which is what can be observed). It is usually reconstructed by applying Mill's Method of Difference (see Annex A).

Counterfactual (in relation to the Boolean minimisation): is a logically possible combination of conditions which is not supported empirically. It is the same as "remainder" and "logical case" (see Section 2.6.2). There is a distinction between "easy counterfactual" and "difficult counterfactual", depending on the directional expectations the researcher sets on the conditions.

Cross-case (comparative) methods: methods aimed at comparing a small or medium number of cases and gaining insight on the cross-cutting features which can be generalised to the entire set of cases; or to create typologies of similar cases. They are logically different from "within-case" methods.

csQCA (or crisp-set QCA): the original version of QCA, where the dataset is a Boolean matrix (a table consisting of 0s and 1s). Cases are represented (or "covered") by combinations of conditions.

Dichotomous (see Boolean)

Disjunction: the logical union of two or more sets, which is larger than any of the sets taken individually. Two sets are united in a disjunction by the logical operator OR, which means that any element is a member of the disjunction if it's a member of one or more sets constituting the conjunction (at least one). For example, the disjunction of people with a PhD and people with at least 5 years of research experience is the set of people who either have a PhD or at least 5 years of research experience (they could have both, too, but what matters for disjunction membership is that they have at least one).

Equifinality: See Section 1.4.2.

External Validity: the degree to which findings obtained in one setting hold for a plurality of cases, situations, contexts, individuals or groups; as opposed to merely the specific cases used to obtain those findings.

fsQCA (or fuzzy-set QCA): the version of QCA handling fuzzy values, or values that can range between 0 and 1 rather than being either 0 or 1. The value assigned to a condition in a given case represents the membership score of the case to that condition, or the degree to which that condition is present in the case (e.g. "1" means full presence, "0" full absence, "0.6" more presence than absence, "0.9" almost full presence, and so on). Instead of being simply "covered" by particular combinations or not, cases have fuzzy degrees of membership to all possible combinations, but a degree of membership higher than 0.5 only to one combination, which will be the combination they are closest to.

Impact Evaluation: an evaluation aimed at establishing a causal connection between one or more interventions and one or more outcomes or effects. DAC defines impact evaluation as "the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended".

Internal Validity: the absence of systematic error or bias in the research findings. Such bias may stem from confounding variables or by particular characteristics of documentary and human sources and resources; informants, or sampling. An internally valid research process is such that the findings are not likely to differ if other researchers replicate the study in the same setting, following the same protocol.

Intersection (see logical intersection)

Intervention: (in this context) one or more policies, programmes, projects, or activities aimed at reaching desirable development objectives. Many interventions have multiple outcomes and are simultaneously implemented. Some are embedded or "multi-level": that is, prepare the ground for other interventions to implemented or to work, and perhaps require other interventions to be implemented or to work, in turn, to be implemented or to work themselves.

INUS cause: literally an "insufficient but necessary component of an unnecessary but sufficient cause" for an outcome, the INUS cause is first a part of a sufficient (but unnecessary) causal package. For example, a condition which is part of a combination (sufficient pathway to an outcome) in a QCA solution. Secondly, it is a special component, because if we take it away, the package loses its sufficiency and no longer leads to the outcome. In other words – even though in itself it is neither necessary nor sufficient for the outcome – it is necessary for the package it is a member of to be sufficient. See Section 2.8 and Annex A.

Large-N (analysis): study analysing 30 or more cases.

Logical case: a logically possible combination of conditions which is not supported empirically and thus excluded from the dataset or the standard Truth Table. It is however included in the complete Truth Table (see Section 2.6.2).

Logical Intersection: in set theory, the logical intersection of two or more sets is a smaller set resulting from the overlap of each set participating in the intersection. In order to be a member of the intersection, being a member of all the sets participating in the intersection is required. For example, the set of blonde tall women is the intersection of the blonde people set, the women set and the tall people set. See also 'conjunction'.

Logical Negation: in set theory, the logical negation of a set is the set comprising all possibly existing elements except those which are members of the first set. In order to be a member of the negation, the only requirement is not being a member of the first set. The union of a set and its negation includes all possibly existing elements. For example, the negation of the AM hours is the set of PM hours: hours are either AM or PM.

Logical Union: in set theory, the logical union of two or more sets is a larger set, comprising all the elements which are members of at least one of the sets constituting the union. In order to be a member of the union, being a member of at least one of its sets is sufficient. For example, the documents which are valid proof of identity is the logical union of driving license, national identity card, and passport. In order to prove identity, it is sufficient to provide at least one type of document, as opposed to all of them (which is what happens with logical intersection, see above; see also 'disjunction').

Medium-N (analysis): study analysing between 10 and 30 cases

Model (see QCA model)

mvQCA: multi-variate QCA is the version of QCA handling multiple values rather than merely 0 or 1 or the intermediate degrees in-between: conditions can take the values or 0, 1, 2, 3, 4 and so on. It is similar to csQCA in that the combinations are "crisp" and cases are unambiguously covered by conditions and combinations (there is no "degree of membership"). However Boolean minimisation is more difficult as the required set of cases for merging combinations is larger than two and requires all cases to be identical on all conditions except one and to cover all possible values of the remaining condition (for example: for a three-value QCA where values can be either 0, 1 or 2, the three combinations A1*B1*C1, A1*B1*C0 and A1*B1*C2 are required for simplification into A1*B1).

Necessary: required for an outcome to materialise. A condition is necessary for an outcome if, whenever we observe the outcome, we also observe the condition. If no single condition is perfectly necessary for an outcome, the larger set of a disjunction of single conditions might be. For example, a passport might not be necessary to prove identity but either a passport, a driving license, or a national identity card most likely are (at least one of these is required, so their disjunction, or logical union, is necessary).

Necessity: in set-theory, if X is necessary for Y, then X is a superset of Y, or fully includes Y. Put differently, Y is fully included by X. Whenever an element is a member of Y, it is also a member of X. In logical terms, Y logically implies X. See also 'necessary' and 'superset'.

Outcome: a desirable or undesirable, intended or unintended observable state, which might be one of the objectives of the intervention; or simply one of its consequences. In QCA, one or more outcomes of interest are analysed and an attempt is made to explain their existence; or to identify a causal connection between the intervention, other causal factors and the outcome. The outcome can be thought of as "the effect" of a causal package. See Section 2.1.1.

Parameters of Fit: measure how well a QCA configuration fit the empirical dataset. There are broadly two categories of parameters of fit: consistency and coverage. They are measured for both the necessity analysis (necessity-consistency and necessity-coverage: see Section 2.5.1.4) and the sufficiency analysis (sufficiency-consistency and sufficiency-coverage: see Section 2.5.2.4).

QCA model: a set of elements comprising one outcome and a list of conditions assumed to affect it/explain it (see Section 2.1).

QCA solution: one or more configurations synthesising the dataset or informing on particular aspects of it. It results from the application of a QCA procedure (supersubset analysis, Boolean minimisation, INUS analysis). QCA solutions come with their own parameters of fit (consistency and coverage). Most often used to refer to the output of the Boolean minimisation: one or more combinations that taken together summarise the information included in the dataset. The combinations in the solution will most likely be simpler and fewer than those included in the Truth Table (see Section 2.7).

Quantitative Methods (see variable-based analysis/methods)

Qualitative Methods (see case-based analysis/methods)

Remainder (see logical case)

Replicability: refers to the extent to which a re-study of a phenomenon repeats the findings of an initial study. It requires reproducibility (which is the ability of the study to be repeated) but goes beyond the former: in order to be replicable, not only the study needs to be reproduced but it also needs to return the same findings.

Rubrics: two-column tables describing the meaning of a scale of values, which can be qualitative (excellent, good, poor, unacceptable) or quantitative (0.2, 0.6, 1.00, etc). Usually the first column will display the values and the second column the corresponding meaning (the meaning assigned to those values).

Sensitivity (Analysis): analysis conducted on the findings obtained towards the end of the research process (or the end of an iteration), aimed at assessing the findings' robustness or reliability. Small changes are made to the parameters or other decisions taken during the research process to see if these have any implications on the findings. Following these changes, the new findings are compared to the old ones and if no change or only minor changes are detected, the findings can be considered robust.

Set theory: the branch of mathematical logic that studies sets, informally known as "collections of objects". The basic logical operations performed on sets are: union, intersection, and negation.

Small-N (analysis): study analysing 10 or fewer cases.

Solution (see QCA solution)

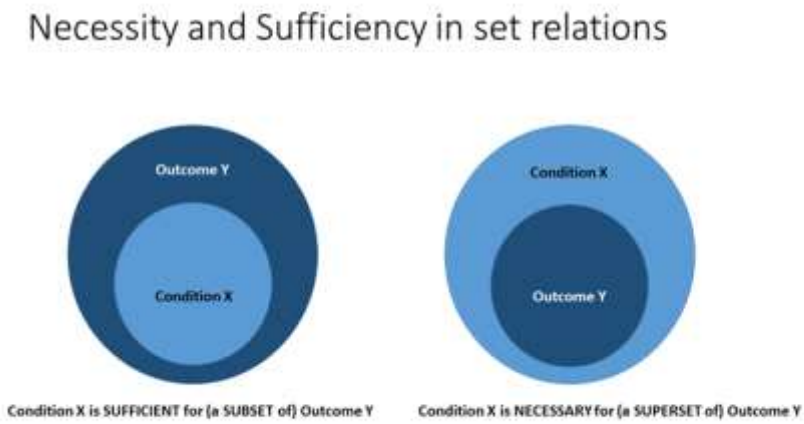
Subset: a set which is smaller than another one is called a subset of the latter if all elements of the former are also included in the latter. Put differently, the first set (the subset) logically implies (is sufficient for) the second (larger) set: being a member of the first implies (is sufficient for) being a member of the second.

Sufficiency: in set-theory, if X is sufficient for Y, then X is a subset of Y, or fully included in Y. Put differently, Y fully includes X. Whenever an element is a member of X, it is also a member of Y. In logical terms, X logically implies Y. See also 'sufficient' and 'subset'.

Sufficient: able to trigger the materialisation of an outcome without requiring the presence of additional conditions. A condition (or combination) is sufficient for an outcome if, whenever we observe the condition (or combination), we also observe the outcome. If no single condition is perfectly sufficient for an outcome, the smaller set of a combination of single conditions might be. For example, flour is not sufficient to make dough, but a combination of flour and water is.

SUIN cause: literally a "sufficient but unnecessary component of an insufficient but necessary cause" for an outcome, the SUIN cause is first a part of a necessary (but insufficient) disjunction. For example, the proof of identity among the documents needed to board a flight is necessary but not sufficient: a boarding card is also needed. Secondly, proof of identity is a disjunction of perhaps 3 different factors: passport, national identity card and driving license. None of the components of this disjunction are necessary because if we don't have a specific one we can replace it with any of the other two; at the same time any component will be sufficient to prove identity. So "passport" is a SUIN cause because it is sufficient but not necessary to prove identity; and proof of identity is only one of the documents required to board a flight. An intervention is a SUIN cause for an outcome if it is one of a set of interchangeable, functionally equivalent factors, at least one of which is necessary (but not sufficient) to achieve an outcome. The fact that at least one of these factors are present satisfies a requirement; it is however not sufficient to achieve the outcome.

Superset: a set which is larger than another one is called a superset of the latter if all elements of the latter are included in the former. The first set (the superset) is logically implied by (is necessary for) the second (smaller) set: it's not possible to be a member of the second without also being a member of the first (the superset). In other words, membership of the first set (the superset) is necessary for membership of the second, and membership of the second set requires membership of the first (the superset).


Necessity and Sufficiency in set relations
Outcome Y
Condition X
Condition X
Outcome Y
Condition X is SUFFICIENT for (a SUBSET of) Outcome Y
Condition X is NECESSARY for (a SUPERSET of) Outcome Y

Union (see logical union)

Variable-based analysis/methods: used to indicate approaches and methods focusing on a set of specific, quantifiable characteristics of cases. The available case based information is converted into numeric values (usually real numbers); and a medium or large number of cases are analysed at the same time. It is used in contrast with case-based analysis/methods (see above).

Within-case methods: methods aimed at analysing one or at most a handful of single cases, gaining in-depth insight on the specific characteristics of each case rather than on the common, cross-cutting features. They are logically different from “cross-case”, comparative methods.

ANNEX A: Causal Frameworks underpinning impact evaluation methods

The methods aimed at establishing causal connections in impact evaluation are based on one or more of three basic causal inference frameworks (Befani B. , 2012): single-cause frameworks, multiple-cause frameworks, and generative frameworks. Single-cause frameworks focus on attributing single causes to effects and include all of Mill's Methods: Mill's Method of Difference, Mill's Method of Agreement, Mill's Method of Concomitant Variation and Mill's Method of Residue. Multiple-cause frameworks aim at attributing or understanding the role of multiple causes: they include a variant of Mill's Method of Concomitant Variation and configurational frameworks. Finally, generative or mechanism-based frameworks aim at explaining in detail the "inner workings" of a causal mechanism as if the analyst could observe the transformation of the cause into the effect with a magnifying lens: as if observing causality at work. They include process or "causal chain" mechanisms as well as complex systemic interrelationships.

Single-cause frameworks: Mill's Method of Difference (MoD) and counterfactual experimentation

So far, the most popular causal inference strategy in impact evaluation has been Mill's Method of Difference, used in experimental evaluation approaches and in many quasi-experiments. In order to attribute the effect to one cause, the method seeks to compare the case where both the cause and the effect have been observed (for example a situation where an intervention has been implemented and an outcome observed) with a situation where both the effect and the cause are missing but every other potential causal factor is the same (Table A1). In evaluation this entails reconstructing the counterfactual, non-intervention situation and estimating the outcome that would have materialised then; eventually subtracting it from the observed outcome to estimate the "net" effect.

Table A1: **Mill's Method of Difference**

Case	Potential Causes	Candidate Cause	Effect
Treatment	A B C D E	X	Y1
Control	A B C D E	-	Y0

Net Effect = Y1 – Y0

While intuitive and easy to understand, the method is philosophically problematic and can be extremely challenging to apply in practice. Philosophically, it is misleading in situations of over-determination and pre-emption: for example, if applied sequentially to each member of a firing squad, none would likely be considered the cause because the prisoner would die anyway (over-determination) if one member of the squad refrains from shooting. This shows that when other factors in addition to the intervention might cause the achievement of the outcome, and the intervention is shown not to make any difference, it doesn't mean that the intervention is not effective: it might just mean that its effect was pre-empted by the effect of other factors.

One example of pre-emption is when two killers try to murder a man on a long walking trip in the middle of the desert, the first by putting poison in his water tank, and the second by puncturing his water tank. The man will either die by drinking poisoned water or by not drinking water at all, but each cause will only act if the other cause has not already acted before. In one reconstructed counterfactual situation where poison is not put in the tank, the man still dies from thirst. In the other reconstructed counterfactual situation where the tank is not punctured, the main is killed by poison. This doesn't mean that either poison or tank puncturing are not effective killers: it just means they were not effective on that day because their effect was pre-empted by another cause. Translated into development evaluation language: if a particular type of intervention does not appear to make any difference in a specific context, it doesn't necessarily mean that it's not effective; it might just mean that something else has influenced the outcome to a point where it can't be improved any further, before the intervention could do it.

In the practice of development evaluation, preserving the equivalence of treatment and control groups and the continuity and exclusivity of the treatment to the treatment group, can be impossible at times (see Hawthorne effect and differential attrition in

Scriven, 2008; 2009). In addition, a long list of threats to the internal validity of experiments has been identified (Campbell D. , 1969; Campbell & Stanley, 1963; Cook & Campbell, 1979). Counterfactual analysis has been accused of being weak on external validity as well (Cartwright, 2012) and it does not provide any indication that guarantees sufficient levels of construct validity.

In terms of evaluation questions, when not used in combination with other methods, counterfactual analysis fails at understanding what made interventions work under given circumstances and more generally, how and why interventions work (or not), also failing to inform prediction of whether and when interventions will work in the future.

Despite these limitations, both Mill's Method of Difference and counterfactual thinking can be fruitfully applied in impact evaluation when the right conditions are met and when the commissioner's interest is focused on the net effect. Imagine an emergency situation, like a food crisis, putting a large number of people at risk of undernutrition and ultimately starvation for a few weeks. Let's assume that we observe two similar groups of people being affected by the crisis in the same way, and that donor agencies have for some reason only limited supplies and can only intervene in one area (area A). If it can be argued that the two groups are equivalent in terms of factors affecting hunger and starvation, and the outcome of area B can be considered a good approximation of what would have happened in the first group (A), had the intervention not been implemented, then the comparison of nutrition-related outcomes (for example starvation-related deaths) between the two groups can be considered a robust estimate of the impact of the intervention.

Following the Method of Difference, in this case we compare two almost identical cases, differing only in terms of whether they have been exposed to the intervention or not, and measure the difference in the outcome of interest, concluding that this difference is attributable to the intervention. The specific impact evaluation question we answer is "how much of a difference did the intervention make" or "how large is the net effect of the intervention". "Net" because – being the two areas practically identical in terms of factors affecting hunger and starvation – it's difficult to think of other factors contributing to that difference.

Single-cause frameworks: Mill's Method of Agreement (MoA) and statistical frequency

An early theory of causation introduced by David Hume (a.k.a. the regularity account of causation, see Brady, 2002) suggested that "cause seekers" focus on the frequency of events: if one effect is constantly observed in conjunction to the candidate cause, while other potential causes change, this constitutes evidence of a causal link between the candidate cause and the effect.

The Regularity account is related to Mill's Method of Agreement, which – similarly to Mill's Method of Difference – seeks to identify a special case to compare with the one where both the candidate cause and the effect have been observed. However, unlike the former, the Method of Agreement seeks a case where the effect is still present while all other potential causes are absent except the candidate cause (see Table A2).

Table A2: **Mill's** Method of Agreement

Case	Potential Causes	Candidate Cause	Effect
Treatment in Context 1	A B C D E	X	Y
Treatment in Context 2	A G C J E	X	Y
...	...		
Treatment in Context N	F G H J K	X	Y

X causes Y
(because it couldn't be any other factor: they all change)

Like the MoD, the MoA is essentially an elimination approach: if candidate causes and effects are not consistently associated, they can be eliminated from the list of possible causes; otherwise, if candidate causes and effects are consistently associated they must be retained (see Scriven 2008).

The challenges for this method are partly similar to the MoD because we never know for certain if all possible causes have been accounted for; and the method does not inherently protect against internal validity or construct validity biases. Another problem is that it is helpful in confirming association but not in understanding why a certain factor (for example the intervention) is affecting or has affected the outcome. However, the MoA is useful to predict that a certain effect will follow from a certain cause in the future or in another context; in this sense it has an external validity advantage.

The Method of Agreement can be applied if a relatively large number of different cases can be observed; for example, back to our Food Aid intervention, when supply meets demand and all areas in need are covered by emergency assistance, it is not possible to compare two similar cases that differ only in the presence or absence of the intervention. However, as data is available across a high number of areas, it might be possible to compare the nutrition-related outcomes, say 3 weeks from the intervention, across all areas affected and assisted. If the nutrition-related indicators are similar across a wide variety of contexts, this uniformity might be due to the intervention because it might be *hard to imagine another factor plausibly contributing to the same outcome, consistently present across all cases*. More tests might be needed but the key point here is that, instead of comparing two almost identical cases, we compare a wide variety of situations that have nothing in common except the intervention and similar nutrition-related outcomes. It is called Method of Agreement because the cases only "agree" on (share) the candidate cause (the intervention) and the effect (nutrition-related outcomes), while all other plausible causal factors are different.

Single-cause frameworks: Mill's Method of Concomitant Variation (MoCV) and correlational analysis

One limitation of the above frameworks is that they only deal with presence and absence of causes and effects, without considering *degrees* or quantities. Mill's Method of Concomitant Variation overcomes this limitation by focusing on the magnitude of change brought about by causes. The basic idea is that, in order to demonstrate causality, there must be some proportion between the magnitude of the candidate cause and the extent of the effect. This logic is visualised in Table A3.

Tabel A3: **Mill's Method of Concomitant Varia**tion

Case	Potential Causes	Candidate Cause	Effect
One	A B C D E	X	Y
Two	A B C D E	X +	Y +

X is correlated (and perhaps causally) related to Y

This framework is particularly useful in its multiple-causality variant, illustrated below.

Multiple-cause frameworks: Mill's Method of Concomitant Variation (MoCV) and multiple correlational analysis

MoCV is particularly useful when the effect can be conceptualised as growing proportionally to not just one, but a series of factors that affect it independently of each other. The effect of each single factor can then be identified by conducting a correlational analysis, illustrated in Table A4. Note that this cannot be considered multiple causality yet, because factors affect the outcome one at a time; and their incremental effects are linearly "added up" to each other in the total value of the outcome, as in a linear regression model with no interaction effects (see also Annex B).

Table A4: **Mill's Method of Concomitant Variation applied to multiple additive** causes

Case	Potential and Candidate Causes	Effect
One	A B C	Y
Two	A B C +	Y +
Three	A C B +	Y +
Four	B C A +	Y +
Five	A B + C +	Y + +
Six	B A + C +	Y + +
Seven	C A + B +	Y + +
Eight	A + B + C +	Y + + +

This method is appropriate when several factors seem to contribute independently to the outcome. For example, in our food aid intervention, we might notice that, in spite of similar interventions being implemented across all areas in need, nutrition-related outcomes are much higher in some areas than in others. There might be a high variation in the outcomes as well as a high variation in factors that could affect the outcomes: perhaps different levels of food supply? Different capacities of resilience to food-related crises? It might be appropriate to undertake a correlational analysis, estimating how much each additional unit of different factors (food supply, pre-intervention ability to withstand crisis, etc.) improves nutrition-related outcomes.

The question we would be answering is still "how much of a difference did the intervention make" or "how large is the net effect of the intervention"; however, we would also be looking at the net effect of other factors. The method is not helpful in understanding how or why each factor made a difference to the outcome and the association might not necessarily be causal (correlation does not imply causation). In addition, like for the MoA, there is nothing in it that necessarily protects from internal validity or construct validity biases. However, like the MoA, it is helpful in predicting future situations and it is strong on external validity.

It is important to notice that the MoCV assesses only independent contextual influence: it sees the context as a variable affecting the outcome in and of itself, "filtering" the action of other factors as an independent "mediator". The MoCV is useful to assess the average impact of a series of factors, including contextual factors and the intervention, but not to understand how these factors intersect and function within particular types of "packages"; and how their influence changes depending on what other factors they're combined with. In other words, it is not appropriate to answer the question "what makes the difference for whom and under what circumstances".

Multiple-cause frameworks: Configurational or Multiple-Conjunctural Causation (MCC)

If MoCV and correlational analysis can be thought of as the quantitative approach to the analysis of multiple causes, one way to think of configurational causality and QCA is as the qualitative approach to the analysis of multiple causation. The differences do not end in the type of data these approaches usually handle (correlations can also be performed on dummy or scale-like variables and configurational analysis can handle quantitative data) but are rooted in the philosophy of causality.

We have seen "regularity" and "counterfactual" accounts of causation above; another theory of causality sees objects as owning specific static properties that make them more likely (or "*disposed*") to produce specific effects or undergo given transformations (Mumford & Anjum, 2013). According to this theory, objects or conditions are "naturally" predisposed to trigger specific effects. In this sense, causality is an intrinsic attribute of those objects and conditions.

In practice, this means that whenever we attempt to causally attribute an outcome, we try to discover the typical conditions under which that outcome is likely (or unlikely) to materialize; and are ready to observe different (levels of) outcomes under different pre-existing circumstances.

In the food-aid intervention situations described above, we were preoccupied with demonstrating a strong, believable association between the intervention and a desired level of nutrition-related outcomes. We were not conceptualizing the intervention as having different roles in different contexts; we were just looking for the average contribution, or the average difference.

But what if correlation between the intervention and the outcomes is weak, and we are unable to find a multiple-variable model that fits the data well? Would this mean that food aid and nutrition-related outcomes are not related? That food aid has not contributed to their improvement?

Before reaching this conclusion, it might be useful to describe and analyse the process leading from delivery of food to actual food intake by local populations in qualitative terms. When planes with food supplies land in airports, the food does not become immediately available to the population. Even if the food is of sufficient quality and quantity to cover the entire population in the area, it still needs to be transported to cities and villages, and some of these locations might be hard to access. They might be reached too late, when deaths have already occurred or health status deteriorated irreversibly. Another possible bottleneck are power dynamics in cities and villages. Food might be delivered to community leaders, who might not distribute it equally and might stock or sell the surplus elsewhere instead. This will also lead to unintended/unexpected outcomes.

These factors do not necessarily affect nutrition-related outcomes in a linear, direct way: they might combine in unexpected packages and produce local effects which are difficult to attribute to the single components of these packages. For example, through the systematic comparison of affected areas as wholes, without isolating the single factors, we might conclude that healthy community dynamics are necessary to achieve high outcomes, because we never observe outcomes unless these are present; but not sufficient, because outcomes are not always achieved in their presence. In order for outcomes to be achieved, good infrastructure might also be necessary, or the vicinity of communities to the landing zones might be. In short, we might identify a limited number of different pathways all equally leading to success (or lack thereof):

- Closeness to airport AND Positive Comm. Dynamics => HIGH Nutrition-related Outcomes
- Good infrastructure AND Positive Comm. Dynamics => HIGH Nutrition-related Outcomes
- Negative Community Dynamics => LOW Nutrition-related Outcomes

One fundamental difference between MCC and the MoCV is that the former draws on the notions of causal **necessity and sufficiency**. We might notice that all affected areas with positive outcomes have either populations living close to the landing zones or relatively better infrastructure in the rural areas, which means that this ***disjunction*** is necessary for the outcome. Moreover, affected areas with positive outcomes appear to require non conflictual social dynamics, but the latter alone are not sufficient for success. We might also notice that, when community dynamics are conflictual, neither infrastructure nor vicinity to airports are relevant: outcomes will invariable be negative.

The causal framework we would be using to answer the question "what difference did the intervention make, for whom and under what circumstances"? is called "**configurational**" or "**multiple-conjunctural**" and allows the identification of patterns of association between packages of causes and effects. This framework shares both fundamental similarities and fundamental differences (Ragin 1987, Mackie 1974) with the ones seen above. The fundamental similarities concern the use of the agreement and difference logic, that is the comparison of either cases with a similar outcome (which will hopefully share the same candidate cause, as in the MoA) or cases with a different outcome (which will hopefully be almost identical and differ only in the candidate cause). The fundamental difference, on the other hand, is that the connection sought is **not between one cause and one effect** (at a time) but rather **between a configuration of causes and the effect**.

The configuration can take the form of a **combination** of causal factors/conditions (all factors need to be observed in order for the combination to be observed) or a **disjunction** of factors/conditions (at least one factor needs to be observed in order for the disjunction to be observed). The configuration in the first bullet point above is a combination of the conditions "closeness to the airport" and "positive community dynamics". If we consider the first two bullet points together (connected together by a logical OR), the result is a disjunction of two combinations leading to the outcome (because only one combination is sufficient to achieve success; we don't need both). The framework logic is illustrated in Tables A5 and A6, the former for necessary but not sufficient causes and the latter for sufficient causes.

Table A5: Configurational causality: identifying necessary (but not sufficient) causes

Case	Potential Causes	Candidate Cause	Effect
Case 1	A B C D E	X	Y
Case 2	A G C J E	X	Y
Case 3	F G H J K	X	Y
...	...		
Case N	F G H J M	X	Z

X is necessary (but not sufficient) for Y

Notice the similarity between the necessity table (A5) and the Method of Agreement (Table A2). The difference between the two causal models is that configurational causality interprets the association as a necessity relation; and implies that in order to determine sufficiency evidence on the absence of the outcome, or cases with negative outcomes, are needed. Another difference is that the Method of Agreement does not explicitly address the association (necessity) of disjunctions, but only of single causes.

Table A6: Configurational causality: identifying sufficient causes

Case	Potential Causes	Candidate Cause	Effect
Case 1	A B C D E	X	Y
Case 2	A B C F G	X	Y
Case 3	A B C H K	X	Y
...	...	...	...
Case N	A B D E F	X	Z

X alone is not sufficient for Y, but the combination A B C X is

The table illustrating sufficiency (A6) also shows similarities with the Method of Agreement (Table A2) (and with the necessity table A5: X is still necessary): the first three combinations are compared and the common elements A B C X retained to causally account for the same outcome Y(note that X alone is not sufficient). However, while MCC would consider the consistent association between the package (A B C X) and Y a satisfactory finding, the MoA (as a single-cause framework) would continue seeking additional cases presenting the conditions included in the package (A B C X) and not presenting Y at the same time, in order to eliminate a higher number of candidate causes, until only one is left. So, for example, if case A B D E F X Z was found, ***the MoA would eliminate A, B and X, leaving C as the only candidate cause***. Another difference is that, in MCC, multiple pathways can be equally sufficient (for example the first three combinations in Table A6 are all sufficient for Y), while in the MoA only C would survive as a candidate cause for Y.

The INUS cause

Grouping and analyzing the cases presenting a similar outcome (say, the successful cases with positive nutrition-related outcomes) allows us to establish which factors are required/necessary for success. However, in order to answer the impact question "**what difference did the intervention make**", we need to compare successful and unsuccessful cases (in this situation, areas with positive and negative nutrition-related outcomes) and identify combinations or "causal packages" which appear sufficient for success (meaning that whenever they are present, we observe success).

To this purpose, let's assume that we find one case with healthy community dynamics, good infrastructure and a negative outcome. This seems to counter what we learned above and our theory that the combination of these two conditions is sufficient for success. How can we explain the contradiction? At a closer look, we might discover that particularly unfriendly weather conditions spoiled the food in some areas before it arrived at destination, while either the friendly weather or proper food preservation facilities in unfriendly weather allowed the food to arrive safely to the rural communities. The second pathway above is then replaced by the following:

1. Good infrastructure AND Positive Comm. Dynamics AND Friendly Weather => HIGH Nutrition-related Outcomes
2. Good infrastructure AND Positive Comm. Dynamics AND Unfriendly Weather AND Preservation Facilities => HIGH Nutrition-related Outcomes
3. Good infrastructure AND Positive Comm. Dynamics AND Unfriendly Weather AND Lack of Preservation Facilities => LOW Nutrition-related Outcomes

Negative nutrition-related outcomes are observed even with good infrastructure and positive community dynamics, when the weather is unfriendly and no food preservation facilities are available. If either the weather is friendly, or there are facilities, the outcome is positive.

How can we then describe the role of the weather in the success of the programme? Is it a cause of success? Can success be attributed to the weather, to some extent? Good weather is not necessary (there are successful cases in unfriendly weather) and is obviously not sufficient for success (irrelevant when community dynamics are conflictual). At the same time, *the weather makes the difference between success and failure when there are no preservation facilities, in a context of good infrastructure, and good community relations* (comparison of 1 and 3 in the box above). In other words, it is an **INUS cause**: **it makes the difference** not in general, in a universal-law kind of way; but **only in a specific context**.

Food preservation facilities have a similar role in another context: they make the difference not in general (they are neither necessary nor sufficient by themselves), but only in the context of unfriendly weather (provided good infrastructure and positive community dynamics are observed – comparison of 2 and 3 in the box above). In other words, preservation facilities are irrelevant in conflictual communities, or when communities live close to airports, or when infrastructure is poor, or in friendly weather; they make the difference only when all these four conditions are missing.

The logic of INUS causality is illustrated in Table A7. Notice the similarity with the Method of Difference: all potential causes are the same except one, and the outcome is different. The only difference between the two causal frameworks is that INUS analysis considers X as determinant for success only *in the context of A B C D E*, while ***the***

MoD (as a single-cause framework) ***would*** simply ***eliminate those causes because they are equally present in the two cases***, ***seeing no role for them in the causal attribution of the different outcome***.

Table A7: Configurational causality: identifying INUS causes

Case	Potential Causes	Candidate Cause	Effect
Treatment	A B C D E	X	Y1
Control	A B C D E	-	Y0

X made the difference between Y0 and Y1 in the context of A B C D E

In sum, the findings from an impact evaluation designed using any of Mills's methods (single-cause frameworks) help us associate the intervention with nutrition-related outcomes. Conversely, the findings from an impact evaluation designed using configurational frameworks reveal that causal factors, among which the intervention, do not work independently, by themselves, to improve nutrition-related outcomes; it is packages of them, which might or might not include the intervention, to be consistently associated to given levels of outcomes. The intervention might not be necessary for success: causal packages not including the intervention could also be successful. And the intervention by itself might not be sufficient, needing help from the right "ingredients" or contextual factors, to achieve success. Despite the inconsistency or irregularity of association between the intervention and the outcome, the intervention might still be a "cause" in the INUS sense (Mackie, 1974), and make the difference in a specific context. Its role would be "conditional" to this context, or "conjunctural"; or in other words not consistently the same across all contexts.

These last findings allow us to understand what conditions or factors are required, under what circumstances, to achieve success, in addition to the intervention (assuming the latter has a role at all). In other words, we learn about the necessary ingredients which, when combined, achieve success. There is not only one good recipe (different combinations of ingredients achieve equal success) and one ingredient in itself is usually not very helpful. This will create **expectations** in terms of results **when the intervention is implemented in the future**, in a context with similar "ingredients". In line with "**dispositional causation**" we would expect some results to materialize only in areas presenting certain "dispositions" or meeting specific requirements.

Which would allow us to answer the question "**where/when will the intervention work in the future**?"

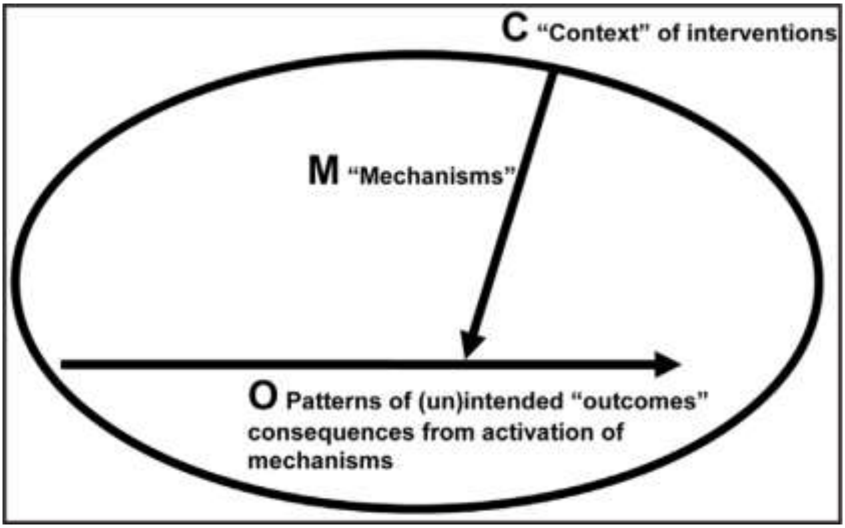
This approach allows some level of external validity, in that the combinations found to be successful might create a "typology" of cases that work in a similar way, which can be to some extent generalised. Construct validity would hinge on how the conditions are identified and internal validity depend, amongst else, on how rigorous and systematic the method used for cross-case comparison is.

Generative Frameworks and causal mechanisms

The third broad category of causal frameworks is based on causality defined as **transference**. Transference theories of causation explain causality through the transfer of properties from one object or condition to another (Mumford & Anjum, 2013). In the natural sciences, these can be energy or impulse: they explain how object states and properties are generated. In this theory, a causal relationship is defined through the detailed description of the causal mechanism or process responsible for the transfer of properties from cause to effect.

At its core, the defining characteristic of the causal explanation under this model is *depth*. The ambition is to reconstruct the event of "effect production" as if the analyst were able to observe causality at work. It is akin to "opening the black box" and learning about the "inner workings" of the process or system. While the complete mechanism can take the form of either a mostly linear causal chain or sequential process, or of a more complicated and complex system, the basic element of the explanation can also be represented with the "realist egg" (Pawson & Tilley, 1997), a self-contained unit where the three components of context, mechanism and outcome are inextricably intertwined. (Figure A1).

Figure A1: The realist "egg"


C "Context" of interventions
M "Mechanisms"
O Patterns of (un)intended "outcomes" consequences from activation of mechanisms

Independently of how we describe the mechanism (see below for process-like or system-like mechanisms), opening the symbolic black box implies describing the behavior of some of the actors, for example in terms of stakes, incentives, beliefs, preferences, skills, resources. Why do some community leaders withhold resources from the poorest groups while others don't? Why are infrastructures and facilities available in some areas and not in others (e.g. what institutional or historical processes created these differences?). How are the routes of supply-carrying flights established and why are some airports not covered?

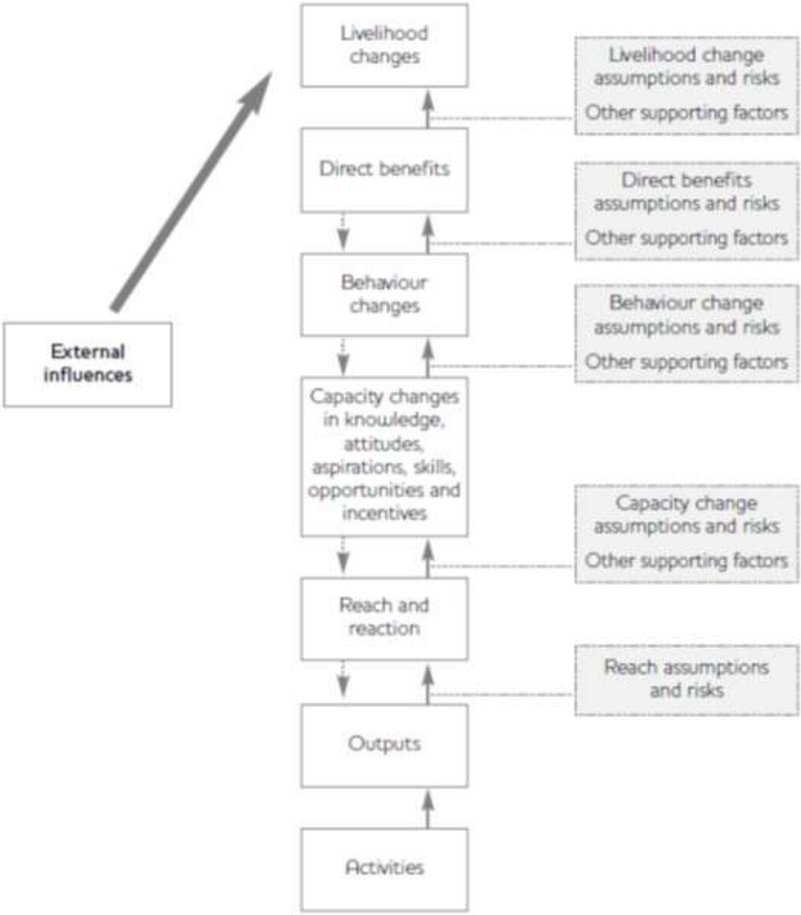
In addition to primary data collection, we would be using previously existing local knowledge or previous evaluations of similar interventions. We would not just observe that some factors made the difference, we would seek fine-grained explanations of why they did; and fine-grained descriptions of those factors and the reasons why certain mechanisms have been triggered under the circumstances. While not providing any particular guarantee of internal or external validity (although both can be strong under particular circumstances), this approach aims to maximize **construct validity**.

Answering the question "**how/why did the intervention make a difference** (or not)?" can help us reverse, or change, some of the conditions that (we might know from other findings) are preventing the intervention from working, ultimately providing a reasonable answer to the "**will the intervention make a difference elsewhere/in the future**" question.

Process-like or "causal chain" mechanisms

Process-like or "causal chain" mechanisms are concatenations of events (sometimes referred to as intermediate outcomes) where each event follows from a previous one, provided some conditions are met. The Theory of Change is described as a typical sequence where resources leads to activities which lead to some kind of behavioural change which in turn leads to the desired outcomes (see Figure A2).

Figure A2: **Causal chain of a typical "contribution story"**


Livelihood changes
Livelihood change assumptions and risks
Other supporting factors
Direct benefits
Direct benefits assumptions and risks
Other supporting factors
Behaviour changes
Behaviour change assumptions and risks
Other supporting factors
External influences
Capacity changes in knowledge, attitudes, aspirations, skills, opportunities and incentives
Capacity change assumptions and risks
Other supporting factors
Reach and reaction
Reach assumptions and risks
Outputs
Activities

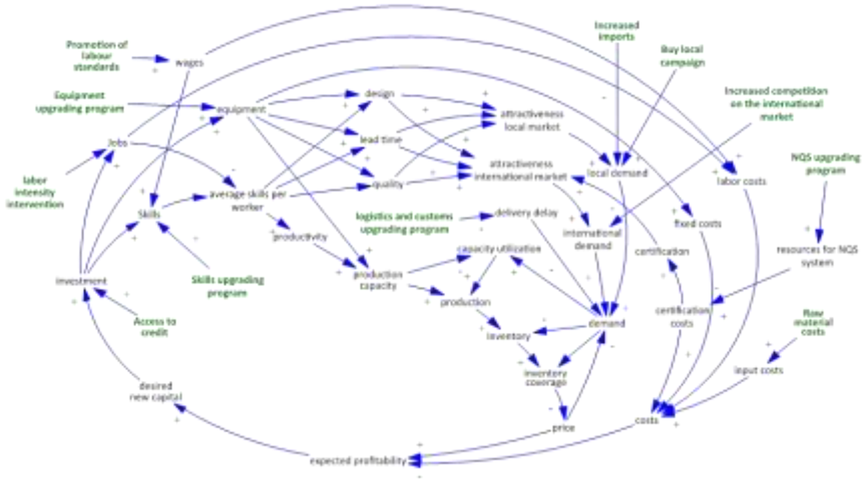
Source: (Befani & Mayne, 2014)

In our food aid example, the combinations can be seen, to some extent, as explanations of the outcomes; or as configurations of explanatory factors for the outcomes. However, as explanations, they are not fully satisfactory because they don't provide any (fine-grained) glimpse into how and why different combinations work. It is only by observing the process of food aid delivery that we can explain, for example, why preservation facilities and unfriendly weather are found in the same combination, and described the role of those facilities; or why community dynamics are so important in all contexts.

Complicated and complex systems

If we gradually move the magnifying lens closer to single areas seeking a better understanding of the details of single cases, a mostly linear causal chain might not describe our observations correctly: we might observe additional layers of complexity. In some areas nutrition-related outcomes might be worsened by the prevalence of medical conditions like HIV and bacterial infections; in others they might be improved by pre-existing structures like community health centres that temporarily take charge of food distribution. These factors might be linked not just to the outcomes, but also to each other, and in complex and unexpected ways: in some areas new health centres might have been recently built following concern related to a high prevalence of certain medical conditions; in others the health centres might have been there for a while and reduced the burden of these conditions. In other words, in-depth study of single areas might surface a complex web of factors that affect the outcomes in a relatively indirect way. These factors might also influence each other, creating reinforcing causal loops: for example, assuming the number of health centres is the same, HIV and infections might increase their workload, which might be too high to focus properly on food distribution; in turn, the lack of food created by improper distribution worsens the consequences of HIV and infections. Any specific area we focus on could be represented as a "system", with multiple arrows and loops connecting the factors involved. Figure A3 illustrates how a system can be represented (in this case the leather shoe sector in Ethiopia).

Figure A3: Representation of a system



Source: Derwisch & Löwe (2015)

No matter how we represent the detailed processes or mechanisms explaining the outcome, we still adopt a generative causation framework describing how and why the “cause”, either a “causal chain” or a complex combination of multiple causes, produces the “effect”.

Table A8 summarises the properties of the causal frameworks underpinning impact evaluation methods we have covered above. The three main categories are single-cause, multiple-cause and generative, with the MoCV that can be either single-cause or multiple-cause. The MoD and MoA are only single-cause, while MCC is only multiple-cause. The overarching evaluation question “did the intervention make a difference” is articulated in a different specific question for each framework. The MoCV and MCC perform relatively well on both internal and external validity, while the best framework for construct validity is generative causation. The frameworks have different requirements and offer different opportunities.

Table A8: An overview of causal frameworks and their properties

Causal Framework		Evaluation Question: Did it make a difference?	Validity	Requirements	Opportunities
Single-Cause	MoD	How much of a difference did it make?	Internal	Case + Control	Net effect, the average difference
	MoA	Did it consistently make the same difference?	External	Diversity of cases	Consistency of causal relation
	MoCV	How much of a difference did it make?	Internal External	Large n, Quantitative data	Net effect, the average difference
Multiple-Cause	MoCV	How much of a difference did the intervention and other factors make?	Internal External	Large n, Quantitative data	Net effect, the average difference
	MCC	For whom/under what circumstances did it make a difference? What other factors made a difference? Which factors are necessary and/or sufficient?	Internal External	Comparable cases	Enabling/necessary conditions; multiple causal packages which are equally sufficient (equifinality); Local, contextual difference (INUS causes).
Generative		How/Why did it make a difference?	Construct	One case for in-depth study	Fined grained explanation

ANNEX B: Differences between QCA and regression analysis

QCA is often compared with regression analysis because both methods attempt to establish an association between a number of causal factors and an outcome (see, for example, Vis, 2012). In regression analysis, these factors are referred to as "variables" because they usually can take any value in an interval of real numbers; while in QCA they are referred to as "conditions" because they denote presence or absence of a certain quality or state in a given case[60]. However, despite some apparent similarities, the differences between QCA and regression are numerous and substantial (Thiem, Baumgartner, & Bol, 2015).

First of all, in regression analysis, association is intended as "concomitant variation" between a single variable and an outcome (see Annex A): if the value of the outcome tends to increase (or decrease) with the value of the independent variable, we observe a positive (negative) correlation between the variable and the outcome. By contrast, in QCA, association is intended as a set relation: union, intersection, inclusion or negation. If the outcome is "included" in the condition, or logically implies the condition, the association will be one of "necessity"; conversely, if the condition is "included" in the outcome and logically implies the outcome, the association will be one of "sufficiency". While correlation is symmetrical (if x is correlated with y, then y is correlated with x), association in QCA isn't: conditions can be necessary but not sufficient, or sufficient but not necessary. This property is also referred to as "causal asymmetry".

The second important difference between QCA and the most common type of regression analysis (that doesn't take interaction effects into account) is that, while in regression analyses associations are established between the outcome and one variable at a time, QCA considers cases "as wholes" or "packages", analysing associations between *combinations* of conditions and the outcome; which makes the emergence of contextual influence easier to spot. While in regression analysis the causal power of one variable, identified by the regression coefficient, is valid "on average" across the entire sample, in

[60] in fsQCA, a value in a 4-point or 6-point scale denotes the case's degree of membership to the set defined by that quality or state

QCA the causal power of one condition is dependent on which other conditions it is combined with. In other words, the association is "conjunctural" (hence the word "conjunctural" in multiple-conjunctural causation, see Annex A), or dependent on a specific context or setting[61].

Thirdly, while regression analysis aims at the identification of *the one single model* that fits the data *best*, QCA allows the identification of *multiple, equally important pathways* to the outcome; for example, two or more conditions that can be equally necessary for an outcome; or two or more combinations of conditions that are equally sufficient (hence the term "multiple" in multiple-conjunctural causality).

Strategies adopted by young farmers to withstand demand decrease

The example that follows should clarify the differences between QCA and regression analysis with no interaction terms. It illustrates the measures adopted by 20 young farmers (Befani B. , 2013) to withstand the decrease of demand in times of crisis. The first consists in decreasing the prices they charge customers for the same products, trying to boost demand for their products; the second in setting up promotional offers, operating discounts for buyers of large quantities of products; the third attempts to make the production process more efficient, sometimes sacrificing product quality, in order to decrease costs and increase revenue for the same level of demand and product price. The outcome indicates whether the farmer has successfully survived the crisis.

Table B1 shows which measures young farmers adopt and whether they successfully withstand the crisis or not. Most farmers (15 out of 20) will adopt some measures but not all. Three farmers (J K and L) adopt all three while two farmers adopt none (S and T).

[61] Regression models can include interaction terms, which explain the residual variation not explained by the coefficients of single terms. However, the regression findings with interaction terms in this case are still less satisfactory than the QCA findings. See the end of the annex for more details.

Table B1: Success of young farmers in relation to the strategies they adopt in times of crisis

Case ID	Measures adopted to withstand decrease of demand in times of crisis			Success?	Risk level of the combination
	Decrease prices (PRICE)	Set up Promotional offers (PROMO)	Decrease costs (COSTS)		
A	1	0	1	1	Low to Medium
B	1	0	1	1	Low to Medium
C	0	1	0	1	Low to Medium
D	0	1	1	1	Low to Medium
E	0	1	1	1	Low to Medium
F	0	1	1	1	Low to Medium
G	0	0	1	1	Low to Medium
H	0	0	1	1	Low to Medium
I	0	0	1	1	Low to Medium
J	1	1	1	0	High
K	1	1	1	0	High
L	1	1	1	0	High
M	0	0	1	0	Low to Medium
N	0	0	1	0	Low to Medium
O	0	0	1	0	Low to Medium
P	0	0	1	0	Low to Medium
Q	0	1	1	0	Low to Medium
R	0	1	1	0	Low to Medium
S	0	0	0	0	High
T	0	0	0	0	High
general average	0.25	0.45	0.85	0.45	
correlation with outcome	-0.058	-0.01	0.099		
Averages for the successful	0.222	0.444	0.889	1	
Averages for the unsuccessful	0.273	0.455	0.818	0	
Average difference	-0.051	-0.01	0.071	1	
Regression coefficients[a]	-0.099	-0.007	0.167		

Note: [a] Model with no interaction effects and constant value of 0.336: the model was run by Lucie Moore.

The general evaluation question is "which measures make farmers successful"? This question can be answered using both a correlational approach and a configurational one, but the two approaches return very different types of information about the farmers, despite being applied on the same dataset.

The correlational approach will take one variable/column at a time and check if success or lack thereof tends to be associated with each column (adoption or lack thereof of each measure). Let's start with the first column/variable "decrease prices". This measure is not very popular across the sample, and is adopted by only 25% of the farmers (5 farmers). If the measure is correlated with success, we would expect the cases where it is adopted to be successful, and the cases where it is not adopted to be unsuccessful. However, the % of farmers adopting the measure is very similar in the two groups: 22% of the successful and 27% of the unsuccessful. There seems to be a slight negative correlation between adopting the measure and success; indeed, when we actually measure it, we discover that the correlation is very low (-0.058). As expected, the regression coefficient is also very low (-0.099).

Now let's move to the second column and look at setting up promotional offers. This measure is more popular than the previous one and is adopted by 45% (9) of the farmers. If this variable is correlated with success, we expect the farmers adopting it to be successful and the farmers not adopting it to be the opposite. But the difference between the two groups is even lower than for the previous strategy: 44% and 46%. Indeed, the correlation between this strategy and success is even weaker (-0.01), together with its regression coefficient (0.007).

The third measure (decreasing costs) is more popular (adopted by 85%, or 17 farmers), and more frequently adopted by the successful (89%) than by the others (82%). We would thus expect a small positive correlation between adopting the third measure and being successful, which we calculate at 0.071 (also evident in the slightly higher regression coefficient of 0.167).

The findings from the correlational analysis allow us to conclude that *no measure is particularly crucial to success* and that there is actually a very small correlation between any of the three and success. In other words, following these findings, *we have no recommendation for the farmers and the policy makers*.

Fortunately, the configurational approach tells a very different story. Instead of taking one column/variable at a time, it cuts across the columns and takes the case, as a whole, as the main unit of analysis. What is of interest here is how each farmer combines the different measures: his portfolio, so to speak. Which measures does s/he choose?

The sufficiency analysis for the successful cases reveals a mixed picture where 4 different combinations of measures, denoting 4 different strategies, appear successful (see Table B2). In the first strategy, the farmer takes some risks, cutting production costs and slightly decreasing product quality; however, as quality does not decrease substantially, demand holds. Therefore, when she decides to decrease prices, customers see the product as value for money, which increases demand substantially. In the second strategy, demand stagnation is solved by the launch of mass promotions, which allows stocks to be cleared and the making of a small profit on the product, while quality and regular prices do not change. The third strategy is similar to the first: changes in the production process make it more efficient and allow savings, with a small decrease in product quality. The ensuing demand decrease is overturned, not by price decrease, but by the launch of aggressive promotional offers, resulting in a demand boost overall. Finally, in the fourth strategy, small savings are made possible by small changes in the production process which leave quality and demand unaffected; the same quantity of the product is sold, at the same price; but revenue increases thanks to the efficiency gains in production.

Note that all these strategies, represented by combinations of measures, are somehow low to medium risk, that is, only some changes are made by the farmers, never all or none. All change carries risk, and the wisest farmers seem to take some risks, but not excessive ones.

This interpretation is confirmed by the sufficiency analysis of the unsuccessful cases, where two completely different sufficient pathways equally guarantee lack of success: in the first, all the measures are adopted, and in the second none. In other words, either making many changes, or making none, is equally ineffective. In the first strategy, the farmer changes the production process substantially, which decreases costs but also alters the quality of the product. As a consequence, she cannot sell the same quantity at the same price and is forced to reduce regular prices. The farmer is now entering uncharted territory because she never sold such a cheap variant of the product, and in spite of these two changes demand does not pick up enough to increase revenue. The response of the farmer is even more risky: she keeps lowering the bar, launching promotional offers that won't bring enough revenue to produce a high-quality product again. The farm business is now caught in a downward spiral and loses its identity, eventually terminating sales (cases J, K and L).

The second pathway describes a situation where the farmer is overly optimistic that demand will soon go back to the previous level and does not make any changes, until production costs are no longer covered and the farmer is forced to discontinue the product. In this case the farmer didn't make any change at all, which in such a crisis situation is shown to be as risky as making too many changes. In other words, the two opposite strategies are both high risk, although for opposite reasons.

Table B2: Findings from the Sufficiency Analysis

Combination of Strategies	Success?	Number of cases covered	% of cases covered
PRICE*promo*COSTS	Y	2	10%
price*PROMO*costs	Y	1	5%
price*PROMO*COSTS	Y	3	15%
price*promo*COSTS	Y	3	15%
PRICE*PROMO*COSTS	N	3	15%
price*promo*costs	N	2	10%

The findings from the configurational analysis are very informative. We learn that adopting some measures turns out to be a successful strategy, in contrast to adopting too many or none: the latter two are opposite strategies but in times of crisis are both high risk, and equally lead to failure. Our recommendation to young farmers, and to policy makers, is to try and change something: either in the production process, in the regular prices, or in temporary promotions; but cautiously, without trying to do too much at the same time, and being careful to avoid implementing all three measures. In addition, we would recommend that the "no action at all" strategy is also avoided.

It should be clearer at this point what the added value of the configurational approach is for evaluation, and the kind of findings it enables. We do not learn whether each measure, taken alone, by itself, is inherently good or bad for success. What we learn is which combinations of measures are so: the low to medium risk strategies, including some but not all measures, are recommended; and the high risk ones, including either all measures or none, are not recommended. Secondly, we learn that there is no single best strategy for success, and no single worst strategy for failure. Different, sometimes even very different to the point of being opposite, strategies, are equally risky for different reasons and equally lead to the same outcome.

Triangulating QCA findings with the interaction effects of regression analysis

The dataset above was analysed[62] with STATA in order to see if a regression model with interaction effects could capture the negative synergies apparently emerging when taking all of the measures or none. In order to keep things simple and test the knowledge we had already acquired on the dataset, priority was given to a model including the three single variables corresponding to the measures and an interaction effect among all three measures. The findings are reported in Table B3.

Table B3: Results of the regression analysis with a triple interaction effect

Variable	Regression Coefficient	Standard Error	t	P>|t|	95% Confidence Interval	
Var1	0.6356589	0.3720715	1.71	0.108	-0.1573927	1.428711
Var2	0.3255814	0.2472856	1.32	0.208	-0.2014954	0.8526582
Var3	0.1395349	0.3028618	0.46	0.652	-0.5059997	0.7850695
Var123	-1.325581	0.4937118	-2.68	0.017	-2.377903	-0.2732595
constant	0.2248062	0.2825508	0.8	0.439	-0.3774365	0.8270489

The regression coefficients of the single variables are much higher than in the model without interaction terms (see Table B1), to compensate a strongly negative regression coefficient of the triple interaction term: -1.326. This model is compatible with one extreme result we observe in the dataset: when a farmer adopts all three measures, 1.326 needs to be subtracted from the sum of the independent effects of the single measures to obtain the outcome. The strongly negative interaction effect signals a strongly negative influence of adopting all three measures at the same time on success.

[62] The author would like to thank Lucie Moore, who conducted the analysis and helped with interpreting the findings.

This finding would support the recommendation to not adopt all three measures at the same time.

The other recommendation, to avoid inaction, would also be supported by the low value of the constant value: 0.225. This means that when no measure is adopted the outcome is predicted to be not far from zero, or 0.225. When a mix of measures are adopted, the interaction effect is reduced to zero and the sum of the regression coefficients predicts the outcome.

The regression analysis with the triple interaction effect thus confirms the patterns spotted with QCA, adding meaningful information on the average effects of the single conditions. However, this type of regression was used to confirm specific hypotheses on patterns formulated after applying QCA: spotting these patterns in the first place would have been much more difficult with regression analysis, given the residual nature of regression terms and the high number of terms required to check all possible interactions of 3 or more variables. A standard initial model with three interactions of two terms in addition to the triple interaction terms provided findings which were much more difficult to interpret and seemed poorly aligned with the QCA findings.

With models of 3 or more conditions, regression analysis with interaction terms is recommended as a means to triangulate QCA findings and obtain sophisticated information on the contribution of single factors. Due to the incremental nature of the effects identified by it, it is not recommended as the first choice for pattern-spotting.

ANNEX C: Case material used in this report

Over 10 studies or evaluations using QCA are mentioned in the report: either fictitious, stylised, or real-life examples. Four evaluations are more heavily referenced than others: in order to allow other evaluators to check the validity of the findings reported in the main text, additional material about these four evaluations is included in this annex.

Making All Voices Count (MAVC)

This real-life study, which is referred throughout the text as "the MAVC study", sought to understand which conditions facilitated the achievement of 3 outcomes related to the performance of mobile phones in affecting repairs of broken water points. While the latter was the ultimate outcome, the two intermediate outcomes of mobile phone use (to report broken water points) and processing of data collected with mobile phones were also considered (Welle, Williams, Pearce, & Befani, 2015). Tables C1, C2, and C3 report data for the models explaining, respectively, outcome 3, outcome 2 and outcome 1; while Table C4 summarises and compares outcome data.

Table C1: Dataset for the model explaining Outcome 3

Outcome 3: Rural water points are repaired based on ICT reports and processing

	Funds are sufficient for carrying out the repair (FUNDSUF)	O&M responsibilities are clear (RESPCL)	Spare parts are available for the repair (SPAREP)	A mechanic is available to carry out the repairs (MECHAV)	Accountability mechanisms in place to ensure that ICT reports are acted on (ACCMEC)	The ICT initiative supports existing sector responsibilities (EXRESP)	Outcome (REPAIR)
Smart Handpumps Kenya (SHP)	1	1	1	1	1	1	1
M4W Uganda (M4W)	0	1	0	0	0	1	0
Maji Matone Tanzania (MM)	0	0	1	0	1	1	1
Maji Voice Nairobi (MV)	1	1	1	1	1	1	1
Next Drop Bangalore (ND)	1	1	1	1	1	1	1
Hum Sensor Web Zanzibar (HSW)	0	1	1	0	0	1	0

Table C2: Dataset for the model explaining Outcome 2

Outcome 2: Local government authority (national sector government, if relevant) or service provider process and follow up on ICT-based reports

	2.1.1	2.1.2	2.1.3	2.1.4	2.2.1	2.2.2	2.2.3	
	GSM reception	Availability of computers and electricity	Access to necessary software to store and process data	Access to ICT-back up support	HR and knowledge to process ICT reports	Clarity of procedures for follow-up on ICT reports	Operational costs largely be met by government/service provider	Outcome
Smart Handpumps Kenya (Oxford University)	1	1	1	1	1	1	1	1
M4W Uganda	0	1	1	1	0	1	0	1
Maji Matone Tanzania	1	1	1	1	1	1	0	0
Maji Voice Kenya	1	1	1	1	1	1	1	1
SIBS Timor Leste	0	1	1	1	1	1	1	1
Re-imagining Reporting, Bolivia	1	1	1	1	1	1	0	0
Next Drop Bangalore, India	1	1	1	1	1	1	1	1
Human Sensor Web Zanzibar	1	1	1	1	1	0	0	0

Table C3: Dataset for the model explaining Outcome 1

Outcome 1: Users or their representatives, including government staff, directly or indirectly, use ICTs in the way specified by the initiative to report rural water supply functionality to the local government authority or relevant stakeholder; this could be either through ad hoc crowdsourcing or through government-led, regular updating mechanisms.

	1.1.1	1.1.2	1.1.3	1.2.1	1.2.2	1.2.3	1.2.4	1.2.5	1.2.6	
	GSM Reception is reliable	ICT devices can be charged	Access to the ICT device	Data collected periodically / related to specific incidents	Reporting requires human interaction / is automatic	Reports crowd-sourcing or government / service provider-led	Preference for using the ICT mechanism	Reporting costs are not a problem	Sufficient information and knowledge for reporting	Outcome
Smart Handpumps Kenya	1	1	1	0	1	1	1	1	1	1
M4W Uganda	0	1	0	1	0	0	0	0	0	1
Maji Matone	0	0	0	1	0	0	0	1	1	0

Tanzania										
Maji Voice Kenya	1	1	1	1	0	0	1	1	1	1
SIBS Timor Leste	1	1	0	0	0	1	1	1	1	1
Re-imagining Reporting Bolivia	1	1	1	0	0	1	1	1	1	1
Next Drop Bangalore, India	1	1	1	1	0	0	1	1	0	1
Human Sensor Web Zanzibar	1	1	0	1	0	0	0	1	0	0

Table C4: Dataset of Outcome Data

	O1USEICT	O2PROCESS	O3REPAIR
Smart Handpumps Kenya	1	1	1
M4W Uganda	1	1	0
Maji Matone Tanzania	0	0	1
Maji Voice Nairobi	1	1	1
SIBS (AusAid)	1	1	-
Re-imagining Reporting	1	0	-
Next Drop	1	1	1
HSW	0	0	0

Evidence-Based Policy for Access to the Health System for the Poor (ATHSP)

This fictitious evaluation, which is referred throughout the text as “the Evidence-Based Policy Evaluation”, sought to understand which conditions led policy makers to ground their decisions on evidence when legislating on access to the health system for the poor. The main dataset is reported in Table C5, while Table C6 describes the conditions included in the models.

Table C5: Main dataset of ATHSP evaluation

Country	INFO	CHAMP	PRES	ALIG	DATA	PROB	SOL	POL	EBPM
Vietnam	0	0	1	1	1	1	0	0	0
Kenya	0	0	1	0	1	1	0	0	0
Zimbabwe	0	1	0	0	1	1	1	0	1
Bolivia	0	1	0	0	1	1	1	1	1
Indonesia	1	1	0	0	1	1	1	0	1
Ethiopia	1	1	0	1	1	1	1	1	1
Laos	1	1	1	0	0	1	1	0	1
Tajikstan	1	0	0	1	1	1	1	1	1

Table C6: Description of the conditions analysed in the ATHSP evaluation

EBPM	Evidence-based policies around ATHSP are put in place
PROB	ATHSP (access to the health system for the poor) is unanimously identified as a problem by stakeholders
SOL	Feasible and acceptable solutions to ATHSP are identified in the course of an active multi-stakeholder dialogue
POL	The political context is favourable to addressing ATHSP issues
PRES	Focused events are organised and other forms of public pressure on ATHSP are put in place
ALIG	Interest groups are generally aligned on policy priorities
DATA	Groups are able to access credible data on ATHSP
INFO	Information-sharing agreements or protocols exist in the multi-stakeholder ATHSP community
CHAMP	Skilled policy entrepreneurs or „champions“ are active in the ATHSP sector

Note that this fictitious example is freely inspired by the evaluation of the Alliance for Transparency in Access to Medicines (MeTA) (Stedman-Bryce, Schatz, Hodgkin, & Balogun, 2016).

Evaluation of Gender-Sensitive Budget Support to Education

This real-life evaluation, which is referred throughout the text as the Budget Support Evaluation, sought to understand which of two policy instruments worked best in improving primary school enrolment of girls (Holvoet & Inberg, 2013). Table C7 reports the main dataset, while Table C8 illustrates the description of the conditions.

Table C7: Dataset of the Budget Support Evaluation example

Country	PAF	GWG	AID	EDU	OUT
Ethiopia	1	1	1	1	1
Mozambique	1	1	1	1	1
Tanzania	1	1	1	1	1
Burkina Faso	1	1	1	0	1
Mali	1	1	1	0	1
Ghana	1	1	0	1	1
Senegal	1	1	0	1	1
Malawi	0	1	1	1	1
Niger	1	0	1	0	1

Zambia	1	0	1	1	0
Gambia	0	0	1	1	0
Kenya	0	0	0	1	0
Lesotho	0	0	0	1	0
Botswana	0	0	0	0	0

Table C8: description of the conditions analysed in the Budget Support Evaluation example

PAF	Presence of sex-disaggregated indicators and targets in Performance Assessment Framework
GWG	Presence of gender working groups
AID	Total aid to basic education per primary school-age child
EDU	Presence of free primary education
OUT	Increase in female net enrolment ratio

Impact Evaluation of GEF/UNDP Support to Protected Areas and Protected Area Systems

The QCA component of this real-life evaluation, which is referred throughout the text as the GEF/UNDP Biodiversity Evaluation, sought to understand the conditions which facilitated the creation of functional protected area systems at the national level and the achievement of a series of objectives, including in relation to biodiversity, at the more local level of the Protected Areas. Table C9, C10 and C11 describe the datasets for the analysis of the biodiversity outcome: they include capacity-, community- and context-related conditions and cover 30 cases/protected areas. Table C12 is the dataset used for the analysis of the functional PA system at the national level and covers 8 national protected area systems. These are followed by Tables C13 and C14, describing the conditions and outcomes used in the analyses. (The Global Environment Facility Independent Evaluation Office GEF IEO, 2015).

Table C9: Dataset for the capacity-related conditions

PA	CAstaff	CAres	CAearn	CAmande	CAhuwiconf	CAleader	CArepcorr	CAbound	CAlocauth	CAcsogov	CAcorpgov	CAotherextsupp	BIOtrend
OA	1	0	1	0	1	1	1	1	1	1	1	1	1
RA	1	0	1	1	1	1	1	1	0	1	1	1	0
TO	1	0	1	0	1	1	1	1	1	1	1	1	1
DB	1	0	1	0	1	1	1	1	0	0	1	1	1
HT	1	1	1	1	1	1	0	0	1	1	0	1	1
WW1	1	1	1	1	1	1	0	0	1	1	0	1	1
WW2	1	1	1	1	1	1	0	0	1	1	0	1	1
MU	1	1	1	1	1	1	0	0	1	1	1	1	1
DH	1	1	1	1	1	1	0	0	1	1	1	1	1
DQ	1	1	1	1	1	1	0	1	0	1	1	1	1
LI	1	1	1	1	1	1	0	1	0	1	0	1	1
RT	0	0	1	0	1	1	1	0	0	1	0	0	1
UA	1	1	0	1	0	1	1	1	1	1	0	0	1
NU	1	1	0	0	0	0	1	0	0	0	0	0	0
AK	1	1	0	1	0	0	1	1	1	1	0	1	1
BO	1	1	0	1	0	0	0	1	1	1	0	1	1
SA	1	1	0	0	0	0	0	0	0	0	0	0	0
ZA	1	0	0	1	0	1	0	1	1	1	0	1	1
HU	1	0	1	1	0	1	0	1	1	1	0	1	1
AO	0	0	0	0	0	0	0	0	1	0	0	0	1
IA	1	0	1	1	1	1	0	0	1	0	0	1	1
EO	1	0	1	1	0	1	0	1	1	1	0	1	1
QG	1	0	1	1	0	1	0	1	1	1	0	1	1
MA	1	0	0	0	0	1	0	0	1	0	0	0	1
KA2	1	1	1	1	0	1	1	1	1	1	1	1	1
AI	1	1	1	1	1	1	0	1	1	1	1	1	1
AU	1	0	1	1	1	1	0	0	1	0	0	1	1
RO	1	1	1	1	1	1	0	1	1	1	1	1	1
NA	1	1	1	1	1	1	0	1	1	1	1	1	1
PZ	1	1	1	1	0	1	0	1	1	1	1	1	1

Table C10: Dataset for the community-related conditions

PA	COMinf	COMsust	COMtrconfres	COMconcrben	COMprovinf	COMcons	COMpart	BIOtrend
OA	1	1	1	1	1	1	1	1
RA	0	0	0	0	0	1	0	0
TO	1	1	1	1	1	1	1	1
DB	0	0	0	0	1	1	1	1
HT	1	1	0	1	1	1	0	1
WW1	1	1	1	1	1	1	1	1
WW2	1	1	1	1	1	1	0	1
MU	1	1	1	1	1	1	1	1
DH	0	1	0	1	1	1	0	1
DQ	1	1	0	1	1	1	0	1
LI	1	1	0	1	1	1	0	1
RT	1	1	0	1	1	0	0	1
UA	1	1	1	1	1	1	0	1
NU	0	0	0	0	0	0	0	0
AK	1	0	1	0	1	1	0	1
BO	1	0	1	0	1	1	0	1
SA	0	0	0	1	1	0	0	0
ZA	1	1	0	1	1	1	0	1
HU	1	1	0	1	1	0	1	1
AO	0	0	0	0	1	0	0	1
IA	1	1	0	0	1	1	0	1
EO	0	0	0	1	1	0	1	1
QG	1	0	0	1	1	1	1	1
MA	0	0	0	0	1	1	1	1
KA2	1	1	1	1	1	1	1	1
AI	1	1	1	1	1	1	1	1
AU	1	1	0	1	1	1	1	1
RO	1	1	1	1	1	1	1	1
NA	1	1	1	1	1	1	1	1
PZ	1	1	1	1	1	1	1	1

Table C11: Dataset for the context-related conditions + Outcome (BIOtrend)

PA	CXTdevpres	CXTaccpa	CXTmand	CXTtourcul	CXTindpop	CXTcommit	CXTpolconf	CXTthreatecval	CXTtenure	BIOtrend
OA	1	1	1	1	0	1	0	1	0	1
RA	1	1	1	1	0	1	1	1	0	0
TO	1	1	0	1	0	1	0	1	0	1
DB	1	1	1	1	0	1	0	1	0	1
HT	1	1	0	1	1	1	0	1	1	1
WW1	1	1	1	1	1	1	0	1	1	1
WW2	0	1	1	1	0	1	0	1	1	1
MU	1	1	1	1	0	1	0	1	0	1
DH	1	0	0	1	0	1	0	1	0	1
DQ	1	0	1	1	1	1	0	1	1	1
LI	1	1	1	1	0	0	0	1	0	1
RT	1	1	1	0	0	0	0	1	0	1
UA	0	0	0	0	0	0	0	1	0	1
NU	1	1	0	0	0	0	0	1	0	0
AK	1	1	0	0	1	0	0	1	1	1
BO	0	0	0	0	1	0	0	1	1	1
SA	0	0	0	1	1	0	0	1	0	0
ZA	1	1	1	1	1	1	0	1	1	1
HU	1	1	1	1	1	1	0	0	1	1
AO	0	1	0	0	1	1	0	0	0	1
IA	1	1	1	1	1	1	0	1	1	1
EO	0	1	1	1	0	1	1	1	1	1
QG	1	1	1	1	0	0	0	1	1	1
MA	1	0	1	0	0	1	0	1	1	1
KA2	1	1	1	1	1	1	1	1	1	1
AI	1	1	1	1	1	1	0	1	1	1
AU	1	1	1	1	1	1	1	1	1	1
RO	1	1	1	1	1	1	1	1	1	1
NA	1	1	1	1	1	1	0	1	1	1
PZ	1	1	1	1	0	1	0	1	0	1

Table C12: Dataset for the PA System analysis

Country	CROSSTRUST	NATGOV	TRANSPFIN	ADQFIN	ADQLEG	TRANSPDEC	CLRMAND	IMPLMAND	COLLAB	CHAMP	SOCATT	CORR	OUT
AE	1	1	1	0	1	1	1	1	1	1	1	1	1
BA	0	1	1	1	1	1	1	0	1	1	1	0	1
DG1	1	0	1	1	1	1	1	0	1	1	0	0	1
MO	1	1	1	0	1	1	1	1	1	0	1	0	1
CE	1	1	1	1	1	1	1	0	1	1	1	0	1
NN	0	1	0	0	1	0	0	0	0	0	0	1	0
NI	0	0	1	0	1	1	0	0	1	0	0	0	0
DG2	0	1	0	0	0	0	1	0	1	0	0	1	0

Table C13: Meaning of presence of the conditions for the PA System analysis

OUT	The PA System is functional: has sufficient resources, adequate staff, useful operational management information system, resilience to catastrophes and shocks
CROSSTRUST	Cross-subsidization across PA system or Trust Fund
NATGOV	National government budget for the PA system
TRANSPFIN	Transparency of financial flows and management
ADQFIN	Adequate financial resources for the PA system
ADQLEG	Adequate legal framework for conservation
TRANSPDEC	Transparency of decision-making procedures
CLRMAND	Clear mandates among institutions (e.g. no overlaps)
IMPLMAND	Coordinated implementation of mandates across government sectors/ scales on natural resources use and conservation (inc. governance structure)
COLLAB	CSO-Corporate sector-Government collaboration within government framework
CHAMP	Presence of champions
SOCATT	Positive societal attitudes towards environment and conservation (national level)
CORR	Reported corruption in government concerning PA system

Table C14: Meaning of presence of the conditions for the PA System analysis

BIOtrend	Decrease in trends in incidents of illegal activities
CAstaff	Professional and trained and dedicated PA staff
CAres	Sufficient operative resources
CAearn	Earning capacity
CAmande	Management Monitoring and evaluation
CAhuwiconf	Human-wildlife conflicts
CAleader	Good leadership

CArepcorr	Corruption in PA reported
CAbound	Clear boundaries
CAlocauth	Effective relation with local authorities
CAcsogov	CSO-Government Partnership within government framework
CAcorpgov	Government-Corporate Sector Partnership within government framework
CAotherextsupp	Other external support e.g. donors
COMinf	Well informed communities
COMsust	Sustainable economic activities
COMtrconfres	Transparent mechanism of conflict resolution
COMconcrben	Concrete Benefits perceived by communities (including projects and financial support)
COMprovinf	Information is provided to the community
COMcons	The community is consulted in decision making
COMpart	The community actively participates in decision-making/ planning/ implimentation
CXTdevpres	Existing Development pressures
CXTaccpa	Easy Access to PA
CXTmand	Unified and clear mandates among institutions (e.g. no overlaps)
CXTtourcul	Tourism asset and cultural values
CXTindpop	Existence of indigenous populations
CXTcommit	Country-International commitments at PA level
CXTpolconf	Political Conflicts (e.g. wars)
CXTthreatecval	Presence of threatened species or High economic value of PA resources
CXTtenure	Tenure issues