cecan Centre for the Evaluation of Complexity Across the Nexus

Choosing Appropriate Evaluation Methods

A Tool for Assessment & Selection

(version 2)

Barbara Befani

November 2020

Acknowledgements

This report[1] accompanies the second/updated edition of the Choosing Appropriate Evaluation Methods tool, originally developed by Barbara Befani with Michael O'Donnell and published on the Bond website https://www.bond.org.uk/resources/evaluation-methods-tool. This update has been funded by CECAN.

The original work, which was an 11-method version of the tool and a shorter report, benefited from the feedback and support of a range of evaluation specialists and prospective users: Hlengani Bhebhe, Zack Brisson, Laura Camfield, Rohati Champan, James Copestake, Rick Davies, Thomas Delahais, Maren Duvendack, Jenny Eddis, Simon Hearn, Sarah Henon, Richard Hummelbrunner, Sofia Jadavji, Matthew Juden, Sebastian Lemire, Sophie Mareschal, Bruno Marchal, Edoardo Masset, Daniel Phillips, Richard Ponsford, Melanie Punton, Saltanat Rasulova, Fiona Remnant, Patricia Rogers, Stanford Senzere, Ethel Sibanda, Tremayne Stanton-Kennedy, Claire Thomas, Alix Tiernan, Giel Ton, Jacques Toulemonde, Jos Vaessen, Marie Weiller, Gill Westhorp, and Bob Williams.

This is the first update and features four new methods which have been added to the previous eleven with the support and contributions of Nigel Gilbert, Stephen Morris, Peter Barbrook-Johnson, Stuart Astill, and Simon Henderson. This expansion has led to the inclusion of several new questions on the second and third worksheet of the Excel tool, which has required the additional input of all the methods experts (mentioned in Annex 1), including the ones who had contributed to the previous version; I would like to thank them all for their continued support of the tool. This update would not have been possible without the intellectual and financial support of CECAN: I am indebted to its director, Nigel Gilbert, and to several members and associates of CECAN who have provided their precious ideas, support, and engagement in various forms throughout this journey: Corinna Elsenbroich, Adam Hejnowicz, Dione Hills, Frances Rowe, Emma Uprichard, Helen Wilkinson.

Last but not least, thanks to all the readers and users who have engaged with the first version: you are many and I am humbled by your interest in the tool. I hope you will continue to enjoy the result of this complex collaboration and continue to send me comments, questions, feedback and suggestions for improvement.

Like the original version, this update is covered by a Creative Commons Attribution-NonCommercial-ShareAlike licence (CC BY-NC-SA 4.0): https://creativecommons.org/licenses/by-nc-sa/4.0/.

Cover image: © blackdovfx/Canva Pro Stock Media, canva.com

Contents

1. Introduction: Why is Appropriateness an Issue? 2
2. Choice and design triangles 6
3. Dimensions of appropriateness 9
4. How the tool works 10
5. Use and limitations 16
6. Annexes 19
References 40

1 Barbara Befani has authored the main text while Annex 3 has been co-written with the methods experts mentioned in Annex 1.

1 | Introduction: Why is Appropriateness an Issue?

In the last few years, a "movement" to explore, develop and test a range of rigorous alternatives to counterfactual methods in impact evaluation[2] has taken an increasingly defined and consistent shape (White & Phillips, 2012; Stern, 2015; Stern, et al., 2012; Befani, Ramalingam, & Stern, 2015; Befani, Barnett, & Stern, 2014; Stern et al. 2012, Stern 2015, Befani et al. 2014 & 2015, etc.)

As a principle, it is now largely accepted that a wide range of methods can be appropriate, under different circumstances, to evaluate the impact of policies and programmes. However, while apparently solving the problem of scarcity of options, this expansion has created a selection problem: as possibilities have expanded, it has become challenging to select or even shortlist methods[3] for particular evaluations.

While unsuitable and unfeasible under many real world circumstances, the rigid "gold standard" hierarchy—which placed experimental and quasi-experimental evidence at the top and qualitative evidence at the bottom—had[4] the (illusory, some might say) benefit of being simple and of leading to clear choices. Now that some policy fields and institutions have expanded their horizons, recognising that the "best" method or combination of methods is dependent on the evaluation questions, intended uses and attributes of the intervention and evaluation process, we are struggling to make and justify choices.

The tool presented in this paper is an attempt to improve this situation and support the process of methodological selection, by helping users make an informed and reasoned choice of one or more methods for a specific evaluation. Its aim is not to necessarily provide a simple answer, but to refine, clarify and articulate the reasoning behind choice and have both commissioners and evaluators weigh pros and cons of possible options in a logical and structured way.

Although it can speed up and improve decision making, this tool is essentially a learning device: it helps users learn about comparative advantages and weaknesses of methods, their specific benefits, and not less importantly their requirements and feasibility. It builds on the "Design Triangle" (Stern et al. 2012) idea that methods need to align with evaluation questions and programme attributes. It expands and reformulates the Design Triangle, by preserving the matching between methods and questions, and by unpacking the matching between methods and "programme attributes" in two different directions: by matching methods with, respectively, requirements/constraints on one hand; and abilities, opportunities, or benefits on the other (see below for definitions and dedicated sections). In addition, while the Design Triangle is a heuristic device for thinking about choices, the tool presented here is also practical and takes the user through the decision-making process in a hands-on, step by step way.

This paper, published together with an excel tool that collects and structures a considerable amount of information on the potential and weaknesses of 15 methods, contains six sections. In the introduction, we cover multiple dimensions of evaluation quality to see where appropriateness fits and how it relates to evaluation quality in general, and we discuss some theories of human choice. Section 2 begins to cover criteria and heuristics that can be used for the choice of methods. Section 3 addresses the dimensions of appropriateness[5] and the conceptual backbone that informs the choice (excel) tool structure. Section 4 illustrates the inner workings of the tool, as well as describing the meaning of the various cells in the excel file. Finally, section 5 tackles the intended users and

2 The title does not refer to impact evaluation specifically since many of the methods considered can have a variety of purposes. However, the selection of methods considered in this exercise can all be used for impact evaluation and are particularly relevant all considered potentially viable solutions to answer impact evaluation questions.
3 For a definition of "methods", see Annex 2
4 For many people and institutions, still "has".
5 Quinn Patton asserts that "methodological appropriateness is the utilization-focused gold standard" (Quinn Patton, 2012, p287)

potential uses of the tool, in addition to discussing the tool's limitations, history, and the potential to develop it further. A series of annexes covers reviewers and contributors (One), the definitions of method, approach and technique on which the tool is based (Two); and provides basic information about the 15 methods included in the tool (Three).

1.1 | Evaluation Quality and human choice

It seems appropriate (!) to start a discussion of methodological appropriateness by tackling the broader notion of evaluation quality and understanding where the particular dimension of appropriateness fits in the evaluation quality landscape.

The notion of evaluation quality is often left implicit in methodological debates. Methods are endorsed on the basis of their "rigour" or "robustness" but the latter are seldom defined as notions and how they relate to other dimensions of evaluation quality is often unclear. It is the author's view that evaluation quality should be assessed on the basis of at least 8 dimensions: conceptual framing, transparency, appropriateness, validity, credibility, transferability, reliability, and structure[6]. It can be argued that, by and large, these are the same criteria defining the quality of scientific research (Bryman 2012). In this section we provide a brief overview of these dimensions and attempt an interpretation of RCTs supremacist arguments in terms of how the notions of rigour and robustness relate to the quality criteria. We also hope to (appropriately!) locate the concept of methodological appropriateness in the wider context of evaluation quality, by covering the dimensions it doesn't account for.

Framing (Conceptual Framing) • The quality of the conceptual or theoretical framework underpinning the evaluation (the policy map, logic model, theory of change, concepts and constructs). It relates to the clear outlining of major assumptions (with appropriate references); the context in which the evaluation is taking place (that might affect understanding and interpretation); the clarity and precision in the definition of concepts and constructs, the minimisation of ambiguity. The evaluation questions and hypotheses need to be formulated considering falsifiability and confirmability within contextual and data-related limitations. This can require management of expectations: the number of issues and questions that can realistically be tackled in a high-quality manner within the time available for the evaluation might be limited. An evaluation presenting a poor conceptual framing might also present a poor allocation of resources to answer questions, ambiguous claims that are difficult to test, and poor construct validity (see below). An additional risk is not incorporating available knowledge that might have otherwise been of use.

Transparency (Replicability/Confirmability) • The different aspects of an evaluation which are required to replicate or reproduce it (required to eventually confirm the findings). In and of itself, transparency does not guarantee that the findings will be confirmed, but it is necessary to check whether they can be confirmed by replicating the evaluation. It concerns the disclosure of designs, methods, and data collection techniques employed to answer the research questions: for example, the specific means and protocols used for data collection (questionnaires, discussion guides for semi-structured interviews, observation of patterns in desk reviews, etc.) and anything a third party might need to potentially replicate data collection and obtain the same data (if the object of analysis has not changed); the processes and tools used for data analysis (software and software settings, for example SPSS syntax or Nvivo; manual and electronic procedures, etc.) and anything a third party using the same software and/or procedures applied to the same data would need to obtain the same findings. It also covers information about the locations and contexts/settings data (including quotations) has been collected in, and reflections on how this might have influenced the findings

6 The content of this chapter builds on work the author has conducted for the UK Department of Business, Environment, and Industrial Strategy, partly in collaboration with Maren Duvendack.

and whether/how much it makes the study difficult to replicate. Finally, it includes sources of funding and considerations on ways funding could potentially bias the findings and undermine independence, for example conflicts of interest and how they've been managed. The main risk of lack of transparency is that the findings might not be checked for reliability/robustness

Appropriateness (Methodological Appropriateness) • It refers to whether the selected methods are appropriate for a specific evaluation. In this tool and user guide, we define methods appropriate if they answer the preferred questions, fulfil other implicit or stated goals, and if their requirements are met in the course of the evaluation. As we will see in more detail in the rest of the guide, there is no single method or design that is always preferable to others: for example, the implementation of an experimental design might not be appropriate (and thus, in and of itself, not necessarily a sign of high quality evaluation) if the questions of interest are not the ones the method excels at answering or if the method's requirements cannot be met in a given evaluation setting.

Validity (Construct/Measurement Validity) is one of three types of validity considered in this framework and concerns whether definitions or operationalisations of a construct, as well as indicators or metrics chosen to measure a concept, are well suited to measure or "indicate" what they're supposed to. For example, is income a valid measure of household welfare or well-being? Or do we need to consider health and happiness? Are IQ tests valid measures of intelligence? In order to assess construct validity, we need ask ourselves whether the proxies we're using adequately represent the notion of interest. The risk is a loss of relevance, or a poor fit between what we measure and what we're interested in measuring.

Credibility (Truth Value of Statements/Findings; Internal Validity) • The extent to which the evaluation findings can be trusted. It can be associated to a formal level of confidence in the truth of the propositions or statements constituting the findings. The latter can be quantitative, qualitative, descriptive or causal. For quantitative statements, for example about the value an indicator takes in a population (inferred from a sample) or the net effect of an intervention in counterfactual analysis, p values or significance levels will be associated to the result and formally represent confidence that the latter is true or the extent to which they can be trusted. For net effect statements, confidence is also affected by the design—the better it is at minimising selection bias, the higher the confidence, as seen in the Maryland Scale. For qualitative statements, confidence in the truth of statements, mechanisms, theories, etc can be formalised using Bayesian updating as in diagnostic evaluation. Credibility of findings is especially important when policy decisions must be based on empirical evidence produced by evaluations.

Transferability (External Validity) refers to whether, how and to what extent the findings can be transferred to other groups, contexts or sectors. How many cases are the findings potentially applicable to or relevant for? At one extreme, it is just one case; at the other, the entire population. Commissioners are often interested in evaluating interventions covering specific areas not only to acquire knowledge concerning that specific area, but also to learn about others not directly investigated.

Robustness (Reliability/Dependability/Consistency/Stability) • refers to whether the findings are reliable, that is dependable or stable/consistent when measurements or data collection and analysis are repeated over time. If (measurement, construct) validity can be associated with "accuracy" (not missing the mark), reliability can be associated with "precision" (we might be very precise and consistently get the same results but be also consistent in missing the mark). One way to thinking about robustness is to ask if the findings are likely to be sensitive or changeable depending on the analytical technique used or a specific researcher collecting, analysing and interpreting the data. If the evaluation is repeated, or if measurements are repeated, it is desirable that they aren't dependent or sensitive to aspects of the evaluation process that we can't control or

repeat. Have any steps been taken to ensure that measurements are consistent? For example, is the sample representative or sufficiently large? Are researchers using standard data collection protocols and analysis procedures? Finally, is a standard process employed to assess the strength or probative value of the evidence?

Structure is about the evaluation report and the logic and clarity of arguments made. The idea is that the final report should allow for easy readability and an easy way to connect different parts and phases of the work. It's both about how the report is presented and how the different phases of the evaluation are actually linked, and whether challenges and limitations are openly discussed. Quality assessors should check for the presence of: a) a clear, logical thread that runs through the entire paper; b) a clear link between the conceptual (theoretical) framework and data collection/analysis; c) a clear link between the conceptual (theoretical) framework and the findings; d) a clear link between data collection/analysis and the findings; e) claims in the conclusions that are clearly backed up by the data and findings; f) signposts guiding the reader through the different sections of the paper; g) self-criticism and clearly identified limitations; and, h) alternative ways of interpreting the same findings or data.

Quality is not strictly related to methods used (except for the appropriateness dimension); however, some methods enjoy comparative advantages on specific quality dimensions. For example, theory-based methods usually have a comparative advantage on Framing; qualitative methods are traditionally strong on construct Validity; while quantitative methods on Transparency and Robustness. As for hybrid methods: QCA compares favourably to other methods on robustness and transferability, while Process Tracing (particularly with Bayesian Updating and more broadly, diagnostic evaluation) aims at enhancing the credibility of qualitative findings.

When the evaluation community demands methodological rigour, they mostly refer to credibility, which can be achieved for quantitative causal statements by removing as much selection bias as possible, and reliability/robustness (the consistency of findings when the evaluation or the assessment is repeated). Other dimensions of validity (construct and external) are mostly ignored, and appropriateness is considered only among quantitative methods—which we will see are not appropriate under several circumstances. Conceptual framing is also mostly ignored.

This tool (and related user guide) focuses on one of these quality dimensions; but by unpacking the logic of appropriateness and considering methods comparatively and in this light, we discover that all quality dimensions can be equally important depending on the preferences and constraints/opportunities of different users.

Factors affecting human choice

Choosing appropriate methods is—first and foremost—choosing: that is, according to the Cambridge dictionary, "to decide what you want from two or more things or possibilities". In choice, "wanting" is not the mere expression of a desire but a decision to select an option out of two or more. In our case, the decision has the practical consequence of drawing up a ToR requesting specific methods (if the person taking the decision is a commissioner) or designing the evaluation in a specific way (if the person taking the decision is an evaluator). In either case, the decision leads to action so we can use sociological theory of action to understand how decisions are usually taken by humans and what implications this has for methodological appropriateness.

In analytical sociology, the DBO theory (Hedstrom 2015) states that intentional actions are explained by an individual's desires, beliefs and opportunities (Davidson 1980). A belief can be defined "as a proposition about the world held to be true (Hahn 1973)"; and a desire as a "wish or want." Opportunities would then refer to the "menu of action alternatives available to the actor"; or "the actual set of action alternatives that exists independently of the actor's beliefs about them." (Hedstrom 2015)

The literature groups together beliefs and desires as both causing an action "in the sense of providing reasons for the action" (Hedstrom 2015); thus being a sort of "motivational force" for it; and stresses how a given combination of desires and beliefs constitutes "a compelling reason for performing an action". A classic example tries to explain why Mr Smith brought an umbrella with him on a given day; the possible reasons being that (1) he believed that it would rain on the day; (2) he desired not to get wet, and (3) there was an umbrella for him to bring. See Box 1 for a detailed outline of the three different explanations.

Box 1 | Explanation of behaviour in the DBO theory (analytical sociology)

- **Belief-based explanation** · Mr Smith desires not to get wet and he had an umbrella that he could have brought, but by mistake he read yesterday's day's weather column in the newspaper which made him believe that it would not rain today. Therefore, he did not bring an umbrella today.
- **Desire-based explanation** · Mr Smith believed that it would rain today and he had an umbrella that he could have brought, but he has somewhat unusual desires: walking in heavy rain always makes him feel like Gene Kelly in Singin' in the Rain, and feeling like Gene Kelly is something he really desires. Therefore he did not bring an umbrella today.
- **Opportunity-based explanation** · Mr Smith believed that it would rain today and he had a strong desire not to get wet, but when he was leaving for work in the morning he found that his son had, once again, taken his umbrella and there were no other umbrellas in the house. Therefore, he did not bring an umbrella today.

(Source: Hedstrom 2015)

It's thus a good idea to think of methodological choices as actions/decisions affected or explained by an interplay of beliefs, desires, and opportunities or constraints. In the tool we can see how good several methods are at fulfilling which desires; and what circumstances (and hence opportunities) they require to be properly applied. We can speak of "constraints" when we're not able to provide methods with the conditions they require.

The DBO theory states that desires, beliefs and opportunities are not only interesting as alternative explanations for action and decision-making; but are perhaps most interesting when they interact. The notion of cognitive dissonance dates back to Festinger (Festinger 1957, Elster 1998) arises when our beliefs about reality (in particular, our opportunities) clash with our desires. There are various strategies of cognitive dissonance reduction, the two most famous possibly being adaptive preferences or sour grapes (when we try to adjust our desires to align with what we believe we can have) and wishful thinking (when we conveniently adjust our beliefs to match our desires).

We will use these metaphors to describe possible states we can find ourselves in when there is a discrepancy between the methods we would like to use and those we can use; and also to explain the disproportionate influence that RCTs and experimental evaluation have enjoyed in the methodological landscape in recent years.

2 | Choice and design triangles

The core idea at the heart of methodological appropriateness is that different methods have different strengths and weaknesses and it's misleading to think in terms of a generic "gold standard". The proponents of counterfactual analysis have historically maintained that—whenever feasible—RCTs and by derivation experimental methods are consistently preferable under all circumstances

or in other words, they're always the optimal choice. This manifests itself as a tendency to choose RCTs whenever possible and to consider alternatives only when RCTs are unfeasible. This state of affairs can be traced to the wrong belief that removing selection bias in case-control comparisons (what RCTs really excel at) is the only rigorous way of testing causality in policy evaluation. But even the Nobel Prize press release, motivating the award to the first proponents of such methods to evaluate interventions aimed at poverty reduction, does not mention selection bias: it mentions the novelty of the approach (in this specific field) and its ability, and I quote, "to obtain reliable answers". It goes on to appreciate the approach for its way of dividing the issue "into smaller, more manageable, questions" and that the Nobel laureates deserve the prize because "they have shown that these smaller, more precise, questions are often best answered via carefully designed experiments among the people who are most affected".

What appears so worthy of praise is then the method's ability to obtain reliable answers to specific questions, not its uniqueness in doing so; in fact, RCTs are the optimal way of obtaining answers to only one type of evaluation questions. Nowhere in the press release it is mentioned that the impact question answered by RCTs is the most interesting one or the one we should always ask; let alone under all circumstances. As many of us now understand, RCTs answer only the "net effect" question but the net-effect one is not the only question that can be made manageable or broken down or that can be answered reliably. There are many other questions that, as we discovered after the explosion of RCT's popularity (Stern et al. 2012), can be equally (if not more) manageable and precise. Finally, they mention that the net effect questions are often best answered with RCTs, which means that, compared to other methods aiming to answer those questions in case-control comparisons, RCTs are the best choice because of their performance in removing selection bias (see the Maryland Scale). In other words, what has been rewarded is a scientific approach to reducing poverty (which is a global policy problem); but thankfully for us and for the world, it is not the only scientific approach that can be implemented in evaluating policies and programmes; its merit is of having perhaps been the first scientific approach to be adopted in this context. Asking questions that can be answered is a requirement for science: but scientific discoveries are made and scientific knowledge is advanced mostly without RCTs.

In summary, what the Nobel Prize recognised and acclaimed was:

1. The introduction of a new, precise way of framing evaluation questions.
2. One method capable of answering some specific evaluations questions (the net effect one) in a reliable way.

Going back to our quality criteria, it seems the Swedish Academy was particularly excited about criteria #1 and #7 (Framing and Reliability). We also know that RCTs are strong on #5 (Credibility) because of the formal assignment of confidence levels to results. But there was no mention of #2, #3, #4, and #6 (Transparency, Appropriateness, Construct Validity, and Transferability).

In hindsight, knowing what happened in the last 5-10 years, it seems clear that having the wrong belief (that RCTs were the only way of establishing causality in a reliable way) triggered an adaptive preference (sour grapes) mechanism in the evaluation community which made them believe the only possibly interesting question in evaluation was the net effect one (because that was the only one RCTs could answer). In other words, the wrong beliefs were held about existing opportunities. We can be glad the preference adaptation didn't go any further than that, although at some point it looked as if part of the community started to believe that the only policy interventions worth funding were the ones that could be evaluated with RCTs. The cognitive dissonance must have been unbearable.

The worst was avoided and—once the wrong belief was rectified and the community realised that causality could be inferred rigorously and reliably without recurring to randomisation, and the opportunities to do so started to increase, the real preferences were freed: the community began to

show interest in other types of evaluation questions again. Because the issues with RCTs are not just about feasibility but also desirability. As this tool and user guide show, there are at least four other equally interesting impact evaluation questions that can be asked (see also Box 2).

Box 2 | The overarching and the specific impact evaluation questions

What difference did the intervention make?
How much of a difference?
(What was the net effect of the intervention)

How did it make a difference?
(What role did the intervention play, what was in it that made it work)

What difference did it make for whom, under what circumstances?
(what conditions facilitated success or lack thereof)

The Stern paper (Stern et al. 2012) introduced the idea that an optimal methodological choice needed to align with evaluation questions: what methods are best suited to answer each question? While it was known that different methods have different comparative advantages and weaknesses, the innovation was that, for the first time, it wasn't taken for granted that evaluators were all most strongly interested in answering net effect questions. In other words, the paper made room for a variety of preferences, or to use the DBO theory terminology, desires.

Continuing to use the DBO theory to frame our narrative, and extending the logic introduced in the Stern paper, we can think of methodological choice as a process that ideally fits our desires, or preferences; but that must also be feasible, or lie within a specific theoretical or practical realm of possibilities: which are the opportunities that the intervention and the institutional evaluation setting provides, as well as its constraints .

Paraphrasing JF Kennedy, we could say that we must not only ask "what can our methods do for us", expressing our desires; but also "what can we do for our methods", acknowledging the need of operating within the constraints they impose on us in terms of what they require to be applied to a certain quality standard (Box 3). We might not always be in a position where every method is feasible or can be used, either for characteristics inherent to the intervention or for limitations related to the institutional setting the evaluation process is unfolding in. In other words, we might have the opportunity to apply the method (to a certain quality standard), but this is not guaranteed and must not be taken for granted.

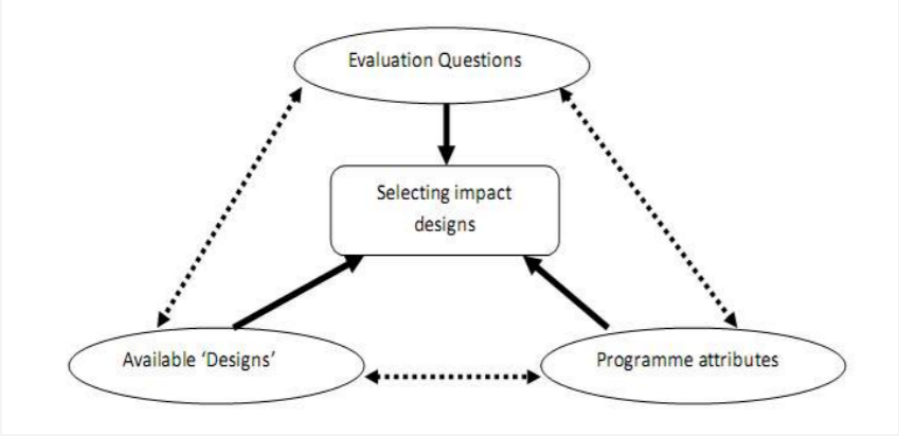
Continuing with the DBO terminology, we can say that this tool is an attempt to shape or refine our beliefs in terms of how different methods can fulfil our desires, and what are realistic circumstances under which we have the opportunity to apply those methods to a high quality standard. Bearing in mind that choice is affected both by what we want to do, and by what we can do: and if the two are not aligned, choice will end up being a balance or compromise between the two.

In methodological choice, desires are not limited to types of evaluation questions: our goals when conducting or commissioning an evaluation might not necessarily be limited to answering evaluation questions. Which is what brings us to introduce, in addition to questions and requirements, a third, residual dimension encompassing other goals we might want our methods to reach or other desires that we would like our methods to fulfil.

3 | Dimensions of appropriateness

The Design Triangle introduced in the Stern paper (Stern 2012) suggests (Figure 1) that methods should be chosen in accordance with availability, ability to answer specific questions, and ability to adapt to specific programme attributes (for example delivery mode or interaction with other programmes in stable or unstable environments). In DBO language, available designs are potential opportunities; evaluations questions desires or preferences; and programme attributes are contextual features that limit choice, thereby affecting actual opportunities.

Figure 1 | Original "Design Triangle" from Stern et al., 2012



However, desires are not only confined to answer evaluation questions: they can extend to other goals that methods allow us to reach. That's why the CHOICE TRIANGLE (figure 2) includes additional evaluation goals in one of its corners. It can be argued that methods perform differently on a series of aspects, which include but are not limited to answering evaluation questions.

Box 3 | Questions governing choice

DESIRES: What can methods do for you? What would you like methods to do for you?
Which evaluation questions would you like methods to answer?
What other evaluation goals would you like to achieve with your methods?

OPPORTUNITIES (and constraints): What can you do for your methods?
Are you able to provide what methods require (to be implemented correctly or to high quality standards?)

The other change from the Design Triangle (Figure 2) is that the relation between programme attributes and available designs is made specific and explicit: no longer an abstract difference between potential and actual opportunities, but a systematic survey of comparative advantages and limitations of a list of methods, in terms of their abilities to adjust to contextual characteristics; which also helps the practicalities of the matching process between opportunities and desires. Methods have different requirements, which have implications for commissioners' ability to use them

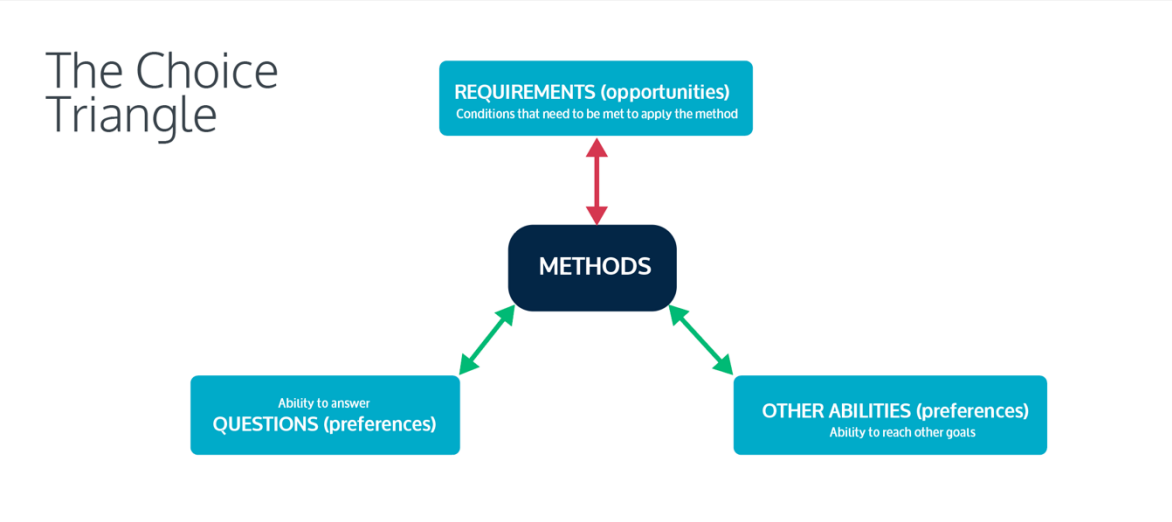
properly in the course of an evaluation. The key question here is not whether a method can or cannot do something, or what possibilities it offers to the evaluation team. It is whether the commissioner and the evaluation team can comply with a series of conditions that need to be met for the method to be implemented properly. In this case, it is not what the method can do for you, but what you can do for the method. This relationship is represented by the red arrow in the figure below.

Notice that the relation between methods and these three entities is symmetrical: while the desire to answer specific questions or to achieve other goals affects methodological choice, it is also true that a limited availability of methods constrains the questions you can answer and the goals you can achieve. Similarly, while having to meet specific requirements in order to apply a certain method reduces your options, you can also try to reduce your constraints and meeting additional requirements once you know which method you would like to use.

The tool's conceptual framework (figure 2) is thus structured around three dimensions of appropriateness:

1. The method's ability to answer a series of specific evaluation questions
2. The method's more general ability to fulfil a series of tasks or reach specific goals
3. The evaluation stakeholders' ability to accommodate the method's specific requirements.

Figure 2 | Revised Design Triangle (the choice triangle)



4 | How the tool works

This section presents the inner workings of the tool and describes how it matches methods to specific planned evaluation exercises. The tool is structured around the three dimensions of appropriateness as described previously.

The summary tab in the excel file returns three sets of results for the eleven methods, relating to the three Stages of the tool:

- the first relates to the ability of each method to answer the user's preferred questions (green = higher, red = lower ability)
- the second to other abilities of interest to the user (green and red have the same meaning as above); and

- the third refers to methods ranked by how many of each method's requirements the user can meet (green for all requirements, red for none, yellow for some).

Methods should be both useful and applicable, in which case they would have high (green) scores on all three rankings.

At a minimum, methods should be applicable: that is, they need to score "0" or "green" in the Stage 3 ranking. Depending on how the applicable methods fare on the other two rankings, the user may decide to use one, all or some. If one method is able to cover all questions and features of interest, it can be used alone, but often the preferred method to answer questions is different from the preferred method for other goals, in which case the best choice would be a combination of the two.

In other situations, more than one method could be able to answer preferred questions or to reach preferred goals. If all these methods are applicable, other considerations will need to determine whether it makes sense to combine all methods, just some, or just one.

Currently, the excel format, while being fully transparent on the inner workings of the tool, imposes limitations on how the findings are communicated and perhaps on user friendliness. While the tool structure and substance is fully included here, future iterations of the tool could potentially improve on these aspects.

4.1 | The method's ability to answer key evaluation questions

In the first section (Stage 1/Worksheet "Questions"), the user is asked to indicate which evaluation questions they are interested in answering. They select at least one option out of the following five:

1. What was the additional/net change caused by the intervention? or How much of the observed outcome(s) can be attributed to the intervention?
    - This is the HOW MUCH question, and usually refers to an average value across a sample or a population.
2. What difference did the intervention make to different population groups, and under what circumstances?
    - This is a WHAT/HOW question: you are interested in understanding which effects have materialised for different groups and contexts—not just an "average" effect.
3. How and why did the intervention make a difference, if any? or What was the process/ mechanism by which the intervention led to or contributed to outcomes?
    - This is the HOW/WHY question, the focus of many theory-based evaluations
4. What other factors needed to be present alongside the intervention to produce outcomes observed? Or Which factors were necessary and/or sufficient for the intervention to work?
    - This is still a HOW/WHY question, but the focus is on the intervention not being the sole cause of change, but working in conjunction with other factors/interventions). An important related question is "Can we expect the intervention to make a difference elsewhere, or in the future?"
5. Which outcomes of the intervention(s) being evaluated do different population groups consider to be the most important?
    - This is the WHAT question, asking which outcomes are relevant for whom.

The tool is applicable to evaluation questions related to impact or effects that can be captured by any of the above five options. The choice of these specific questions was informed by the relative clarity of their link with specific methods. Evaluation questions not included in these five are, for the moment, outside the scope of this framework; but additional or different questions could be added in future iterations of the tool (see section on uses and limitations).

Out of these five questions, almost no single method can answer more than three well (Agent-Based Modelling being the only exception). Some methods answer only one question well, while most answer either two or three well. The hidden columns in the excel tool (columns E to S) show the link between each question and the eleven methods considered. The coloured cells indicate that the methods have scored either 80/100 or 100/100 and are thus appropriate to answer the question.

When the user selects the questions they want to answer, the related scores appear under every method for each of those questions (columns U to AI). In the excel file, every score is illustrated with a different colour, with green representing the highest scores and red the lowest. Experts on the different methods have assigned the scores based on the following rubrics (Table 1):

Table 1: Descriptors of scores indicating methods' ability to answer questions

Score	Description: the method received the score on the question IF
100/100	The question is the primary question answered by the method, which is fully "self-sufficient" to answer it
80/100	The method greatly helps answering the question, but it needs to be combined with other methods to answer some formulations of the question (sub-questions)
60/100	The method helps answer the question, but it needs substantial input from other methods to answer the question properly
40/100	The method is useful to answer the question only very indirectly, rarely or under special circumstances
20/100	The question is substantially different from the primary question answered by the method, which provides little to no help with it.

When the user selects more than one question, the tool also returns an overall measure of the ability of each method to answer the group of questions the user is interested in, which is obtained as the average ability of the method to answer all the questions the user is interested in (row 11). The colour-coding highlights the methods most suited to answering all selected questions (green) and the least (red). However, it is important to be aware that the single highest-scoring method may still not be suited to answering all questions, and thus the overall result should also be considered alongside the results for each individual evaluation question of interest.

4.2 | The method's ability to carry out specific tasks/ achieve specific goals

In the second section (Stage 2, Worksheet "Other Preferences"), each method has been assigned a score on a three-point scale (Low, Medium or High), measuring its ability to achieve each single goal (rows 4 to 25). These scores can be found in the hidden columns E to S, where LOW is indicated with 0.33, MEDIUM with 0.67 and HIGH with 1.00. The methods experts made these assessments based on the following definitions:

- HIGH: the method is ideally suited to do X;
- MEDIUM: the method is able to do X under specific/limited circumstances;
- LOW: the method is not well suited to do X and is mostly unsuitable for it.

The user expresses their preferences about what they want to achieve with their evaluation in addition to answering questions. They indicate their level of interest in achieving each of 22 possible goals (see excel tool) on the following 4-point scale:

- Not desired
- Slightly desirable
- Desirable
- Very desirable

After the user inputs their level of interest in a specific goal, the tool returns a score measuring both the ability of the method to achieve that goal and the user's interest in it (columns U to AI). If the user is very (maximally) interested in the goal, the tool simply returns the ability of the method to reach that goal as calibrated above (0.33, 0.67 or 1.00). If the user is less interested in the goal, the tool returns a lower score, taking account of both the method's ability and the level of user interest.

The score returned by the tool for each ability is highest when the method is fully able to do what the user is very interested in, and lowest when it has a low ability to do what the user has a low interest in. The middle scores can signal either methods fully able to do something the user is not very interested in, or methods poorly able to do something the user is very interested in.

In order to understand what the middle scores mean, the user can unhide columns E to S and compare them with columns U to AI. If the latter scores are lower than the former, it means that the user's level of interest in the specific ability is lower than the maximum. The larger the difference, the lower the level of user's interest. The ability of method to achieve the specific goals can be found in columns E to S.

The summary row (row 29) indicates the average ability of each method to achieve the overall set of goals selected by the user, weighed by the level of user's interest in the different goals. It can be considered an overall measure of how useful the method will be for the user. If the score is high, it means the method is very capable of achieving the whole set of the user's most desired goals; if it is low, it means the method is not capable of achieving the user's least desired goals. The middle scores can mean either that the method is fully capable of achieving the user's not so highly desired goals, or that it is not fully capable of achieving the user's mostly desired ones.

Finally, row 23 indicates the number of goals the user is very interested in ('very desirable') that are fully achieved by each single method. You can see which method offers which possibility by focusing on the green cells.

As a general principle, it is important to read the scores in connection with the results of section one on questions. A method might have a high ability to answer your desired questions but score lower on your other interests, or vice versa. The "summary results" tab in the spreadsheet compares the two overall rankings.

4.3 | The team's ability to accommodate the method's requirements

The third section (Stage 3/Worksheet "Requirements") addresses constraints that can be imposed on methodological options by the nature of intervention to be evaluated and other real-life constraints like the natural of the evaluation process or context. Discovering a method that is perfectly capable of achieving all our goals does not guarantee that we can actually apply it. Methods cannot be applied unless a series of conditions are met, which differ for different methods.

Usefulness and feasibility are independent: we can create a ranking of methods based on their feasibility or applicability for a specific evaluation, which might be completely different from the ranking of our preferred methods.

The expert group has identified a number of conditions that are required for the applicability of each method, as well as others that are desirable but not required. The requirements for one method may also be irrelevant or not required for others (e.g. control groups are required for RCTs, but are not

required for Contribution Analysis). The reviewers have made their assessments based on the following definitions:

- X a requirement for the method if an acceptable application of the method cannot be produced unless X is met (indicated with "1" in the excel tool, hidden columns F to T).
- Y is a desirable condition for the method if an acceptable application of the method can be produced even when Y is not met; however, Y is likely to increase the quality of such application (indicated with "0.5" in the excel tool, hidden columns F to T).

These assessments are used to inform the user about which requirements they need to satisfy if they are to use a given method; and which other conditions they are encouraged to ensure. A method cannot be implemented properly unless all requirements are met. However, the tool does not just indicate whether the user can meet all requirements or not, it also provides information about which requirements are still unmet, or for which information is unavailable. Ultimately, this will give users an idea of which methods are feasible in the particular evaluation they have in mind, but also what needs to be changed in order for other methods to be applicable.

The requirements for the 15 methods are captured in the sets of questions in this section: the first ten questions relate to requirements of experimental and quasi-experimental methods, and the following fifteen questions relate to requirements of the non-experimental/ theory-based methods in the tool. Which questions/requirements are relevant for each method can be seen by unhiding columns F to T.

When using the tool, the user is asked to indicate whether they can meet a series of conditions in a specific evaluation process. More precisely, they indicate the degree to which they can meet either of 25 conditions (Stage 3 "Requirements" tab, rows 4 to 28), on the following 4-point scale:

1. Fully
2. To Some Extent
3. Poorly
4. Not At All

A "don't know" option is also included. This highlights information that might be missing in order to make a decision on methods. It is recommended that users seek the necessary information to answer all questions so that they can understand in full which methods may be feasible to use.

The tool is not meant to be used at any particular phase of the evaluation process: it works as long as the information the user is asked to input is available.

Once the user has provided the above answers to the 25 questions, the tool returns the following information:

- A row at the end of the table (row 31) indicating the number of essential requirements for each method that the user cannot meet. This gives the user an idea of how far they are from being able to apply the method and what actions can potentially be taken to open up this possibility.
- A row (row 32) describing the number of requirements that the user does not know if they can meet. This tells the user that further investigation is needed in order to determine whether the method can be applied or not.
- A final row (row 33) illustrating the number of desirable requirements that the user cannot meet. This is not directly or generally relevant to the possibility of applying the method; however, if the method is used, it affects the quality of its application.

When the user selects the degree to which they can meet a specific condition, a corresponding score is assigned to the methods for which that condition is relevant. For the 'essential' requirements, the score is computed as follows: 100/100 or 1 for "fully", 67/100 for "to some extent", 33/100 for "poorly" and 0/100 or 0 for "not at all", with shades of green indicating high and shades of red indicating low

scores in the excel file. For the desirable conditions, the score is divided by 2 (or multiplied by 0.5); so for example if the user can meet a desirable feature "fully" the score will be 50/100; if they can meet it "poorly" it will be 17/100, and so on.

A summary row at the end of the table (row 35) returns an overall score for every method, representing the average degree to which its requirements can be met by the evaluation. This value is obtained by calculating the average score for each requirement, weighed by its importance (1 for an essential requirement and 0.5 for a desirable feature). The value will be highest (100/100 or 1) when all requirements and desirables can be fully met; and lowest (0/100 or 0) if none of the requirements or desirables can be met. The intermediate scores can indicate a wide variety of situations, where the user can for example meet some requirements fully while others not at all, or all requirements partially. The hidden columns F to T clarify how the overall score is calculated in each specific case.

4.4 | The tool's output

The summary spreadsheet of the tool provides an overview of how methods have fared in the three different stages, using the same colours for ease of reference: green if the method scores high in answering the desired questions, in fulfilling other stated goals, or if there are no requirements that cannot be met. Red if the method scores low at answering questions or fulfilling goals, or if there are essential requirements that cannot be met, with a number indicating how many essential requirements cannot be met.

A mostly green column indicates that the method of that column is both desirable and feasible; a mostly orange or red column a method that is neither; while contrasting colours will indicate some degree of "cognitive dissonance", or of conflict between reality and desires. Figure 3 below[7] shows a situation where the preferred methods (the methods scoring the highest in terms of abilities) are unfeasible: in particular, where the users are mostly interested in the net effect question and in other features best embodied by experimental methods, while they're unable to realise the conditions that would enable a correct implementation of these methods, failing the fulfilment of several requirements.

Figure 3 – typical gold standard influenced situation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**SUMMARY RESULTS - ALL STAGES**	RCT (Randomised Contol Trial)	Difference-in-Difference	Statistical Matching	Instrumental Variables (RDD)	Outcome Mapping	Most Significant Change	Soft Systems Modelling	Causal Loop Diagram	PSM (Participatory Systems Mapping)	BBN (Bayesian Belief Networks)	ABM (Agent Based Modelling)	Realist Evaluation	QCA (Qualitative Comparative Analysis)	Process Tracing/ Bayesian Updating	Contribution Analysis
2	**Stage 1: Which Method is Best Suited to Answering My Key Evaluation Question(s)?**	1	80/100	60/100	1	20/100	40/100	20/100	20/100	20/100	60/100	80/100	20/100	40/100	20/100	20/100
3	**Stage 2: Which method is most able to address my other interests?**	81/100	60/100	69/100	74/100	47/100	43/100	57/100	59/100	55/100	59/100	62/100	62/100	50/100	59/100	62/100
4	**Stage 3: Which Method has the fewest essential methodological requirements that cannot be met by my intervention? (Which method is most feasible to use?)**	6	6	6	4	0	0	0	0	0	0	0	0	0	0	0

This situation can evolve into a "sour grapes" or *adaptive preferences* (Elster, 1998) case where the user adapts their preferences to reduce cognitive dissonance and achieve alignment along the columns (see figure 4 below)

Figure 4 – reduction of cognitive dissonance through a change in preferences…

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**SUMMARY RESULTS - ALL STAGES**	RCT (Randomised Contol Trial)	Difference-in-Difference	Statistical Matching	Instrumental Variables (RDD)	Outcome Mapping	Most Significant Change	Soft Systems Modelling	Causal Loop Diagram	PSM (Participatory Systems Mapping)	BBN (Bayesian Belief Networks)	ABM (Agent Based Modelling)	Realist Evaluation	QCA (Qualitative Comparative Analysis)	Process Tracing/ Bayesian Updating	Contribution Analysis
2	**Stage 1: Which Method is Best Suited to Answering My Key Evaluation Question(s)?**	60/100	60/100	60/100	47/100	40/100	60/100	60/100	80/100	67/100	80/100	87/100	87/100	87/100	80/100	80/100
3	**Stage 2: Which method is most able to address my other interests?**	44/100	48/100	48/100	41/100	60/100	49/100	85/100	78/100	68/100	60/100	88/100	85/100	65/100	74/100	64/100
4	**Stage 3: Which Method has the fewest essential methodological requirements that cannot be met by my intervention? (Which method is most feasible to use?)**	6	6	6	4	0	0	0	0	0	0	0	0	0	0	0

7 Notice that the screenshot was taken on a previous version of the tool with 11 methods – the logic and substance of the argument doesn't chance and is valid for the newer version as well.

Another possible conflict can arise when the user has the option of implementing experimental methods but they prefer methodologies that excel at answering other kinds of questions and fulfil goals that experimental methods are best suited for. Figure 5 below indicates that the user is failing to meet the requirements presented by their preferred methods.

Figure 5 – I can do RCTs but I don't want to!

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**SUMMARY RESULTS - ALL STAGES**	RCT (Randomised Contol Trial)	Difference-in-Difference	Statistical Matching	Instrumental Variables (RDD)	Outcome Mapping	Most Significant Change	Soft Systems Modelling	Causal Loop Diagram	PSM (Participatory Systems Mapping)	BBN (Bayesian Belief Networks)	ABM (Agent Based Modelling)	Realist Evaluation	QCA (Qualitative Comparative Analysis)	Process Tracing/ Bayesian Updating	Contribution Analysis
2	**Stage 1: Which Method is Best Suited to Answering My Key Evaluation Question(s)?**	60/100	60/100	60/100	47/100	40/100	60/100	60/100	80/100	67/100	80/100	87/100	87/100	87/100	80/100	80/100
3	**Stage 2: Which method is most able to address my other interests?**	44/100	48/100	48/100	41/100	60/100	49/100	85/100	78/100	68/100	60/100	88/100	85/100	65/100	74/100	64/100
4	**Stage 3: Which Method has the fewest essential methodological requirements that cannot be met by my evaluation / intervention? (Which method is most feasible to use?)**	0	1	0	1	2	1	4	3	1	3	1	1	1	1	1

5 | Use and limitations

5.1 | Uses and users of the tool

The primary expected uses of this tool are to:

- Promote further learning about the range of evaluation methods available and the logic of appropriate methodological choice
- Contribute to informing the design of interventions to improve their evaluability; by helping users sense-check the compatibility between their evaluation options/intervention attributes, intended evaluation questions and other interests/intended uses, and alter some of those if necessary to ensure evaluability
- Contribute to informing the design and/or commissioning of evaluations of specific interventions; by helping users understand how and to what extent evaluation questions, other interests and concrete evaluation opportunities/constraints align.

The tool is particularly intended for users with some knowledge of evaluation methods and issues, and can help commissioners become more "intelligent customers" when engaging with evaluators. It can also be an aid to negotiation between evaluation stakeholders about what may be desirable and feasible in planning interventions and evaluations, including when particular stakeholders have a strong preference for using a method that may not be appropriate. Experts may find simplifications in the tool limiting, although some of these can potentially be overcome in future iterations of the tool (see next section).

The tool can inform the choice of evaluation methods for a wide variety of interventions. Although some of the methods are more suited to, for example, advocacy and policy-influencing interventions or to service delivery interventions, there are no hard and fast rules on this. This is why the tool checks the applicability of all methods against a long list of questions aimed at understanding the specifics of each evaluation/intervention. Similarly, the tool can inform evaluation method choice for interventions in any sector or thematic area (governance, health, livelihoods, energy, food, etc.).

5.2 | Limitations and potential for further development

The tool is not prescriptive: it does not provide definitive answers and needs to be used with judgement and flexibility to address the interests of evaluation stakeholders. It should not be used, for example, by those with no knowledge of evaluation to determine the method to be used in an evaluation, although it can be used by the same people to learn more about opportunities and constraints related to given methods. Furthermore, it does not envisage all the conditions and

requirements of real-world evaluations and we expect additional information as well as common sense to play a role in the final decision of methodological selection.

This tool has a number of potential limitations, all of which can be overcome under the right conditions. The first is the particular selection of experts that are responsible for the assessments of the methods' abilities and requirements. These judgements are dependent on the specific group of experts selected and could differ if a larger or different group were involved.

Although world class experts were involved, engaging with different experts could potentially not only change the scores, but also the types of requirements and the comparative advantages considered. The second and third level elements (requirements and other abilities) were in fact selected based on expert advice: a broader or different group of experts might have perhaps selected different requirements and proposed different abilities.

To my knowledge, no structured sensitivity tests have been formally carried out on the tool, and at this stage, it is currently unknown how much the results of particular selection would change following small changes in the ability measurements and the requirements' assessments. At the same time, however, the first version of the tool has been available and in use for four years and no alarming incident of this kind has been flagged during this time. In addition, the spreadsheet format the tool comes in makes such tests quite easy to conduct.

The specific selection of methods and questions also comes with its own list of pros and cons. While an attempt has been made to cover both the most well-known and innovative/promising methods, covering counterfactual-based as well as theory-based and systems-based "families" (see Annex 2); and formulate questions in a way that would clearly link them to such methods, adding more questions or methods might expand the relevance of the tool. It would not necessarily improve its utility, though, as the tool might become too dense and complicated and hence less user-friendly.

The tool does not attempt to incorporate feasibility limitations related to the budget available for evaluation. Budget constraints may have a significant bearing on methods choices under some circumstances, but we have not attempted to incorporate this for two reasons:

- Variations in how methods may be applied and in key evaluation costs (such as consultant fees) mean that it is not possible to identify useful ranges for the cost of using each method. In brief, with the exception of RCTs, none of the methods come with a price tag attached.
- Ideally, evaluation budgets should reflect the cost of delivering a quality, useful evaluation, rather than methods being chosen to fit a given budget. This tool could be used to inform negotiations about evaluation budgets, e.g. if users want a wide range of evaluation questions answered but do not appreciate that this may require a variety of methods to be used.

We also did not consider qualifications or skills as requirement for the most part, because a high quality application of all methods requires specialist expertise. Suggesting that this is not the case and some methods can be applied by less qualified evaluators or evaluators with lower fees might have led to some methods being seen as more sophisticated or more desirable than others just because the skills required are rarer or more expensive. This is contrary to the value of "methodological equality" and the idea that "all methods have equal dignity" that this tool supports. Appropriateness never means automatic superiority. You have to go through the structured reasoning proposed by the tool, and match your opportunities and constraints with methods and their characteristics, before ranking methods against each other. And beware that the ranking will usually be conjunctural and tied to particular circumstances.

Finally, we were very keen for the tool not to produce a final, single best ranking of methods, or - even worse - one single best method. We preferred taking the user through the selection process step by step, letting them in on the reasoning and logic behind the assessment of options.

The most definitive result returned by the tool is a series of three rankings (in the "summary results" tab) indicating:

1. which methods are best suited to answer your questions of interest (row 2);
2. which methods are the most able to achieve your preferred goals (row 3).
3. which methods have the fewest requirements that cannot be met (row 4);

If the user is to make a final selection, they need to take all these three rows into consideration. There are different ways to do it. One can start from the feasible methods (third bullet point) and see later which ones are most useful in terms of answering the preferred questions and achieving the preferred goals. Or one can start from preferences, about questions or goals – and see whether their preferred methods are feasible. And if not, what actions can be taken in order to implement those methods that would allow the user to answer their preferred questions or achieve their selected goals.

One last critique that has been moved to the tool is its supposed "orthogonality", or consideration of all its dimensions as presumably independent from each other. One question that is sometimes asked is, how do the different factors interact in a particular evaluation and what are the consequences of intersections and combinations? Given the current scope of this work, we judge it as very difficult to expand the systematic comparisons we perform here to take interactions into account. The combinatorial complexity is very likely to quickly become unmanageable. We are confident that the multiple health warnings and encouragements that the user exercises an overall judgement that takes into account the peculiarities of a specific situation, including the consequences of different factors combining and interacting, do not mislead the reader or the user to misread or misinterpret our intentions. As we mention above, this work is intended to fill a gap in the breadth of methodological knowledge that is often necessary to design and to commission an evaluation, and is intended to empower decision makers against "sellers" of methodological ideologies that want to adjust evaluations to specific methods instead of adapting methods to specific evaluations. It structures a wealth of diverse methodological knowledge that is usually preserved in scattered repositories and aims at constructing an insightful overview. The purpose is to increase understanding of the potential and weaknesses of different options in a "hands-on" and hopefully engaging way.

6 | Annexes

Annex 1 | List of Reviewers

Method	Reviewers
1. Randomised Controlled Trials	Maren Duvendack, Edoardo Masset, Daniel Phillips, Matthew Juden
2. Difference in Differences	
3. Statistical Matching	
4. Instrumental Variables	Stephen Morris
5. Outcome Mapping	Simon Hearn
6. Most Significant Change	Rick Davies
7. Soft Systems Methodology	Bob Williams, Richard Hummelbrunner
8. Causal Loop Diagrams	
9. Participatory Systems Mapping	Peter Barbrook-Johnson
10. Agent-Based Modelling	Nigel Gilbert
11. Bayesian Belief Networks	Stuart Astill, Simon Henderson
12. Realist Evaluation	Gill Westhorp, Bruno Marchal
13. Qualitative Comparative Analysis	Barbara Befani
14. Process Tracing & Bayesian Updating	
15. Contribution Analysis	Thomas Delahais, Sebastian Lemire, Jacques Toulemonde
Overall tool	Laura Camfield, James Copestake, Rick Davies, Sebastian Lemire, Saltanat Rasulova, Patricia Rogers, Giel Ton, Jos Vaessen

Annex 2 | Defining "method" as opposed to "approach" or "technique"

The meanings of evaluation approach, method and technique are often fuzzy and overlap with each other. Within the evaluation community, there is no agreed definition of these terms and of the distinctions between them, and it is not our intention to reach agreement on that. For the purposes of this tool, we have used certain definitions of our own that informed the selection of "methods" to be included.

For the specific purpose of this study, we define the three terms as follows:

- Technique is a procedure for data collection and/or analysis, and comes with quality criteria aimed at minimising researcher bias and maximising internal validity.
  - Examples of techniques: surveys, questionnaires, interviews, desk reviews, (critical) observation, Nvivo or similar, SPSS, other data processing software.
- Method is a short description of the process used to answer research questions, including but not limited to techniques; and can aim at external validity.
  - For example, difference in difference, propensity score matching, and the form taken by the theory of change (what elements need to be included). Examples of methods include Contribution Analysis (a causal chain with intermediate outcomes with risks and assumptions); realist evaluation (with context-mechanism-outcome (CMO) configurations); Systems-Based Evaluation (with representation of systemic relations between a variety of causal factors and intermediate outcomes, usually with loops and descriptions of the relations; or Agent-Based Modelling).
- Approach is broader than method (it can include method but is not limited to it) and describes or represents the ontological and/or epistemological foundations of the method. It is inspired by principles such as equity, justice, empowerment; but also the nature of causal inference, for causal questions. "Approach" sits either at the ontological level (about the "nature of reality") or at the normative one (about what is "right" and "valuable"), incorporating political or "boundary" considerations (e.g. whose perspectives are included).
  - Examples of approaches are the realist ontology (often combined with the realist method but not always); constructivism and pluralism (at the basis of many systems-based methods); those aspects of Critical Systems Heuristics exploring boundaries and power relations; the ideas behind capturing multiple perspectives with Soft Systems Methodologies; and for causal relations, models of causality and causal inference (Mill's Methods, Hume's account of causality, Mackie's INUS and SUIN causes, and generative or mechanism-based approaches—see also Befani, 2012).
  - "Approach" may also sometimes be used to emphasise a particular focus for the evaluation, such as gender- or conflict-sensitive evaluation or utilization-focused evaluation. These sorts of focus do not necessarily determine particular choices of methods, and may or may not be associated with some techniques or tools

Some "methods" (or "denominations" are underpinned by more abstract philosophical principles, while others come with precise indications on how to collect and analyse data (Table 2). For example:

- Realist Evaluation is both underpinned by an ontology (critical realism), produces a specific representation of the Theory of Change (CMO configurations) and provides a technique for data collection (the "realist interview"), even though is compatible with many others.
- QCA is based on configurational causality (approach), comes with a set of procedures that can answer different questions (method), and algorithms for data analysis (technique).
- Counterfactual-based methods stem from Mill's Method of Difference (approach); and comprise a series of different designs used to reconstruct the counterfactual in different ways (method).

Table 2 | locating "denominations" across approach, method and technique boundaries

	Approach	Method	Technique
1. Randomised Controlled Trials		X	
2. Difference in Differences		X	
3. Statistical Matching		X	
4. Instrumental Variables		X	
5. Outcome Mapping		X	X
6. Most Significant Change		X	X
7. Soft Systems Methodology	X	X	
8. Causal Loop Diagrams		X	
9. Participatory Systems Mapping	X	X	
10. Agent-Based Modelling		X	
11. Bayesian Belief Nets		X	
12. Realist Evaluation	X	X	X
13. Qualitative Comparative Analysis	X	X	X
14. Process Tracing and Bayesian Confidence Updating		X	X
15. Contribution Analysis	X	X	

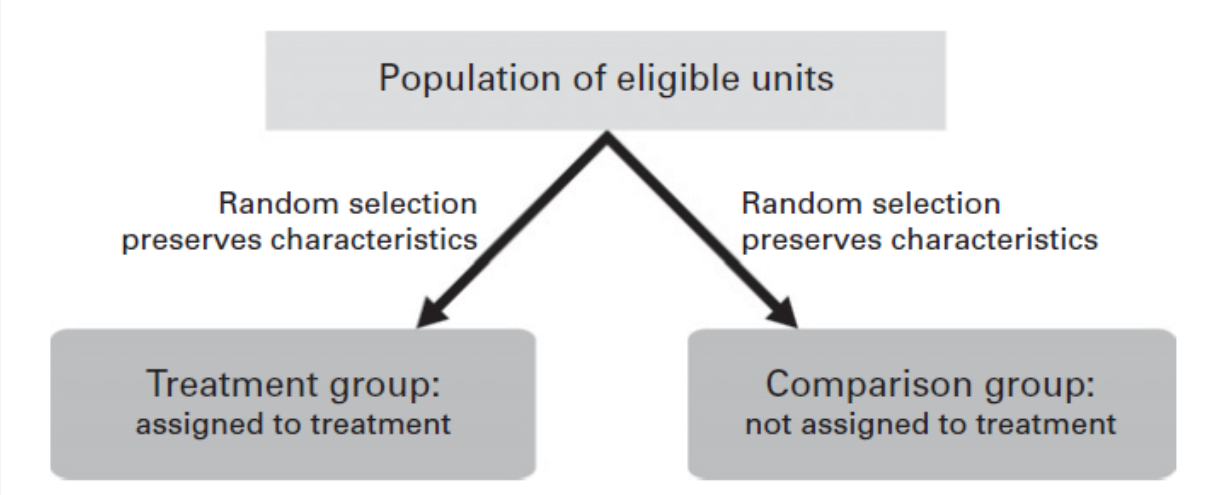
The "objects" handled in the tool are methods, which means that—even for those "denominations" crossing boundaries to approach and/or technique, we tried to focus on the "method" dimension. So, for example, when we mention "Realist Evaluation" we mostly mean a specific way of representing the theory of change, not the realist ontology per se nor the realist interview.

Annex 3 | Basic Characteristics of Methods

Annex 3.1 | Randomised Controlled Trials (RCTs)

The main purpose of Randomized Controlled Trials is to compare the outcome observed in the population exposed to the intervention to a counterfactual outcome, representing the alternative outcome that would have been achieved without the intervention (Figure 6). If the reconstruction of the non-intervention outcome is plausible, the difference between the observed outcome and the counterfactual outcome can be reliably taken to estimate the "net effect" or added value of the intervention (Figure 7).

Figure 6 | Logic of randomisation in RCTs

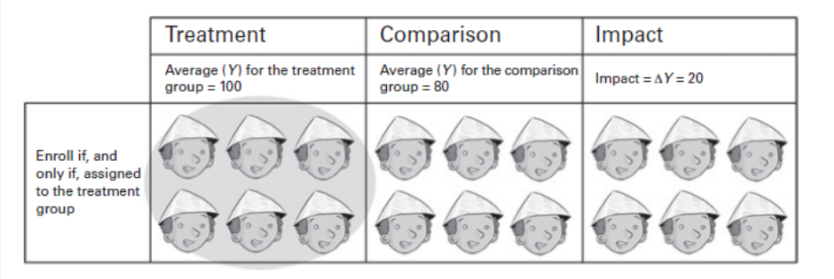


Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

In RCTs, the non-intervention outcome is estimated by designing a control group, which is virtually identical to the treatment group (figure 6). In addition, a series of precautions should be taken to maximize internal validity, like keeping the two groups separate in an attempt to prevent one influencing the other, and other strategies that protect against the development of different conditions in the two groups after group selection (differential attrition, etc. see Campbell, Donald T. Campbell, 1969, "Reforms as experiments" in American Psychologist, vol. 24 n. 4).

The similarity between treatment and control groups is guaranteed by randomization (figure 6): the treatment and control states are randomly assigned to the larger group of beneficiaries, all equally eligible for the treatment. Only a part of those eligible will then receive the treatment, while others will be part of the control group.

Figure 7 | how impact is estimated in RCTs



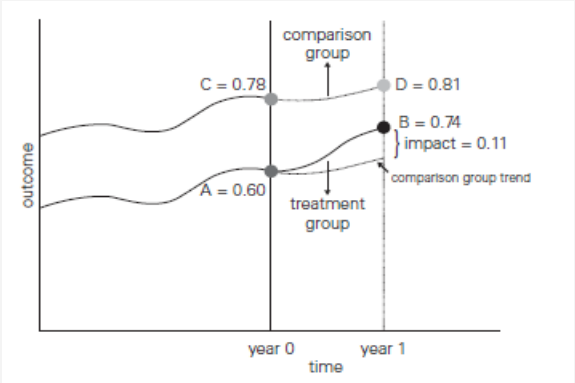
Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

Annex 3.2 | Difference in Differences

Quasi-Experiments encompass a wide range of counterfactual-based designs that, like RCTs, aim at reconstructing a control case providing a plausible estimate of the alternative, non-intervention outcome. However, unlike RCTs, quasi-experiments are observational studies and do not randomly assign exposure to treatment and control status; and most use previously existing theory on what factors affect the outcome in order to construct a plausible control, although this theory is mostly quite rudimentary and often synthesized with the probability of being assigned to the treatment group. The main variants use so-called Difference in Differences (DID), Pipeline methods, Statistical Matching (SM) , Interrupted Time Series (ITS) and Instrumental Variables (IV).

Difference-in-Difference (DID) (Figure 8) is possibly the variant making the most hypotheses: it assumes that treatment and control groups are subject to the same external influence (and internal dynamics) during treatment, and that the difference observed between the post-treatment time and the baseline in the control group faithfully represents how much the outcome would have changed in the treatment group without the intervention. In order to apply this variant, a lot needs to be known on what makes the two groups equivalent with respect to (factors influencing) the outcome.

Figure 8 | Representation of a DID design



Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

DID attempts to mimic an experimental research design using observational study data. It calculates the effect of a treatment (i.e., an explanatory variable or an independent variable) on an outcome (i.e., a response variable or dependent variable) by comparing the average change over time in the outcome variable for the treatment group to the average change over time for the control group. This method may be subject to certain biases (mean reversion bias, etc.), although it is intended to eliminate some of the effect of selection bias. In contrast to a within-subjects estimate of the treatment effect (which measures differences over time) or a between-subjects estimate of the treatment effect (which measures the difference between the treatment and control groups), the DID measures the difference in the differences between the treatment and control group over time" (from the Wikipedia Entry for "Difference in differences")

Annex 3.3 | Statistical Matching

Matching is a statistical technique that is used to evaluate the effect of a treatment by comparing the treated and the non-treated units in an observational study or quasi-experiment (i.e. when the treatment is not randomly assigned). The goal of matching is, for every treated unit, to find one (or more) non-treated unit(s) with similar observable characteristics against whom the effect of the treatment can be assessed (Figure 9). By matching treated units to similar non-treated units,

matching enables a comparison of outcomes among treated and non-treated units to estimate the effect of the treatment reducing bias due to confounding" (Wikipedia Entry for "Matching (statistics)")

Figure 9 | the logic of Statistical Matching

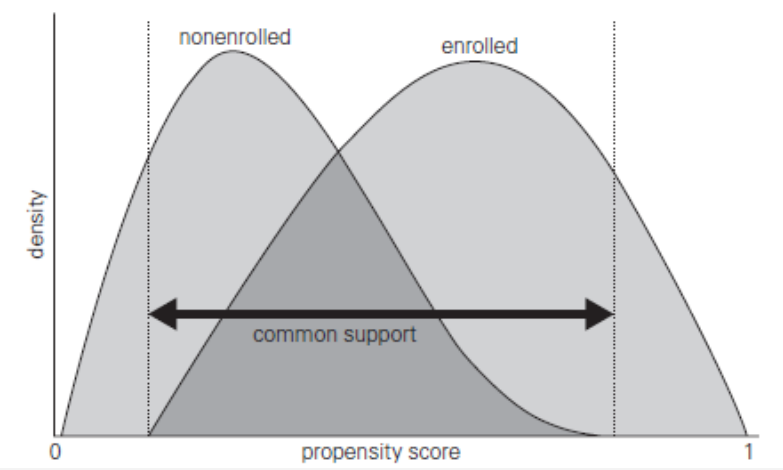
Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	2	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

In practice, statistical matching might be very complicated due to the high number of relevant factors involved. This problem is solved with a technique called "Propensity Score Matching", whereby an overall score is calculated, summarizing the (observable) characteristics of the population which are believed to "bias" the selection; namely because they affect the probability of participating to the intervention. The units are then matched based on their propensity to be enrolled. The "common support" (or overlapping) interval represents the range of propensity scores for which both enrolled and non-enrolled units are available, which constitutes the basis for the creation of treatment and control groups (Figure 10).

Figure 10 | illustration of how the treatment and control groups overlap in the "common support"



Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

Annex 3.4 | Instrumental Variables

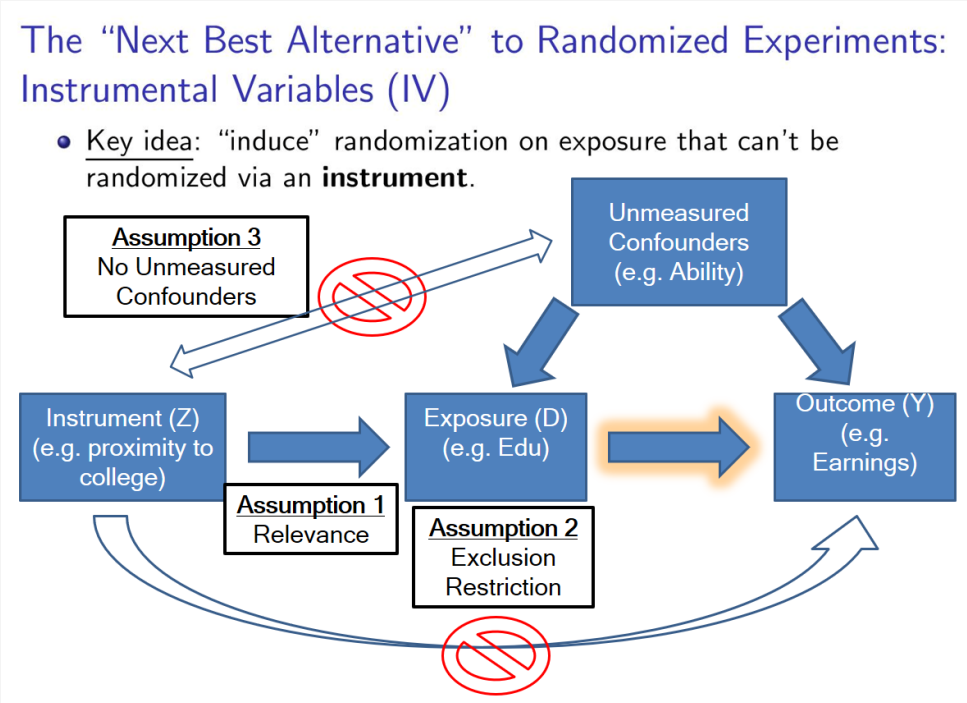
In statistics, econometrics, epidemiology and related disciplines, the method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Intuitively, IVs are used when an explanatory variable of interest is correlated with the error term, in which case ordinary least squares and ANOVA give biased results. A valid instrument induces changes in the explanatory variable but has no independent effect on the dependent variable, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable (figure 11).

In linear models, there are two main requirements for using IVs:

1. The instrument must be correlated with the endogenous explanatory variables, conditionally on the other covariates. If this correlation is strong, then the instrument is said to have a strong first stage. A weak correlation may provide misleading inferences about parameter estimates and standard errors.
2. The instrument cannot be correlated with the error term in the explanatory equation, conditionally on the other covariates. In other words, the instrument cannot suffer from the same problem as the original predicting variable. If this condition is met, then the instrument is said to satisfy the exclusion restriction.")"

(source: https://en.wikipedia.org/wiki/Instrumental_variables_estimation)

Figure 11 | visual representation of a regression model using an instrument



Source: https://www.r-bloggers.com/instrumental-variables

Annex 3.5 | Outcome Mapping

Outcome Mapping is a participatory approach to planning, monitoring and evaluation. It was designed for programme managers and practitioners to build evaluative thinking into implementation and focuses on results within the programme's sphere of influence. It is especially useful when the sphere of influence is significant and complex (e.g. not for programmes with direct control over outcomes, nor for programmes with little influence on outcomes).

Outcome Mapping is more appropriate for developmental and formative purposes than for summative. It emphasises the importance of behaviour change in all processes of development and focusses its attention on understanding this and on outlining different programme priorities, goals and activities in relation to a number of different boundary partners (actors with which the programme interacts, directly and indirectly), setting out how to promote progress towards anticipated results. It consists mainly of two phases, a design phase and a record-keeping phase. For the purposes of this project, the Design Phase is the most important one.

The design stage is an approach for developing a theory of change which is actor-centred and focussed on behavioural change. It is about clarifying the intent of the programme and expressing it in a way which makes it easier and more systematic to monitor. It involves setting monitoring priorities and creating a framework for data collection. Project leaders draw a map of which parties will likely be influenced by the project in any way (direct boundary partners), and which parties will in turn be influenced by those parties (indirect boundary partners). Project leaders select three or four "primary" boundary partners upon which they focus additional activities (e.g. the direct recipients or beneficiaries of the project's deliverables).

For each primary boundary partner, project leaders write a statement of desired overall behavioural change (called an outcome challenge) and a list of specific behavioural changes or actions the project would like the boundary partner to exhibit by the end of the project (called progress markers). There are three types of progress markers, namely expect-to-see, like-to-see and love-to-see. These mark progress from simple behaviours which are reactive to the programme activities (expect to see) to behaviours which are the boundary partners own initiative (like to see), to more complex, transformative change (love to see).

Earl, S., Carden, F., & Smutylo, F. (2001). Outcome mapping: building learning and reflection into development programs. Ottawa, ON: International Development Research Center. Accessible from: https://www.idrc.ca/en/book/outcome-mapping-building-learning-and-reflection-development-programs

Smutylo, T. (2005). Outcome mapping: A method for tracking behavioral changes in development programs. ILAC Brief 7. Accessible from: https://cgspace.cgiar.org/bitstream/handle/10568/70174/ILAC_Brief07_mapping.pdf?sequence=1&isAllowed=y

Jones, H., & Hearn, S. (2009). Outcome mapping: A realistic alternative for planning, monitoring, and evaluation. Accessible from: https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/5058.pdf

Annex 3.6 | Most Significant Change

MSC was first developed to help NGOs monitor the impacts of participatory development projects. It is flexible enough to identify a diversity of development outcomes across a variety of locations and emphasises the need to respect participants' own judgement regarding the changes that an initiative has made to their lives (Davies, 1998). Davies and Dart set out clear steps for using the approach in their MSC guide (Davies & Dart, 2005).

The central element of MSC involves the systematic collection and selection of a sample of significant change stories. The stories themselves are elicited from programme participants by asking them to relate what significant changes (positive or negative) have occurred in their lives in the recent past, and enquiring why they think that these changes occurred and why they regard them as being significant. Stories can be written down or video- or audio- recorded and can be obtained through interviews or group discussions or can simply be written reports from field staff.

"It is participatory because project stakeholders are involved in deciding the sorts of changes or stories of significant change to be recorded and in analysing the data collected. It is a form of monitoring because it occurs throughout the programme cycle and provides information to help people manage the programme. It contributes to evaluation by providing data on short-term and long-term outcomes that can be used to help assess and improve the performance of the programme as a whole" (Davies and & Dart, 2005).

A key step in MSC is the process by which the most significant of the "significant change stories" are selected. After stories of significant change have been collected, they are then subject to a structured selection process involving panels of stakeholders. How this is designed will depend on which stakeholder views need to be solicited and used. Panels of designated stakeholders systematically review the stories. The intention is for stakeholders to engage in in-depth discussion at each stage of the selection process regarding the significance of each story, the wider implications of the changes that they relate, and the quality of evidence that they contain. The use of multiple levels of selection enables large numbers of significant change stories to be reduced to a smaller number of stories viewed as being most significant by a majority of stakeholders. It also allows comparisons of stories coming from different locations and/or stakeholder groups. Selection is important primarily because it involves forced choices and attention to underlying values that might help make such choices: much of this process is about values clarification.

MSC was originally developed as an approach for impact monitoring, rather than as an evaluation approach designed to generate summative statements about aggregate change. As an approach to impact monitoring, it is designed to report on the diverse impacts that can result from a development programme and participants' perceived significance of these changes. It is intended to be an ongoing process occurring at regular intervals during the programme cycle, with the information gathered fed back into the programme to improve its management and running.

MSC has since been adapted for use in impact evaluations, by expanding the scale of story collection, extending the range of stakeholders engaged in story selection, and using it alongside other evaluation methods, such as in tandem with a log-frame/theory-of-change approach (for example, Van Ongevalle et al., 2012). The stories of significant change which MSC generates can provide useful sources of information for the specification and subsequent assessment of a theory of change.

Annex 3.7 | Soft Systems Methodology

SSM is about solving problems through the comparison of different perspectives of how systems work (Figure 12). (Checkland & Scholes, 1999; Checkland and Poulter 2006) proposed to use SSM when a problematic situation that people are trying to improve is perceived differently by people with different worldviews. The idea is to map out one system for each perspective, and then use the comparison of these systems to stimulate a discussion about which changes are both desirable and culturally feasible (Williams & Hummelbrunner 2010) according to stakeholders with different worldviews, eventually/hopefully identifying a framework which is compatible with all.

Checkland describes the approach in the following way (Checkland & Poulter 2006):

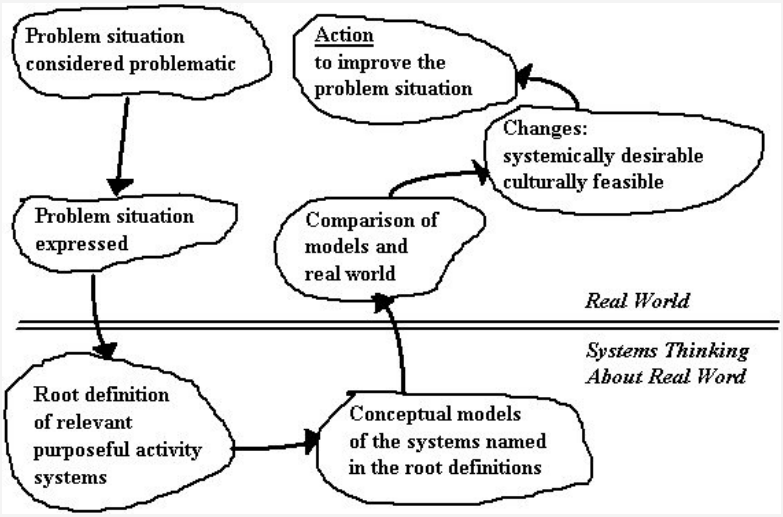
1. You have a perceived problematic situation that;
   a. will contain people trying to behave purposefully
   b. will be perceived differently by people with different worldviews
2. So make models of purposeful activity as perceived by different worldviews
3. Use these models as a source of questions to ask of the problematical situation: thus structuring a discussion about changes which are both desirable and culturally feasible
4. Find versions of the to-be-changed situation which different worldviews could live with
5. And implement changes to improve the situation
6. Be prepared to start the process again.

Checkland (1993) also developed a mnemonic checklist (the CATWOE) to help guide the process. Following CATWOE, every systemic perspective needs to include information on:

- WHO benefits from the problem being solved (**C**ustomers)
- WHO provides the enabling environment for the problem to be solved (**A**ctors)
- WHAT changes (**T**ransformation)
- WHAT **W**orldview/value basis makes solving the problem important/meaningful
- WHO **O**wns the systems and controls its existence (e.g., holders of key resources, elected representatives, sponsors)
- Context/**E**nvironment: important factors that must be taken as "given"

Figure 12 | the 7-stage model of SSM

(adapted from Checkland, available at http://cci.drexel.edu/faculty/sgasson/ssm/Process.html)



Annex 3.8 | Causal Loop Diagrams

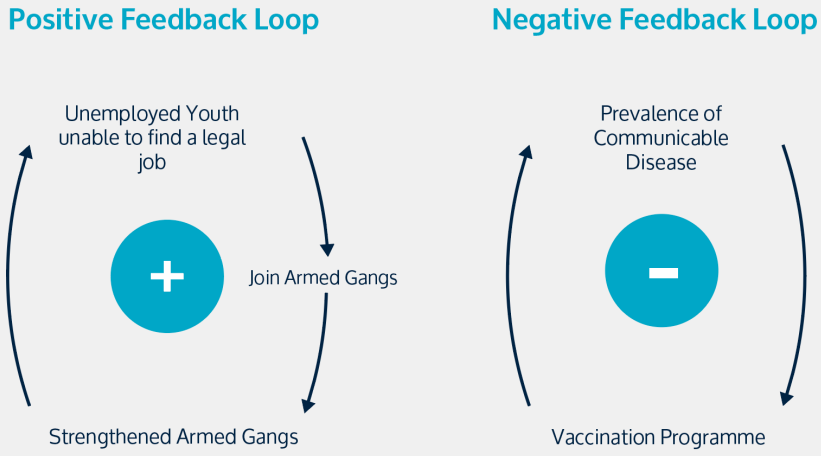
Causal loop diagrams are tools that test the assumptions of linearity and independence behind causal relationships. In an independent, linear relationship the causal factor x contributes to the effect e; but e does not affect the state of x. In many real situations, however, the effect also reinforces or balances the cause: there is often a double-loop relationship between cause and effect where a change in the effect influences the state of the cause of that effect.

For example, unemployment may lead some young people in poor, underserviced neighbourhoods to join armed gangs (Lind & Mitchell, 2013) (Lind and Mitchell 2013). This will strengthen the ranks of local gangs and reinforce their control of the territory, further reducing opportunities of legal employment for young people (see Figure 13). Another example is unpaid care (Chopra, 2015)

(Chopra 2013Chopra 2015). In cultures where women are expected to perform care duties towards other family members, women who cannot enter the labour market cannot pay others to perform those duties and are forced to spend the majority of their time in unpaid care activities. This further reduces their opportunities to get a job. Yet another, perhaps classic example is the relation between policy change and cultural change: policy formation is influenced by culture and policies are usually compatible with ingrained attitudes and beliefs; but policy change also contributes to cultural change.

The above examples refer to positive causal loops that give rise to reinforcing dynamics, where more of the effect contributes to more of the cause. However, loops can also be negative, originating so-called balancing dynamics, where more of the effect contributes to less of the cause. For example, effective vaccination policies that succeed in eradicating the infective disease that originated them, lose relevance with time when there is no longer need to fight the disease (Figure 13).

Figure 13 | Positive and negative feedback loops



Annex 3.9 | Participatory Systems Mapping

Overview

The participatory systems mapping method involves teams of people collaboratively constructing a causal map of their system (e.g. policy area, industry) of interest. At first, they do this in a workshop setting, around a table with post-its, 'white-board paper', and pens, before the map is converted into a digital format for analysis and sharing. An example map is shown below (Figure 14).

The map is made up of 'factors' and their causal connections. Factors can represent anything as long as they are expressed as variables (i.e. they can go up and down). Connections represent causal relationships, in a mathematical sense either: positive (i.e. an increase in one factor causes an increase in the next, or a decrease in one factor causes a decrease in the next); negative (i.e. an increase in one factor causes a decrease in the next, or a decrease in one factor causes an increase in the next); unclear (i.e. we believe there is a causal relationship but we are unsure of its nature); or complex (i.e. the relationship depends on other third party factors, or is non-linear).

The map of factors and connections is built in a workshop setting, guided by a facilitator, with typically no more than twelve people in anyone mapping session. The map produced is an intersubjective object, it reflects the beliefs of the group of people that built it. It should not be

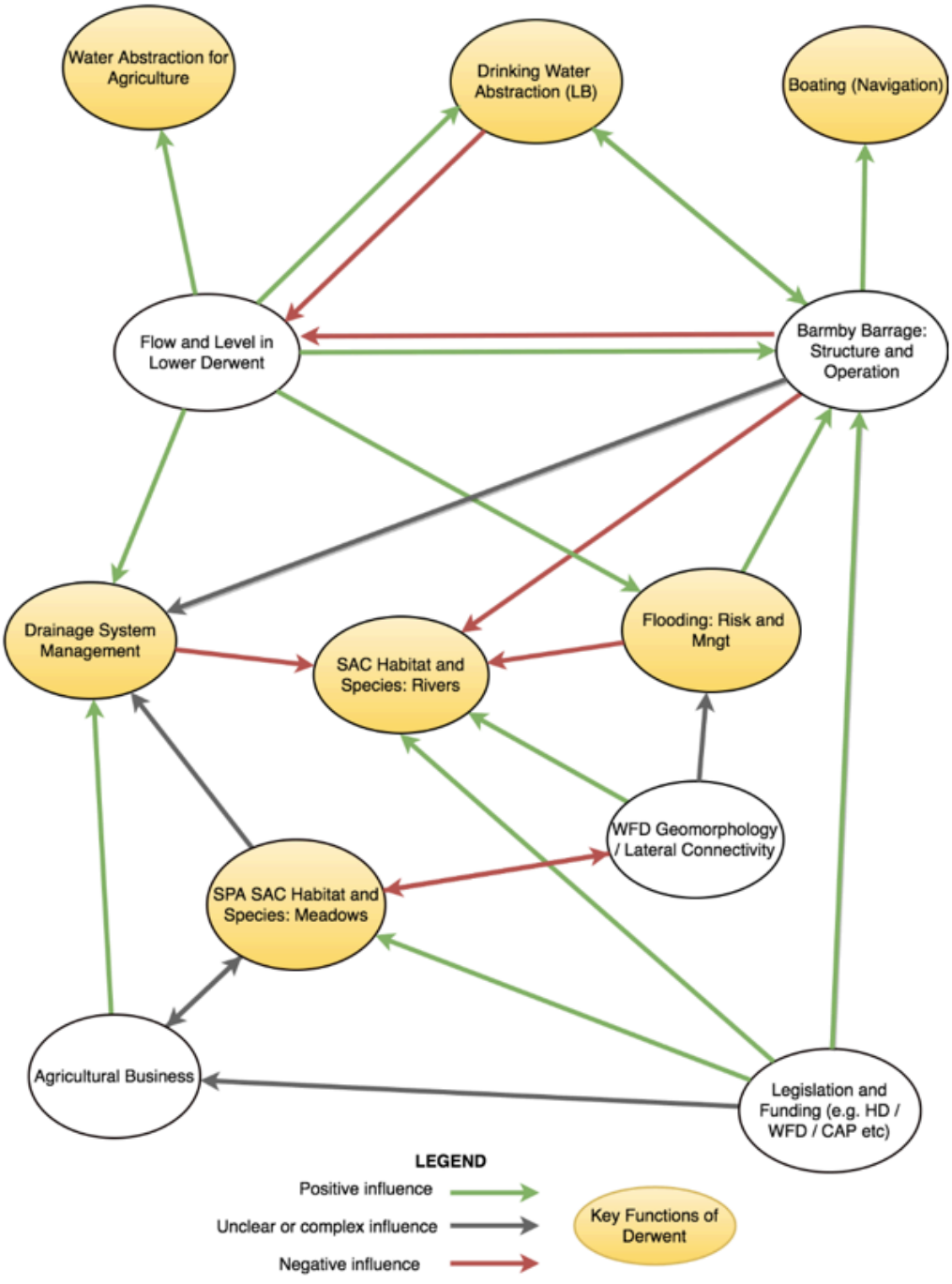
assumed to be objective or comprehensive. A mapping process can give great value to those involved in its creation as the act of building a map often leads to important conversations and the development of shared understandings and consensus.

The map can be digitised in many ways, using software as 'simple' as Microsoft PowerPoint, all the way through to specialist network software; we recommend Draw.io for ease of use and quick visualisations, and Gephi for network analysis and large maps.

Once digitised, the map can be analysed in a number of ways. In all the approaches to analysis described here, the participatory nature of the analysis is key; the emphasis is on using the map and associated analysis as a discussion and thinking tool with a varied group of stakeholders, in cycles of map building, analysis, and reflection; rather than as a model to generate 'answers'. Analysis options include:

- Network analysis: Standard network analysis (e.g. degree, betweenness, centrality) can be conducted on a map to aid reflection on the nature of the map stakeholders have built.
- Combined network structure and subjective factor information: The network structure of the map can also be combined with subjective information from participants about the nature of factors, to analyse the map. For example, we may look at subsets of the map which are 'downstream' (i.e. in a causal sense) from factors that are controllable (from a stakeholder perspective), we may look for factors that we control (e.g. policy interventions) and whether any of them contradict or complement one another's influence on the map, or we may look at those factors which are 'upstream' of a factor which we believe is a key outcome or function of the system.
- Interrogating the whole map: The whole map can also be used to facilitate discussion amongst groups, or be used to search for new research and evaluation questions. For example, the map can be laid out in different ways (e.g. as originally laid out when built, in a left to right form with factors with high out-degree on the left, and high in-degree on the right), to see if this prompts stakeholders to see new things. The map can also be presented by stakeholders to other stakeholders, as way to lead into discussion and information sharing.
- Workshop observation: The discussion in the map building workshop itself can also be observed with a note taker present or recording taken; this is a valuable piece of data for analysis to generate understanding of discussion and points of conflict and consensus.
- Scenarios: The map can also be used to run through potential scenarios of change. For example, if there is a factor stakeholders have high control over, we may explore how a range of different changes in that factor affect the rest of the map, using the map to walk through these changes step by step.

Figure 14 | An example participatory map showing the causal connections in a river catchment in the UK



Annex 3.10 | Agent-Based Modelling (ABM)

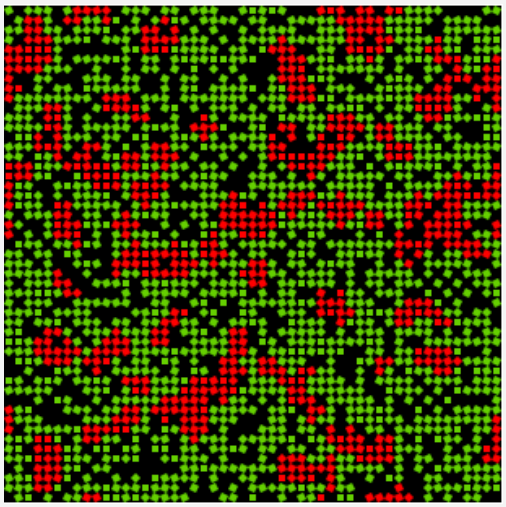
An agent-based model (ABM) consists of a number of software objects, the 'agents', interacting within a virtual environment. The agents are programmed to have a degree of autonomy, to react to and act on their environment and on other agents, and to have goals that they aim to satisfy. In such models, the agents can have a one-to-one correspondence with the individuals (or

organisations, or other actors) that exist in the real social world that is being modelled, while the interactions between the agents can likewise correspond to the interactions between the real world actors. With such a model, it is possible to initialise the virtual world to a preset arrangement and then let the model run and observe its behaviour. Emergent patterns of action (e.g. 'institutions') may become apparent from observing the simulation.

Agents are generally programmed using either an object-oriented programming language or a special-purpose simulation library or modelling environment, (e.g. NetLogo) and are constructed using collections of condition-action rules to be able to 'perceive' and 'react' to their situation, to pursue the goals they are given, and to interact with other agents, for example by sending them messages. Agent-based models have been used to investigate the bases of leadership, the functions of norms, the implications of environmental change on organizations, the effects of land-use planning constraints on populations, the evolution of language, and many other topics.

The Schelling model is an early and simple example of an ABM. Schelling (1971) used a simulation to show that high degrees of residential segregation could occur even when individuals were prepared to have a majority of people of different ethnicity living in their neighbourhood.

Figure 15 | The pattern of clusters that emerge from Schelling's model



Schelling modelled a neighbourhood in which homes were represented by squares on a grid. Each grid square was occupied by one simulated household (in Figure 15, either a green or a red household), or was unoccupied (black). When the simulation is run, each simulated household in turn looks at its eight neighbouring grid squares to see how many neighbours are of its own colour and how many of the other colour. If the number of neighbours of the same colour is not sufficiently high (for example, if there are fewer than three neighbours of its own colour), the household 'moves' to a randomly chosen unoccupied square elsewhere on the grid. Then the next household considers its neighbours and so on, until every household comes to rest at a spot where it is content with the balance of colours of its neighbours.

Schelling noted that when the simulation reaches a stopping point, where households no longer wish to move, there is always a pattern of clusters of adjacent households of the same colour. He proposed that this simulation mimicked the behaviour of whites fleeing from predominantly black neighbourhoods, and observed from his experiments with the simulation that even when whites were content to live in locations where black neighbours were the majority, the clustering still

developed: residential segregation could occur even when households were prepared to live among those of the other colour.

While most agent-based simulations have been created to model real social phenomena, it is also possible to model situations that could not exist in our world, in order to understand whether there are universal constraints on the possibility of social life. (For example, can societies function if their members are entirely self-interested and rational?) These are at one end of a spectrum of simulations ranging from those of entirely imaginary societies to those that aim to reproduce specific settings in detail.

Annex 3.11 | Bayesian Belief Networks

In an evaluation context, a Bayesian belief network (BBN) can be seen essentially as a graphical model of a chain of causal interactions, represented as a network of nodes that are linked by probabilities. The nodes represent factors that affect outcome(s) of interest to the evaluation, and the links represent how a previous factor affects the next. A BBN is an acyclic graph, that is, a network with no feedback loops, where the "predictor" nodes are direct or indirect causal factors of the outcome variable(s).

A BBN has a qualitative and quantitative element. The qualitative element relates to its structure, which involves mapping the factors considered relevant to the outcome(s) of interest and the dependencies or links between them, (including the order or direction of causality). Depending on exact context, this may be equivalent to graphically representing the theory-of-change for the intervention being evaluated. The quantitative element is the inclusion of probabilities that quantify the relationships between the factors. Probabilities need only be specified for factors that are linked (i.e. direct relationship). Probabilities that quantify the relationship between a factor and its indirect causes or effects are computed automatically by inference algorithms. This makes a BBN efficient in terms of the data required to populate the network and a very powerful reasoning tool for evaluation purposes.

Depending on the context, probabilities may be derived from observed ('objective') data about the relationship between factors. More commonly, they are subjective probabilities reflecting the beliefs of key informants. A BBN can, however, accommodate both objective and subjective data.

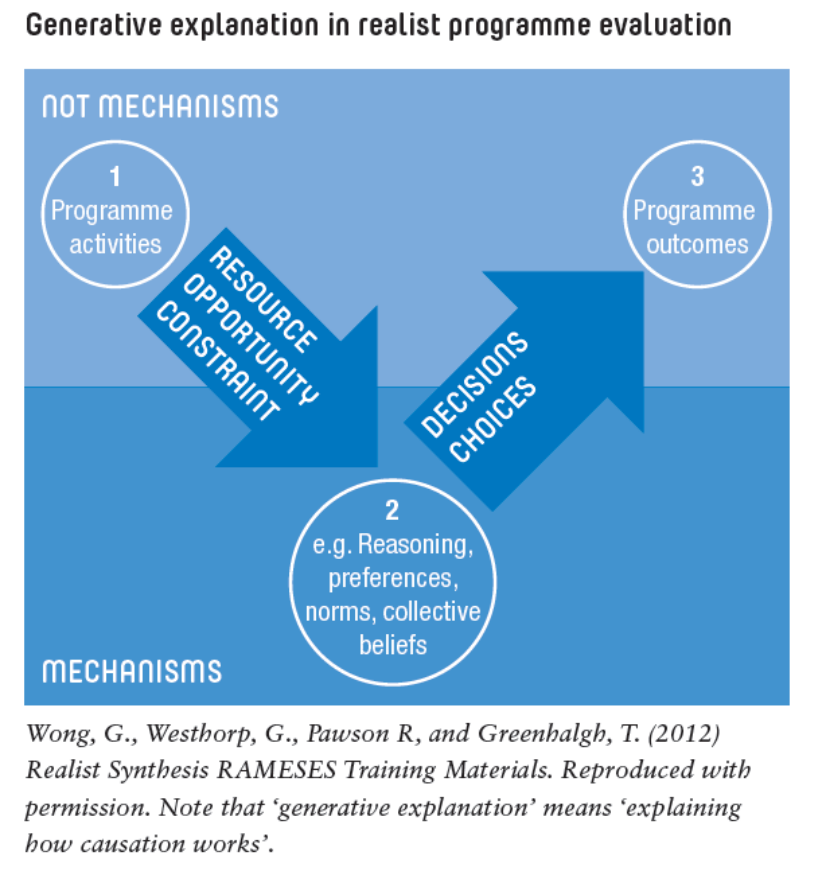
BBN is a method that is very useful for theory-based evaluations, when (a) the cause-and-effect being modelled is subject to significant uncertainty and hence must be described probabilistically; and (b) data on the causal relationship between factors of interest are not readily available and hence significant reliance must be placed on informants' views.

Annex 3.12 | 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 (Figure 16).

Figure 16 | Generative explanation in realist programme evaluation

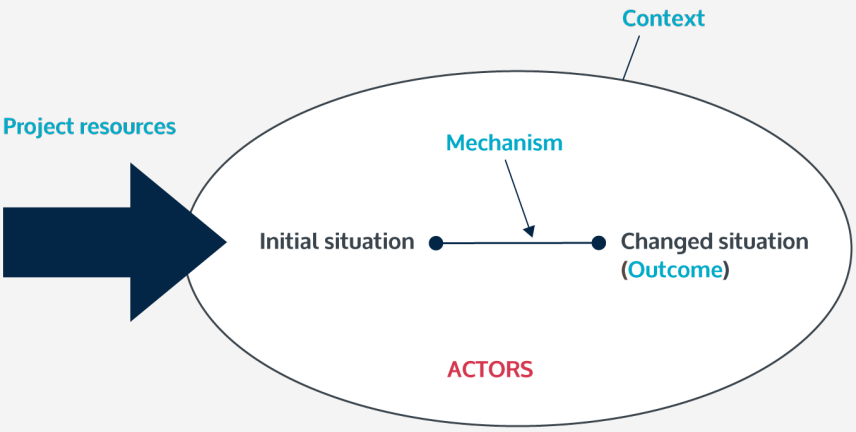


(Reproduced from Gill Westhorp (2014) Realist Evaluation: An Introduction. Methods Lab Overseas Development Institute London (Westhorp, 2014))

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 individuals take based on the resources available in their context; and outcomes are the product of individuals' behaviour and choices. CMO configurations are often represented with the "realist egg" (Figure 17).

For this particular report, we will not consider realist evaluation as an approach (or an ontology/philosophy); we will focus on its methodological aspects.

Figure 17 | the realist "egg"



Annex 3.13 | Qualitative Comparative Analysis (QCA)

[From Befani 2016]

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" (de Meur, Rihoux, & Yamasaki, 2002) because of its mathematical grounding.

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, 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, 2013). Historically, the method has always been very popular with political scientists and other scholars interested in cross-country generalisation.

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 labour 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 "variables" to emphasise the distinction between QCA and statistical methods.

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. 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 for the outcome to occur and which "recipes" or combinations of factors are most likely to "trigger" the outcome.

Annex 3.14 | Process Tracing/Bayesian Updating

[From Befani & Stedman-Bryce, 2016]

Process Tracing has been referred to as a method (Collier, 2011; Beach & Pedersen, 2013) but also as a tool (Collier, 2011; Bennett, 2010) and a technique (Bennett & Checkel, 2014) for data collection and analysis. This reflects its focus on theory development as much as on the search and assessment of evidence for a causal explanation (also reflected in the distinction between the two "deductive" and "inductive" variants (Beach & Pedersen, 2011; Bennett & Checkel, 2014)). Its purpose is to draw causal inferences from 'historical cases', broadly intended as explanations of past events. It is based on a

mechanistic understanding of causality in social realities, and starts from the reconstruction of a causal process intervening between an independent variable and an outcome, which could for example be a Theory of Change, a complex mechanism or a CMO configuration.

The method operates a clear distinction between:

a. the process described in the Theory of Change, considered a possible "reality", or an ontological entity which might or might not exist or have materialised; which is usually unobservable;
b. the evaluator's hypothesis on the existence of that reality (which is an idea in 'our head' (Bennett & Checkel, 2014) rather than a reality "out there"; and
c. the observable and therefore testable implications of the existence of such reality.

This tripartite conceptual framework stems from the awareness that mechanisms in the social sciences are usually not directly observable; we can never attain perfect certainty of their existence but nevertheless we formulate hypotheses about their existence and look for evidence in an attempt to increase or decrease our confidence in such hypotheses. Put differently, the aspiration of Process Tracing is to minimise the inferential error we risk making when producing statements about an ontological causal reality.

The backward perspective takes advantage of the fact that, at the time of the investigation, the mechanism has presumably had enough time to leave traces, which are able to provide a strong indication of its existence. Process Tracing recognises that not all these traces are equally informative, and as a consequence focuses on assessing the quality, strength, power, or probative value that select pieces of evidence hold in support of (or against) the causal mechanism.

One of its advantages is that it allows a clear distinction between 'absence of evidence', which has no inferential power and does not add any value to what the researcher already knows, and 'evidence of absence' which on the contrary can strongly challenge a hypothesis, if it contradicts observable implications stemming from such hypothesis.

In Process Tracing, four well-known metaphors are often used to describe the different ways evidence affects our confidence about a certain mechanism or Theory of Change: the Hoop test, the Smoking Gun test, the Straw-in-the-Wind test and the Doubly-Decisive test (Bennett, 2010; Van Evera, 1997). See Box 4 for the properties of these tests.

Box 4 | the Four Process Tracing tests and their properties

Smoking Gun (confirmatory): If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is not confirmed; but this is not sufficient to reject the hypothesis.

Hoop Test (disconfirmatory): If the evidence is not observed, the hypothesis is rejected. If the evidence is observed, the hypothesis is not rejected (it "goes through the hoop", passes the test); but this is not sufficient to confirm the hypothesis.

Doubly Decisive (both confirmatory and disconfirmatory): If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is rejected.

Straw-in-the-Wind (neither confirmatory nor disconfirmatory): If the evidence is observed, this is not sufficient to confirm the hypothesis. If the evidence is not observed, this is not sufficient to reject the hypothesis.

One possibility offered by Process Tracing, attempted in social science (Fairfield, 2015), law (Friedman, 1986; Kaye, 1986; Edwards, 1986), and most recently also in evaluation (Befani 2020a), is its combination with a rigorous mathematical formalisation. While the concepts of Process Tracing can be modelled with different mathematical concepts and tools, one branch of mathematics we find very useful in connection with the method is Bayesian Updating—(see also (Bennett, 2008; Beach & Pedersen, 2013; Bennett, 2014; Befani & Mayne, 2014). This is also referred to "Bayesian Confidence Updating", "Contribution Tracing", or—when it's applied more generally to Theory-Based Evaluation, "Diagnostic Evaluation". In this formalisation, the Process Tracing tests in Box 4 are linked to the Confusion Matrix used for diagnosis (Befani 2020) and the inferential power or probative value of a piece of evidence E for a theory T can be measured in ways (Kaye, 1986; Bennett, 2014; Friedman, 1986), that are related to the difference between the true positives rate or "sensitivity" and the false positives rate or "Type I error". The larger the difference between the true positives rate and the false positives rate, the higher the probative value of evidence E for theory T (see also (Befani, D'Errico, Booker, & Giuliani, 2016—note that these values also appear in the Bayes formula, used to update the level of confidence that the theory is true, from before to after observation of empirical evidence (from the "prior" to the "posterior" confidence)).

Intuitively, this means that if an observed piece of evidence has a higher chance of being observed if theory T holds true (sensitivity), than if theory T does not (Type I error), this will strengthen the theory. If the opposite is true, and the evidence has a higher chance of being observed if the theory does not hold, compared to if the theory holds, observation of that evidence will weaken the theory. Finally, if the evidence has a similar chance of being observed whether the theory holds true or not (sensitivity is roughly the same as Type I error), observing it will not significantly alter our confidence in the theory.

In Bayesian confidence updating, different pieces of evidence have different values of sensitivity and specificity, hence different likelihood ratios, and thus different abilities to alter the evaluator's initial confidence in the Contribution Claim. The evaluator is thus forced to be transparent about their assumptions affecting confidence that the claim is true, and to 'declare' its observable implications: if the claim holds true—or doesn't—what should I expect to observe? With what probability? Making these assumptions—mostly left out or at best left implicit with other methods—transparent means making them open to scrutiny; if they aren't challenged, this will increase their legitimacy and credibility. Bayesian Updating potentially presents additional challenges when dealing with multiple pieces of evidence; suggestions on how to deal with these are provided in Befani, 2020b.

Annex 3.15 | Contribution Analysis

[From Befani & Mayne 2014]

Contribution Analysis (Mayne 2001, 2008, 2012) is based on a theory of change for the intervention being examined in detail. Depending on the situation, the theory of change may be based on the expectations of the funders, the understandings of those managing the intervention, the experiences of the beneficiaries and/or prior research and evaluation findings. The theory of change may be developed during the planning for the intervention—the ideal approach—and then revised as implementation occurs or be built retrospectively at the time of an evaluation. Good practice is to make use as much as possible of prior research on similar interventions. The analysis undertaken examines and tests the theory of change against logic and the data available from results observed and the various assumptions behind the theory of change, and examines other influencing factors. The analysis either confirms the postulated theory of change or suggests revisions in it where the reality appears otherwise. The overall aim is to reduce uncertainty about the contribution an intervention is making to observed results through an increased understanding of why results did or did not occur and the roles played by the intervention and other influencing factors.

Six key steps in undertaking a CA are set out as shown in Box 5, adapted and expanded from Mayne 2012a and Mayne 2011. These steps are often part of an iterative approach to building the argument for claiming that the intervention made a contribution and exploring why or why not.

CA argues that if one can verify or confirm a theory of change with empirical evidence—that is, verify that the steps and assumptions in the intervention theory of change were realized in practice, and account for other major influencing factors—then it is reasonable to conclude that the intervention in question has made a difference, i.e., was a contributory cause for the outcome. The theory of change provides the framework for the argument that the intervention is making a difference, and the analysis identifies weaknesses in the argument and hence where evidence for strengthening such claims is most needed.

Causality is inferred from the following conditions and evidence illustrated in Box 5.

Box 5 | Key Steps in Contribution Analysis

Step 1 | Set out the cause-effect issue to be addressed

- Acknowledge the causal problem for the intervention in question
- Scope the problem: determine the specific causal question being addressed; determine the level of confidence needed in answering the question
- Explore the nature and extent of the contribution expected from the intervention
- Determine the other key factors that might influence the realization of the results
- Assess the plausibility of the expected contribution given the intervention size and reach

Step 2 | Develop the postulated theory of change and risks to it, including other influencing factors

- From intervention documents, interviews and relevant prior research, develop the postulated theory of change of the intervention, including identifying the assumptions and risks for the causal links in the theory of change
- Identify the roles other key influencing factors may play in the theory of change
- Determine how contested is the postulated theory of change to better understand the strength of evidence needed

Step 3 | Gather the existing evidence on the theory of change

- Gather the evidence that exists from previous measurement, past evaluations, and relevant research to assess the likelihood (1) of the expected results, assumptions and risk being realized, (2) for each of the causal links in the results chain occurring, and (3) for the other influencing factors making a significant difference.

Step 4 | Assemble and assess the contribution claim, and challenges to it

- Set out the contribution 'story' on the likelihood that the intervention 'worked': the causal claim based on the analysis of logic and evidence so far
- Assess the strengths and weaknesses in the postulated theory of change in light of the available evidence, and the relevance of the other influencing factors; which links seem reasonable and which look weak and need more evidence
- If needed, refine or update the theory of change

Step 5 | Gather new evidence from the implementation of the intervention

- With a focus on the identified weaknesses, gather data on the ToC results that occurred, the assumptions and risks associated with the causal links and the other identified influencing factors

Step 6 | Revise and strengthen the contribution story

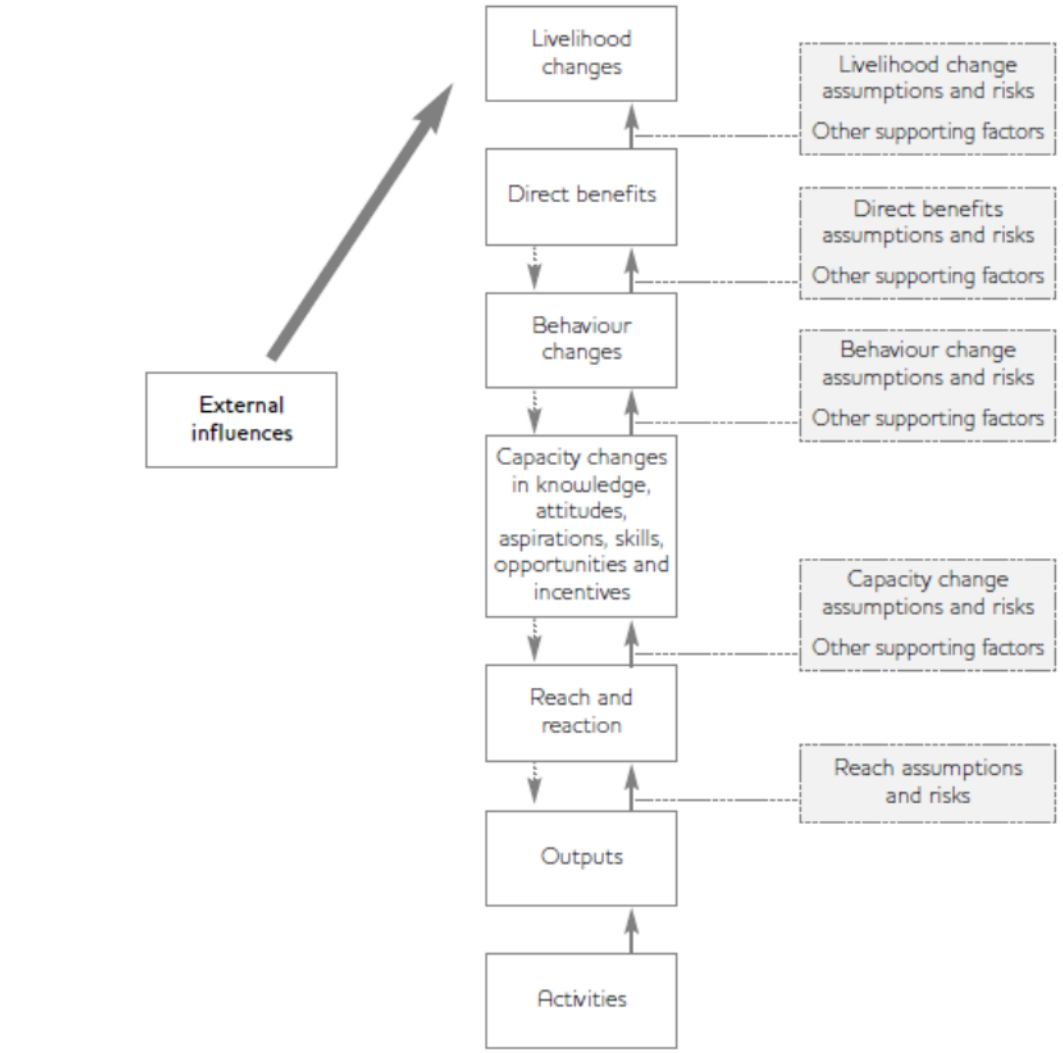
- Build a more credible contribution claim based on the new data gathered
- Reassess its strengths and weaknesses, i.e., the extent to which the results, assumptions/risks and other influencing factors occurred
- Conclude on the strength of the ToC and the role played by other influencing factors, and hence on the contribution claim
- If the evidence still is weak, revisit Step 5

In the end, conclusions are reached—a contribution claim about whether the intervention made a difference, and on how the results were realized.

In Contribution Analysis, the Theory of Change is represented as a series of intermediate outcomes, linked by assumptions that need to hold and risks that need to be avoided, for the process of change to be able to progress to the next step (figure 18).

Figure 18: Representation of a Theory of Change in Contribution Analysis

Figure 1 **Basic generic intervention ToC**



Source Authors' own.

(reproduced from Befani & Mayne, 2014)

References

Baumgartner, M. (2012). Detecting Causal Chains in Small-n Data. Field Methods, 25(1), 3-24.

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—Annex to Stern et al., 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). Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA). Stockholm: EBA.

Befani, B. (2020a). Quality of quality: A diagnostic approach to qualitative evaluation. Evaluation, 26(3), 333–349. https://doi.org/10.1177/1356389019898223

Befani, B. (2020b). Diagnostic Evaluation and Bayesian Updating: Practical Solutions to Common Problems. Evaluation. 2020;26(4):499-515. doi:10.1177/1356389020958213

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) Available "Online First" at http://evi.sagepub.com/content/early/2016/06/24/1356389016654584.abstract

Befani, B., Barnett, C., & Stern, E. (2014). Introduction: Rethinking Impact Evaluation for Development. IDS Bulletin, 45(6), 1-5.

Befani, B., D'Errico, S., Booker, F., & Giuliani, A. (2016). Clearing the fog: new tools for improving the credibility of impact claims. IIED Briefing. London: International Institute for Environment and Development. Available at: http://pubs.iied.org/17359IIED.html

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. (2008). Process Tracing: a Bayesian Perspective. In J. Box-Steffensmeier, H. Brady, & D. Collier, The Oxford Handbook of Political Methodology. OUP.

Bennett, A. (2010). Process Tracing and Causal Inference. In H. Brady, & D. Collier, Rethinking Social Inquiry. Rowman and Littlefield.

Bennett, A. (2014). Disciplining our conjectures. In A. Bennett & J. Checkel (Eds.), Process Tracing: From Metaphor to Analytic Tool (Strategies for Social Inquiry, pp. 276-298). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139858472.015

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.

Berg-Schlosser, D., De Meur, G., Rihoux, B., & Ragin, C. (2009). Qualitative Comparative Analysis (QCA) As An Approach. In 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.

Bryman, A. (2012). Social Research Methods. 4th Edition. Oxford University Press: Oxford

Campbell, D. (1969). Reforms as experiments. American Psychologist, 24, 409-429.

Caren, N., & Panofsky, A. (2005). TQCA: A Technique for Adding Temporality to Qualitative Comparative Analysis. Sociological Methods & Research, 34(2), 147-172. https://doi.org/10.1177/0049124105277197 page 42

Checkland, P. and Holwell, S. (1993), Information management and organizational processes: an approach through soft systems methodology. Information Systems Journal, 3: 3-16. doi:10.1111/j.1365-2575.1993.tb00111.x

Checkland, P., & Poulter, J. (2006). Learning for Action: A Short Definitive Account of Soft Systems Methodology, and Its Use Practitioners, Teachers and Students. Wiley.

Checkland, P., & Scholes, J. (1999). Soft Systems Methodology in Action. Wiley.

Chopra, D. with A. Kelbert & P. Iyer (2013) 'A Feminist Political Economy Analysis of Public Policies Related to Care: A Thematic Review', IDS Evidence Report 9, Brighton: IDS page 34

Chopra, D. (2015). Balancing Paid Work and Unpaid Care Work to Achieve Women's Economic Empowerment. IDS Policy Briefing 83. IDS.

Collier, D. (2011). Understanding Process Tracing. Political Science and Politics, 44(4), 823-830.

Davidson D. (1980). Essays on Actions and Events. Oxford: Clarendon Press.

Davies, R. (1998). An evolutionary approach to facilitating organisational learning: an experiment by the Christian Commission for Development in Bangladesh. Impact Assessment and Project Appraisal, 16(3), 243-250.

Davies, R., & Dart, J. (2005). The 'Most Significant Change' (MSC) Technique: A Guide to Its Use.

De Meur, G., Rihoux, B., & Yamasaki, S. (2002). L'analyse quali-quantitative comparée (AQQC-QCA): approche, techniques et applications en sciences humaines. Louvain-la-Neuve: Academia-Bruylant.

Edwards W (1986) Summing up: The Society of Bayesian Trial Lawyers. Boston University Law Review 66: 937-41.

Elster, J. (1998). A plea for mechanisms. In P. Hedström & R. Swedberg (Eds.), Social Mechanisms: An Analytical Approach to Social Theory (Studies in Rationality and Social Change, pp. 45-73). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511663901.003

Fairfield T and Charman A (2015) Applying formal Bayesian analysis to qualitative case research: An empirical example, implications, and caveats. SSRN. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2647184.

Festinger, L. (1957). A Theory of cognitive dissonance. Stanford, CA: Stanford University Press

Friedman R (1986) A close look at probative value. Boston University Law Review 66: 733–59.

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.

Hedstrom, P. (2005). Dissecting the Social: On the Principles of Analytical Sociology. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511488801

Kaye D (1986) Quantifying probative value. Boston University Law Review 66: 761–6.

Lind, J., & Mitchell, B. (2013). Understanding and Tackling Violence Outside of Armed Conflict Settings. IDS Policy Briefing 37. IDS.

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). Contribution analysis: Coming of age? Evaluation, 18(3), 270-280. https://doi.org/10.1177/1356389012451663

Pawson, R., & Tilley, N. (1997). Realistic Evaluation. Sage.

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.

Rihoux, B., Ragin, C., & (eds). (2009). Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques. Sage.

Schneider, C., & Wagemann, C. (2012). Set-Theoretic Methods for the Social Sciences. Cambridge University Press.

Schelling, T. (1971) Dynamic models of segregation, The Journal of Mathematical Sociology, 1:2, 143-186, DOI: 10.1080/0022250X.1971.9989794

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.

Van Evera, S. (1997). Guide to Methods for Students of Political Science. Cornell University Press.

Van Ongevalle, Jan et al. (2012) Dealing with complexity through "actor-focused" Planning, Monitoring & Evaluation (PME). From results-based management towards results-based learning

Vis, B. (2012). The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses. Sociological Methods Research, 41(1), 168-198.

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 Evaluation: An Introduction. London: Overseas Development Institute (ODI).

Westhorp, G. (2014). Realist impact evaluation: an introduction. London: Overseas Development Institute (Methods Lab).

White, H., & Phillips, D. (2012). Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. Working Paper 15. International Initiative for Impact Evaluation.

Williams, B., & Hummelbrunner, R. (2010). Systems Concepts in Action: a practicioner's toolkit. Stanford University Press.