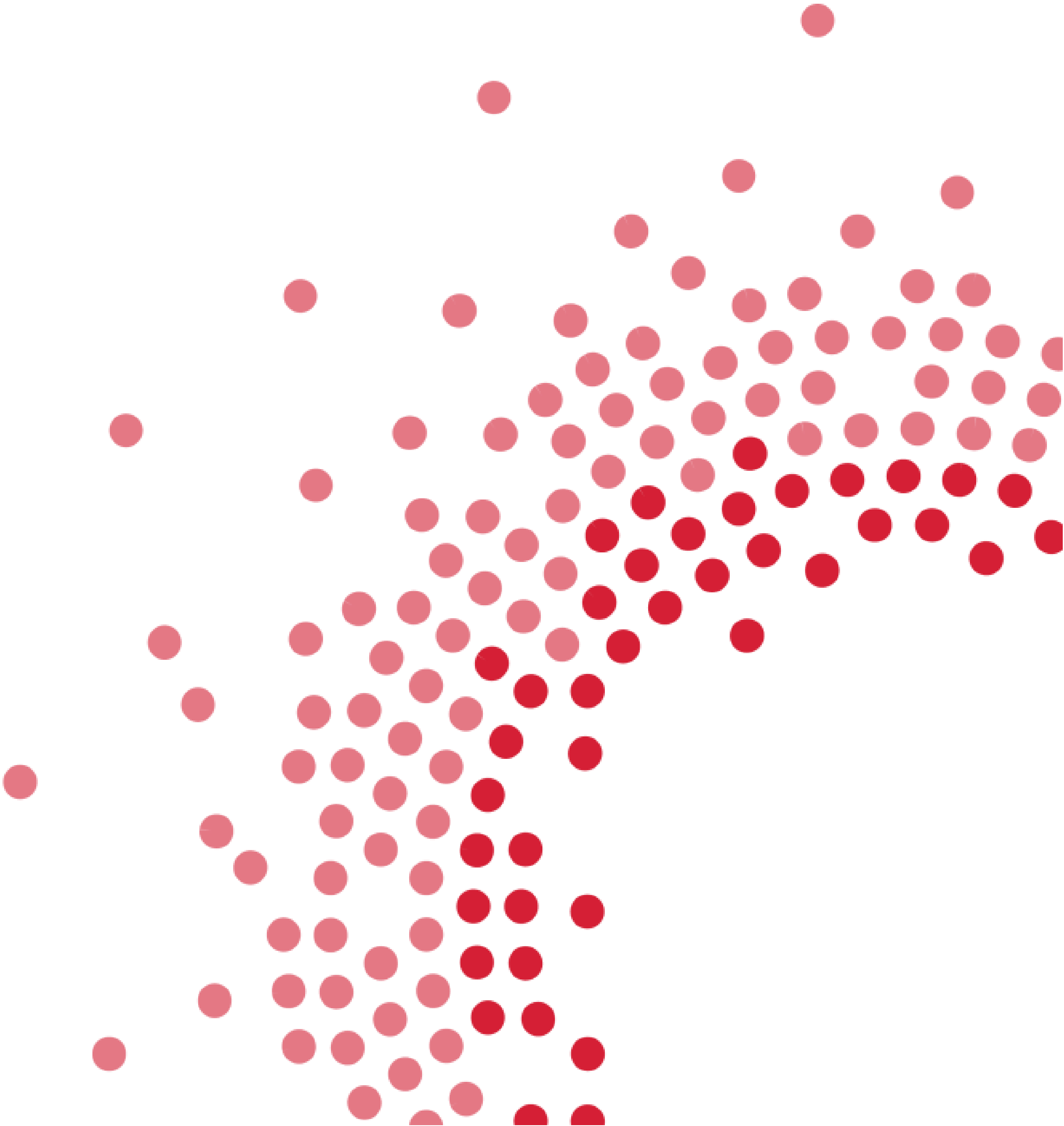
Choosing Appropriate Evaluation Methods

A Tool for Assessment & Selection

Barbara Befani



Contents

About Bond

Bond is the civil society network for global change. We bring people together to make the international development sector more effective. bond.org.uk

Acknowledgements

This report was written by Barbara Befani, and accompanies the Choosing Appropriate Evaluation Methods tool developed by Barbara Befani with Michael O'Donnell.

The authors are grateful for feedback on previous drafts of this narrative and the accompanying tool from a range of evaluation specialists and prospective users: Hlengani Bhebhe, Zack Brisson, Laura Camfield, Rohati Champan, James Copestake, Rick Davies, Jenny Eddis, Sarah Henon, Sofia Jadavji, Sebastian Lemire, Sophie Mareschal, Richard Ponsford, Melanie Punton, Saltanat Rasulova, Fiona Remnant, Patricia Rogers, Stanford Senzere, Ethel Sibanda, Tremayne Stanton-Kennedy, Claire Thomas, Alix Tiernan, Giel Ton, Jos Vaessen, and Marie Weiller. Any interpretations, errors or omissions made are the author's own.

This work was completed with funding for the UK's Department for International Development and Comic Relief.

The Choosing Appropriate Evaluation Methods tool can be downloaded from

https://www.bond.org.uk/resources/evaluation-methods-tool

Choosing Appropriate Evaluation Methods: A Tool for Assessment and Selection, October 2016
Published by Bond, Society Building, 8 All Saints Street, London N1 9RL, UK
Registered Charity No. 1068839 Company Registration No. 3395681 (England and Wales)

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[1] 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)

As a principle, it is now largely accepted that a wide range of methodological options are appropriate, under different circumstances, to evaluate the impact of development programmes. However, while apparently solving the problem of scarcity of options, this expansion has created a selection problem.

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[2] the (illusory, some might say) benefit of being simple and leading to inevitable and 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[3].

The tool presented in this paper is an attempt to improve this situation and the process of methodological choice, by helping users make an informed and reasoned choice of one or more methods[4] 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 the user learn about comparative advantages and weaknesses of methods, their specific abilities, and not less importantly their requirements. 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 dimensions: requirements and abilities (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 takes the user through the decision making process step by step.

This paper contains four sections. Following this introduction, section 2 addresses the dimensions of appropriateness and the conceptual backbone that informs the tool structure. Section 3 illustrates the inner workings of the tool, as well as describing the meaning of the various cells in the excel file. Finally, section 4 tackles the intended users and potential uses of the tool, in addition to discussing the tool's limitations and the potential to develop it further. A series of annexes cover reviewers (One), the definitions of method, approach and technique on which the tool is based (Two), and provide basic information about the 11 methods included in the tool.

[1] 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.

[2] For many people and institutions, still "has".

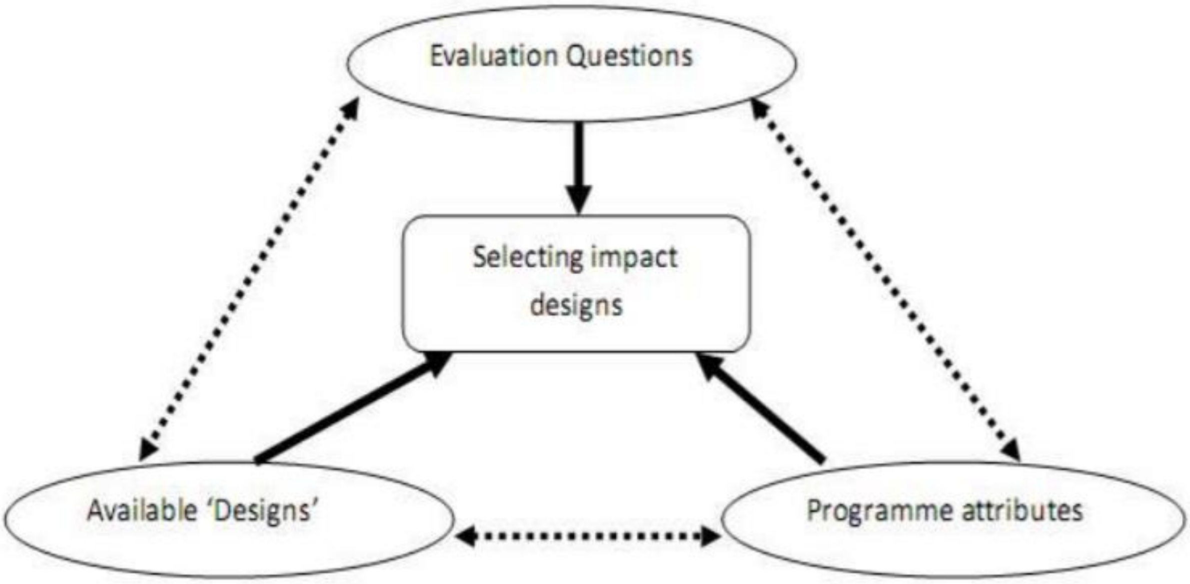
[3] Patton asserts that "methodological appropriateness is the utilization-focused gold standard" (Patton, 2012, p287), and "appropriateness" is one of Bond's five Evidence Principles (https://www.bond.org.uk/effectiveness/monitoring-and-evaluation)

[4] For the definition of "methods", see Annex 2

2. Dimensions of appropriateness

As the Design Triangle suggests (Figure 1), methods need to be aligned with evaluation questions. Not all methods are equally able to answer the same question, and the tool measures the ability of methods to answer specific questions. It does so by identifying five different (impact) evaluation questions and presenting an assessment of the different methods' abilities to answer each one of them[5].

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



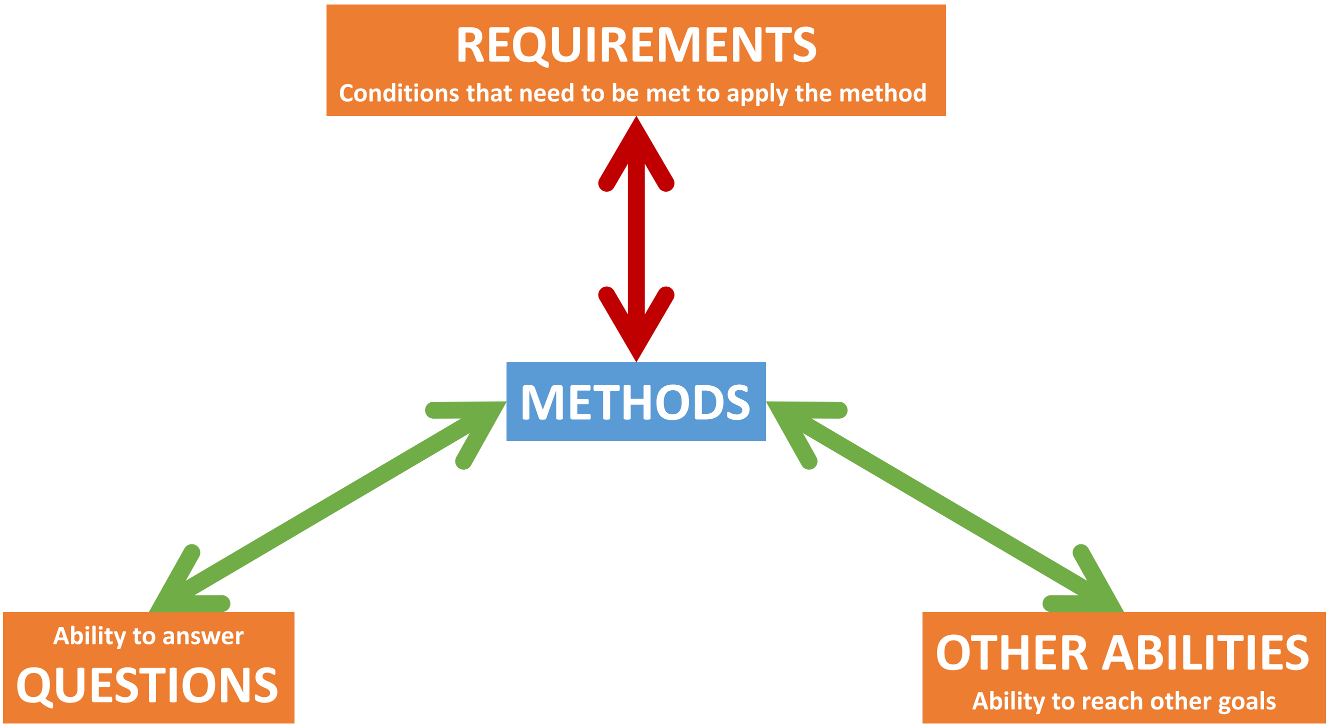
Compared to the current version of the Design Triangle, the main conceptual innovation of the tool lies in unpacking the relation between programme attributes and methods into the two dimensions of 'requirements' and 'abilities' (Figure 2). It can be argued that methods have different abilities in general, not just different abilities to answer questions. They have comparative strengths and weaknesses that commissioners of evaluations might find more or less interesting for a wide variety of reasons, and that can be more or less helpful in reaching a variety of goals. The green arrows in the figure below represent these abilities.

In addition, 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.

[5] More details are presented in the next section.

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.

Figure 2: Revised Design Triangle


REQUIREMENTS
Conditions that need to be met to apply the method
METHODS
Ability to answer
QUESTIONS
OTHER ABILITIES
Ability to reach other goals

The tool's conceptual framework 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[6].

[6] Note that the evaluation stakeholders as defined here include commissioners, the evaluation team, and other stakeholders that affect methodological choice.

3. 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.

3.1 The method's ability to answer key evaluation questions

In the first section (Stage 1), 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, no single method can answer more than three well. 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 O) show the link between each question and the eleven methods considered. The coloured cells indicate that the methods are 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 Q to AA). 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:
5	The question is the primary question answered by the method, which is fully "self-sufficient" to answer it
4	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)
3	The method helps answer the question, but it needs substantial input from other methods to answer the question properly
2	The method is useful to answer the question only very indirectly, rarely or under special circumstances
1	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.

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

In the second section (Stage 2), 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 18). These scores can be found in the hidden columns E to O, where LOW is indicated with 0.33, MEDIUM with 0.67 and HIGH with 1.00[7]. The methods experts made this assessment 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 15 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 Q to AA). 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[8].

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 O and compare them with columns Q to AA. 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 O.

The summary row (row 22) 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

[7] The use of percentages rather than categories is an artificial construct for the purposes of initial development of the excel-based tool. Further development of the tool – particularly as an online resource – should further review the feasibility of replacing percentages with categories.

[8] Namely, the basic measure of ability multiplied by 0.67 if the user selects "desirable" and 0.33 if the user selects "slightly desirable".

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.

3.3 The team's ability to accommodate the method's requirements

The third section (Stage 3) considers constraints that can be imposed on evaluation method options by the nature of intervention to be evaluated and other real life constraints. 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 P).
- 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 P).

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 11 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 nine 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 P.

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 19 conditions ("Stage 3" tab, rows 4 to 23), 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. Users are recommended to 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 19 questions, the tool returns the following information:

- A row at the end of the table (row 25) 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 26) describing the number of requirements that the user *does not know* if they can meet. This tells the user what further action needs to be taken in order to determine whether the method can be applied or not.
- A final row (row 27) 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 its application quality.

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% for "fully", 67% for "to some extent", 33% for "poorly" and 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%; if they can meet it "poorly" it will be 17%, and so on.

A summary row at the end of the table (row 29) 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%) when all requirements and desirables can be fully met; and lowest (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[9]. The hidden columns F to P clarify how the overall score is calculated in each specific case.

[9] Again, as per footnote 7, the use of percentages rather than categories is an artificial construct for the purposes of initial development of the excel-based tool. Further development of the tool – particularly as an online resource – should review the feasibility of replacing percentages with categories.

4. Use and limitations

4.1 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 issue of appropriateness
- Contribute to informing the design of interventions to improve their evaluability (by helping users sense-check the compatibility between their planned 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 the best fit between evaluation questions, interests and programme attributes)

The tool is particularly intended for users with some knowledge of evaluation methods and issues, but not experts in evaluation. Many commissioners of evaluations fall into that space, and this tool 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 sections).

This 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 to each intervention. Similarly, the tool can inform evaluation method choice for interventions in any sector or thematic area (governance, health, livelihoods, etc.).

4.2 Limitations and potential for further development

It is stressed that 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 little existing 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.

Involving different experts could not just change the scores, but also the types of requirements and abilities considered. The second and third level elements (requirements and other abilities) were in fact selected based on expert advice: a broader group of experts might have perhaps selected different requirements and proposed different abilities.

No structured sensitivity tests have been carried out on the tool, and at this stage, it is currently unknown how much the results of particular selections would change following small changes in the ability measurements and the requirements' assessments. At the same time, the nature of the tool as a spreadsheet 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 complicated and less user-friendly.

The tool does not attempt to incorporate limitations on feasibility related to the budget available for evaluation. In reality, budgetary constraints may have a significant bearing on methods choices, 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 state useful ranges for the cost of using each method
- 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 a requirement, 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.

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 final 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, it needs to take these three rows into consideration. There are different ways to do it. One can start from the feasible methods (third bullet point) and see which ones are most useful in terms of answering the preferred questions and achieving the preferred goals – later. Alternatively, 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.

Annexes

Annex 1: List of Reviewers

Method	**Reviewers**
Randomised Controlled Trials	Maren Duvendack, Edoardo Masset, Daniel Phillips, Matthew Juden
Difference in Differences	
Statistical Matching	
Outcome Mapping	Simon Hearn
Most Significant Change	Rick Davies
Soft Systems Methodology	Bob Williams, Richard Hummelbrunner
Causal Loop Diagrams	
Realist Evaluation	Gill Westhorp, Bruno Marchal
Qualitative Comparative Analysis	Barbara Befani (no additional external review)
Process Tracing & Bayesian Updating	
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

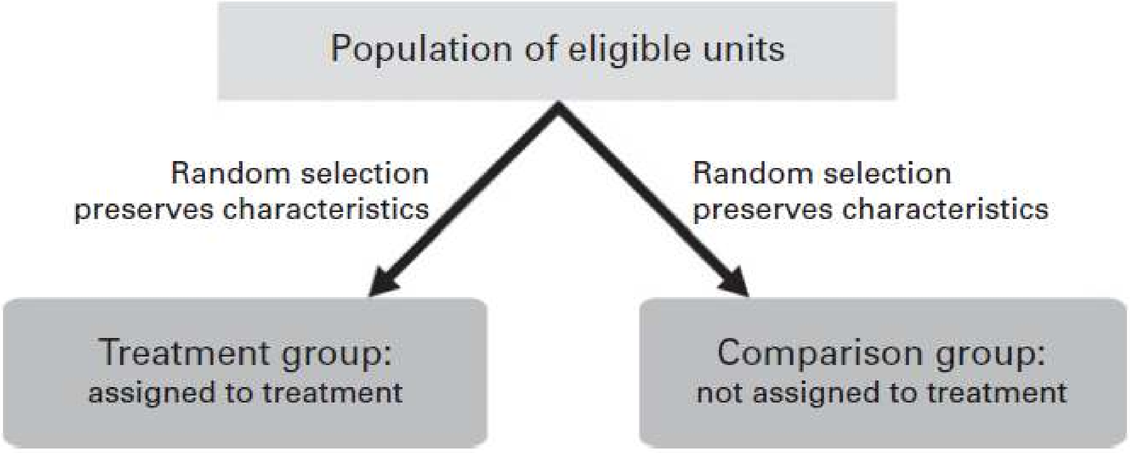
	Approach	Method	Technique
Randomised Controlled Trials	X	X	
Difference in Differences		X	
Statistical Matching		X	
Outcome Mapping		X	X
Most Significant Change		X	X
Soft Systems Methodology	X	X	
Causal Loop Diagrams		X	
Realist Evaluation	X	X	X
Qualitative Comparative Analysis	X	X	X
Process Tracing and Bayesian Confidence Updating		X	X
Contribution Analysis		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)

Figure 3: Logic of randomisation in RCTs


Population of eligible units
Random selection preserves characteristics
Random selection preserves characteristics
Treatment group: assigned to treatment
Comparison group: not assigned to treatment

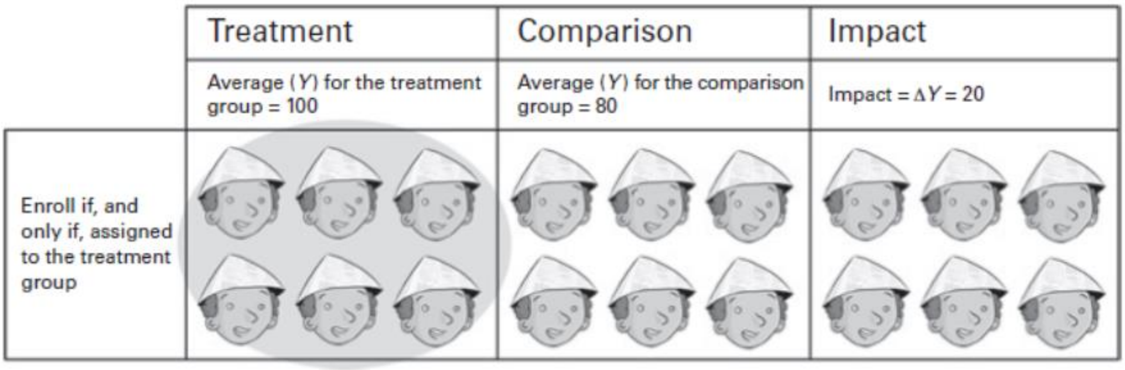
Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

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. 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 4).

In RCTs, the non-intervention outcome is estimated by designing a control group, which is virtually identical to the treatment group. In addition, ideally, 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, 1969).

The similarity between treatment and control groups is guaranteed by randomization (figure 3): 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 4: how impact is estimated in RCTs

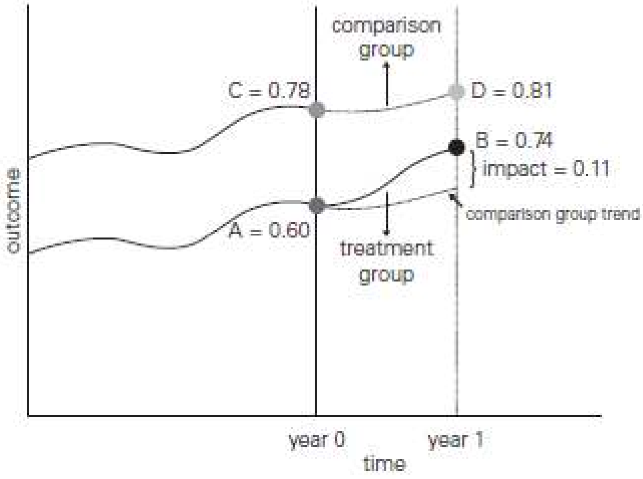

Treatment
Comparison
Impact
Average (Y) for the treatment group = 100
Average (Y) for the comparison group = 80
Impact = ΔY = 20
Enroll if, and only if, assigned to the treatment group

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[10], Statistical Matching (SM)[11], Interrupted Time Series (ITS) and Instrumental Variables (IV).

Figure 5: Logic of Difference in Differences



Difference in Differences (DID) (Figure 5) 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

[10] The best known of which is Regression Discontinuity Design (RDD).
[11] Including well-known variant Propensity Score Matching (PSP).

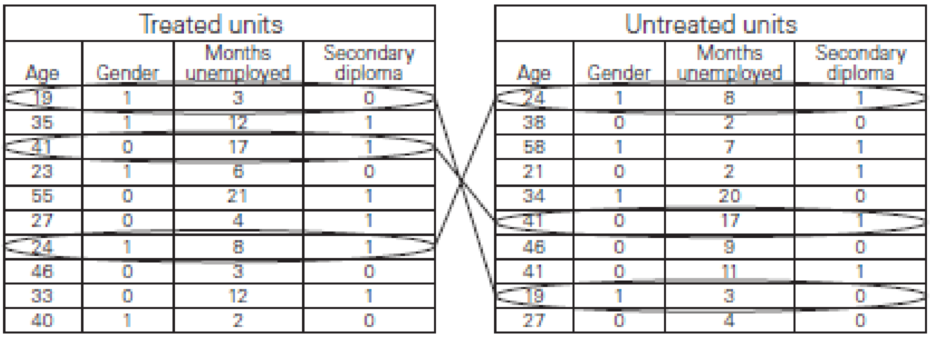
the two groups equivalent with *Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011* respect to (factors influencing) the outcome[12].

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 6). 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 6: 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

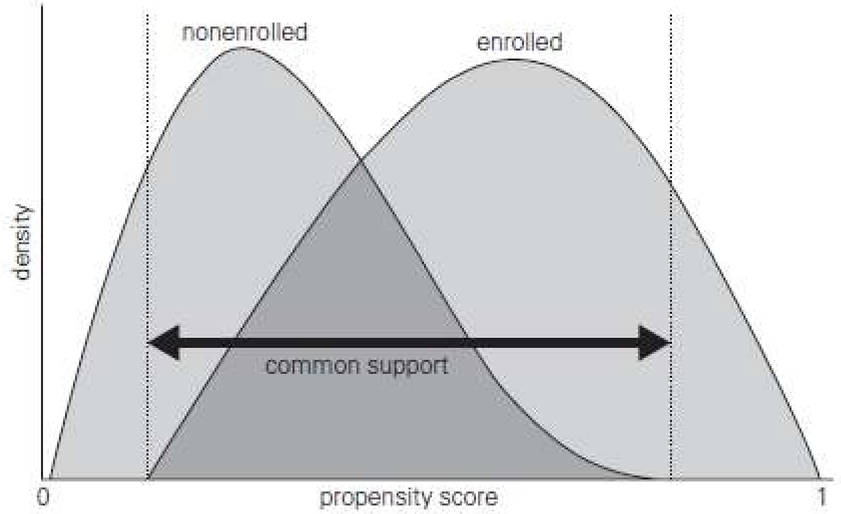
Source: Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011

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

[12] Another variant requiring a lot of knowledge is statistical matching. Here the treatment sample is matched with the control sample on the basis of similarity with regard to factors that are known (or assumed) to affect the outcome (in addition to the intervention). For small samples, the matching is done individually one by one.

non-enrolled units are available, which constitutes the basis for the creation of treatment and control groups (Figure 7).

Figure 7: The common support in Propensity Score Matching



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

Annex 3.4 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).

Annex 3.5 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 & 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. The stories of significant change that MSC generates can provide useful sources of information for the specification and subsequent assessment of a theory of change.

Annex 3.6 Soft Systems Methodology

SSM is about solving problems through the comparison of different perspectives of how systems work (Figure 8). (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):

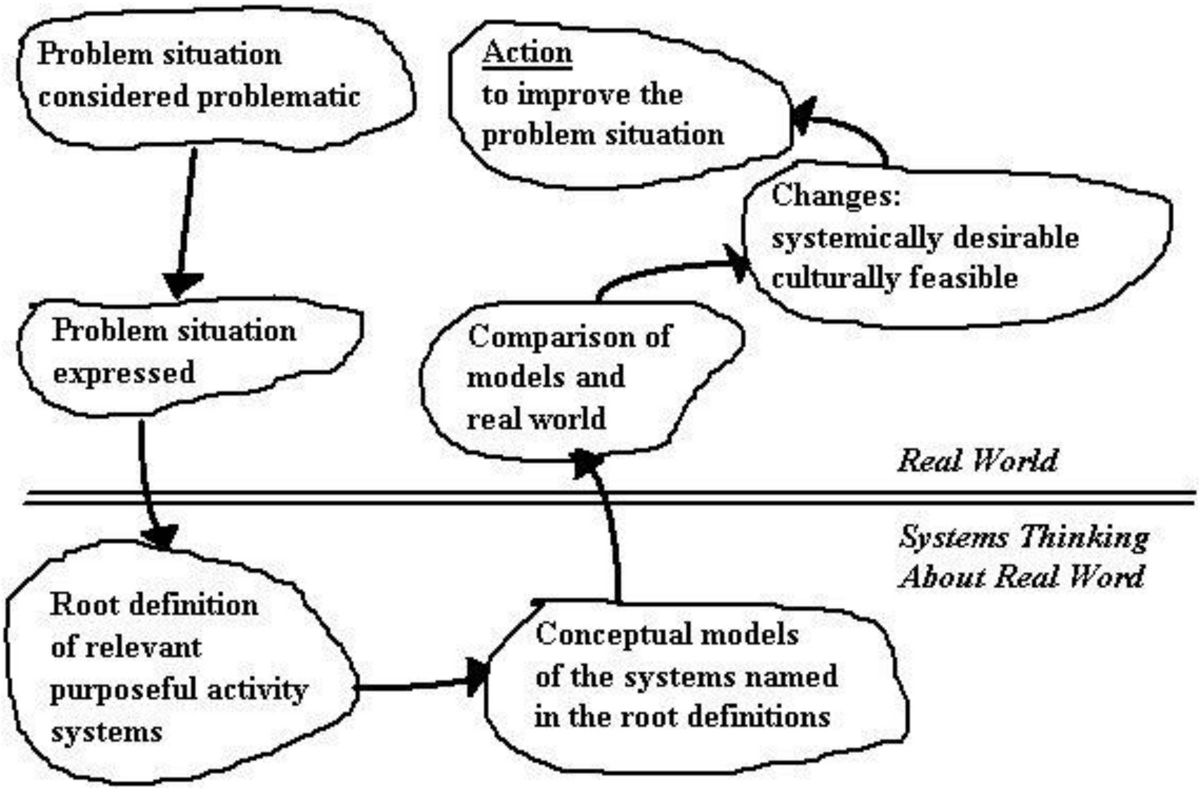
1. You have a perceived problematic situation that;
    - will contain people trying to behave purposefully
    - will be perceived differently by people with different worldviews
2. So, make models of purposeful activity as perceived by different worldviews
3. Use these models as a source of questions to ask of the problematical situation: thus structuring a discussion about changes that 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 also developed a mnemonic checklist (the CATWOE[13]) 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"

[13] Note that CATWOE is only a technique for one of the steps in SSM (root definition).

Figure 8: The Steps in Soft Systems Modelling


Problem situation considered problematic
Action to improve the problem situation
Changes: systemically desirable culturally feasible
Problem situation expressed
Comparison of models and real world
Real World
Systems Thinking About Real Word
Root definition of relevant purposeful activity systems
Conceptual models of the systems named in the root definitions

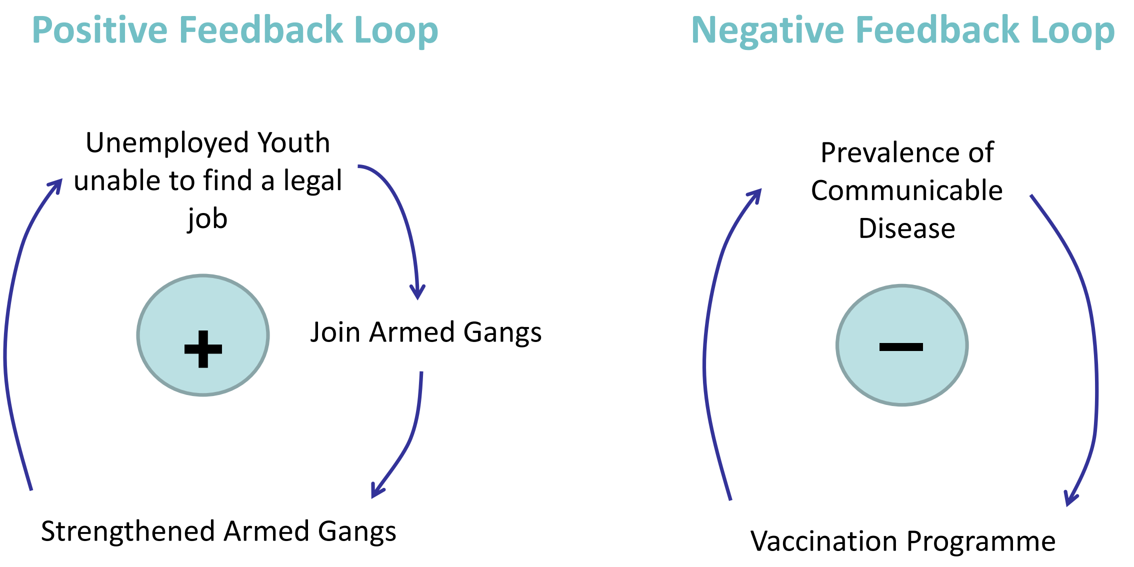
Annex 3.7 Causal Loop Diagrams

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.

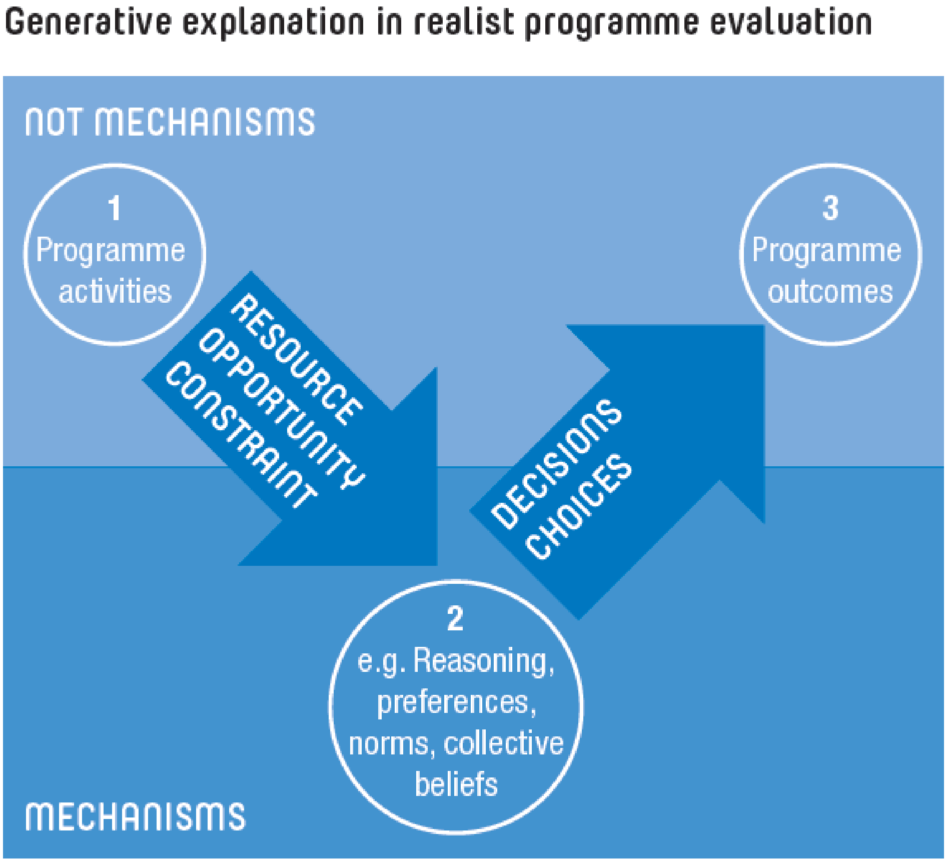
For example, unemployment may lead some young people in poor, underserviced neighbourhoods to join armed gangs (Lind & Mitchell, 2013). This will strengthen the ranks of local gangs and reinforce their control of the territory, further reducing opportunities of legal employment for young people (see Figure 9). Another example is unpaid care (Chopra, 2015) Chopra 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 9).

Figure 9: positive and negative feedback loops



Annex 3.8 Realist Evaluation



Wong, G., Westhorp, G., Pawson R, and Greenhalgh, T. (2012) Realist Synthesis RAMESES Training Materials. Reproduced with permission. Note that 'generative explanation' means 'explaining how causation works'.

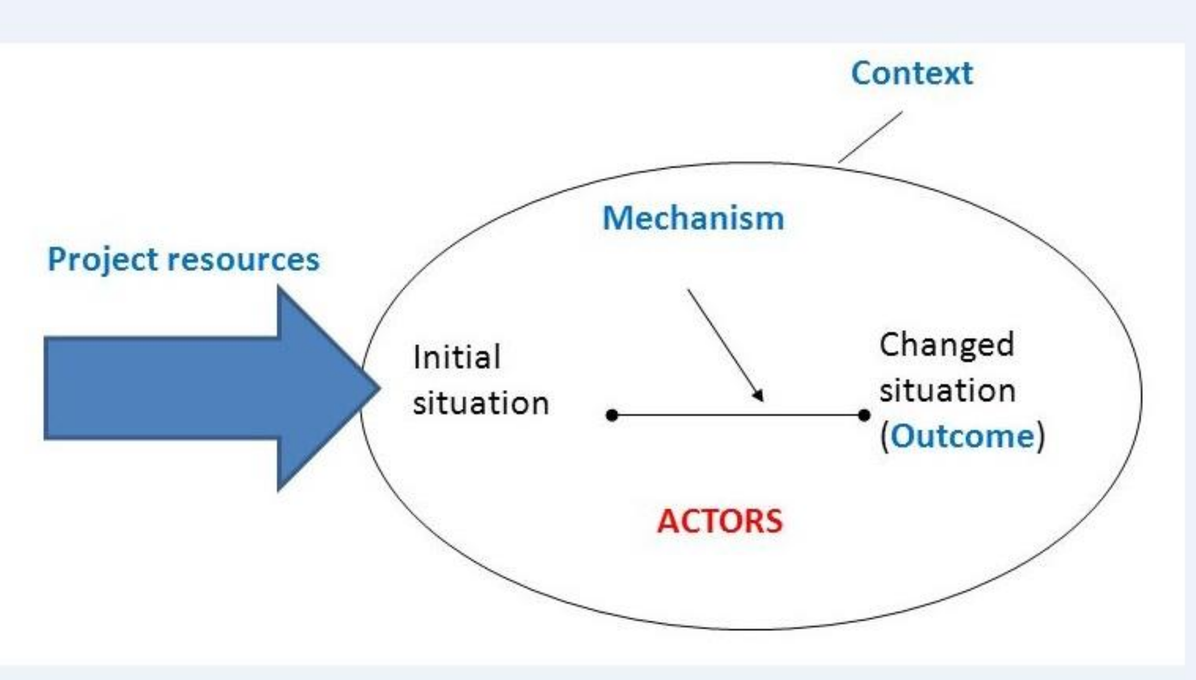
Realist evaluation is an application of scientific realism to evaluation. Scientific realism (Bhaskar, 2009) is an ontology framing reality as a stratified object made of nested layers, sometimes represented as an onion, where action is entirely embedded and as such dependent on the context. As an approach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

(reproduced from Westhorp (2014))

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

Technically, Realist Evaluation entails identifying one or more Context-Mechanism-Outcome (CMO) configurations, where contexts are made of resources, opportunities and constraints available to the beneficiaries; mechanisms are choices, reasoning or decisions that 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 10).

Figure 10: The realist “egg”



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

Annex 3.9 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), 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 basis (see Glossary for a definition of the terms “qualitative” and “quantitative”).

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 statistics.

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 and which ones are most effective for the outcome to occur.

Annex 3.10 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 materialized; 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 we nevertheless 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 little inferential power and does not add much 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 1 for the properties of these tests.

Box 1

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 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 that is 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 as "Bayesian Confidence Updating". In this formalisation of Process Tracing, the inferential power or probative value of a piece of evidence E for a theory T can be measured in a number of ways, all related to the difference between the true positives rate or "sensitivity" (the probability that the evidence confirms that the theory holds when this is in fact the case) and the false positives rate or "Type I error" (the probability that the evidence confirms that the theory holds when this is actually not the case). 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).

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 constitutes a confirmation of 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 weakens the theory. Finally, if the evidence has a similar chance of being observed whether the theory holds 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 and confidence on the existence of the claim, 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 challenge; if no major objections are offered, this will increase their legitimacy and credibility. Just like in a judicial trial where evidence is produced in favour or against a defendant and the jury is left to assess the probative value of that evidence, if the prosecution cannot produce any significant evidence of guilt or if the defence finds proof that the suspect is innocent, then the suspect is considered innocent by the jury.

Annex 3.11 Contribution Analysis

[From Befani & Mayne 2014]

Contribution Analysis (Mayne 2001, 2008) 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 it may 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 3. 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 conditions and evidence illustrated in Box 2.

Box 2: Conditions needed to infer causality in Contribution Analysis

1. ***Plausibility.*** The intervention is based on a reasoned theory of change: the chain of results, and the assumptions behind why the intervention is expected to work are plausible, sound, informed by existing research and literature and supported by key stakeholders.
2. ***Fidelity.*** The activities of the intervention were implemented as outlined in the theory of change.
3. ***A verified Theory of Change (ToC).*** The theory of change is verified by evidence: the chain of expected results occurred, and the causal assumptions held.
4. ***Accounting for other influencing factors.*** Context and other factors influencing the intervention are assessed and are either shown not to have made a significant contribution or, if they did, their relative contribution is recognized and included in the ToC, as part of a larger causal package that the ToC captures as faithfully as possible.

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 11).

Box 3: 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

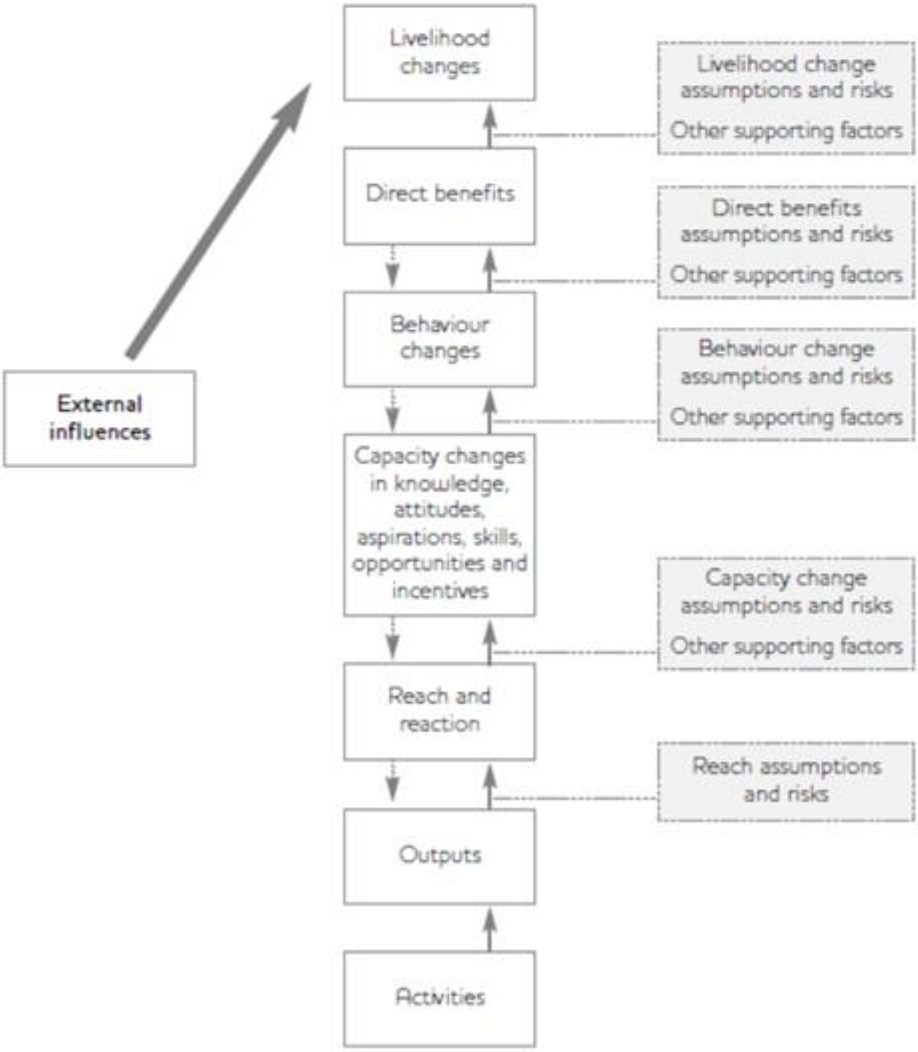
- With a focus on the identified weaknesses, gather data on the ToC results that occurred, the assumptions and risks associated with the causal links and the other identified influencing factors

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

.

Figure 11: Representation of a Theory of Change in Contribution Analysis


Livelihood changes
Livelihood change assumptions and risks
Other supporting factors
Direct benefits
Direct benefits assumptions and risks
Other supporting factors
Behaviour changes
Behaviour change assumptions and risks
Other supporting factors
External influences
Capacity changes in knowledge, attitudes, aspirations, skills, opportunities and incentives
Capacity change assumptions and risks
Other supporting factors
Reach and reaction
Reach assumptions and risks
Outputs
Activities

(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., & 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., & 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.

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

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. (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.

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.

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.

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).

Patton, M.Q. (2012). Essentials of Utilization-Focused Evaluation. Sage

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.

Stern, E. (2015). Impact Evaluation: a Guide for Commissioners and Managers. Bond.

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.

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 practitioner's toolkit. Stanford University Press.

Bond
Society Building
8 All Saints Street
London
N1 9RL, UK

+44 (0)20 7837 8344
bond.org.uk

Registered Charity No. 1068839
Company registration No. 3395681 (England and Wales)

bond