# Description of evaluation methodologies

## Realist evaluation

### Overview

Pawson and Tilley's (1997) starting point for setting out the realist approach to evaluation is to argue that the 'traditional' experimental evaluation is flawed because its attempt to reduce an intervention to a set of variables and control for difference using an intervention and control group strips out *context*. Instead evaluators need a method which "seeks to understand what the program actually does to change behaviours and why not every situation is conducive to that particular process." (Pawson and Tilley 1997: 11). They assume a different, 'realist' model of explanation in which "causal outcomes follow from mechanisms acting in contexts" (Pawson and Tilley 1997: 58).

A mechanism explains what it is about a programme that makes it work. Mechanisms are not variables but accounts that cover individual agency and social structures. They thus should 'reach down' to individual reasoning and 'reach up' to the collective resources embodied within a social programme that is being evaluated (Pawson and Tilley 1997). For Pawson and Tilley "A mechanism is thus a theory – a theory which spells out the potential of human resources and reasoning." (Pawson and Tilley 1997: 69). Astbury and Leeuw's (2013) definition of mechanisms as "...underlying entities, processes, or [social] structures which operate in particular contexts to generate outcomes of interest"

For Pawson and Tilley (1997) causal mechanisms and their effects are not fixed, but contingent on context. A programme will only 'work' if the contextual conditions into which it is inserted are conducive (Pawson and Tilley 1994). Programs are always introduced into a pre-existing social context and pre-existing structures enable or disable the intended mechanism of change (Pawson and Tilley 1997). This is to recognise the complexity of social interventions, using 'complexity' in its sociological sense to include the principle of non-linearity (small changes in inputs may, under some conditions but not others, produce large changes in outcome); the contribution of local adaptiveness and feedback loops; the phenomenon of emergence; the importance of path dependence; and the role of human agency (Marchal et al. 2012).

A substantial part of Pawson and Tilley's key texts (1994, 1997) in which they set out the case for scientific realist evaluation are given over to a discussion of causation. For Pawson and Tilley (1997) the model of causation adopted in 'traditional' experimental evaluation design is external, successionist causation. This is the idea that causation itself is unobservable but that it can be inferred from the basis of observation. Scientific realists prefer a model of generative causation that sees causation acting internally as well as externally. Scientific realists do not therefore make predictions about the probability of an intervention leading to an outcome. This is because complex interventions are only semi-predictable (Lawson 1997, Marchal et al. 2012). Lawson's (1997) concept of demi-regularity is that human choice or agency is only semi-predictable because variations in patterns of behaviour are attributable partly to context. Human behaviour is not determined, but neither is it completely haphazard. There will be some patterning and therefore the best realist evaluation can offer is plausible explanations of for whom, in what

circumstances and in what respects an intervention is more likely to succeed (Wong et al. 2013).

To be clear, Pawson and Tilley argue in favour of an experimental method. However, they reject the model of experiment based on a similar intervention and control group. They argue instead, following philosophers such as Bhaskar that the two essential elements of an experiment are triggering the mechanism being studied to make sure that it is active and preventing interference with the operation of the mechanism. In this model, rather than simply activating an independent variable and observing the outcome, the experimentalist's task is to manipulate the entire experimental system.

A particular concern has been about the ability of the approach to deal with complexity (Pederson and Rieper 2008, Blamey and Mackenzie 2007, Davis 2005). Critics argue that scientific realism was developed partly in the field of crime reduction where programmes under evaluation were relatively small scale, operating at a local level, targeted on distinct groups and involving relatively few stakeholders. However, Pederson and Rieper in their own work and through referencing that of others (e.g. Davis 2005) demonstrate how the scientific realist approach can be adapted to regional and national level policies and programmes that are more complex.

## Resources required for an evaluation

### *Skill set for evaluators*

In the scientific realist paradigm the evaluator is a researcher and theorist, with a detailed understanding of the programme being evaluated and able to construct mid-level theories (groups of context-mechanism-outcome configurations) for subsequent testing.

Although scientific realist evaluation can incorporate both quantitative and qualitative data collection, qualitative data collection, and in particular the semi-structured interviews tend to be most common (Manzano 2016). However, scientific realism involves a particular conception of the interview as a teacher-learner cycle Pawson and Tilley 2004, see above). The approach to interviewing required in a scientific realist interview may contradict the standard research methods training that an evaluator has received. Whereas the prevailing approach to interviewing is to maintain neutrality and avoid leading questions, a realist evaluation interview in which exploration of theory is the aim is likely to be led by the interviewer and aims at 'assisted sensemaking' (Manzano 2016). With a focus on the interview in realist evaluation, Manzano emphasises, not so much particular training or skills, but the idea of the researcher learning the 'craft' of interviewing, suggesting that this is an approach that takes time and experience to develop.

### *Resource implications*

While there are no set rules for how much data should be collected, it is accepted that a realist evaluation should aim to collect large amounts of data (Manzano 2016: 348). Manzano explains that "substantial amounts of primary or secondary data are needed – even when the sample is small – to move from constructions to explanation of causal mechanisms." This is because the unit of analysis is not the person, but the events and processes around them and "every unique programme participant uncovers a collection of micro events and processes, each of which can be explored in multiple ways to test theories"

(ibid.). It is not possible to quantify what amounts to ‘large amounts of data’ in the abstract, but collecting multiple sources of data on each case suggests days, rather than hours of work.

# Process tracing

## Overview

Process tracing can be defined as:

> "the analysis of evidence on processes, sequences, and conjunctures of events within a case for the purposes of either developing or testing hypotheses about causal mechanisms that might causally explain the case." (Bennet and Checkel 2015: 7)

Process tracing is a methodology that combines pre-existing generalisations with specific observations from within a single case to make causal inferences about that case (Mahoney 2012). It involves the examination of 'diagnostic' pieces of evidence within a case to support or overturn alternative explanatory hypotheses (Bennett 2010). Identifying sequences and causal mechanisms is central (Bennett 2010).

Process tracing can provide leverage on several aspects of causal inference that are difficult to address in traditional, counterfactual impact evaluations:

- The challenge of establishing causal direction: process tracing that is focused on sequencing who knew what, when and what they did in response can help establish causal direction (Bennett 2010).
- The challenge of spuriousness: where X and Y are correlated but it is unclear whether X caused Y or a third factor caused both X and Y, process tracing can help establish the causal chain of steps connecting X and Y and whether there is evidence for other factors causing both X and Y (Bennett 2010).
- The inductive element of process tracing provides opportunities for evaluators to uncover unforeseen or unexpected causal explanations (Bennett and Checkel 2015).
- As Bennett and Checkel (2015) note, because causal mechanisms are operationalised in specific cases and process tracing is a within-case method of analysis, generalisation can be problematic. However, conversely, the use of process tracing to test and refine hypotheses of causal mechanisms can clarify the conditions under which a hypothesis is generalisable.

Process tracing is subject to two particular criticisms:

- Infinite regress: The fine-grained level of detail involved in process tracing can potentially lead to an infinite regress (Bennett 2010, citing King, Keohane and Verba 1994). Bennet and Checkel (2015) argue that while a commitment to explanation by mechanisms means that explanations are always incomplete and provisional and every explanation can be called into question if it can be shown that its hypothesised processes are not evident at a lower level of analysis, researchers can and do make defensible decisions about when and where to begin and stop in constructing and testing explanations.
- Degrees of freedom: Qualitative research on a small number of cases, or even a single case, with a large number of variables that are free to vary could lead to a form of indeterminacy in analysis (Bennett 2010).

In process tracing these challenges are overcome by recognising that not all data is of equal probative value in discriminating between alternative explanations and a series of tests outlined below are used to establish causal explanations that recognise this.

## Resources required for an evaluation

### *Evaluator skills and experience*

Process tracing can draw on a range of data collection and data analysis approaches. Commonly, these will include reviewing documents, interviewing key informants, undertaking observations and analysing performance management and programme monitoring data. The evaluator should therefore have a postgraduate level of research methods training in a range of commonly used qualitative and quantitative research skills.

Process tracing also requires other knowledge and experience:

- Evaluators will need to acquire a deep knowledge of the case from which evidence is drawn and for sufficient evidence to be gathered from the case to distinguish between competing and incompatible hypotheses. Whether this is historical, archival data or data collected through interviews, observations or ethnographies, the evaluator using process tracing will require the skills appropriate to the chosen methods of data collection.
- The evaluator will need knowledge of the pre-existing evidence base relevant to the case being evaluated, which in turn implies good knowledge of wider practice in the sector because this will provide context for the case.

Actors, whether historical or contemporaneous may go to great lengths to obscure their actions and motivations, so biasing available evidence (Bennett 2010). The literature on process tracing often uses the metaphor of a detective solving a case or a doctor diagnosing a medical condition. These metaphors hint at the ‘soft’ skills that evaluators using process tracing require if they are to successfully sift the evidence and discover ‘whodunnit’. As well as the analytical skills of a Holmes, they might also need the guile and resilience of a Columbo or a Poirot such that they are confident to cast the net widely for alternative explanations, be relentless in gathering diverse and relevant evidence and can consider potential bias in different evidentiary sources (Bennett and Checkel 2015). In some instances, this might point to the need for an external evaluator.

### *Resource implications*

Although process tracing is a within-case method it also requires both diverse and deep evidence. Where decisive evidence is not available and straw-in-the-wind tests are being relied on process tracing can be very time consuming (Bennett and Checkel 2015). There is no prescribed amount of evidence gathering required for process tracing. However, data will be drawn from multiple sources and Bennett and Checkel (2015) suggest that data collection should continue on any given stream of evidence until it becomes repetitive i.e. a saturation point is reached.

# Contribution analysis

## Overview

“Contribution analysis explores attribution through assessing the contribution a programme is making to observed results.” Mayne (2008: 1) Four conditions are needed to infer causality in Contribution Analysis (Befani and Mayne 2014, Mayne 2008):

- **Plausibility:** The programme is based on a reasoned Theory of Change.
- **Fidelity:** The activities of the programme were implemented.
- **Verified Theory of Change**: The Theory of Change is verified by evidence such that the evaluator is confident that the chain of expected results occurred.
- **Accounting for other influencing factors:** Other factors influencing the programme were assessed and were either shown not to have made a significant contribution or, if they did, the relative contribution was recognised.

Theory of Change is thus key to undertaking Contribution Analysis and a specific understanding of causality underpins the analysis. Causation is multiple (multiple factors can be responsible for the outcome) and conjectural (factors combine in complex ways to produce outcomes) (Befani and Mayne 2014).

## Resources required

### *Skill set for evaluators*

In steps 1 – 4 of contribution analysis a range of mostly qualitative research skills are required including interviewing stakeholders, workshops to develop theories of change, analysis of documentary evidence and literature reviews which might include the use of systematic review and meta-analysis techniques.

A broader range of research and analysis skills might be required in steps 5 and 6. Mayne suggests a range of research activities including surveys and using existing monitoring and performance management or administrative data as well as more in-depth qualitative research, implying that the evaluator(s) will also have some skill and experience in quantitative research methods.

The emphasis, particularly in steps 2 and 4 on engaging with stakeholders and in step 1 in understanding the needs of decision-makers and funders implies evaluators with good people skills and the confidence and authority to engage with a wide range of people from front-line staff to senior managers and organisational leaders.

### *Scope of evaluation*

Contribution Analysis requires in-depth engagement with the case and an iterative approach that will require repeated engagement with key stakeholders through the development of the Theory of Change, building contribution stories, gathering data to test them and repeating the process as required.

## Qualitative Comparative Analysis (QCA)

### Overview

QCA (Ragin, 1987) is a method that changes qualitative data into Boolean logic. It is "conjunctural in its logic, examining the various ways in which specified factors interact and combine with one another to yield particular outcomes. " (Cress and Snow, 2000:1079).

QCA as we know it today emerged from the social sciences, notably through the work of Ragin (1987). It has become increasingly accepted as a method enabling systematic comparisons. In the social sciences csQCA was mostly used as a tool for macro-level studies (e.g. state or country level), although some organisational researchers have used it at the meso-level (e.g. firm-level) (Rihoux and Ragin, 2009; Ott et al., 2018), and it is becoming increasingly used at a micro-level (e.g. individuals) (Berg-Schlosser et al., 2009).

Its use as an evaluation method is more recent and came mainly through the field of international development (Pattyn et al., 2019). It is deemed particularly relevant for dealing with complexity as it offers an alternative to traditional statistical methods that are often linear (Ott et al., 2018)

QCA is a "synthetic strategy" (Ragin, 1987, p. 84) that allows for "multiple conjunctural causation" across observed cases. This means that different (i.e. multiple) causal pathways can lead to the same result and that each pathway consists of a combination of conditions (i.e. they are conjunctural) (Berg-Schlosser et al., 2009). The method draws on the assumption that it is often the combination of multiple causes that has causal power (Befani, 2016). Furthermore, the same cause can have different effects depending on which other cause it is combined with and therefore lead to different outcomes. It builds on set theory to consider causal asymmetry and determine whether the conditions (or causes) are sufficient and/or necessary. QCA aims to achieve parsimonious (i.e. short) explanations whilst still accounting for causal complexity (Berg-Schlosser et al., 2009).

QCA is a method gaining traction with evaluators as a way to "unravel explanatory patterns for "success" and "failure" of existing cases, with the possibility to inform potential future cases" (Pattyn et al., 2019: 56). The method is also attractive to evaluators as it works with a small number of cases and aims to generate some form of generalisation (Befani, 2013). QCA bridges quantitative and quantitative methods. It integrates the strengths of both methods whilst overcoming key concerns such as the lack of generalisability often associated with qualitative methods (Ragin, 1987).Some consider its level of generalisation to be modest as it can only be applied to similar cases (Rihoux and Ragin, 2009; Ott et al., 2018).

> "The degree of maturity and robustness of a generalization will strongly depend on the quality of the empirical data set constructed by the researcher, and it will generally be a long and hard job to produce it, with many trials and errors, new questionings, and assessments." (Berg-Schlosser et al., 2009, 11)

It is nevertheless considered as a rigorous method due to its replicability and transparency (Rihoux and Ragin, 2009).

QCA is an iterative process where the researcher “engages in a dialogue between cases and relevant theories” (Berg-Schlosser et al., 2009: 2). The technique is both deductive as the choice of variables (i.e. conditions and outcome) is theoretically driven and inductive as insights emerge from case knowledge (Rihoux, 2006; Rihoux & Lobe, 2009).

In QCA, each case is changed into a series of features, including some condition variables and one outcome variable (Berg-Schlosser et al., 2009). The method generally starts with a Theory of Change identifying ‘conditions’ (factors) that may contribute to the anticipated outcomes. QCA analysis is a reiterative process that requires in-depth knowledge of cases. It also requires data to have a certain granularity. For instance, it needs to include cases where the outcome was negative. Similarly, the conditions need to include cases where the condition is present as well as cases where the condition is absent. Therefore, the quality and granularity of the data is paramount.

There are three main techniques: crisp set QCA (csQCA), fuzzy set (fzQCA), and multi-value QCA (mvQCA). They differ in how they code / consider membership of the cases. csQCA was initially developed drawing on Boolean logic. In csQCA, membership is dichotomous (e.g., 1= member, 0 = non member). However, its dichotomous nature is not always adapted to real life situations. fsQCA was developed in response to this limitation as a means to assign gradual values to conditions such as quality or satisfaction (Ragin, 2000). In fsQCA and mvQCA, membership is multichotomous and partial (e.g., 1 = full member, 0.8 strong but not full member, 0.3 = weak member, 0 = non member). Fuzzy set theory can be an interesting tool to capture the fuzzy nature of some conditions and allows for variance of the observations. It therefore overcomes one of the challenges of crisp set csQCA which entails a dichotomous analysis (Ott et al., 2018). However, csQCA allows for a more transparent process as calibration (i.e., setting of thresholds) are done manually and explained theoretically. In the case of fsQCA, thresholds are set by the programme at a later stage. Here we will focus on csQCA as a good introduction to the methodology. The logic underpinning the technique is then extended to fsQCA and mvQCA.

***Necessary and sufficient conditions***

A cause is defined as necessary if it must be present for an outcome to occur (e.g., there must be a cloud for rain to occur). A cause is defined as sufficient if by itself it can produce a certain outcome (e.g., a cloud is NOT sufficient to know that it is raining, but rain is sufficient to know that there is a cloud). Necessity and sufficiency are usually considered together because all combinations of the two are meaningful (Rihoux, 2017: 36):

- “A cause is both necessary and sufficient if it is the only cause that produces an outcome and it is singular (that is, not a combination of causes).
- A cause is sufficient but not necessary if it is capable of producing the outcome but is not the only cause with this capability.
- A cause is necessary but not sufficient if it is capable of producing an outcome in combination with other causes and appears in all such combinations
- A cause is neither necessary nor sufficient if it appears only in a subset of the combinations of conditions that produce an outcome.”

There are different types of analysis that can be run to determine whether conditions are necessary (i.e., superset analysis), sufficient (i.e., subset analysis), or both (INUS analysis). They can be conducted manually on Excel using the filter function.

## Resources required for an evaluation

### *Skill set for evaluators*

QCA draws on Boolean algebra and set theory. The evaluator will therefore have to be quantitatively orientated. The method requires understanding (or learning) main conventions of Boolean algebra such as:

- An uppercase letter represents the [1] value for a given binary variable. Thus [A] is read as: “variable A is large, present, high, …”
- A lowercase letter represents the [0] value for a given binary variable. Thus [a] is read as: “variable A is small, absent, low, …”
- A dash symbol [−] represents the “don't care” value for a given binary variable, meaning it can be either present (1) or absent (0). This also could be a value we don't know about (e.g., because it is irrelevant or the data is missing). It is not an intermediate value between [1] and [0].

Boolean algebra uses a few basic operators, the two chief ones being the following:

- Logical “AND,” represented by the [*] (multiplication) symbol. NB: It can also be represented with the absence of a space: [A*B] can also be written as: [AB].
- Logical “OR,” represented by the [+] (addition) symbol.

Beyond competencies, QCA is a method that works with a small number of cases but requires a large amount of information about these cases to inform calibration and resolve contradictions. It may not be appropriate for programmes where the understanding of cases is limited. It is also time consuming, especially if not familiar with the method. Whilst books and online resources are available, formal training will be recommended in most cases. QCA training is regularly available through the UK Evaluation Society and the Centre for Evaluation of Complexity Across the Nexus. They will usually offer a good introduction, geared towards evaluation work, in a one day workshop.

Furthermore, QCA is best conducted with the support of software. Most of them are available for free (such as R). Specialist software include:

- fsQCA freeware includes CRISP and FUZZY QCA.
- TOSMANA freeware is used for crisp-set QCA and for a multi-valued outcome QCA.

### *Resource implications*

QCA can be run with a small number of cases, typically between 10 and 50. It is, however, the depth and quality of the data required to run the analysis that makes the method challenging. The iterative process between data and theory is time consuming. In some cases (i.e. where the evaluator does not have an in-depth knowledge of individual case), the process relies on the participation of external stakeholders. QCA analysis can be conducted alone, using a quality assurance checklist (see Befani, 2016: 182, Schneider and Wagemann, 2010), but collaborative work will provide opportunities to increase the quality of the evaluation. An evaluation using a QCA can take between three (if data is QCA ready)

and six months (if data needs cleaning and clarifying). QCA relies on data being available on the outcome of choice. Therefore, the evaluation can take longer if this data is not available from the onset.