TASO
Transforming Access
and Student Outcomes
in Higher Education

Impact Evaluation With Small Cohorts: Methodology Guidance

June 2022

Manchester
Metropolitan
University

PERU Policy Evaluation & Research Unit

Acknowledgements

TASO would like to acknowledge:
Chris Fox, Sandor Gellen, Stephen Morris, Jessica Ozan, with Julian Crockford.

CONTENTS

1. INTRODUCTION

Increasingly, public bodies aspire to draw on evidence that supports a causal interpretation to inform their work. The need for such evidence is beyond question – public bodies want to engage in activities that are effective and proven to be so – but the search for causal evidence also leads them into the difficult and contested area of 'causal' inference and the related paradigm wars in the social sciences.

One increasingly common challenge is the need to reconcile the nature of the programmes and policies for which causal evidence is required with the conventional 'what works' toolkit, comprising systematic review, randomised controlled trial and quasi-experimental design. Many interventions are not compatible with these forms of causal inference, due to their complex nature, resource issues, emergent or developmental features or small scale. Faced with these challenges, evaluators, service-deliverers and policymakers are increasingly seeking alternative understandings of causality to fill the perceived gaps in actionable evidence.

Some of these alternatives are the focus of this report, specifically designed for impact evaluation that can be used with small cohorts: so-called 'small n' impact evaluations.

While these alternatives promise to address gaps in our understanding, evaluators and policymakers must be aware of the often subtle but important differences between them and the standard 'what works' toolkit. For example, many of the approaches offered up as alternatives address a different set of impact questions and have a different perspective on the problem of 'causality' when viewed alongside the standard 'what works' toolkit. This guide, therefore, includes a discussion of the philosophies that underpin different causal strategies (see Appendix One for a more detailed discussion of philosophy), the types of questions regarding cause and effect prioritised in different approaches and, thus, the nature of the evidence that will emerge from resulting studies. It then sets out a taxonomy of the various methodologies before describing each one in detail, including an overview of the methodology, the key stages in using it and an indication of the resources, skills and experience that evaluators may require to implement the methodology.

It is important to note that no approach removes the fundamental challenge of uncertainty, nor ultimately the requirement for policymakers to exercise their judgement, recognising that all judgements are provisional and highly contingent. At the heart of the scientific method is a recognition of uncertainty and an acknowledgement that gaining knowledge is an incremental and dynamic process. Each of the methods we discuss approaches the problem of uncertainty in a different way, but all are explicit in their acknowledgement of this issue.

1.1 IMPROVING ACCESS AND WIDENING PARTICIPATION

Improving access and widening participation activities in Higher Education (HE) are delivered through a range of activities. Drawing on a pre-existing analysis of university access plans, Harrison et al. (2019) divided these into four broad categories:

- Campus visits
- Mentoring and Tutoring
- Summer Schools
- Academically Focused Activities

Such activities are designed to produce a range of outcomes, including the development of individuals' confidence, motivation and self-efficacy, positive attitudes to school or higher education, increased academic attainment and social and academic capital (Harrison et al., 2019; Harrison and Waller, 2017).

1.2 OUTREACH EVALUATION IN POLICY

Effective evaluation of these widening participation (WP) activities has assumed increasing importance in policy over the last 15 years. The first guidance from the Office for Fair Access (OFFA) focussed primarily on monitoring, to ensure activities were reaching the intended cohort and investment was correctly targeted (OFFA, 2004). The period of austerity beginning in 2008 heightened ministerial and regulatory concerns over the value-for-money aspects of these activities, prompting calls for evaluation to focus on identifying the effectiveness of interventions. By 2012, at the same time as the trebling of tuition fees (and consequently a ramping up of WP-spend by universities), OFFA was stressing the importance of evaluation, both to demonstrate to the government the value of WP investment and as a means of identifying best practice to guide the sector towards the interventions with the greatest impact (OFFA, 2013). The National Strategy for Access and Student Success, published jointly by OFFA and the Higher Education Funding Council England further emphasised the importance of taking an evidence-informed approach to WP work (BIS, 2014). The current regulator, the Office for Students (OfS), has intensified this focus on evaluation approach and methodology, issuing guidance on three standards of evaluation:

- Type 1 (narrative) – theory informed
- Type 2 (empirical) – change across time
- Type 3 (epistemological) – counterfactual (OfS 2019)

As the OfS guidance on evaluation has evolved, it has combined a focus on positivist trial-based evaluation designs with the philosophy and approach of the *What Works* initiative, which aims to 'create, share and use [...] high quality evidence' to improve outcomes for public sector organisations[1] and theory-informed approaches, rooted in an understanding of how interventions are expected to work (OfS, 2019). The evaluation typology set out in the pages below reconfigures this distinction into those approaches focusing on the *effects of causes* and those focusing on the *causes of effects*.

1.3 SECTOR PROGRESS IN DEVELOPING EVIDENCE OF 'WHAT WORKS'

Despite this regulatory pressure, the HE sector has often struggled to generate effective and robust evaluation outcomes for WP outreach interventions, perhaps indicating the challenges of evaluating such interventions. Harrison and Waller's (2017) research, drawing on interviews conducted in the mid-2010s, reveals that the majority of outreach evaluation was conducted through participant questionnaires, time-series analysis of data, teacher questionnaires and longitudinal tracking. This reliance on familiar and established monitoring and evaluation approaches led to criticism of the sector's ability to evidence impact. The evaluators of the national *Aimhigher* outreach programme, which ran in England between 2004 and 2011, for example, pointed to an over-reliance on qualitative data and small scale quantitative studies that identified correlations rather than causal relationships (Passy et al., 2009; Passy & Morris, 2010).

Many of the methodological criticisms levelled at UK WP evaluation originate from a post-positivist-inflected position.[2] Following a review of WP research, Gorard and Smith (2006), for example, criticised the validity of existing UK WP research on the grounds that it frequently lacked sufficient detail to make a judgement on quality and robustness or to test replicability, that links between the evidence presented and the conclusions drawn were weak, that most studies lacked an adequate control or robust comparator groups, and that they showed no consensus on how to compare differences over time and space. Similarly, a Sutton Trust report (Torgerson et al., 2014, p. 4) drew a distinction between robust US evaluations using 'systematic review, meta-analysis, experimental, regression discontinuity and other quasi-experimental designs', with UK access strategies 'with limited evidence of their promise from studies using weaker experimental designs'. From a similar standpoint, Younger et al. (2019) also noted the lack of robust evaluation of UK-based interventions. In a review of sector-submitted outreach evaluation reports, Robinson and Salvestrini (2020) noted a tendency for WP evaluation to focus on intermediate rather than long-term HE progression outcomes and observed that much of the evidence included in reviews did not adequately demonstrate causality.

[1] https://whatworks.blog.gov.uk/about-the-what-works-network/

[2] See Appendix One for a discussion of post-positivism.

1.4 THE CHALLENGES FACED BY WP EVALUATORS

Several studies have explored the challenges frequently faced by practitioners in delivering effective evaluation of WP interventions. These challenges can be divided into practical and logical factors and epistemological challenges.

Some of these challenges stem from capacity issues. Reporting the outcome of interviews with a range of HE-based WP evaluators, Crawford et al. (2017) report that interviewees expressed concerns about the lack of resources allocated to evaluation by their institution, the prevalence of split roles in which individuals acted as both practitioner and evaluator, limited access to academic guidance, and frequent staff turnover leading to difficulties in retaining relevant expertise and experience (pp. 12–17). In a similar study of the experiences of WP practitioner-managers, Harrison and Waller (2017) found that 91% were seeking to improve their evaluation practices, while Harrison et al. (2015) noted that respondents felt that Higher Education Providers did not adequately resource the evaluation function.

In these studies, evaluation practitioners also highlight the restrictions caused by data access issues to their ability to carry out effective outreach evaluation, particularly where the intervention's impacts are deferred over a long timeframe or where it is run in collaboration with a range of partners (usually schools and colleges). In their interim report on the impact of the *Aimhigher* network, Passy et al. (2009) reported issues in the evaluators' ability to collect consistent data about individual students and their progression across a complex and changeable education ecology (Passy et al., 2009). Similarly, Gorard and Smith (2006) found issues with the large-scale datasets used to evaluate the impact of WP interventions: 'they include only participants; have incomplete coverage; have substantial proportions of missing data or cases; have changed key definitions over time; or are incompatible in range or aggregation with other datasets' (p. 591). This was supported by the manager-practitioners in Harrison and Waller's (2015) working paper, who noted 'two closely related challenges around data – availability and quality' (p. 19). Interviewees in a study by Crawford et al. (2017, p. 1) described similar issues...

> *"... around getting access to data, having all of the data that is needed, having data in the form that you need it, getting data at the time that you need it"*

In a report focusing on the evaluation of activities for school pupils aged 16 and below, Harrison et al. (2019) note practitioners' practical concerns around data collection, including the need to observe the requirements of data protection legislation, the disruptive impact of collecting data during activities, the time taken to clean and analyse datasets and the challenges of working with large cohorts, small cohorts and sample sizes, and young participants. Some respondents were also concerned about whether the data available (e.g. grade outcomes) was sufficiently sensitive to reflect activity impacts appropriately.

Evaluation practitioners also reported issues around measuring the longitudinal HE progression impacts of WP interventions, particularly when delivered to younger participants (Crawford et al., 2017; Harrison et al., 2019). While these are partially due to the data issues discussed above, practitioners observed that they had no reliable way of disentangling the causal impact of individual activities from a wide range of confounding factors (Crawford et al., 2017, p. 5) since the potentially long gap between intervention and HE progression outcome provides an opportunity for a range of such confounding factors to intervene and impact on outcomes (Harrison & Waller 2017b, p. 4). It is for this reason that Harrison and Waller (2017b) propose the adoption of a 'small steps' evaluation approach, focusing on intermediate impact outcomes that are likely to support HE progression, rather than the eventual progression outcomes themselves. By significantly reducing the evaluation timeframe, a more robust foundation is created for causal impact claims.

Many outreach interventions are designed as a component of a wider programme of interrelated activities, or to have multiple impacts. This makes it difficult for evaluators to focus on particular intervention outcomes. For this reason, many of the commentators critical of the sector's evaluation progress refer to these interventions as 'black box' activities, because they combine two or more outcomes (Robinson & Salvestrini, 2020; Torgerson et al., 2014; Younger et al., 2019). Harrison and Waller (2017b, p. 157) observe that this uncertainty about outcomes can lead to the 'absence of a clear epistemology for assessing success'. In other words, the intended outcome of a particular intervention is not always clear to the evaluator. Moreover, Harrison et al. (2019) observe that where outcomes were articulated, this tended to be in the form of wideranging conceptualisations such as 'raising aspirations', which were often loosely defined and opaque (p. 27). This lack of clarity is at least partially due to the slow progress of embedding theory-informed evaluation approaches in WP outreach evaluation. As Harrison and Waller (2017b, p. 5) observe:

> *Outreach activities are, at their heart, about causing change within individuals. If practitioners expect to cause change, then they need to have a clear articulation of the mechanisms by which they expect this to occur at the individual level – a Theory of Change.*

Based on their interviews with WP managers and practitioners, Harrison and Waller (2015, 2017a, 2017b) identify the epistemological challenges faced by outreach evaluation practitioners relying on conventional evaluation methods (such as surveys, focus groups and interviews) as a reliance on self-reported data from participants with the risk of introducing attendant biases (e.g. priming, social desirability and Dunning Kruger effects) (Harrison & Waller, 2017b).

Nonetheless, in recent years, some outreach evaluators have started to draw on a wider range of evaluation practices, including quasi-experimental evaluation designs and mixed-method approaches (Horton & Hilton, 2020) and realist approaches (Formby et al., 2020a; Formby et al., 2020b; Gibson Smith et al., 2021; Pickering 2021).

The small *n* approaches described below help to address several of the issues discussed above, particularly those around an opaque black box approach, small sample sizes, data challenges and the relative paucity of theory-informed evaluation methodologies.

2. TAXONOMY OF METHODOLOGIES FOR SMALL COHORTS

In this section, we set out a taxonomy of methodologies for small cohorts. To construct a taxonomy, it is necessary to address the various understandings of causal inference inherent in these alternative approaches. This section, therefore, starts with a discussion of different ways of understanding causality and causal inference (some of the philosophical issues raised are addressed in more detail in Appendix One). A taxonomy is then described. The part of the taxonomy that is the focus of the remainder of this report – and the wider project of which it forms part – concerns those approaches that are not based on counterfactuals and statistical controls, and in which large sample sizes are not required: so-called 'small n' impact evaluation (White & Phillips, 2012). However, as will be clear in the discussion of the taxonomy and the different conceptions of causal inference embedded within small *n* methodologies, it would be a mistake to see the choice of methodology as informed purely by pragmatic considerations and sample size. To support evaluators considering the various small *n* methodologies, this section finishes with a decision tree to guide choices.

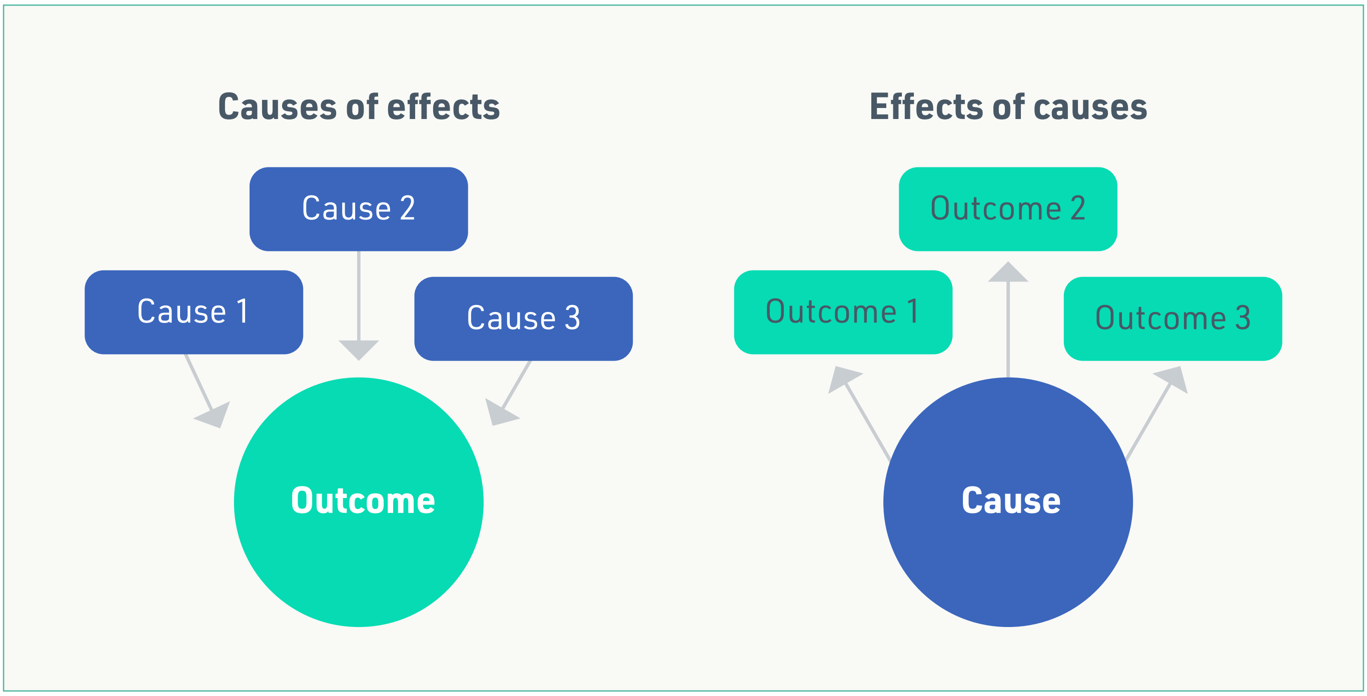
2.1 CAUSAL INFERENCE IN IMPACT EVALUATION

It is important to start with an understanding of the different perspectives of evaluators regarding the problem of causal inference. These perspectives determine the approaches that researchers use, the methods they advocate and why. Causal inference considers the assumptions, study designs and estimation strategies that allow researchers to draw causal conclusions based on data (Hill & Stuart, 2015).

To help advance our understanding, it is worth starting from the fundamental distinction between two types of question that social scientists ask when they use the tools and techniques of social science in evaluation. It is important to add that these two conceptualisations of causality are not always made explicit by researchers in their work.

First, social scientists may ask: What are the effects of a causal factor (i.e. an intervention or treatment)? Second, and in contrast, they may ask: What are the causal factors that give rise to an effect? These questions are what Dawid (2007) calls 'effects of causes' and 'causes of effects' type questions (Figure 1).

Figure 1: Causes of effects versus effects of causes



2.1.1 EFFECTS OF CAUSES

Imagine we wish to evaluate the effects of an intervention introduced to reduce student drop-out among BAME students from undergraduate courses over the first year of study. Following the OfS (2019) standards of evidence, we may opt for an impact evaluation that provides evidence of a causal effect of the intervention through an evaluation design that measures pre/post-treatment change on a treated group, relative to an appropriate control or comparison group. This is what the OfS (2019) refers to as a 'Type 3' evaluation. An evaluation design that embodies a Type 3 evaluation would be a randomised controlled trial (RCT) design, assuming that we wish to know the effects of a particular causal factor – the specified intervention. Due to the nature of an RCT, other factors causing drop-out (as expected over the long term) are distributed equally over two groups – an intervention and a control group. Only the intervention group receives the intervention; the control group does not. At the point at which the two groups are created, through randomisation, they are statistically equivalent to one another. Any average difference in a pre-specified outcome(s) that we subsequently observe between the two groups can be attributed to the intervention and not the other causes of drop-out, subject to standard statistical thresholds for sampling uncertainty. This is a classic 'effects of causes'-type approach. The other factors that influence the outcome are treated as exogenous or given. Their effects are not removed or, as some critics have claimed, bracketed out. Instead, the research design holds other causal factors in balance across the two groups, enabling attention to be focused on the causal factor that can be manipulated by the policymaker – that is the intervention itself.

The most important feature of randomisation is that while no prior knowledge of these other or extraneous causal factors is required, knowledge of them may be extremely useful. Randomisation distributes extraneous causal factors across groups created at random, whether we are cognisant of their existence or not.

The formal development of the idea of randomisation, by Sir Ronald Fisher in the 1920s, was a breakthrough. It enabled experiments to be conducted in field settings (that is outside laboratory conditions) based on the fundamental properties of statistical distribution and statistical inference. Fisher's pioneering work took place in the agricultural sciences (Fisher, 1925). His influence extended quickly into the health sciences where the RCT developed as a methodology and its use became widespread (Armitage, 2003). In the 1970s and 1980s, statisticians further refined and developed Fisher's ideas (Holland, 1986; Rubin, 1977) and found new audiences in policy studies, psychology and economics. Particularly influential was the work of the psychologist Donald T. Campbell, who – through his 'threats to validity' framework – advocated strongly in favour of 'effects of causes'-type questions and randomised or experimental designs as the best way of addressing such questions (Campbell & Stanley, 1963).

2.1.2 CAUSES OF EFFECTS

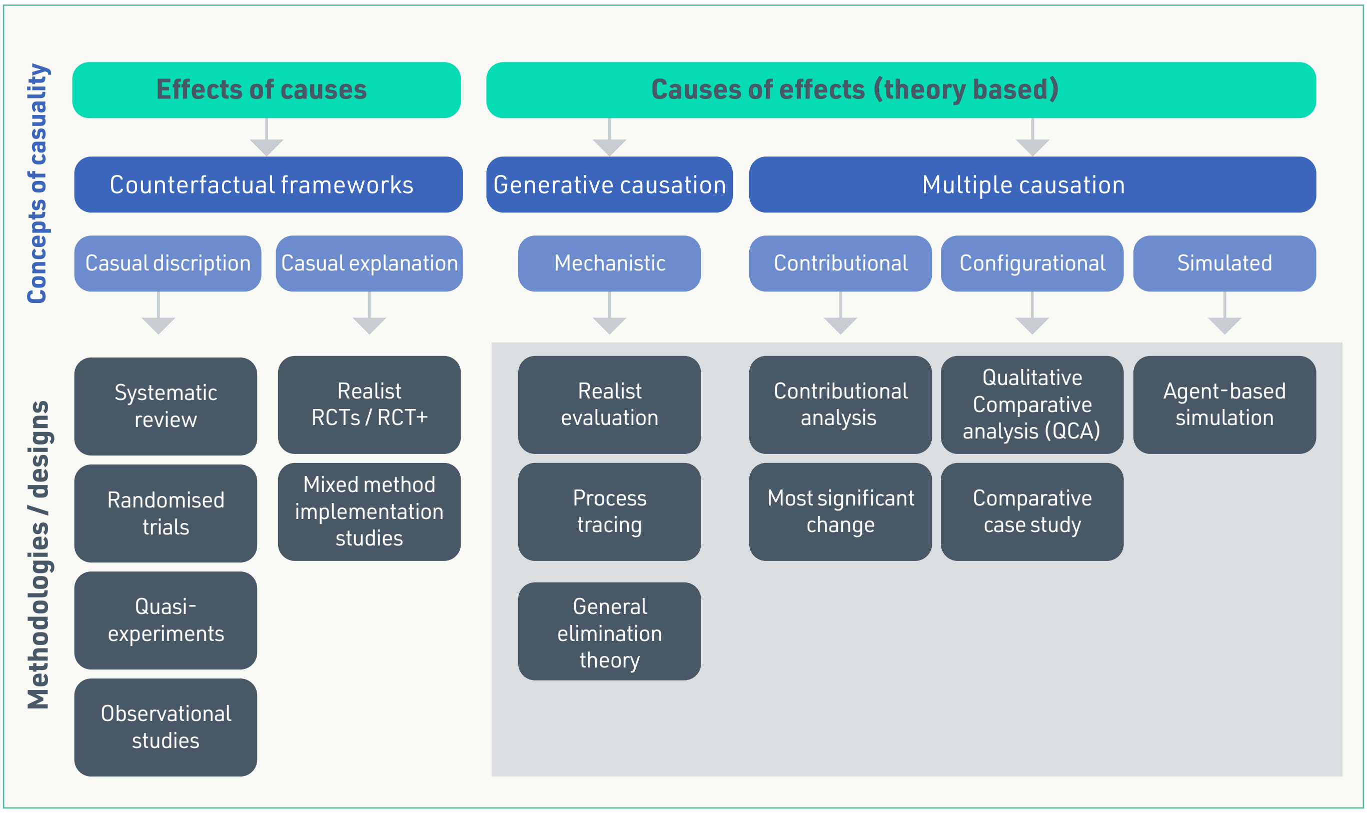
Alternatively, we may evaluate the intervention seeking to reduce drop-out among BAME students by using a realist evaluation approach. In many ways, 'realist' evaluation can be seen as antithetical to the 'effects of causes' stance and, in particular, randomisation. Scientific realism primarily understands an intervention as operating within a context, with multiple factors and influencing outcomes at play. Realist approaches seek to articulate these various factors through specifying different context/mechanism/outcome configurations and testing these constructs empirically. The intervention is understood as producing effects through mechanisms which, as Cartwright and Hardie (2012) describe, act as just one ingredient in a 'causal cake'. The various relevant ingredients need to be identified, defined and explored before the causal question can be fully addressed, and causation or otherwise inferred. Here, the emphasis is on the explanation of the effects or outcomes that are altered through the triggering of causal mechanisms. Specifically, evaluation is concerned with the role of the intervention in these processes. So, for example, if we wish to understand the causal effect of an intervention that seeks to reduce drop-out rates among BAME students, we need to consider its effects in conjunction with all (or all relevant) other causes of student drop-out. In other words, causes work in conjunction or together in causal packages to produce effects (Cartwright & Hardie, 2012).

In some senses, the 'causes of effects'-style question is a more comfortable ground for many social scientists and has traditionally been the focus for many social scientists. It is more consistent with those types of research that have been the mainstay of the social sciences over the past 20 to 30 years. Typically, social science research is largely naturalistic or observational. There is no attempt to manipulate a particular factor through randomising it; instead, policymakers and practitioners observe interventions in a more naturalistic setting as one factor among many that influence outcomes.

2.2 A TAXONOMY OF IMPACT EVALUATION

The taxonomy of impact evaluations set out in Figure 2 is organised around different concepts of causality. As a starting point, we have distinguished two groups of impact evaluation designs based on whether the focus of the evaluation is 'causes of effects' or 'effects of causes'. This initial distinction leads to further distinctions between different concepts of causality and, hence, to different evaluation methodologies and designs. The ultimate focus of this project is the methodologies and designs to the right of the diagram, in the grey shaded box, that sit apart from the more common, standard 'what-works'-type toolkit.

Figure 2: A taxonomy of impact evaluation designs



2.3 EFFECTS OF CAUSES AND COUNTERFACTUAL EVALUATION

These designs involve the concept of manipulation (Shadish et al., 2002); that is, some causal factor, treatment or intervention, is manipulated – it is introduced, scaled up, scaled down or ended. Moreover, evaluators either have some knowledge of how the causal factor is manipulated or can intervene directly in its manipulation, as is the case in a randomised design.

2.3.1 CAUSAL DESCRIPTION AND EXPERIMENTATION

The process of manipulation enables the researcher to uncover the counterfactual. In essence, an intervention or programme is targeted at some clearly defined population. For example, an intervention that is designed to address the problem of student drop-out among BAME students from undergraduate courses over the first year of study would target the first year BAME undergraduate student body in a particular university. For each member of this population, two outcomes are understood to be possible, and these are known as potential outcomes (Gerber & Green, 2012; Holland, 1986). The first is the potential outcome that would be observed for a member of the population under the intervention or under 'treatment', all else being equal. The second potential outcome is the one that would hold if the same member of the population remained 'untreated', again all else being equal. Thus, in the simplest case, there are two potential outcomes for every member of the population. The average treatment effect (ATE) is defined as the sum of the difference in the two potential outcomes for every member of the population, divided by the total number of people or units in the population. It will be obvious that this quantity cannot be observed or known. We cannot observe potential outcomes for members of the population in two different states simultaneously.

Randomisation seeks to address this problem by intervening directly and manipulating exposure to the intervention by randomly allocating members of the population to intervention or control groups. Through this process of randomisation, two sample potential outcomes are generated: average outcomes from a control group and average outcomes from an intervention group, in other words, random samples of outcomes under treatment and without treatment. Since these samples of potential outcomes are generated randomly, their averages are unbiased, and the difference in these averages – the average treatment effect – is also unbiased.

In many applications, however, manipulating the intervention or treatment in this way, at random, is not possible. Some other set of factors, usually choice, in our example, by either the student or an administrator, determines whether a member of the population is exposed to an intervention or not. Thus, the outcomes we see are not necessarily good estimates of the potential outcomes.

We may be able to intervene in some other way (that is, other than randomising) to determine who is exposed to the intervention (based on an explicit rule, for example) or collect enough information about how choices are made to model exposure. In both these situations, we use statistical techniques to adjust our analysis to try to improve our estimates of the potential outcomes. These approaches are known as quasi-experimental, in that they attempt to mimic an experiment by using some other form of manipulation and/or statistical techniques to adjust the analysis to take account of likely biases. In some rare circumstances, it may be possible to find an element in the way an intervention or programme is designed and implemented that may lead us to consider exposure to it as effectively at random (so-called natural experiment), even though this has not been explicitly designed into the study.

Generally, under the counterfactual approach, policymakers and those designing programmes are encouraged not to make decisions or to act based on results from a single study or a small group of studies. Proper inference can only be made when many studies of a treatment, programme or intervention have been amassed and have been summarised through systematic review or meta-analysis. Ideally, these studies would span multiple settings, different populations and be conducted at different time points. This would permit nuanced and fine-grained inferences, enabling researchers and policymakers to understand the context, types of populations and periods over which a given treatment or intervention 'works'. This concept is reflected in the operation of the 'What Works' centres, including TASO, which collate the evidence available from rigorous impact evaluations and pull it together in systematic reviews and meta-analyses. It is also reflected in evidence standards such as that developed by Nesta (Puttick & Ludlow, 2012) in which demonstrating causality using a control group in a single study is not at the top of the 'evidence hierarchy' but, instead, mid-way, with replication studies and meta-analyses above it in the hierarchy.

All the cases we have discussed in this section, however, are interested in the causal effects of an intervention, that is, the 'effects of a cause'. An attempt is made, either through design and conscious manipulation or through statistical adjustment, or both, to identify the effects of the cause without explicitly attempting to account for or model all the other causal factors that may influence the outcome. These other causal factors are considered extraneous or exogenous (i.e. having an external origin). They are only of interest in that they enable statistical estimation to be more precise than it might otherwise be – they are not the substantive focus.[3] The focus is solely on the intervention, programme or treatment – the substantive phenomenon over which the policymaker or practitioner has some influence.

This broad approach has several benefits:

- Through design and/or statistical adjustment the challenge of causal inference is simplified. The policymaker or researcher does not require full or perfect information about the full range of causal factors that influence the outcome, nor are they required to rely on the theory that potentially lacks credible falsification.
- Due to its reliance on statistical theory, the 'effects of causes' approach will provide not only evidence of whether a causal effect is present but also the magnitude and direction of any effect, as well as a measure of sampling uncertainty associated with that estimate.
- The statistical estimates obtained through this approach can form an input into cost-benefit analysis, cost-effectiveness estimates or break-even analysis. When combined with this information, the policymaker can compare the cost-effectiveness of the different uses to which scarce public resources may be allocated.

There are also limitations to the "effects of causes" approach:

- The simplification of the causal problem may gloss over important relationships between causal factors that act together to produce outcomes. While the influence of all causal factors that affect an outcome is not 'removed' from estimates obtained from, for example, a randomised experiment, they are not specifically accounted for. This could be important in circumstances where policymakers are attempting to determine whether an intervention that appears to be effective in a particular setting might 'work' in other settings, or among other populations.
- Experimental or quasi-experimental designs, along with many other approaches in the social sciences, may be considered too static – they provide a snap-shot picture not only of what was happening in a particular setting, for a specified population, but also at a particular point in time. Relevant circumstances may change rapidly, rendering results from a particular study misleading.

2.3.2 CAUSAL EXPLANATION, REALIST AND MIXED METHOD RCTS

The limitations described in the preceding section can be addressed *within* the 'effects of causes'/'counterfactual frameworks' approach to impact evaluation. Attempts to address these problems tend to recognise a distinction between 'causal description' and 'causal explanation', as described by Shadish, Cook and Campbell (2002, p. 9):

> *The unique strength of experimentation is in describing the consequences attributable to deliberately varying treatments. We call this causal description. In contrast, experiments do less well in clarifying mechanisms through which, and the conditions under which, that causal relationship holds – what we call causal explanation.*

[3] We a set aside here for the sake of brevity extensions to the effects of causes approach that do seek to elaborate on interactions between programmes/interventions/treatments and other variables initially considered extraneous. We have in mind here moderator and mediator analysis often incorporated into randomised and quasi-experimental designs (Baron & Kenny ^{1986}, and many others). Whilst these strategies certainly muddy the waters when it comes to the distinctions we are making, mediator/moderator analysis is still quite rare in experimental policy evaluations, due to sample size restrictions and the required assumptions.

At least three broad types of question remain unanswered in the results from standard experimental analysis: 1) questions relating to the processes or mechanisms that generate or lead to the observed effects; 2) questions addressing the factors present in the *context* in which the study took place that may enable or constrain the operation of the intervention; and 3) questions connected with the implementation of the intervention and how far implementation fidelity was achieved (Moore et al., 2015). Attempts to design experiments that address these three types of question more fully can be seen as striving to strengthen causal explanation within experimental approaches to impact evaluation.

To address the first group of questions relating to causal processes or mechanisms, researchers have resorted to estimating structural statistical models and forms of mediator and moderator analysis (Imai et al. 2011). Sample sizes in many randomised interventions, however, tend to underpower moderator analyses (Frazier et al., 2004). Moreover, the assumptions required for the valid analysis of mediators are often highly restrictive (Keele, 2015; Suzuki & VanderWeele, 2018). These limitations have been a significant motivating factor in leading many researchers to the use of mixed-method and qualitative research to explicate causal processes or mechanisms (Bamberger, 2015; Bonell et al., 2012; Jamal et al., 2015; White, 2013).

This takes us to the second type of question that remains unanswered in standard experiments: those about context. Context can refer to differences in 'treated' populations but also geography, historical trends, political institutions, physical environment, available resources or alternative opportunities. Cartwright and Hardie (2012) refer to these influences as supporting factors. Knowledge of the factors prevailing within the context in which a specific study is conducted is important in understanding the extent to which the results might hold elsewhere and indeed to improve the quality of inferences that can be drawn from meta-analysis and research synthesis. Qualitative approaches have been advanced as means of understanding and incorporating knowledge of wider contextual factors into randomised designs (Bonell et al., 2012; White, 2009).

Understanding the third type of question about implementation, and an intervention's fidelity to the intended design, has been a key concern in evaluations using randomised designs for many years. To interpret results from randomised studies, knowledge of the nature of the intervention, its implementation, its fidelity to the intervention design and nature of exposure is essential. In the field of health services research, process evaluations have been widely integrated into randomised studies (Moore et al., 2015; Oakley et al., 2006). More recently, in education, formal methods of implementation process evaluation have been proposed, drawing on the development of process evaluation approaches and techniques in health research and elsewhere (Humphrey et al., 2016).

In international development, for example, what are variously termed mixed-method, RCTs or RCT+ designs have been discussed and implemented (Bamberger et al., 2016; White, 2013). In health research, there is a long tradition of promoting mixed-method intervention studies (Boeije et al., 2015; Hansen & Jones, 2017; Johnson & Schoonenboom, 2016; Moore et al., 2015; Oakley et al., 2006). In education, the growing use of randomised designs has been accompanied by an increased emphasis on studies that combine randomisation with mixed-method implementation process evaluation (Humphrey et al., 2016; Lendrum & Humphrey, 2012).

2.4 CAUSES OF EFFECTS AND THEORY-BASED, SMALL *N* IMPACT EVALUATION

All the small *n* methodologies we describe have certain elements in common: the importance of starting an impact evaluation by specifying mid-level theory and the use of cases. They can be distinguished by the different understandings of causation that they draw on: generative or multiple causation.

2.4.1 Mid-level theory

All the approaches to impact evaluation associated with uncovering 'causes of effects' involve specifying mid-level theory or a theory of change together with alternative causal hypotheses. Causation is established beyond reasonable doubt by collecting evidence to validate, invalidate or revise hypothesised explanations (White & Phillips, 2012). In some of the impact evaluation designs and methodologies that fall within this part of the taxonomy, theory-building and testing are explicit elements of the approach; in others, they are present but less central to the methodology.

2.4.2 CASES

Key to all small *n* approaches is the concept of the 'case'. These relatively recent methodologies and designs for impact evaluation are to be distinguished from traditional understandings of 'case studies' (Stern et al., 2012). The tradition within the evaluation of naturalistic, constructivist and interpretive case studies that generally focus on the unique characteristics of a single case may contribute to a richer understanding of causation but cannot itself support causal analysis (Stern et al., 2012). In contrast, these approaches that use small numbers of cases are interested in generalising beyond a single case but distinguish 'generalising' from 'universalising' (Byrne, 2009, p. 1).

Cases are generally seen as complex systems where:

> *[T]rajectories and transformations depend on all of the whole, the parts, the interactions among parts and whole, and the interactions of any system with other complex systems among which it is nested and with which it intersects.*

(Byrne 2009: 2)

A key distinction between case-based approaches and experimental designs is the rejection of analysis based on variables (Byrne, 2009). Advocates of case-based approaches reject the 'disembodied variable' (Byrne, 2009, p. 4). The case is a complex entity in which multiple causes interact:

> *It is how these causes interact as a set that allows an understanding of cases… This view does not ignore individual causes of variables but examines them as 'configurations' or 'sets' in their context.*

(Stern et al. 2009: 31)

Small n, case-based methodologies are varied, but Befani and Stedman-Bryce (2017) suggest that case-based methods can be broadly typologised as either between-case comparisons (such as qualitative comparative analysis) or within-case analysis (for example, process tracing). Generally, quantitative and qualitative data are used and hard distinctions between quantitative and qualitative methods are rejected (Stern et al., 2012).

Some of the 'small *n*' methodologies can only be used with small numbers of cases, while others, such as Qualitative Comparative Analysis (QCA) can involve either small or relatively large numbers. However, it is important to recognise that the choice of a small *n* methodology is not simply a pragmatic one. Occasionally, the choice to use a small *n* methodology may be driven by budgetary or political constraints that prevent the recruitment of a sufficient sample size or comparison group to make a counterfactual impact evaluation design possible (White & Phillips, 2012). However, the choice to use a small *n* approach is more likely to be driven by methodological considerations such as the nature of the cases (for example, where cases are institutions and only a small number exist) or where there is significant heterogeneity within the treatment population or the wider context for the intervention, or where the intervention is itself emergent or developmental (White & Phillips, 2012).

2.4.3 GENERATIVE AND MULTIPLE CAUSATION

Within what might broadly be classified as 'causes of effects' approaches to impact evaluation, Stern et al. (2012) make a broad distinction between approaches based on the concept of generative causation and those based on multiple causalities which depend on combinations of causes that lead to an effect, whereas generative causation is closely associated with identifying mechanisms that explain effects.

Generative causation

The generative conception of causation 'sees the matter of causation "internally". Cause describes the *transformative potential* of phenomena' (Pawson & Tilley, 1994, p. 293, original emphasis). Generative causation depends on identifying the 'mechanisms' that explain effects. This is the inferential basis for 'realist' approaches to impact evaluation (Stern et al., 2012), but is also important in approaches such as Process Tracing and the General Elimination Method where identifying and tracing mechanisms is also a central task of the evaluation design. Generative causation can be contrasted with successionist causation, which is associated with the ideas of Hume. In successionist models of causation, 'Causation is "external" in that we do not and cannot observe certain causal forces at work' (Pawson & Tilley, 1997, p. 33).

Using the concept of generative causation means that realists do not make predictions about the probability of an intervention leading to an outcome, 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 (Wong et al., 2013).

Mechanisms

Key to causal inference is the idea of 'mechanisms', which generate outcomes but do so in particular contexts (Pawson & Tilley, 1997).

> *A mechanism consists of individual 'parts', each with its own properties, and these components act, or interact, to bring about outcomes. Each of these parts could be said to be individually insufficient but collectively necessary for outcomes to occur. The outcomes that are produced will depend on the properties of these parts – their structure, duration and temporal order – but also on the context in which they operate.*

(White and Phillips 2012).

A mechanism explains what it is about a programme that makes it function effectively. Mechanisms are not variables but accounts that encompass agency and structure. They, thus, 'reach down' to individual reasoning and 'reach up' to the collective resources embodied within the social programme being evaluated (Pawson & Tilley, 1997): 'A mechanism is thus a theory – a theory which spells out the potential of human resources and reasoning' (p. 69).

Mechanistic approaches to causal inference can be clearly distinguished from approaches based on counterfactuals, as observed by White and Phillips (2012), 'Whereas experimental approaches infer causality by identifying the outcomes resulting from manipulated causes, a mechanism-based approach searches for the causes of observed outcomes.' (White and Phillips 2012, p.22)

As White and Phillips (2012) note, mechanism-based explanations are not merely historical narratives, but require detail to understand the mechanisms in play and critically reconstruct each link in the causal chain. The best examples should not only ascertain whether the evidence supports a theorised explanation of cause and effect, but also whether the observed effects might be produced by other (known or unknown) mechanisms. In this sense, they argue, mechanism-based explanations include implicit counterfactuals.

Multiple causation, contributions, configurations and simulations

Stern et al. (2012) note that, in classical approaches to causal inference, causality is established by seeking a strong association between a single cause and a single effect. This occurs either by observing a regular combined presence of cause and effect in a number of diverse cases (Hume's regularity and Mill's Method of Agreement) or through the observation of quasi-identical cases in which only the cause and the effect are different (Mill's Method of Difference). Stern et al. characterise this approach as conceiving of the cause as being both necessary and sufficient for the effect. They note that this approach is not usually able to untangle the complexities of causal relations when causes are interdependent and affect outcomes as 'causal packages' rather than independently. The focus of this approach to causation is on attribution (Stern et al., 2012).

However, in small *n* methodologies, when multiple causes are recognised, the focus tends to switch to understanding the *contribution* of an intervention to an observed outcome. Thus, the notion of a 'contributory' cause recognises that effects are produced by several causes at the same time, none of which may be necessary nor sufficient for impact. This, in turn, leads to several impact questions that go beyond attribution to develop an understanding of *how* an intervention contributes to an observed effect (Stern et al., 2012):

> *If a causal 'package', i.e. the intervention plus other factors, is the concept most likely to be relevant in the impact evaluation of complex… projects, this focuses attention on the role of the intervention in that package. Was it a necessary ground-preparing cause, a necessary triggering cause or something that did not make any difference and a similar effect would have occurred without the intervention?*

(Stern et al., 2012, p. 40)

Questions of contribution often also lead to questions of *configuration*, for instance, the sequencing of and relationships between multiple causes. Thus, for Pawson (2008):

> *Configurationists begin with a number of 'cases' of a particular family of social phenomenon, which have some similarities and some differences. They locate causal powers in the 'combination' of attributes of these cases, with a particular grouping of attributes leading to one outcome and a further grouping linked to another. The goal of research is to unravel the key configurational clusters of properties underpinning the cases and which thus are able to explain variations in outcomes across the family.*

(p. 1)

This leads to 'causal imagery' which evokes the notion of interventions as part of complex systems (Pawson, 2008).

2.5 DECIDING WHICH METHODOLOGY TO USE

2.5.1 THINKING ABOUT EVALUATION

Stern et al. (2012) and Befani (2020) identify three key elements to be considered when making decisions regarding the most appropriate evaluation approach:

Resources and constraints:

- Does the institution have sufficient resources (e.g. direct or in-kind funding) to conduct an evaluation? Issues to consider may include the burden associated with data collection for both students and practitioners, ethics, likely response rate and level of engagement, and the possible impact that data collection may have on the intervention and student outcomes.
- When is the evaluation report needed (i.e. what is the time available)? In the context of higher education, the academic calendar (i.e. holidays) be considered alongside the nature of the data collection required and the timeframe of the evaluation.
- Do the institution's staff have the necessary expertise (i.e. knowledge and skills)?

The nature of the programme:

- Are elements of the programme simple, complicated or complex?
- Are there gaps in the logic or evidence available that the evaluation should focus on?
- How diverse is the target group?
- Is the programme about to be launched, running or completed?

The nature of the evaluation:

- What type of question does it want to answer?
- What are the requirements of key stakeholders?

It is important to note that a negative or unclear response to some of these questions may suggest that the time is not yet right to undertake an evaluation and that further preparatory work is required. For instance, when choosing the most appropriate methodological approach, questions relating to resources and constraints are particularly important. If the practitioner does not have sufficient resources and/or expertise, or the timeframe is not realistic, it is better to reconsider the purpose of the evaluation and the potential harm caused if methods were poorly employed. Some of the methods outlined here may require institutional investment (i.e. staff training and time) before they can be used effectively.

2.5.2 A COMPREHENSIVE TOOL FOR CHOOSING IMPACT EVALUATION DESIGNS

Befani (2020) has developed a tool to select appropriate impact evaluation methodologies. It covers a wide range of 'small *n*' impact methods as well as 'traditional' counterfactual evaluation designs and was developed through a structured consultation with industry experts, building on the work of Stern et al. (2012). The 'Choosing Appropriate Evaluation Methods Tool, Version 2.1' is an Excel spreadsheet that can be downloaded from the CECAN website at: https://www.cecan.ac.uk/news/choosing-appropriate-evaluation-methods-a-tool-for-assessment-and-selection-version-two/. We strongly recommend that you familiarise yourself with this tool and use it as an aid to decision-making. However, below we set out a simpler set of questions and suggestions for methodologies, based in part on Befani's tool.

2.5.3 CHOOSING A SMALL *N* IMPACT EVALUATION METHODOLOGY

The questions below draw in part on Befani's (2020) tool and in part on the issues mentioned above to provide some relatively simple guidance when selecting an appropriate small *n* impact methodology.

A useful starting point for selecting a methodology is to think about the main evaluation question to be addressed. Figure 3 sets out some broad evaluations and indicates the methodologies that might be most appropriate to address them. It is important to recognise that there will be an element of judgement in the selection of methodologies, so this table should not be read as a definitive or categorical statement of methodology suitability.

Figure 3: Appropriate small *n* methodologies to address different evaluation questions (based in part on Befani, 2020)

Evaluation question	Appropriate small *n* methodologies
'What was the additional/ net change caused by the intervention?' *or* 'How much of the observed outcome(s) can be attributed to the intervention?' *(Note: this is a core question of experimental and quasi-experimental impact evaluations.)*	Agent-Based Modelling
'What difference did the intervention make to different population groups, and under what circumstances?' (i.e. you are interested in effects in different groups and contexts, not just an 'average' effect).	Qualitative Comparative Analysis, Agent-Based Modelling, Realist Evaluation, Contribution Analysis
'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?'	Agent-Based Modelling, Realist evaluation, Process Tracing, General Elimination Theory, Contribution Analysis
'What other factors needed to be present alongside the intervention to produce outcomes observed?' (Which factors were necessary and/or sufficient for the intervention to work?) *(Note: this is a focus area of some evaluations where the intervention is not assumed to be the sole cause of change, but works in conjunction with other factors/ interventions.)*	Agent-Based Modelling, Realist Evaluation, Qualitative Comparative Analysis, Process Tracing, General Elimination Theory, Comparative Case Study
'Which outcomes of the intervention(s) being evaluated do different population groups consider to be the most important?' *(Note: this seeks to understand the relevance of the outcomes to different population groups or stakeholders.)*	Most Significant Change

Befani (2020) suggests a second set of questions that may guide evaluation design choices, relating to additional interests and preferences of the evaluator or other stakeholders. These are set out in Figure 4.

Figure 4: Features of interest that may guide the choice of small *n* (based closely on Befani, 2020)

Desirability of evaluation addressing each of the following areas of interest?	Appropriate small *n* methodologies
I want to be able to extrapolate or generalise the evaluation findings outside the cases or sample used for the analysis (external validity).	Realist Evaluation, Qualitative Comparative Analysis, Comparative Case Study
I want to allow the community(ies) in which the intervention was carried out to produce a collective evaluation of the most relevant changes at a community level.	Most Significant Change
I want to make a distinction between the achievement of minimum /expected goals and ideal/more ambitious programme goals.	Agent-Based Modelling
I want to explore the higher-order goals or values of the participants (such as attitudes, norms, values and laws shaping their worldview, probing why certain results mattered to participants).	Most Significant Change
I want the evaluation to capture a broad, systemic view of the situation (e.g. understanding how historical forces or path dependency or power relations or the economic system affect results, and seeing how those factors interact).	Agent-Based Modelling

Desirability of evaluation addressing each of the following areas of interest?	**Appropriate small *n* methodologies**
I want the evaluation to make explicit different perspectives about results or the causes of results, particularly between programme participants representing different groups or households within the community, including the weakest.	Most Significant Change, Agent-Based Modelling
I want the evaluation to identify and explain unintended changes and consequences, both positive and negative.	Most Significant Change, Agent-Based Modelling, Realist Evaluation, Qualitative Comparative Analysis, Process Tracing, General Elimination Theory, Comparative Case Study
I want to analyse complicated/complex mechanisms, including outcomes of non-linear relationships, vs. a predominantly linear description of the programme theory of change.	Agent-Based Modelling, Realist Evaluation
I want to bring together varied groups of stakeholders to build consensus on their system; uncover misunderstandings, assumptions, and differences of opinion; develop management and decision-making capacity; build and digitise causal maps, focusing on what is most useful to users, utilising subjective information from stakeholders and network structure.	Agent-Based Modelling
I want to communicate results and processes easily to non-specialists or others not involved in the analysis/evaluation.	Most Significant Change, Contribution Analysis
I want to simulate emergent properties and dynamics, and explain changes over time.	Agent-Based Modelling, Realist Evaluation
I want to grow a system and change its macro properties starting from micro rules of behaviour.	Agent-Based Modelling, Realist Evaluation
I want to generate estimates of the contribution of different factors and conduct 'what-if' analyses.	Agent-Based Modelling
I want to handle uncertainty and scarce/different types of data in complicated/complex contexts.	Agent-Based Modelling, Realist Evaluation, Qualitative Comparative Analysis, Process Tracing, Contribution Analysis, General Elimination Theory, Comparative Case Study
I want to obtain insights into the behaviour, attitudes and thinking of stakeholders.	Realist Evaluation, Process Tracing, General Elimination Theory, Contribution Analysis
I want to identify the various conditions that enable change in different contexts, as opposed to seeking a universal, population-wide or average explanation.	Realist Evaluation, Qualitative Comparative Analysis, Comparative Case Study
I want the evaluation to investigate the factors that are necessary and/or sufficient for the intervention to produce results.	Agent-Based Modelling, Qualitative Comparative Analysis, Comparative Case Study, Process Tracing, General Elimination Theory
I want the evaluation to measure confidence in one or more causal claims and, for example, determine whether the evaluation evidence is strong and conclusive for such claims.	Process Tracing, General Elimination Theory
I want the evaluation to provide a detailed description of the process, leading from programme activities to outputs, to intermediate outcomes and finally impacts.	Contribution Analysis

Finally, Befani (2020) suggests that certain features of the intervention or evaluation context may need to be present for different methods to be correctly applied. These are set out in Figure 5.

Figure 5: Feasibility of each small *n* method to evaluate the intervention, given the requirements of those methods and the context of the evaluation or intervention (based closely on Befani, 2020)

How feasible is it to use each method to evaluate the intervention, given the requirements of those methods and the evaluation context/intervention attributes	Appropriate small *n* methodologies
To what extent is information on (at least a small number of) factors that are assumed to affect the outcome consistently available across at least 5 or 10 cases? *(Note: each 'case' may refer to an application of the intervention in different locations/contexts, or among different individuals, institutions or groups.)*	Agent-Based Modelling, Qualitative Comparative Analysis, Comparative Case Study
To what extent are excellent facilitation skills available in the evaluation team?	Most Significant Change
To what extent are those conducting the evaluation and those who will be consulted for the evaluation likely to be open to airing different perspectives on the intervention, its outcomes and how change happened (e.g. the power dynamics are such that multiple worldviews could be expressed, rather than only one dominant worldview being voiced)?	Most Significant Change, Agent-Based Modelling, Realist Evaluation, Contribution Analysis
When examining many different cases, to what extent do you expect the evaluation team to be able to consistently gain an understanding of the contextual factors that affected the outcomes of your intervention (e.g. if your intervention addresses different locations, population groups or institutions that may affect how the mechanisms between your intervention and the outcomes work)?	Realist Evaluation
To what extent is the evaluation team able to formulate, test and refine theoretical assumptions about the behaviour, attitudes and thinking of stakeholders (i.e. 'identifying the mechanisms generating the outcomes'; this may require insights into political science, psychology, social sciences, or other specific domains and may, therefore, be easier to achieve in multidisciplinary teams)?	Agent-Based Modelling, Realist evaluation, Process Tracing, General Elimination Theory
To what extent is the evaluation team able to map or understand complicated/ complex mechanisms, including outcomes of non-linear relationships? *(in contrast to following a predominantly linear description of the programme's theory of change).*	Agent-Based Modelling, Realist Evaluation, Process Tracing, General Elimination Theory
To what extent are evaluators able to access the broad range of detailed and high-quality data necessary to answer your evaluation questions, including hard-to-find data? *(Note: this could include – for example – data on sensitive issues, data from conflict-affected or hard-to-reach areas/ populations, minutes of private meetings or personal emails.)*	Process Tracing, General Elimination Theory
To what extent can you be confident that your chosen evaluator is able to set up a theory of change with a causal chain, and risks and assumptions for each step that will help shape complementary or alternative explanations for observed changes, either from scratch or by adapting an existing theory of change for the intervention?	Contribution Analysis
To what extent are you able to access technical skills to write and use models and software?	Agent-Based Modelling, Qualitative Comparative Analysis, Comparative Case Study
To what extent are the available theoretical assumptions solid, ideally validated and tested for sensitivity? For example, are agent behaviour rules and the agent environment sufficiently understood? Is the structural model valid?	Agent-Based Modelling, Realist Evaluation, Contribution Analysis
To what extent do you have access to relatively large amounts of data on environments, contexts and individual or group behaviour?	Agent-Based Modelling, Realist Evaluation
To what extent are you able to quantify estimated relationships between factors in terms of probabilities?	Agent-Based Modelling, Contribution Analysis
To what extent are you experienced in eliciting unbiased subjective assessments from individuals and groups?	Process Tracing, General Elimination Theory

3. METHODOLOGIES FOR SMALL COHORTS

In this section, we describe several methodologies.

Theory of change is a precursor to undertaking an impact evaluation and is the basis of most small *n* impact evaluations as well as several other evaluation approaches.

The small *n* methodologies we describe are:

- Realist evaluation
- Process tracing
- General elimination theory
- Contribution analysis
- Most significant change
- QCA
- Comparative case study
- Agent-based simulation

3.1 THEORY OF CHANGE: A PRECURSOR TO IMPACT EVALUATION

Theory of change is not a small *n* impact evaluation; rather, it is a precursor to undertaking most small *n* impact evaluations

3.1.1 OVERVIEW

The Theory of change method emerged from discussions on the evaluation of complex programmes and was fully articulated in the 1990s at the Aspen Institute Roundtable on Community Change. Weiss (2000) hypothesised that complex community initiatives and other complex programmes are difficult to evaluate primarily because the theories of change that underpin them are poorly articulated. According to Rogers et al., (2000, pp. 7–8):

> *[A]t its simplest, a program theory shows a single intermediate outcome by which the program achieves its ultimate outcome. [...] More complex program theories show a series of intermediate outcomes, sometimes in multiple strands that combine to cause the ultimate outcomes.*

In turn, Wholey (1987, p. 78) states that programme theory identifies 'program resources, program activities, and intended program outcomes, and specifies a chain of causal assumptions linking program resources, activities, intermediate outcomes, and ultimate goals.' Programme theory, therefore, emerged from the need to better understand the rationale of programmes and, more importantly, the chain of causality that leads to their outcome(s).

A useful theory of change must set out clearly the causal mechanisms by which the intervention is expected to achieve its outcomes (HM Treasury, 2020). The Magenta Book (HM Treasury, 2020) details how more sophisticated theory of change exercises produce a detailed and rigorous assessment of the intervention and its underlying assumptions, including the precise causal mechanisms that lead from one step to the next; alternative mechanisms to the same outcomes; the assumptions behind each causal step; the evidence that supports these assumptions; and how different contextual, behavioural and organisational factors may affect how, or if, outcomes occur.

Developing a theory of change often starts with articulating the desired long-term change that a programme intends to achieve, based on several assumptions that hypothesise, project or calculate how such change can be enabled. Assumptions are crucial:

> *The central idea in theory of change thinking is making assumptions explicit. Assumptions act as 'rules of thumb' that influence our choices, as individuals and organisations. Assumptions reflect deeply held values, norms and ideological perspectives. These inform the design and implementation of programmes.*
>
> **(Vogel, 2012, p. 4)**

The theory of change is fundamentally participatory in its process of development; it includes a variety of stakeholders and, therefore, perceptions. The process of developing a theory of change should be based on a range of rigorous evidence, including local knowledge and experience, past programming material and social science theory, all of which are brought together in an iterative process (Stein & Valters, 2012).

First articulated as an evaluation tool, the theory of change developed into an approach to programme planning as well as a tool for evaluation (Fox et al., 2017). Thus, it is increasingly common for an evaluator to be presented with an existing theory of change at the start of the evaluation process. The challenge is then to decide whether to accept this theory of change at face value or whether to start the evaluation with a fresh theory of change exercise. Many evaluators will wish to develop their own theory of change, particularly where the theory is a precursor to the use of a small *n* methodology because it is through the theory of change that the evaluator starts to develop a deep understanding of the case(s) that is so important to most small *n* methodologies.

Clearer distinction from logic models

While the theories of change are sometimes referred to as though they are interchangeable with logic models, it is important to note that they are different, and recent guides to evaluation and academic discussion have clarified this distinction. Asmussen et al. (2019) argue that while logic models are primarily concerned with *how* an intervention will achieve its outcomes, theories of change identify *why* these outcomes are fundamentally important. The key elements of a logic model are inputs, outputs and short-to-long-term outcomes. In contrast, a theory of change starts with questions about *why* an outcome is important, raising questions about why an intervention is important and what it will achieve. This, in turn, leads to important questions about what the intervention does. Thus, to the distinction between theories of change and logic models, outlined by Asmussen et al., we can add that the former, with their interest in causal explanation and articulation of programme theory and what an intervention does, will tend towards a deeper understanding of mechanisms of change than the latter.

3.1.2 KEY ELEMENTS OF METHODOLOGY

Theories of change may be developed at different points in the life-cycle of a programme. They can be prospective and developed at the initial phase – conceptualisation, planning and design. Alternatively, they can be retrospective and 'reconstructed' or pieced together after the programme is fully underway (Fox et al., 2017).

There is no single process to develop a theory of change. Fox et al. (2017) note that, over the years, many different processes that arrive at a programmatic theory of change have been conceptualised. These can broadly be grouped into one of the two following categories, or a combination of both:

- **Researcher-led:** Developing a theory of change follows a rigorous research-like process, because certain relevant elements are researched and investigated, e.g. the context. Assumptions may also be formulated more like research hypotheses that can, therefore, be tested in the future in greater depth.
- **Stakeholder-led:** Researchers/programme managers facilitate a process in which stakeholders are central. Stakeholders are provided with the basic information, such as the context, but their own perceptions are taken into account. This configures a collective induction exercise whose objective is to generate the collective vision underlying the programme.

There are various influential guides on undertaking a theory of change. The steps below draw in particular on Connell et al. (1995), Fulbright-Anderson et al. (1998), Blamey and Mackenzie (2007) and Asmussen et al. (2019). They are expressed as a series of questions, which should generally be addressed in the order set out below:

What is the intervention's primary intended outcome? The focus here is on the long-term vision of an initiative and is likely to relate to a timescale that lies beyond the timeframe of the initiative (Blamey & Mackenzie, 2007). It should be closely linked to the existence of a local or national problem. Asmussen et al. (2019) recommend focusing on one or two primary outcomes.

Why is the primary outcome important and what short and long-term outcomes map to it? Blamey and Mackenzie (2007) suggest that, having agreed the ultimate aim of the programme, stakeholders should consider the necessary outcomes that will be required by the end of the programme if such an aim is to be met in the longer term. These might be broken down into shorter and longer-term outcomes (Asmussen et al. 2019).

Who is the intervention for? Asmussen et al. (2019) observe that while programme developers often assume their intervention will be of benefit for everyone, in reality, this is rarely the case. Understanding precisely who might benefit is a useful precursor to understanding why the intervention is necessary, what value it will add and what it will do.

Why is the intervention necessary? Asmussen et al. (2019) note that most interventions are developed to fulfil a need. A theory of change should, therefore, be able to justify the need for an intervention. This may draw both on specific analysis of the need and the wider scientific literature.

Why will the intervention add value? In order for an intervention to have an impact, it needs to provide measurable value over what is currently available. In other words, the intervention needs to fill a gap. Again, an identification of needs may draw on the analysis of a particular population or place as well as findings from the wider scientific literature that help to explain why gaps exist.

What outputs are needed to deliver the short-term outcomes? A detailed consideration of outputs belongs in a logic model, but identifying key outputs provides an important sense-check between the intended outcomes and the intervention.

What will the intervention do? Asmussen et al. (2019) are clear that no theory of change is complete without specifying what the intervention will do; however, this does not need to contain substantial detail because the detail is set out in a logic model. Nevertheless, a useful theory of change must set out clearly the causal mechanisms through which the intervention is expected to achieve its outcomes (HM Treasury, 2020). This rich description of mechanisms is comparable to understandings of mechanisms in realist evaluation, where mechanisms are not variables but accounts that encompass agency and structure. They thus should 'reach down' to individual reasoning and 'reach up' to the collective resources embodied within the social programme being evaluated (Pawson & Tilley, 1997). An understanding of causal mechanisms will come from engagement with key informants and an understanding of the scientific evidence base (Asmussen et al., 2019).

What inputs are required? What are the resources committed and which activities are undertaken to deliver the programme?

According to Connell and Kubisch (1998), the theory of change should be:

- **Plausible** – The available evidence must sustain the assumptions and, hence, support the change potential of the activities to be implemented.
- **Doable** – The necessary resources – from financial to institutional – must be in place to ensure that the initiative can be operationalised.
- **Testable** – It must be sufficiently specific and complete for the evaluator to assess progress and evaluate contribution to change.

Participatory approach

The construction of a programme's theory of change should be participatory or 'co-produced' (Asmussen et al., 2019; Blamey & Mackenzie, 2007; Fox et al., 2017). The evaluator may start with programme documentation such as funding bids, project plans or steering group minutes. Often, the evaluator needs to conduct a series of structured and semi-structured interviews with key informants and stakeholders to piece together reasoning that was never consciously, or at least structurally, articulated (Fox et al., 2017). Other techniques such as workshops can also be used. The final step is to validate the theory of change.

There is no problem, nor should there be, with different stakeholders bringing different perspectives to bear in the process of developing a theory of change. If anything, theories of change are strengthened by a diversity of perceptions that ground the project in its complexity and work with it. Moreover, a consensus is not always present in reality and power relations permeate all social relations. Rather, the challenge often arises from the different assumptions and the difficulty in assessing which are critical to the overall success of the initiative. Valters (2014, p. 10) argues that:

> *Appreciating the difficulties inherent in this task is important, as ignoring them may encourage discussion of arbitrary assumptions or allow people to uncover only those assumptions that they are comfortable defending.*

3.1.3 MULTI-METHOD APPROACHES

Theories of change differ from the other small *n* methods covered in this guide. While some commentators see it as a methodology, it is probably better understood as a method that can be used in combination with all the other methodologies covered in this guide. In most cases, it will precede those methodologies, helping to clarify research questions and hypotheses that can be addressed in the methodology.

Nevertheless, theories of change share some important similarities with realist evaluation, namely an emphasis on understanding causality and a recognition that an understanding of context is key to attributing cause (Blamey & Mackenzie, 2007).

3.1.4 RESOURCES REQUIRED FOR AN EVALUATION

Skill set for evaluators

When developing a theory of change, the evaluator acts as both researcher and theorist, requiring a detailed understanding of the programme being evaluated and a good understanding of the wider evidence and theory relevant to the programme.

Although theories of change can incorporate both quantitative and qualitative data collection, qualitative data collection – in particular, facilitating workshops and semi-structured interviews – tend to be most common.

Resource implications

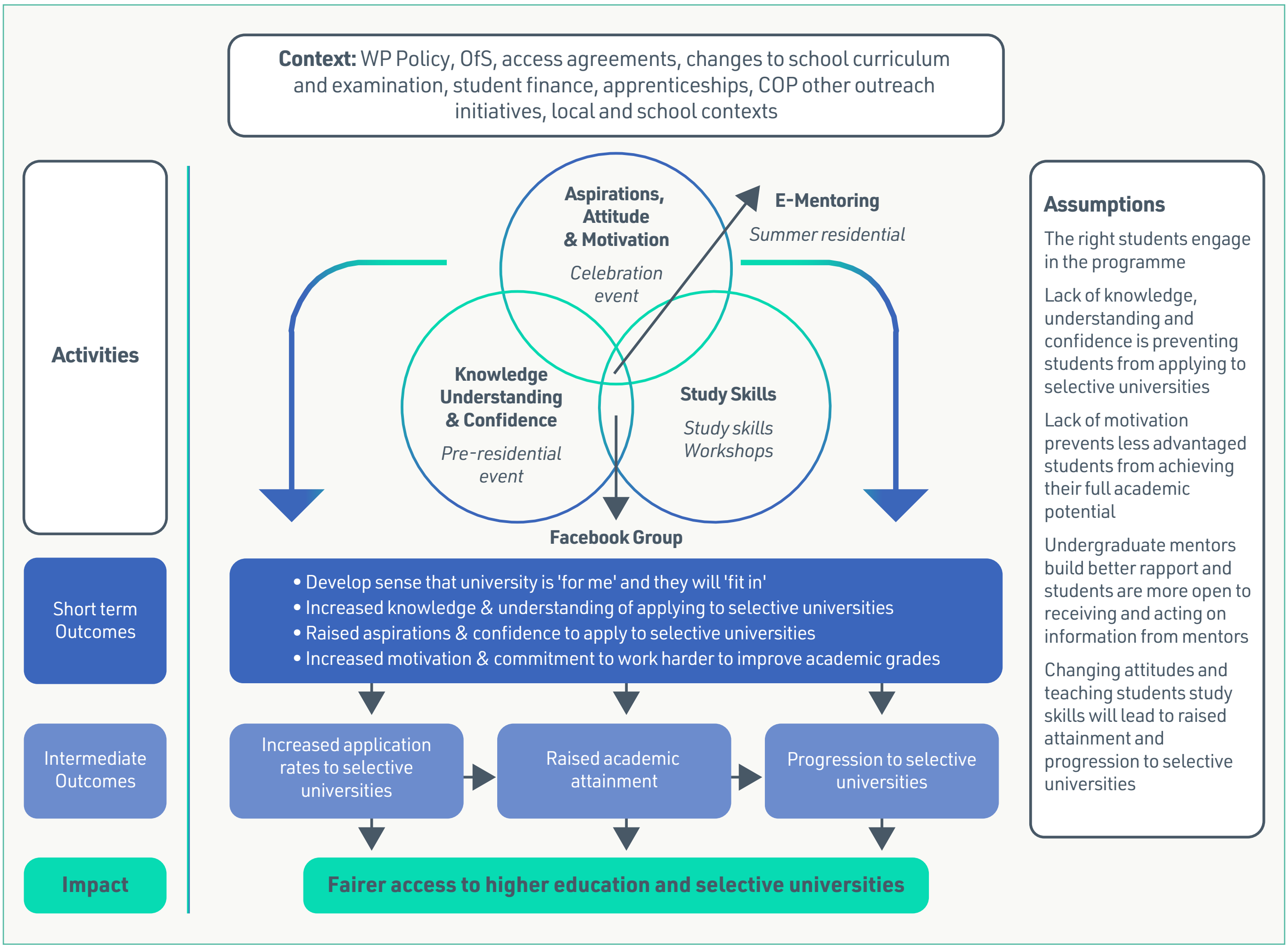
Reconstructing a programme's theory of change involves substantial work. The process is likely to be iterative and participatory, meaning that the evaluator moves from analysing programme documentation, such as funding bids, project plans or steering group minutes, to semi-structured interviews and workshops with a wide range of participants. The iterative process means that multiple engagements with informants are typically required. Developing a theory of change thus requires several days of work for the evaluator and involves engagement with multiple stakeholders: it is not an exercise that can be completed within a single workshop.

3.1.5 CASE STUDY

Barkat (2019) describes the application of a theory of change approach as a framework to plan and design the evaluation of the Academic Enrichment Programme (AEP) at the University of Birmingham. The programme aimed to support underrepresented students in securing places at selective universities. The process of developing the theory of change for the AEP was based on interviews with staff at the University of Birmingham who deliver the programme, including the programme lead, as well as a review of programme documents and a general literature review.

As described by Barkat (2019), the theory of change suggests supporting students in three key areas – (a) targeting aspirations, attitudes and motivation; (b) developing knowledge, understanding and confidence; (c) supporting attainment through study skills – would positively influence students and lead to change by raising their aspirations and confidence to apply to selective universities. In turn, this would raise individuals' attainment by motivating them to achieve the required grades, leading to students progressing to selective universities. The theory of change also suggests that these intermediate outcomes may ultimately lead to a longer-term impact by supporting fairer access at selective institutions. The five strands of activity within the programme include a five-day summer residential, a Facebook group, study skills sessions, e-Mentoring by undergraduate students and a celebratory event. The theory of change also includes wider political, local and national contextual factors that may support or hinder change and articulates the key assumptions that underpin it: that less-advantaged students lack knowledge, understanding and confidence in applying to selective universities and that a lack of motivation is preventing them from achieving their full academic potential. The theory of change is represented diagrammatically by Barkat and reproduced here in Figure 6.

Figure 6: Theory of Change of an Academic Enrichment Programme (reproduced from Barkat 2019: Figure 1).



Context: WP Policy, OfS, access agreements, changes to school curriculum and examination, student finance, apprenticeships, COP other outreach initiatives, local and school contexts

- Develop sense that university is 'for me' and they will 'fit in'
- Increased knowledge & understanding of applying to selective universities
- Raised aspirations & confidence to apply to selective universities
- Increased motivation & commitment to work harder to improve academic grades

Increased application rates to selective universities → Raised academic attainment → Progression to selective universities

Fairer access to higher education and selective universities

Assumptions

The right students engage in the programme

Lack of knowledge, understanding and confidence is preventing students from applying to selective universities

Lack of motivation prevents less advantaged students from achieving their full academic potential

Undergraduate mentors build better rapport and students are more open to receiving and acting on information from mentors

Changing attitudes and teaching students study skills will lead to raised attainment and progression to selective universities

The paper goes on to describe how the resulting evaluation was undertaken and how the theory of change was refined during the evaluation.

Reference

Barkat, S. (2019) Evaluating the impact of the Academic Enrichment Programme on widening access to selective universities: Application of the theory of change framework, *British Educational Research Journal* 45(6) pp. 1160–1185.

3.1.6 RESOURCES

Web resources

TASO has produced guidance on developing a theory of change as part of their wider evaluation guidance. This is supported by videos discussing theory of change and how to run a theory of change workshop.

Key reading

Carol Weiss was instrumental in developing the theory of change approach as we now understand it. A good introductory article is:

Weiss, C. (1995) Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In Connell, J. P., Kubisch, A. C., Schorr, L. B. and Weiss, C. H. (eds), *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts*. Washington, DC: Aspen Institute.

Many organisations have produced guidance on the theory of change, including:

Asmussen, K., Brims, L. and McBride, T. (2019) *10 steps for evaluation success*, London: Early Intervention Foundation, pp. 15–26. https://www.eif.org.uk/resource/10-steps-for-evaluation-success

Noble, J. (2019) *Theory of change in ten steps*, London: New Philanthropy Capital. https://www.thinknpc.org/resource-hub/ten-steps/

Rogers, P. (2014) *Theory of Change, Methodological Briefs: Impact Evaluation 2*, UNICEF Office of Research, Florence. https://www.unicef-irc.org/publications/pdf/brief_2_theoryofchange_eng.pdf

A helpful worked example of how to build a theory of change produced by ActKnowledge and the Aspen Institute Roundtable on Community Change is available here.

Further references

Asmussen, K., Brims, L. & McBride, T. (2019) *10 steps for evaluation success*, London: Early Intervention Foundation.

Blamey, A. & Mackenzie, M. (2007) Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? *Evaluation* 13 pp. 439–455.

Connell, J. P. & Kubisch, A. C. (1998) Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects, and problems. In: Fulbright-Anderson K., Kubisch, A. C. & Connell J. P. (eds.) *New Approaches to Evaluating Community Initiatives: Theory, Measurement, and Analysis*. Washington, DC: The Aspen Institute.

Fox, C., Grimm, R. & Caldeira, R. (2016) *An Introduction to Evaluation*, London: Sage.

HM Treasury (2020) *The Magenta Book: Central Government Guidance on Evaluation*, London: HM Treasury.

Pawson, R. & Tilley, N. (1997). *Realistic evaluation*. London: Sage Publications.

Rogers, P. (2014). *Theory of Change, Methodological Briefs: Impact Evaluation 2*, UNICEF Office of Research, Florence. https://www.unicef-irc.org/publications/pdf/brief_2_theoryofchange_eng.pdf

Stein, D. & Valters, C. (2012). *Understanding Theory of Change in International Development*. London: The Justice and Security Research Programme, LSE.

Valters, C. (2014) *Theories of Change in International Development: Communication, Learning or Accountability*. London: The Justice and Security Research Programme, LSE.

Vogel, I. (2012) *Review of the use of 'Theory of Change' in international development*. DFID Research Paper, Department for International Development.

Weiss, C. (2000) Which links in which theories shall we evaluate? In: Rogers, P. J., Hacsi, T., Petrosino, A. and Huebner, T. A. (eds.) *Program Theory in Evaluation: Challenges and Opportunities, New Directions for Evaluation*, San Francisco: Jossey-Bass.

3.2 REALIST EVALUATION

3.2.1 OVERVIEW

Pawson and Tilley's (1997) starting point for setting out the realist approach to evaluation is to argue that 'traditional' experimental evaluation is flawed because, in its attempt to reduce an intervention to a set of variables and control for difference using an intervention and control group, it 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 & Tilley 1997, p. 11). They assume a different, 'realist' model of explanation in which 'causal outcomes follow from mechanisms acting in contexts' (Pawson & Tilley, 1997, p. 58).

A mechanism explains precisely what it is about a programme that makes it successful. Mechanisms are not variables, but accounts that encompass individual agency and social structures. They should, thus, 'reach down' to individual reasoning and 'reach up' to the collective resources embodied in the social programme being evaluated (Pawson & 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 & Tilley, 1997, p. 69). Astbury and Leeuw (2013) define 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 to success (Pawson & Tilley, 1994). Programmes are always introduced into a pre-existing social context, and pre-existing structures enable or disable the intended mechanism of change (Pawson & Tilley, 1997). This recognises 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, 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, is devoted to a discussion of causation. For Pawson and Tilley (1997), the model of causation adopted in 'traditional' experimental evaluation design is external, successionist causation: the idea that causation itself is unobservable but can be inferred from observation. Scientific realists prefer a model of generative causation that sees causation as acting internally as well as externally. They do not, therefore, make predictions about the probability of an intervention leading to an outcome, 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 that realist evaluation can offer is a plausible explanation of what works 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 intervention and control groups. Instead, following philosophers such as Bhaskar, they argue that the two essential elements of an experiment are to trigger the mechanism being studied to ensure that it is active and to prevent 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.

Particular concerns surround the ability of the approach to address complexity (Blamey & Mackenzie, 2007; Davis, 2005; Pederson & Rieper, 2008). Critics argue that scientific realism was developed to some degree in the field of crime reduction, where the programmes under evaluation were relatively small scale, operating at a local level, targeted at 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 more complex regional and national level policies and programmes.

3.2.2 KEY ELEMENTS OF METHODOLOGY

There is still much debate about exactly how to undertake a realist evaluation. Marchal et al. (2012), in their review of realist evaluation in health systems research, found significantly different approaches used. These included different ways of treating mid-range theory, different understandings of 'mechanisms' and 'context' and different approaches to handling context-mechanism-outcome configurations. While, at the time of writing, a similar review did not exist for WP in higher education, it is likely that different approaches to realist evaluation are also used within this sector. It is not, therefore, possible to set out a precise set of agreed steps that an evaluator should follow when undertaking a realist evaluation. Nevertheless, it is possible to lay out some key elements of a realist evaluation.

Mechanisms

Wong et al. (2013) suggest that one way to identify a programme mechanism is to reconstruct, in the imagination, the reasoning of participants or stakeholders. They also note that:

- Mechanisms cannot be seen or measured directly (because they happen in people's heads, or at different levels of reality than the one observed).
- There will potentially be many mechanisms, and the role of the realist researcher is to identify the 'main mechanisms', which they define as 'those that are common and significant enough to contribute to the pattern of outcomes of the intervention' (Wong et al., 2013, p. 6).
- The 'causes' of outcomes are not simple, linear or deterministic. This is partly because programmes often function through multiple mechanisms and partly because a mechanism is not inherent to the intervention but is a function of the participants and the context. Consequently, the same intervention can trigger different mechanisms for different participants, even in the same location.
- Mechanisms are context-sensitive.

Context

Wong et al. (2013) note that context can take a multitude of forms including:

- Broad social or geographical features (for example the country in which an intervention operates and its cultures)
- Features affecting the implementation of programmes (for example whether the programme occurs in a prison, hospital or health service, whether there is adequate funding, the qualifications of staff)
- The make-up of the participants on a programme or the different population profiles of locations receiving an intervention

- Conditions in which subjects seek to enact their choices (graduates of a vocational training programme will find it easier to find work in a context of high employment; recipients of a housing subsidy will find it harder to use that subsidy in a context of housing shortages).

However, Wong et al. (2013, p. 9) are also clear that, simply because context can take a multitude of forms, it is not necessary to list the 'infinite potential "surrounds" to an intervention'. Instead:

> *What matters is developing an understanding of how a particular context acts on a specific program mechanism to produce outcomes – how it modifies the effectiveness of an intervention.*
>
> **(Wong et al. 2013, p. 9)**

Context-Mechanism-Outcome configurations

Pulling these elements together, scientific realist evaluators construct their explanation around the three vital ingredients of context, mechanism and outcome, which Pawson and Tilley refer to as context-mechanism-outcome configurations:

> *The basic task of social inquiry is to explain interesting, puzzling, socially significant regularities...Explanation takes the form of positioning some underlying mechanism... which generates the regularity and thus consists of propositions about how the interplay between structure and agency has constituted the regularity. Within realist investigation there is also an investigation of how the workings of such mechanisms are contingent and conditional, and thus only fired in particular local, historical or institutional contexts...*
>
> **(Pawson & Tilley, 1997, p. 71)**

For Pawson and Tilley (1994, 1997), traditional experiments misunderstand what makes programmes work: *'Programmes cannot be considered as some kind of external, impinging "force" to which subjects "respond"' (Pawson & Tilley, 1994, p. 294).* Instead, social programmes are social systems involving an interplay between individual and institution, or – in the language of Giddens (1984) – agency and structure. Thus, it is not that programmes work but, rather, that people co-operate and choose to make them work. However, scientific realists do not adopt the same formulation as constructivists; rather, they see people's choices as constrained by social structures:

> *[P]rogrammes 'work', if subjects choose to make them work and are placed in the right conditions to enable them to do so. This process of 'constrained choice' is at the heart of social and individual change to which all programmes aspire...*
>
> **(Pawson & Tilley, 1994, p. 294)**

Undertaking a scientific realist evaluation

The starting point is theory, and 'empirical work in programme evaluation can only be as good as the theory which underpins it' (Pawson & Tilley, 1997, p. 83). Thus, a scientific realist evaluation should not be data-driven, but theory-driven. To give a practical example: the subject discussed in an interview would be the evaluator's theory, which the interviewee is asked to confirm, reject or refine (Pawson & Tilley 1997, p. 155). The interview relationship has been described as a teacher-learner cycle, in which the interviewee is taught the programme theory being tested and, having learned the theory, is able to teach the evaluator about components of the programme (Pawson & Tilley, 2004). For initial theory-gleaning, interviews are likely to focus on practitioners.

Although both quantitative and qualitative data have a place, and scientific realism is 'method neutral' (Marchal et al., 2013), more emphasis is generally placed on qualitative data that allows theory to be explored, and semi-structured interviews are particularly common (Manzano, 2016).

To build explanations, data collection should be iterative (Manzano, 2016). The ideal empirical evaluation would therefore collect 'before' and 'after' data to give an overall picture of outcomes but, thereafter, more attention would be given to gaining data that tapped mechanism and contextual variation (Pawson & Tilley, 1994). Thus, there is a strong assumption that data collection will be formed of multiple strands, with, for example, semi-structured interviews supplemented by observations and/or analysis of quantitative data.

The standard scientific realist data matrix would compare variations in outcome patterns across groups, but these groups would not be experimental and control groups. Instead, they would be defined by the mechanism/context framework, with the evaluator running a systematic range of comparisons across a series of studies to understand which combination of context and mechanism is most effective (Pawson & Tilley, 1994).

3.2.3 MULTI-METHOD APPROACHES

Scientific realism can provide a framework within which some of the other small *n* methodologies sit. Process tracing, for instance, draws on many of the same underlying assumptions about causality, the nature of the social world (ontology) and the status of knowledge of the social world (epistemology).

Commentators have also noted that realist evaluation, as a form of theory-led evaluation, closely resembles the theories of change approach. However, Blamey and Mackenzie (2007, p. 452) argue that 'Theories of Change and Realistic Evaluation may both be from the same stable, [but] they are in practice very different horses.' For Blamey and Mackenzie, one of the key differences is that in a 'theories of change' approach 'theory' is articulated by a wide range of stakeholders, whereas in realist evaluation it is the evaluator who articulates the theory.

3.2.4 RESOURCES REQUIRED FOR AN EVALUATION

Skill set for evaluators

In the scientific realist paradigm, the evaluator is both researcher and theorist, with a detailed understanding of the programme being evaluated and the ability 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 semi-structured interviews – tends to be most common (Manzano, 2016). However, scientific realism involves a particular conception of the interview as a teacher-learner cycle (Pawson & Tilley, 2004, see above). The interview approach required in a scientific realist interview may contradict the training in standard research methods 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 does not emphasise particular training or skills but, rather, 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 (Manzano, 2016, p. 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' (p. 348). It is not possible to quantify 'large amounts of data' in the abstract, but the collection of multiple sources of data on each case suggests days, rather than hours, of work.

3.2.5 CASE STUDY

Formby et al. (2020) describe the realist evaluation of Go Higher West Yorkshire (GHWY) Uni Connect – an initiative to reduce educational inequalities through collaborative WP outreach. It contributes to wider debates on WP policy by demonstrating how Higher Education Progression Officers (HEPOs) normalised 'progression' based on community and learners' needs. The evaluation aim was to understand the differences in approach of HEPOs working in community settings.

Three programme theories were developed around the ideal practice of HEPO staff, based on a series of initial interviews and focus groups conducted with GHWY staff involved in the Uni Connect programme. These were:

- High-quality continuing professional development (CPD) will equip school/college-based staff with the skills and information needed to support young people to make informed choices.
- Dedicated progression staff in schools/colleges will have more time to invest in young people and support them in planning for their future.
- Facilitating the delivery of outreach activity aimed at helping young people to make informed choices.

The programme theories were then split into several hypothetical context-mechanism-outcome (CMO) configurations through analysis of programme documentation, literature reviews on effective WP outreach and discussion with stakeholders. The CMOs suggested the need for individual responses from HEPO staff and management to uncover wider cultural models relating to the normalisation of WP in school/college environments, as well as wider societal/community factors that may impact the HEPO. The evaluators undertook realist interviews with HEPOs managers and conducted focus groups with HEPO front-line staff to determine and refine valid CMO configurations. Both data collection methods were chosen for their effectiveness in identifying contexts and mechanisms that produce variant outcomes.

The evaluation showed that HEPOs both complemented existing arrangements in settings that already practiced WP, and introduced new WP activities that shifted the wider cultural practice in settings where WP resources had been introduced for the first time.

Reference

Formby, A., Woodhouse, A. & Roe, F. (2020) 'A Presence in the Community': Developing Innovative Practice through Realist Evaluation of *Widening Participation in West Yorkshire, Widening Participation and Lifelong Learning* 22(3) 173–186. doi:10.5456/WPLL.22.3.173

3.2.6 RESOURCES

Web resources

An interesting interview about realist evaluation with Ray Pawson is available on the Vimeo platform. The interview is divided into four parts and the first one is available here: https://vimeo.com/84215487

The RAMESES II project, funded by the NIHR, developed quality and reporting standards and resources and training materials for realist evaluation. These are available online at: https://www.ramesesproject.org/Home_Page.php

A Supplementary Guide on Realist Evaluation, issued as part of the Magenta Book 2020, is available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879435/Magenta_Book_supplementary_guide._Realist_Evaluation.pdf

Key reading

Pawson and Tilley's influential and widely cited book on scientific realist evaluation is a good starting point for exploring scientific realism:

Pawson, R. & Tilley, N. (1997) *Realistic evaluation*. London: Sage.

Ana Manzano explores the craft of realist interviewing and, in doing so, helps to illuminate key elements of conducting a realist evaluation:

Manzano, A. (2016) The craft of interviewing in realist evaluation. *Evaluation* 22(3) 342–360. doi:10.1177/1356389016638615

Pederson and Rieper explore a key criticism of realist evaluation, namely its ability to deal with complexity:

Pederson, L. H. & Rieper, O. (2008) 'Is Realist Evaluation a Realistic Approach for Complex Reforms?', *Evaluation*, 14(3) 271–93.

Realist evaluation is closely related to theories of change, and Blamey and Mackenzie's article exploring the similarities and differences also highlights important elements of the realist approach to evaluation:

Blamey, A. and Mackenzie, M. (2007) Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? *Evaluation* 13 439–455.

Further references

Blamey, A. & Mackenzie, M. (2007) Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? *Evaluation* 13 439–455.

Davis, P. (2005) The Limits of Realist Evaluation: Surfacing and Exploring Assumptions in Assessing the Best Value Performance Regime, *Evaluation* 11, 275–95.

Giddens, A. (1984) *The Constitution of Society*, Cambridge: Polity Press.

Lawson, T. (1997). *Economics and Reality*. London: Routledge.

Manzano, A. (2016) The craft of interviewing in realist evaluation, *Evaluation* 22(3) 342–360. doi:10.1177/1356389016638615

Marchal, B., Belle, S., Olmen, J., Hoerée, T. and Kegels, G. (2012) Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research, *Evaluation* 18(2) 192–212.

Pawson, R. & Tilley, N. (1994) What works in evaluation research? *British Journal of Criminology* 34, 291–306.

Pawson, R. & Tilley, N. (1997) *Realistic evaluation*. London: Sage.

Pawson, R. & Tilley, N. (2004) *Realistic evaluation*. British Cabinet Office. Available at: http://www.communitymatters.com.au/RE_chapter.pdf [accessed 6th September 2021]

Pederson, L. H. & Rieper, O. (2008) Is Realist Evaluation a Realistic Approach for Complex Reforms?, *Evaluation*, 14(3) 271–93.

Wong, G., Westhorp, G., Pawson, R. and Greenhalgh, T. (2013) Realist Synthesis RAMESES Training Materials.

3.3 PROCESS TRACING

3.3.1 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 & Checkel, 2015, p. 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 are 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 evidence exists that other factors caused both X and Y (Bennett, 2010).
- The inductive element of process tracing provides opportunities for evaluators to uncover unforeseen or unexpected causal explanations (Bennett & 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 although 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; a series of tests outlined below is used to establish the causal explanations that recognise this.

3.3.2 KEY ELEMENTS OF METHODOLOGY

Ricks and Liu (2018) set out a series of steps involved in process tracing:

1. Identify hypotheses
2. Establish timelines
3. Construct a causal graph
4. Identify alternative choices or events
5. Identify counterfactual outcomes
6. Find evidence for the primary hypothesis
7. Find evidence for rival hypotheses.

Each of these is described in more detail below.

Step 1: Identify hypotheses

The evaluator draws on broader generalisations and evidence from within the case to generate a series of (preferably competing) testable hypotheses about how an intervention may connect to an outcome. A theory of change exercise preceding the process tracing may provide a useful starting point for the generation of hypotheses. Broader generalisations that the evaluator draws upon may include elementary understandings of associations that are nearly universally accepted as true or inferences derived from previous research (Mahoney, 2012).

When developing hypotheses, it is important to cast the net widely for alternative explanations (Bennett & Checkel, 2015). Bennett and Checkel (2015) suggest that a useful criterion in this respect is to ask whether any major theoretical categories of social explanation have been omitted, giving examples such as actor's use of power, institutional constraints, social norms and legitimacy. Similarly, they suggest considering whether both agent-based and structural explanations have been considered.

Ricks and Liu (2018) emphasise the importance of constructing competing or rival hypotheses. For each hypothesis, the evaluator should specify what should be observed if the hypothesis is true or false (White & Phillips, 2012). Hypotheses must, therefore, be very specific. The level of granularity required will often be much greater than is found in many theories of change. In other words, the theory of change must include a level of detail sufficient to generate testable hypotheses.

Process tracing can involve both deduction and induction. The deductive approach tests theory by examining the observable implications of hypothesised causal mechanisms within a case. The inductive approach uses evidence from within the case to develop hypotheses and, hence, theory, which may then be tested deductively (Bennett & Checkel, 2015). The particular mix in a project depends on pre-existing knowledge and theory and whether the case selected for study is similar to a defined population or an outlier (Bennett & Checkel, 2015). Where there is little prior knowledge or a case is an outlier, process tracing proceeds primarily through induction, and the researchers will amass significant amounts of information that may or may not later become part of the hypothesised explanation (Bennett & Checkel, 2015). They will try out many proto-hypotheses and be open to multiple possible causal explanations, the more promising of which may then be subject to deductive processes. If theories exist that already seem to offer potential explanations, process tracing can proceed more deductively (Bennett & Checkel, 2015).

Step 2: Establish timelines

Evaluators should then sequence events in a timeline. The conclusion of the timeline will be at or shortly after the outcome of interest and the start far enough back to capture the emergence of the theorised causal variable (Ricks & Liu, 2018). Ricks and Liu (2018) identify several purposes for the timeline: first, it clarifies the researcher's thought processes; secondly, it establishes temporal precedence; thirdly, it provides a 'face-validity' test for the hypotheses being tested and, fourthly, it helps to identify major events that could have shaped the outcome of interest.

Step 3: Construct a causal graph

A causal graph depicts visually the causal process through which X causes Y and follows the timeline. It identifies the independent variable(s) of interest and provides structure to the process of enquiry by showing all the moments at which the actor concerned (an individual, organisation or group) made a choice that could have affected the result (Ricks & Liu, 2018).

Step 4: Identify alternative choices or events

At each relevant moment in the causal graph, a different choice could have been made. These alternatives should be identified and theoretically grounded (Ricks & Liu, 2018).

Step 5: Identify counterfactual outcomes

Counterfactuals are vital to process tracing, as Ricks and Lui (2018, p. 844) explain:

> *When treating hypothetical predictions, it is imperative that another outcome was possible. If there is no plausible theory-informed alternative outcome, then no real choice or event has taken place. Thus, the link between the input and the outcome was predetermined; hence, process tracing provides little value added.*

Ricks and Liu (2018) emphasise the importance of conducting Steps 1 to 5 before data collection.

Step 6: Find evidence for the primary hypothesis

No single type of data collection method is specified for Process Tracing. Data collection should be designed to match the evidence specified in the hypotheses being tested. Data collection involves in-depth case study analysis and, thus, is likely to be predominantly qualitative, including historical reports, interviews and observations, but quantitative data may also be used. Evaluators should be relentless in gathering diverse and relevant data (Bennett & Checkel, 2015). It is particularly important that data is diverse and that independent streams of evidence are found. Diversity supports the triangulation of data, but it is also important to ensure independent streams of data because if all or most of the data derives from one source, triangulation can be misleading and disguise selection bias (Bennett & Checkel, 2015).

Through the analysis, the evaluator develops an explicit chronology of events, setting out the hypothesised causal link between each stage. The evidence gathered is then used to overturn or substantiate the rival hypotheses, with the aim of establishing whether the causal mechanisms at work match those predicted (White & Phillips, 2012).

It is important to examine each alternative hypothesis equally rigorously (Bennett & Checkel, 2015). This does not mean exploring all explanations in equal depth: if a hypothesis can be quickly and decisively dismissed, it should be, and the remaining hypotheses then investigated in more detail. Bennett and Checkel warn against confirmation bias, for instance, when undue privileged status is given to one hypothesis, with process tracing being performed on that hypothesis first and others only introduced to address anomalies facing the privileged first mover (Bennett & Checkel, 2015). For example, it may be hypothesised that an intervention operates by increasing self-confidence in participants, and the evaluation effort focuses on measuring and testing self-confidence without also considering alternative hypotheses, such as the intervention increasing self-efficacy or creating a different organisational culture.

Researchers should distinguish between unavailable evidence and evidence that is contrary to a hypothesis:

- Where evidence is unavailable, this lowers the upper limit of probability that one can attach to the likely truth of the explanation.
- Where evidence is contrary to a hypothesis and the hypothesis is reformulated or modified, the greater the modification, the more important it is to generate and test new observable implications. This is to 'guard against "just so" stories that explain away anomalies one at a time' (Bennett & Checkel, 2015, p. 19).

Where evidence is missing it is also important to make a judgement about whether 'absence of evidence' constitutes 'evidence of absence' (Bennett & Checkel, 2015). Bennett and Checkel highlight the importance of this in relation to evidence expected to be available and decisive in testing a hypothesis. They use the example of feeling for change in our pocket to make the point that, sometimes, failure to find something constitutes strong evidence that it does not exist (Bennett & Checkel, 2015).

Process tracing involves several different kinds of empirical tests of causation, which distinguish evidence with differing probative value. Van Evera (1997) distinguished four tests based on the degree to which a hypothesis uniquely predicts the evidence, and its reliability in doing so (Figure 7).

Figure 7: Four tests for causation used in Process Tracing (based on Van Evera 1997 and Bennett 2010)

		Is evidence sufficient (uniquely able) to establish causation?	
		No	**Yes**
Is evidence necessary (certain) to establish causation?	**No**	**Straw in the Wind** Passing affirms the relevance of hypothesis but does not confirm it. Failing suggests the hypothesis may not be relevant, but does not eliminate it.	**Smoking gun** Passing confirms hypothesis Failing does not eliminate it.
	Yes	**Hoop** Passing affirms relevance of hypothesis but does not confirm it. Failing eliminates it.	**Doubly Decisive** Passing confirms hypothesis and eliminates others. Failing eliminates it.

- Hoop tests involve evidence that is certain but not unique (Bennett & Checkel, 2015; Van Evera, 1997). Hypotheses must 'jump through the hoop' to remain under consideration; therefore, hoop tests provide a necessary but not sufficient criterion for accepting an explanation (Bennett, 2010; Van Evera, 1997). Failing a hoop test eliminates a hypothesis but passing it does not greatly increase confidence in it (Bennett & Checkel, 2015). Hoop tests are, therefore, most useful in excluding alternative hypotheses.
- Smoking gun tests are unique, but not certain (Bennett & Checkel, 2015; Van Evera, 1997). Thus, passing a smoking gun test strongly affirms a hypothesis, but failure to pass such a test does not eliminate it. Smoking gun tests, therefore, provide a sufficient but not necessary criterion for confirmation.
- Doubly decisive tests use evidence that is both unique and certain (necessary and sufficient) to confirm one hypothesis and eliminate all others (Bennett & Checkel, 2015; Bennett, 2010; Van Evera, 1997).
- Straw in the wind tests provide weak or circumstantial evidence that is neither unique nor certain (Bennett & Checkel, 2015; Van Evera, 1997). They are, therefore, neither necessary nor certain (Bennett, 2010; Van Evera, 1997).

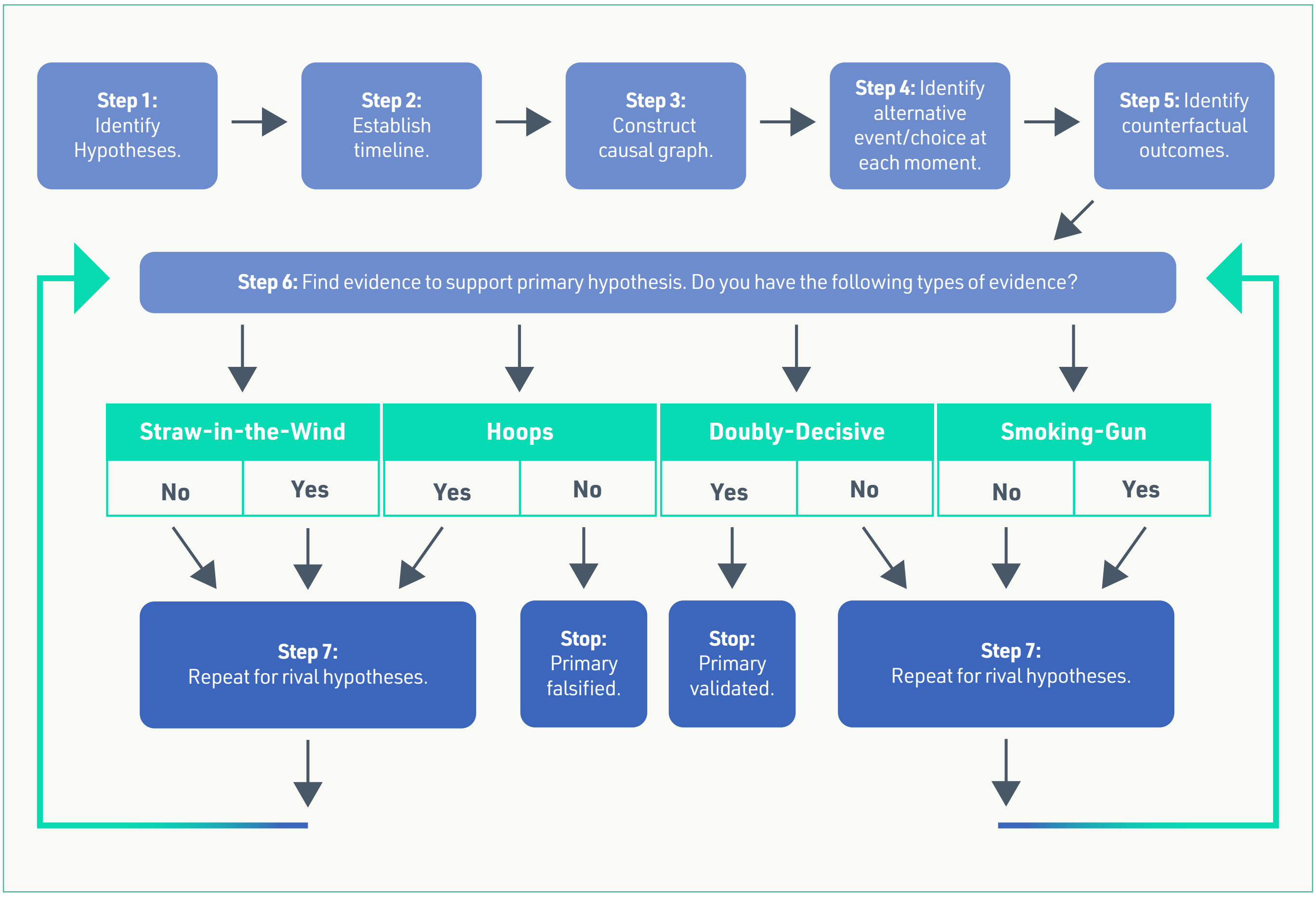
The tests associated with process tracing can help a researcher establish that: (1) a specific event or process took place; (2) a different event or process occurred after the initial event or process; and (3) the former was a cause of the latter (Mahoney 2012). Doubly decisive tests are rare; thus, the two types of test most commonly used to achieve these goals are hoop tests and smoking gun tests (Mahoney 2012). Hoop tests eliminate hypotheses and smoking gun tests confirm hypotheses. Taken together, they can therefore both confirm a hypothesis and eliminate alternative ones. Where they cannot decisively do one or other of these things, they become straw in the wind tests (Mahoney, 2012).

Step 7: Find evidence for rival hypotheses

The final step is to repeat step 6 for each alternative explanation.

Ricks and Lui (2018) summarise the different steps in process tracing in a diagram (Figure 8).

Figure 8: Process Tracing flow chart (reproduced from Ricks and Lui 2018: Appendix, Figure 0)


Step 1: Identify Hypotheses.
Step 2: Establish timeline.
Step 3: Construct causal graph.
Step 4: Identify alternative event/choice at each moment.
Step 5: Identify counterfactual outcomes.
Step 6: Find evidence to support primary hypothesis. Do you have the following types of evidence?
Straw-in-the-Wind
No
Yes
Hoops
Yes
No
Doubly-Decisive
Yes
No
Smoking-Gun
No
Yes
Step 7: Repeat for rival hypotheses.
Stop: Primary falsified.
Stop: Primary validated.
Step 7: Repeat for rival hypotheses.

Some commentators have suggested that hoop, smoking gun and straw in the wind tests can provide greater insight than is sometimes assumed:

- Hoop tests are most useful in excluding alternative hypotheses. However, Mahoney (2012) notes that where a hypothesis passes a very rigorous hoop test, this does provide some positive evidence in favour of the hypothesis.[4] The difficulty relates to how frequently the condition necessary for the hypothesis to be valid is found: if it is always present or common, the test is easy to pass; if it is rare, the test is hard to pass (Mahoney, 2012).
- Passing a smoking gun test strongly affirms a hypothesis, but failure to pass such a test does not eliminate it. However, Mahoney (2012) notes that where a hypothesis fails an easy smoking gun test, this does provide some evidence supporting its elimination. An easy test is one in which the condition whose presence is sufficient to prove the validity of the hypothesis is frequently present (Mahoney, 2012).
- Although no one straw in the wind test is decisive, a series of such tests that all or mostly point in the same direction can increase or decrease confidence in a hypothesis (Mahony, 2012).

[4] Mahoney (2012, p. 576) argues that 'Just as some hoops are smaller than others, and thus more difficult to jump through, some hoop tests are more demanding and thus harder to pass. While failing a hoop test will eliminate a hypothesis regardless of whether the test is easy or hard, passing a hoop test will lend positive support for a hypothesis in proportion to the degree that it is a difficult test.'

3.3.3 MULTI-METHOD APPROACHES

Process tracing hypothesises causal mechanisms to arrive at causal explanations. Due to its mechanistic approach and assumption that explanation should combine social and institutional structure and context with individual agency and decision-making, it is closely related – epistemologically and ontologically – to scientific realism (Bennett & Checkel, 2015) and could be seen as a specific analytical process that fits within the broader scientific realist framework. However, process tracing could align with approaches to evaluation grounded in other epistemologies and ontologies, such as pragmatism or constructivism (Bennett & Checkel, 2015).

Although process tracing is a within-case methodology (i.e. it takes place within a single case) it can be combined with case comparisons, where feasible (Bennett & Checkel, 2015). For example, in a most similar case comparison, process tracing can help establish the role of the single independent variable that differs between the cases in explaining the outcome.

Befani and Mayne (2014) have noted that contribution analysis and process tracing are similar, both seeking to make causal inferences using non-counterfactual approaches, based on causal mechanisms and theories of change. They also note that a potential limitation of contribution analysis is that 'it is an approach and does not spell out detailed steps to follow in data collection or discuss explicitly the types and strength of evidence used' (Befani & Mayne, 2014, p. 25). They, therefore, suggest combining the two approaches so that the evaluator follows the logic of contribution analysis but uses the various tests developed in process tracing to provide an indication of what evidence to look for and what criteria to use to judge the strength of the evidence.

Process tracing can be combined with quantitative approaches in mixed-method designs, for example where a few cases from statistical analysis are selected to clarify the direction of causal inference (Bennett & Checkel, 2015). Another example would be where agent-based modelling is used to check the plausibility of inferences about mechanisms derived from process tracing (Bennett & Checkel, 2015).

3.3.4 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 be trained to a postgraduate level 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 data is historical, archival or 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 which will provide context for the case.

Actors, whether historical or contemporaneous, may go to great lengths to obscure their actions and motivations, thus adding bias to the 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 may also need the guile and resilience of a Columbo or a Poirot, to the extent that they are confident to cast the net widely for alternative explanations, relentless in gathering diverse and relevant evidence and able to consider potential bias in different evidentiary sources (Bennett & Checkel, 2015). In some instances, this may indicate 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 relied on, process tracing can be very time-consuming (Bennett & Checkel, 2015). No prescribed amount of evidence gathering is 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. until saturation point is reached.

3.3.5 CASE STUDY

The modernisation of public action (MAP) was an endeavour to make evaluation the primary instrument of the reform of public policies at a national level in France. Eighty evaluations were launched between 2012 and 2017 by the French government, and Delahais and Lacouette-Fougère (2019) led an evaluation of the impact of the programme. Having undertaken an initial assessment of 65 of the evaluations, focusing on their quality, the authors chose eight 'best case' evaluations for in-depth analysis. The study described here focused on those eight evaluations and assessed the probability that they had an impact on the evolution of the policy that they evaluated. Delahais and Lacouette-Fougère started by constructing a theory of change to explain how the evaluations might impact policy. They then used process tracing to build operational tests to support an assessment of causal inference. One of the challenges in process tracing is envisaging how to 'operationalise' the four types of empirical test: 'straw in the wind', 'hoop', 'smoking gun' and 'double-decisive'. Delahais and Lacouette-Fougère's approach is interesting because they developed 'real-world' indicators that would match the empirical tests. In all, Delahais and Lacouette-Fougère constructed 10 such tests that combined to answer three sub-questions: Is the contribution of the evaluation to the observed changes possible, probable or intense? Considering intensity was a way of avoiding giving too much importance to real, but marginal, contributions. For each case study, 5 to 13 stakeholders were interviewed. The findings showed the diversity of impact pathways leading to reform (or lack thereof), including some unexpected ones, and stressed the importance of context and attitude of stakeholders in the impacts that could be expected. Also of interest is how elements of contribution analysis were used alongside process tracing.

Reference

Delahais, T. and Lacouette-Fougère, C. (2019) Try again. Fail again. Fail better. Analysis of the contribution of 65 evaluations to the modernisation of public action in France, *Evaluation*. doi: 10.1177/1356389018823237

3.3.6 RESOURCES

Web resources

Ricks and Lui have placed a number of worked examples online to accompany their 2018 article on process tracing. These follow the same steps as set out in their article and can be accessed at: https://static.cambridge.org/content/id/urn:cambridge.org:id:article:S1049096518000975/resource/name/S1049096518000975sup001.pdf

Key reading

For a detailed overview of the development of process tracing and its application, a good starting point is:

Bennett, A. & Checkel, J. (2015) Process tracing: from philosophical roots to best practice. In Bennett, A. & Checkel, J. (Eds.) *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

For a shorter overview of process tracing see:

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

For a practical guide to undertaking process tracing:

Ricks, J. I. & Liu, A. H., 2018. Process tracing research designs: a practical guide. *PS: Political Science & Politics*, 51(4), 842–846.

An important paper that has progressed thinking about the different tests used in process tracing:

Mahoney, J. (2012) The Logic of Process Tracing Tests in the Social Sciences, *Sociological Methods and Research*, 41(4) 570–597.

Further references

Bennett, A. & Checkel, J. (2015) Process tracing: from philosophical roots to best practice. In Bennett, A. & Checkel, J. (Eds.) *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

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

Mahoney, J. (2012) The Logic of Process Tracing Tests in the Social Sciences, *Sociological Methods and Research*, 41(4) 570–597.

Ricks, J. I. and Liu, A. H. (2018) Process-tracing research designs: a practical guide. *PS: Political Science & Politics*, 51(4), 842–846.

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

White, H. and 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.

3.4 GENERAL ELIMINATION METHODOLOGY

3.4.1 OVERVIEW

General Elimination Methodology (GEM) is a theory-driven qualitative evaluation method that improves our understanding of cause and effect relationships by systematically identifying and then ruling out causal explanations for an outcome of interest (Scriven, 2008; White & Phillips, 2012). Often used as a post-hoc evaluation method, it supports a better understanding of complexity.

Scriven, often cited as the originator of GEM, refers to it as 'inference to the best explanation' (Cook et al., 2010, p. 109) and explains:

> *"We claim that the intervention has causes that are visible, and we do that by eliminating other possible causes in relatively systematic ways, a complicated but perfectly feasible process."*
>
> **(Cook et al., 2010, p. 109)**

Also called Modus Operandi Approach, the GEM is compared to detective work, where suspects are ruled out based on the presence or absence of motive, means and opportunity. As such, it describes an approach to thinking about causation that we all use, subconsciously, on a regular basis.

Scriven (2008) proposed the GEM in a broader paper that discussed the limitations of RCTs and the concept of causal inference in the social sciences. His broader argument is that RCTs do not have a monopoly on making causal claims in the social sciences and are a specific tool for addressing causal inference in a narrow set of circumstances, whereas the GEM lies within the skills of every expert practice and 'is the underlying logic of RCTs and all quasi-experimental approaches as well' (Scriven, 2008, p. 21). In this sense, Scriven (2008) argues that the GEM is the basis for all causal claims. However, despite its ubiquity, there is relatively little literature on this approach.

The GEM has some similarities to process tracing, which could be seen as a more complex version of GEM. As such it may be a useful introduction to some of the more complex small *n* methodologies.

3.4.2 KEY ELEMENTS OF METHODOLOGY

GEM involves three primary steps (Scriven, 2008; White & Phillips, 2012).

Step 1 Establish a 'List of Possible Causes'

First, the evaluators identify all the possible causes for the impact of interest. Scriven (2008, p. 21) states that a List of Possible Causes:

> *...usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualization, and will vary depending on these parameters; of course, the context of the investigation determines the appropriate distance parameters.*

Second, they identify the necessary conditions for each possible cause and assess whether these are present. This work can be based on secondary data analysis such as a review of reports, articles, websites and other sources generally used to build a theory of change.

The evaluators then need to identify rival explanations for the outcome of interest. This is generally achieved by engaging stakeholders in interviews or workshops.

Step 2 List the modus operandi for each cause

A modus operandi (MO) is a sequence of events or set of conditions that need to occur/be present for the cause to be effective. In investigative terms, detectives (i.e. evaluators) construct a list of means, motives and opportunities which are considered against each suspect (i.e. cause). Scriven (2008, p. 21) states that:

> *Each cause has a set of footprints, a short one if it's a proximate cause, a long one if it's a remote cause, but in general, the MO is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective.*

The list of MO helps evaluators decide whether certain conditions should be included or rejected.

Step 3 Assess each case against the evidence available

For each possible cause, the evaluator will consider the presence or absence of the factors identified in the MO, and only keep those where all factors are present.

As White and Philip (2012) discuss, the logic here is two-fold. Identifying the elements of an MO that are present provides evidence that a possible cause may be an actual cause, whereas identifying the elements of an MO that are not present allows any possible cause that does not fit the evidence to be eliminated, leaving only those that do have a causal link. This reduces the number of potential causes and, ideally, leaves very few causal pathways. There are parallels here with the 'hoop tests' and 'smoking gun' tests of process tracing, which go further in formalising the logic of this approach.

3.4.3 RESOURCES REQUIRED

Skill set for evaluators

The GEM methodology is a form of case study analysis that uses predominantly qualitative methods, including interviews, observations, ethnography and document analysis.

Scale of the undertaking

Scriven (2008) emphasises that much work presented as a 'case study' is of poor quality: 'In the past a great deal of hopelessly unscientific work has been put forward as "qualitative methodology", including many anecdotal reports described as case studies' (p. 19).

It is important, therefore, that case studies are undertaken in-depth and to a high standard, and significant resources are therefore likely to be required.

3.4.4 CASE STUDY

There are relatively few published examples of the use of General Elimination Theory. The majority come from the conservation field.

Salazar et al. (2019) evaluated the impact of a social marketing campaign in the conservation sector using GEM. They wanted to assess the long-term impacts of a 1998 Rare Pride campaign on the island of Bonaire designed to increase the population of the Lora (Amazona barbadensis), a threatened species of parrot. Salazar et al. interviewed stakeholder groups to determine their perceptions of the drivers of changes in the Lora population over time. They used this data to develop an overall theory of change to explain changes in the Lora population by looking at the overlap in hypotheses within and between stakeholder groups. They then triangulated that theory of change with evidence from government reports, peer-reviewed literature, and newspapers. The increase in the Lora population was largely attributed to a decrease in illegal poaching of Loras and an associated decrease in local demand for pet Loras. They concluded that decreases in poaching and demand were likely driven by a combination of law enforcement, social marketing (including the Rare Pride campaign), and environmental education in schools. GEM helped to demonstrate how multiple interventions influenced a conservation outcome over time.

Reference

Salazar, B., Mills, M. & Verissimo, D. (2019) Qualitative impact evaluation of a social marketing campaign for conservation, *Conservation Biology*, 33(3) 364–644.

3.4.5 RESOURCES

Key reading

There is very little published material on General Elimination Methodology. The key paper in which GEM was first described in the terms set out here was

Scriven, M. (2008) A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research, *Journal of MultiDisciplinary Evaluation*, 5(9) 11–24. Available at: https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/160

For an interesting discussion about causation between Scriven (the originator of GEM) and Cook, a proponent of 'traditional' models of counterfactual impact evaluation, see:

Cook, T.D., Scriven, M., Coryn, C. L. S. & Evergreen, S. D. H. (2010) Contemporary Thinking About Causation in Evaluation: A Dialogue with Tom Cook and Michael Scriven, *American Journal of Evaluation* 31(1) 105–117. doi:10.1177/1098214009354918

Further references

Cook, T. D., Scriven M., Coryn C. L. S. & Evergreen S. D. H. (2010) Contemporary Thinking About Causation in Evaluation: A Dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, 31(1) 105–117. doi:10.1177/1098214009354918

Scriven, M. (2008) A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. *Journal of MultiDisciplinary Evaluation*, 5(9) 11–24. Available at: https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/160

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.

3.5 CONTRIBUTION ANALYSIS

3.5.1 OVERVIEW

'Contribution analysis explores attribution through assessing the contribution a programme is making to observed results' Mayne (2008, p. 1). Four conditions are needed to infer causality in contribution analysis (Befani & 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; either they were shown not to have made a significant contribution or their 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 may be responsible for the outcome) and conjectural (factors combine in complex ways to produce outcomes) (Befani & Mayne, 2014).

3.5.2 KEY ELEMENTS OF METHODOLOGY

Mayne (2008) sets out six steps of contribution analysis.

Step 1: Set out the attribution problem to be addressed

Mayne (2008) emphasises the importance of acknowledging the 'problem' of attribution and recognising that there are often legitimate questions about the extent to which a programme has brought about the results observed. It is therefore important to:

- Determine the specific cause-effect question being addressed, ensuring that it is a reasonable question to ask in the context.
- Determine the level of confidence required, by looking at how evaluation findings will be used and the kinds of decision that will be based on the findings.
- Explore the type of contribution expected, asking questions such as 'What would show that the programme made a difference?' and 'What kind of evidence would funders/decision-makers accept?'
- Determine other key influencing factors that will influence outcomes.
- Assess the plausibility of the expected contribution in relation to the size of the programme and, if it is not plausible, consider whether further work on cause and effect should be pursued.

Step 2: Develop the theory of change and the risks to it

Contribution analysis is based on a well-developed theory of change that makes clear the results chain that links the programme to outcomes. The theory of change should not be excessively detailed and can be refined later (Mayne, 2008). When determining the expected contribution of the programme, Mayne (2008) draws on Montague et al. (2002) to identify three circles of influence:

- **direct control** – where the programme has fairly direct control of the results, typically at the output level
- **direct influence** – where the programme has a direct influence on the expected results, typically the immediate outcomes and perhaps some intermediate outcomes, and
- **indirect influence** – where the programme has less influence on the expected results due to its lack of direct contact with those involved and/or the significant influence of other factors.

For each element of the theory of change, it is then necessary to identify the assumptions behind the theory e.g. what conditions have to exist for A to lead to B, and the risk of those conditions (Mayne, 2008). The theory of change should also consider how external factors influence outcomes. Finally, there may be alternative or competing theories of change amongst those involved in the programme; these should be assessed with evidence gathered to confirm or discard alternative theories (Mayne, 2008).

Step 3: Gather existing evidence on the theory of change

The evaluator should next assess the logic of links in the theory of change. Evidence will need to be gathered in three areas (Mayne, 2008):

- Evidence on results and activities: evidence should cover the occurrence (or not) of key results including outputs and outcomes as well as evidence that the programme was implemented as intended.
- Evidence on assumptions: evidence should demonstrate that assumptions in the theory of change are valid, or at least plausible, and this is likely to involve reviewing existing research evidence. Depending on the scale of the evaluation, this could potentially include some form of systematic review or rapid evidence assessment.
- Evidence on other influencing factors: evidence should examine other significant factors that may have had an influence.

Mayne (2008) describes this as an iterative process.

Step 4: Assemble and assess the contribution story, and challenges to it

The contribution story can now be assembled and assessed critically (Mayne, 2008). This will involve examining links in the results chain and assessing which are strong and which are weak, assessing the overall credibility of the contribution story and ascertaining whether stakeholders agree with the story. Mayne (2008) stresses that, so far, no 'new' data has been gathered other than from discussions with programme individuals and perhaps experts and/or a literature search.

Step 5: Seek out additional evidence

Based on the robustness of the contribution story, the evaluator next identifies what new data is needed to address challenges to its credibility (Mayne, 2008). At this stage, it may be useful to update the theory of change or look at certain elements of it in more detail. Generally, gathering evidence from multiple sources and triangulating findings are preferable. Contribution analysis does not specify particular data collection and analytical strategies. Mayne (2008) gives examples that include surveys, case studies and using monitoring data to track variations in programme implementation over time. He also suggests the possibility of conducting a component evaluation on a particular issue or area where performance information is weak and synthesising research and evaluation findings.

At this point in the process:

> *[I]f one can verify or confirm a ToC [Theory of Change] with empirical evidence – that is, verify that the steps and assumptions in the intervention ToC were realised 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.*

(Befani & Mayne, 2014, p. 21)

Step 6: Revise and strengthen the contribution story

Contribution analysis is most effective as an iterative process and should, ideally, be seen as an ongoing process that incorporates new evidence as it emerges (Mayne, 2008). This may include responding to evolution within the theory of change as well as new monitoring data or the results of a longitudinal survey.

3.5.3 MULTI-METHOD APPROACHES

Contribution analysis is closely related to theory of change, with a theory of change being the starting point for the use of contribution analysis.

Befani and Mayne (2014) have noted that contribution analysis and process tracing are similar: both seek to make causal inferences using non-counterfactual approaches, based on causal mechanisms and theories of change. They also note that a potential limitation of contribution analysis is that 'it is an approach, and does not spell out detailed steps to follow in data collection or discuss explicitly the types and strength of evidence used' (Befani & Mayne 2014, p. 25). They, therefore, suggest combining the two approaches so that the evaluator follows the logic of contribution analysis but uses the various tests developed in process tracing to provide an indication of what evidence to look for and what criteria to use to judge the strength of the evidence. This strengthens Step 5 of the contribution analysis approach in particular by encouraging the evaluation to ask specific questions related to data collection such as: 'What kind of evidence is (mostly) necessary and/or (mostly) sufficient to confirm/disconfirm a causal explanation?' Thus, the evaluator would make use of three of the tests developed by Van Evera (1997): the 'smoking gun', 'hoop' and 'doubly decisive' tests, used to assess evidence for hypotheses that are being tested.

In practical terms, this involves introducing an additional three-stage procedure at Step 5, which may be repeated in Step 6 as required. The three stages involve testing the intervention's main mechanism, other causal mechanisms external to the intervention and the comprehensive theory of change, including the intervention and external factors. The intervention's main mechanism and other causal mechanisms external to the intervention are tested using 'hoop tests' designed to disconfirm the causal mechanisms under analysis and then 'smoking gun' tests designed to confirm them. The combination of these two tests maximises certainty in the case of the hoop test and uniqueness in the case of the smoking gun test. The final stage, if the evaluator is confident that they have good evidence on all relevant causal factors, is to attempt to test the whole theory of change using a doubly decisive test, described in 3.3.2 (Befani & Mayne, 2014).

3.5.4 RESOURCES REQUIRED

Skill set for evaluators

In Steps 1 – 4 of a contribution analysis, a range of mostly qualitative research skills are required, including the abilities to interview stakeholders, hold workshops to develop theories of change, analyse the documentary evidence and conduct literature reviews, which may include the use of systematic review and meta-analysis techniques.

A broader range of research and analysis skills may be required in Steps 5 and 6. Mayne suggests a range of research activities, including surveys and the use of 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 on understanding the needs of decision-makers and funders, implies a need for evaluators to have 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 need 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.

3.5.5 CASE STUDY

There are relatively few published examples of the use of contribution analysis. Delahais and Toulemonde (2012) published an article in the journal *Evaluation* that draws on five evaluations that apply contribution analysis in the context of EU policies in development aid, agriculture, employment and governance. Each of the five evaluations explores different elements of the methodology, illustrating its strengths and weaknesses and describing how the authors applied key steps in the methodology to real-world situations. For example, one evaluation looks at two programmes funded by the European Commission, aimed at encouraging citizens to debate European issues through 'citizen consultations' run by selected non-governmental organisations. The debates were intended to make citizens' voices heard in EU policy-making processes. Contribution analysis was used by the authors to answer a question about 'the contribution of the programme to citizens' debates on the future of the EU and the impact of the EU on their daily lives'. They developed a logic model based on a literature review and expert views on what could reasonably be expected from the programme. The model was tested through 21 case studies that attempted to balance evidence confirming and refuting each intended contribution in the logic model, and to explore all other contributory factors. As the draft contribution story was somewhat negative, it was strongly challenged by the programme managers. This meant that several findings had either to be consolidated by further evidence or reformulated. The findings showed that, despite enthusiasm among most of the participants about the deliberative process itself, the programmes failed to trigger any debates outside their own small audience, contrary to the expectations of the promoters of the programme. They failed to gain mass media coverage and, therefore, had no influence on public opinion.

Reference

Delahais, T. & Toulemonde, J. (2012). Applying contribution analysis: Lessons from five years of practice. *Evaluation*, 18(3) 281–293. https://doi.org/10.1177/1356389012450810

3.5.6 RESOURCES

Web resources

An interesting lecture that includes John Mayne talking about contribution analysis (19 minutes in) and starts with a worked example (3 mins 25 seconds in) can be found at: https://www.youtube.com/watch?v=VkBbr8dOisk

Key reading

The originator of contribution analysis, John Mayne, has published several articles. The most commonly cited, that sets out the key elements of contribution analysis in a concise and accessible form is:

Mayne, J. (2008) *Contribution Analysis: An approach to exploring cause and effect*. Brief 16, Institutional Learning and Change (ILAC) Initiative.

A few years later, Mayne took stock of developments in contribution analysis and wrote another, widely cited, article on the topic:

Mayne, J. (2012) Contribution analysis: Coming of age? *Evaluation*, 18(3) 270–280. https://doi.org/10.1177/1356389012451663

More recently, he has again revisited contribution analysis:

Mayne, J. (2019) Revisiting Contribution Analysis, *Canadian Journal of Program Evaluation*. 34(2) 171–191

The following paper provides practical examples of the use of contribution analysis and a discussion of key concepts, such as contribution claim, causal mechanism and strength of evidence:

Delahais, T. & Toulemonde, J. (2012) Applying contribution analysis: Lessons from five years of practice. *Evaluation*, 18(3) 281–293. https://doi.org/10.1177/1356389012450810

Further references

Befani, B. & Mayne, J. (2014) Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation, *IDS Bulletin* 45(6).

Mayne, J (2008). *Contribution Analysis: An approach to exploring cause and effect. Brief 16*, Institutional Learning and Change (ILAC) Initiative.

Montague, S., Young, G. & Montague, C. (2003) Using circles to tell the performance story. *Canadian Government Executive* 2 12–16.

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

3.6 MOST SIGNIFICANT CHANGE

3.6.1 OVERVIEW

The Most Significant Change (MSC) technique is a participatory, qualitative method: a dialogical, story-based *monitoring and evaluation* technique that involves the collection and selection of *significant change stories* which have occurred in the field. Significant change stories are, in most cases, elicited directly from programme participants. They are then passed upwards in the organisational hierarchy to panels of stakeholders who assess their significance, discuss how they relate to wider implications of changes and review the available evidence that supports them. This process helps reduce the number of stories to those identified as being the most significant by the majority of stakeholders.

MSC is designed to run throughout the programme and, as an evaluation technique, its primary aim is 'to facilitate program improvement by focusing the direction of work toward explicitly valued directions and away from less-valued directions' (Mathison, 2005). It may also be useful in informing decision-makers about performance through success stories, promoting the recognition of different values among stakeholders and identifying unintended outcomes. In MSC, the stories themselves reveal the causal patterns (even if implicitly) and storytellers interpret these causal links through the construction and interpretation of stories. Thus, MSC has the potential to facilitate a dynamic dialogue between designated stakeholders and enable participants to reflect on what the programme really wants to achieve and how best to do so.

While MSC was originally designed as an impact monitoring – rather than an evaluation – approach, it has since been adapted for use in impact evaluation by expanding the scale of story collection and the range of stakeholders involved. MSC is useful in producing illustrative arguments to support the evaluation; it can capture and aggregate the views of different stakeholders and facilitate improvements by enabling the programme to work in explicitly articulated directions. Its obvious limitations are that it invites selection and social desirability biases (i.e. stories may be selected as most significant because they align with organisational views and visions) (Lennie, 2011). For instance:

- The voices of those good at telling stories may dominate over others who are less articulate.
- Story-selection processes are inevitably subjective and will mediate the views of those included in the selection panels.
- Majority votes in the selection process may silence minority voices and unpopular views.

For these reasons, MSC is not typically used alone for generating summative judgements about overall programme outcomes. Instead, it usually precedes or complements summative evaluation and may best be used alongside more rigorous approaches, such as contribution analysis or process tracing, when tackling causal inference.

Davies and Dart see the stories produced through MSC as mini-case studies and argue that:

> *It is quite conceivable that the stories could be a rich source of hypotheses about how things work in programmes. MSC could be used, in part, to identify causal relationships between particular activities and outcomes in stories and to then recommend systematic surveys of the incidence of these activities and their relationship to the outcomes.*

(2005, p. 62)

This suggestion, however, still presumes the inevitability of introducing statistical control at some point in the evaluation process. In contrast, Stern et al. (2012) stress that participatory methods – including MSC – can support causal inference by focusing on the 'agency' of the stakeholders. Such a perspective is consistent with Ellerman's (2009) claim that development is only possible through self-directed actions. As MSC and other participatory approaches move away from seeing beneficiaries as passive recipients, the focus of evaluation shifts to beneficiary and stakeholder perspectives in demonstrating contributors to change. This raises questions regarding voice and power that more traditional (quasi) experimental evaluations sometimes struggle to address. Thus, MSC provides 'a greater voice to those at the bottom of the organisational hierarchy' (Davies & Dart, 2005, p. 71) and MSC's focus on stories and narratives may allow the disruption of dominant discourses by giving voice to those on the margins (Dinh et al., 2019).

3.6.2 KEY ELEMENTS OF METHODOLOGY

MSC is described by the developers as:

> *...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, p. 8)

The evaluators collect significant stories of change and engage in the dialogical interpretation of these stories. This process requires the sample and reporting period to be defined and the 'domains of change' to be established. Although the method can be applied to identify a negative impact, it is most typically used to explore positive, exemplary cases rather than negative (or average) ones (Davies & Dart, 2005). The collection of stories is then reviewed by stakeholders, who are guided by facilitators to select the most significant. The selection criteria are determined by the stakeholders: this can happen through informal discussions or by using a formal rating process. The selected cases and the selection process (i.e. the reason for selection) are recorded by facilitators. Such reflectivity may also feed into the interpretation of underlying values and preconceptions of the various stakeholders.

A detailed guide is provided by Davies and Dart (2005), who discuss the steps involved in implementing MSC. The manual includes 10 steps, of which Steps 4, 5 and 6 are deemed fundamental, while the others are discretionary:

1. How to start and raise interest

Evaluators should be clear about the purpose of using MSC within the organisation, and use past programme examples to demonstrate how the method can be effective. It should be noted that MSC is easy to implement and – for most practitioners involved – does not require deep theoretical knowledge.

It may be useful to identify people excited by the notion of MSC who could act as catalysts in the process. These 'champions' can play a key role in designing and implementing MSC across the organisation.

2. Establishing 'domains of change'

Domains of change are fuzzy categories that must be defined to guide which significant change stories are sought. They should be broad and non-prescriptive, allowing participants to interpret what constitutes a change within the given domain (e.g. 'changes in the quality of people's lives', 'changes in the nature of people's participation in development activities').

There are examples of domains being developed by top-down or bottom-up processes (i.e. by senior managers or beneficiaries). Domains can be formed around individuals, organisations, communities or partnerships – depending on the level of interest.

Domains are not essential; it is possible to proceed without them, particularly in small organisations, where the number of stories is likely to be smaller. It is also possible to identify them after the stories are collected as a means to sort these into meaningful categories.

3. Defining the reporting period

The frequency of collecting significant change (SC) stories can vary. Higher frequency reporting (e.g. fortnightly) allows people to integrate the process more quickly, but increases the cost of the process and the risk of the participants running out of SC stories. Low-frequency reporting (e.g. yearly) requires fewer resources but also means a slower learning process. There is also a risk that participants may forget how the process works and what the aims are. There are examples of organisations decreasing the frequency of reporting over time (e.g. a monthly selection eventually evolving into a three-monthly reporting; see for example Dart, 2000).

4. Collecting stories of change

Data collection should start with a central open question, such as, 'Looking back over the last month, what do you think was the most significant change in the quality of people's lives in this community?'

This captures a specific period ('last month'), empowers participants ('what do you think'), asks them to be selective and focus on change rather than static events ('most significant change'), and defines the 'domain of change' ('quality of people's lives'), and establishes boundaries ('in this community').

SC stories can be captured in different ways:

- Unsolicited stories documented by fieldworkers in the course of their work
- Interviewing
- Group sessions
- Stories written directly by beneficiaries

Key information about the stories should be documented:

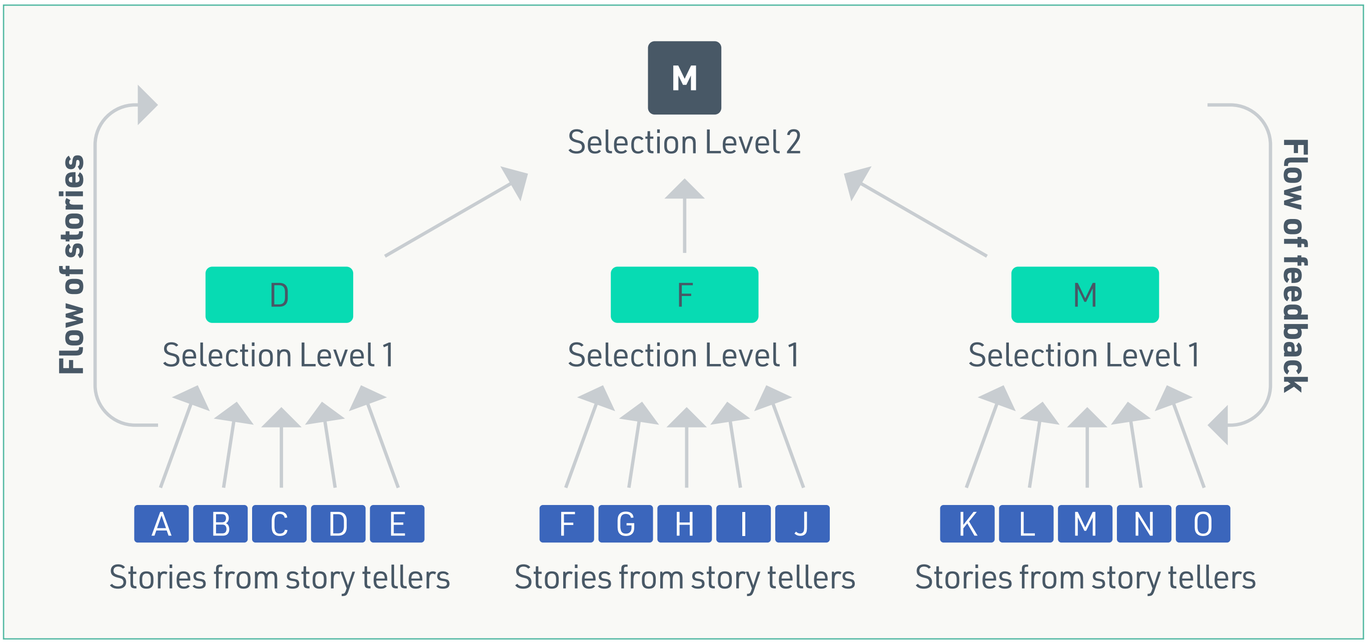
- Who collected the story and when the events occurred
- Description of the story itself – what happened
- Significance (to the storyteller) of the events described in the story

SC stories should be recorded as they are told. The description of the stories should be a simple narrative description of what happened, to whom and in what circumstances. From reading the stories, it should be clear why the storyteller identifies the story as significant. Stories should be short and comprehensible for all stakeholders.

5. Reviewing the stories within the organisational hierarchy

MSC uses an iterative process to select the most significant stories. Storytellers discuss their SCs and identify and submit the most significant to a level above. The same process is then run at mid-levels, where stories are selected and submitted to the next level. This process is called the 'summary of selection' and allows widely valued stories to be distinguished from those that have only local importance (see Figure 9).

Figure 9. Flow of stories and feedback in MSC (Davies & Dart, 2005, p. 29)



The structure of the process can follow pre-existing organisational hierarchies or can be set up for the purposes of MSC.

Depending on the evaluation aims and the scale of the project, the different levels may involve beneficiaries, field workers, managers, donors and investors. The selection of criteria for significant SCs can also vary: the decision can be reached using majority or iterative voting, scoring or a secret ballot. However, Davies and Dart (2005) suggest that the identification of selection criteria should not take place in advance but should emerge through the discussions of those involved as a way of opening up the process to new experiences.

6. Providing stakeholders with regular feedback about the review process

As in every learning-oriented system, results in MSC must be fed back to storytellers. Communicating the reasoning behind the choice of the most significant accounts can aid people during the next reporting period and move the focus of attention to relevant ideas and away from marginal ones.

7. Setting in place a process to verify the stories if necessary

The verification of accounts can be beneficial, as it may identify deliberately fictional stories, real events that have been misunderstood or misrepresented, and those whose significance has been exaggerated.

8. Quantification

While MSC is essentially qualitative, the quantification of surrounding information may be useful, including:

- counting the number of people involved and the number of events that took place
- retrospectively measuring (usually at the feedback stage) whether a significant event occurred in other instances besides the one already recorded
- counting the number of times that a specific type of change is noted (see next step).

9. Conducting secondary analysis and meta-monitoring

It may be useful to classify and examine the topics identified in SC stories using thematic coding, analyse positive and negative changes (e.g. a growing number of negative incidents may signal negative developments), analyse the differences between selected and non-selected stories and investigate patterns in selection criteria (e.g. do criteria vary over time? Do different groups use different selection criteria?).

10. Revising the MSC process

MSC should not be used unreflectively; rather, implementation should be adapted throughout the process. This may involve changing the frequency of reporting or the sampling population during and after the introductory phase. Revising the system suggests organisational learning and reflection – key underlying features of the MSC method.

3.6.3 MULTI-METHOD APPROACHES

Given that MSC has different biases to those present in more conventional techniques, it is a particularly useful addition to other evaluation methods (to offset inherent biases) and is generally a good complementary piece in the evaluation of complex participatory programmes with numerous stakeholder groups and multiple organisational layers, especially those producing diverse, emergent outcomes (Dart & Davies, 2003). For example, it may be combined with a more 'technical' and formalised methodology such as process tracing or QCA to give a greater emphasis to the user's voice.

Dart and Davies (2005) highlight the complementary function of MSC in deductive approaches to (1) improve understanding of the logic of an intervention, (2) enhance contextual knowledge about the success of the outcome or (3) complement studies whose main focus is on the 'average' experience of people.

Yet MSC is also useful as an inductive approach to generate hypotheses in exploratory studies (e.g. Pimentel et al., 2020) and to identify the unintended consequences of an intervention or programme.

3.6.4 RESOURCES REQUIRED FOR AN EVALUATION

Evaluator skills and experience

To build the capacity of a program evaluation team in MSC, Dart and Davies recommend two options:

- 1–3 days in-house training led by an external consultant or an internal evaluator
- Practice and improvement: given that MSC is a reflective process with an inbuilt improvement cycle, implementing it through trial and error may be feasible when training is not an option, using revisions and feedback loops, for example.

Resource implications

MSC is not a quick option. The analysis takes a significant amount of time and requires advanced project infrastructure. Maintaining the engagement of the different groups involved can be challenging, so too many review cycles are not recommended.

Running MSC evaluation requires the ability to identify priorities alongside good facilitation skills. Hence, the appointment of 'champions' who have the necessary theoretical and practical skills together with hands-on knowledge of the organisational structure is beneficial in raising organisational interest and identifying how MSC can be implemented in a given environment (see Step 1 to implementing MSC).

According to Kotvojs and Lasambouw (2009), MSC can be maximised by the following points:

- 'Using it over the life of the Program so that it is able to support change.
- Providing formal and on-the-job training to story collectors and selection panel facilitators [...]
- Only introducing the concept of domains when those involved are confident in MSC.
- Specifically stating you are seeking 'good and bad' changes [...]
- Managing data collected with a database, which facilitates effective management of data and efficient analysis of data [...]
- Consider having one person facilitate and document all selection panels, otherwise provide significant support in this process.
- Analysing all the data collected, not just the MSC story, or the stories selected by the selection panel.
- Using reporting which meets each stakeholder's needs, this may require a variety of reporting approaches.' (2009: 9).

MSC should not be used where:

- 'The evaluation is looking for typical cases. MSC focuses on extreme cases.
- There is no sense of 'mystery' about the outcomes, they are known, well defined and measurable.
- The evaluation focus is accountability rather than learning.
- Training of story collectors and selection panel facilitators cannot be well resourced.
- There is an expectation that it is 'easy' and needs little support.
- There is an expectation that it will require few if any additional resources. MSC is very time-intensive.
- Feedback will not be given.' (2009: 9).

3.6.5 CASE STUDY

Dahmen-Adkins and Peterson (2019) describe an application of MSC reported at the end of a European gender-equality change project. The four-year project involved 20 change agents who worked towards implementing action plans to tackle gender inequality in seven research institutions. MSC stories were collected from beneficiaries, change agents and other stakeholders via questionnaires and interviews that asked them to reflect upon their experiences of the most significant change that emerged during the project. Stories reflected on both personal and institutional changes. On an individual level, changes were categorised into three types: changes occurring in the realms of knowledge/awareness, behaviour or daily lives. The types of institutional change were categorised as referring to either cultural, policy or structural/management changes.

The authors argue that the MSC technique allowed evaluators to 'systematize the changes that contributed to closing the gender gap' (Dahmen-Adkins & Peterson, 2019, p. 157). It effectively complemented more traditional evaluation tools by gathering evidence and generating knowledge around unique dimensions of change involved in the project, such as

- tangible and intangible changes
- changes in behaviour and attitude
- expected and unexpected changes
- changes in collective and organisational character

This case study provides an example of the complementary usefulness of MSC. MSC can act as a prelude to other – more rigorous – methods as an inductive approach to explore underlying mechanisms contributing to change as well as – in this example – providing participatory or emancipatory perspectives on projects in their final stages to elaborate on the unique and otherwise unobserved experiences of beneficiaries and stakeholders and how micro-, meso- and macro-level changes are perceived by beneficiaries and different stakeholders.

Reference

Dahmen-Adkins, J. & Peterson, H. (2019). Most Significant Change: Closing the Gender Gap in Research. In Paoloni, P., Paoloni, M. and Arduini, S. (Eds) 2nd *International Conference on Gender Research*. Academic Conferences and Publishing Ltd. 151–158. Available to download at: https://www.researchgate.net/publication/332180557_Most_Significant_Change_Closing_the_Gender_Gap_in_Research

3.6.6 RESOURCES

Web resources

A website providing a brief overview and signposting translations, training opportunities and software regarding MSC:

Davies, R. *Most Significant Change (MSC) – Monitoring and Evaluation NEWS*. [online] Available at: https://mande.co.uk/special-issues/most-significant-change-msc/ [Accessed 24 Aug. 2021].

This site collates and makes available papers of any kind written on the subject of MSC technique:

Davies, R. *Most Significant Change technique. Group Library* [online] Available at: https://www.zotero.org/groups/266453/most_significant_change_technique/ [Accessed 24 Aug. 2021].

Key reading

For a detailed description of the methodology and its application:

Davies, R. & Dart, J. (2005). *The 'most significant change'(MSC) technique. A guide to its use*. PDF available at: https://www.researchgate.net/publication/275409002_The_'Most_Significant_Change'_MSC_Technique_A_Guide_to_Its_Use

And, more recently:

Lennie, J. (2011). *The Most Significant Change technique. A manual for M&E staff and others at Equal Access*. Washington, DC: USAID.

Further references

Bateson, G. (1979). *Mind and nature: A necessary unity* (Vol. 255). New York: Bantam Books.

Baú, V. (2016). A narrative approach in evaluation: 'Narratives of Change' method. *Qualitative Research Journal* 16(4), 374-387

Costantino, T. E. & Greene, J. C. (2003). Reflections on the use of narrative in evaluation, *American Journal of Evaluation*, 24(2), 35–49.

Dart, J. J. (2000), *Target 10 Evaluation stories*, Department of Natural Resources and Environment, Victorian State Government, Melbourne. PDF available at: http://www.clearhorizon.com.au (site posted April 2005).

Dart, J. & Davies, R. (2003). A dialogical, story-based evaluation tool: The most significant change technique. *American Journal of Evaluation*, 24(2), 137–155.

Davies, R. & Dart, J. (2005). *The 'most significant change'(MSC) technique. A guide to its use*. PDF available at: https://www.wikifplan.org/WIKIPLAN/1%201%20151%20-%20Most_significant_change_methodology_pa_abril%202005.pdf

Dinh, K. Worth, H. & Haire, B. (2019) Buddhist evaluation: Applying a Buddhist world view to the most significant change technique. *Evaluation*, 25(4), 477–495.

Ellerman, D. (2009) *Helping people help themselves: From the World Bank to an alternative philosophy of development assistance*. University of Michigan Press.

Kotvojs, F. & Lasambouw, C. (2009) MSC: *Misconceptions, Strengths and Challenges*. Presented at the Australasian Evaluation Conference, September 2009. Available at: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=027FFD5FCB9F459CD6EC276FA93DA99E?doi=10.1.1.624.7803&rep=rep1&type=pdf

Lennie, J. (2011). The *Most Significant Change technique. A manual for M&E staff and others at Equal Access*. Washington, DC: USAID.

Mathison, S. (2005). *Encyclopedia of evaluation* (Vols. 1-0). Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781412950558

Pimentel, J., Kairuz, C., Merchán, C., Vesga, D., Correal, C., Zuluaga, G., ... & Andersson, N. (2020). The experience of Colombian medical students in a pilot cultural safety training program: a qualitative study using the most significant change technique. *Teaching and Learning in Medicine*, 33(1), 58–66.

Riley, T. & Hawe, P. (2005). Researching practice: The methodological case for narrative inquiry. *Health education research*, 20(2), 226–236.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations*.

3.7 QUALITATIVE COMPARATIVE ANALYSIS (QCA)

3.7.1 OVERVIEW

Qualitative Comparative Analysis (QCA) is a 'synthetic strategy' (Ragin, 1987, p. 84) that allows 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 comprises 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. QCA builds on set theory to consider causal asymmetry and determine whether the conditions (or causes) are sufficient and/or necessary. It aims to achieve parsimonious (i.e. short) explanations while still accounting for causal complexity (Berg-Schlosser et al., 2009).

QCA, as we know it today, emerged from the social sciences, notably through the work of Ragin (1987) and has become increasingly accepted as a method enabling systematic comparison. In the social sciences, QCA is most frequently 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. organisational level) (Ott et al., 2018; Rihoux & De Meur, 2009), and it is increasingly now used at a micro-level (e.g. individual level) (Berg-Schlosser et al., 2009). Its use as an evaluation method is more recent and is derived primarily from 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 gaining traction with evaluators to 'unravel explanatory patterns for "success" and "failure" of existing cases, with the possibility to inform potential future cases' (Pattyn et al., 2019, p. 56). The method is also attractive to evaluators as it works with a small number of cases. QCA bridges quantitative and quantitative methods, integrating the strengths of both methods while 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 (Ott et al., 2018; Rihoux, Rihoux & De Meur, 2009):

> *The degree of maturity and robustness of a generalisation 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, p. 11)

It is nevertheless considered a rigorous method due to its replicability and transparency (Rihoux, Rihoux & De Meur, 2009).

QCA is an iterative process in which the researcher 'engages in a dialogue between cases and relevant theories' (Berg-Schlosser et al., 2009, p. 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 transformed 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 is an iterative process that requires in-depth knowledge of cases, as well as data to have a certain granularity. For instance, it needs to include cases where the outcome was negative as well as positive. Similarly, the conditions need to include cases where the condition is present as well as those where it is absent. Therefore, the quality and granularity of the data are paramount.

There are three main QCA techniques: crisp-set (csQCA), fuzzy set (fzQCA) and multi-value (mvQCA). They differ in how they code and consider membership of the cases. In csQCA, initially developed from Boolean logic, membership is dichotomous (e.g. 1= member, 0 = non-member). However, this dichotomous nature is not always adapted to real-life situations. In response to this limitation, fsQCA was developed 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 a variance in the observations, thus overcoming one of the challenges of csQCA which involves dichotomous analysis (Ott et al., 2018). However, csQCA allows for a more transparent process, as calibration (i.e. setting of thresholds) is conducted manually and explained theoretically. In 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 the presence of rain is sufficient to show that there is a cloud). Necessity and sufficiency are usually considered together because all combinations of the two are meaningful:

- '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.' (Rihoux, 2017, p. 36)

Different types of analysis can be conducted to determine whether conditions are necessary (superset analysis), sufficient (subset analysis), or both (INUS analysis).

3.7.2 KEY ELEMENTS OF METHODOLOGY

Rihoux and De Meur (2009) identify steps for csQCA:

Step 1 Building a dichotomous data table

Drawing on the theory of change, data is coded dichotomously for each condition and outcome (i.e. 1 = member, 2 = non-member). Thresholds need to be clearly justified and recorded when defining the presence and absence of conditions. This process of assigning numerical values to empirical manifestations of conditions is called calibration (Befani, 2006). Good practice for dichotomising conditions in a meaningful way can be found in Rioux and De Meur (2009, p. 41).

Step 2 Constructing a 'Truth Table'

Using software such as FsQCA or R, a first 'synthesis' of the raw data is produced in what is called a truth table. This is a table of configurations (i.e. a number of combinations of conditions associated with a given outcome). The truth table provides five types of configuration:

- Those with a [1] outcome
- Those with a [0] outcome
- Those marked by '-', which indicates an indeterminate outcome
- Those marked by 'C', which indicates contradictory outcomes (i.e. the configuration leads to a [1] outcome in some cases, but a [0] outcome in other cases). These contradictions need to be resolved
- Those marked 'L' or 'R', which indicate logical remainders (i.e. combinations that are theoretically possible but were not observed in empirical cases)

Figure 10: Example of Truth Table of Boolean Configurations from Rihoux and De Meur (2009, p. 44)

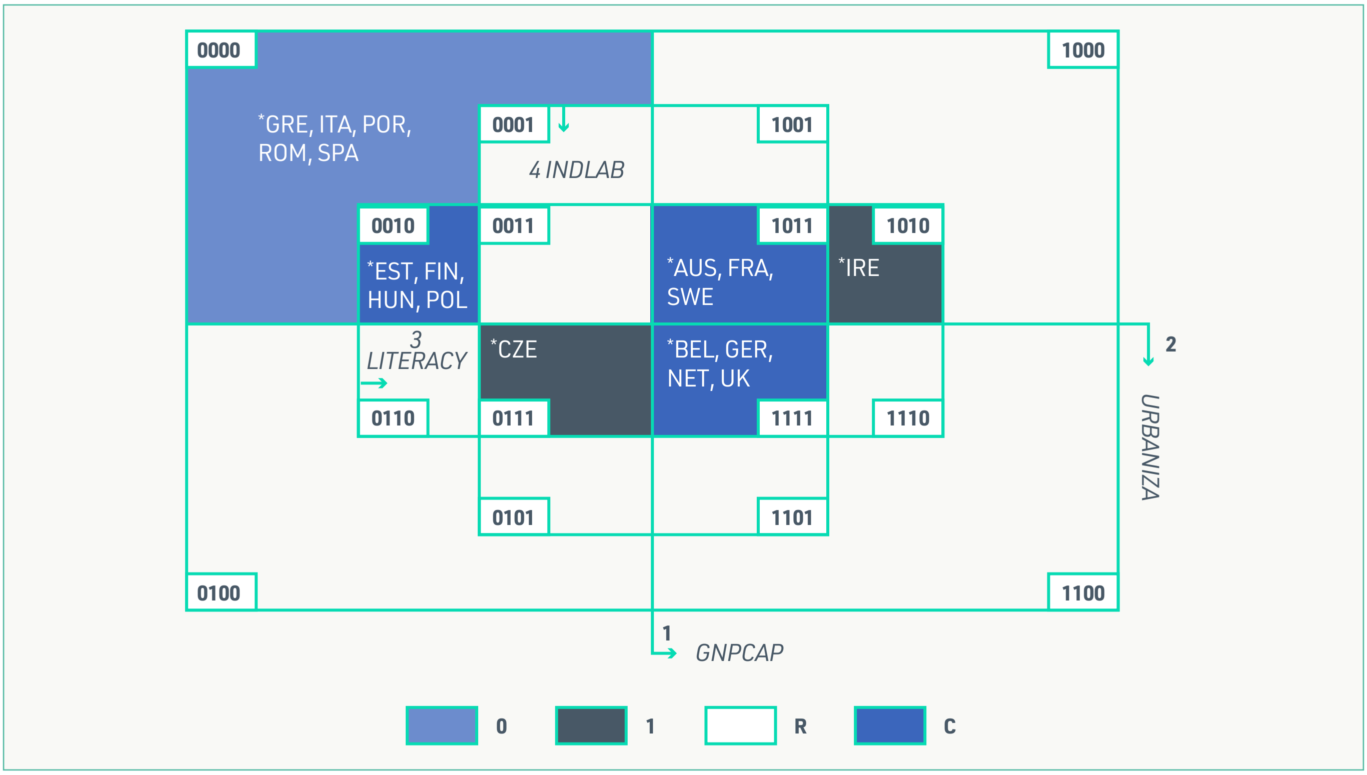
CASEID	GNPCAP	URBANIZA	LITERACY	INLAB	Survival
SWE. FRA. AUS	1	0	1	1	C
FIN. HUN. POL. EST	0	0	1	0	C
BEL. NET. UK. GER	1	1	1	1	C
CZE	0	1	1	1	1
ITA. ROM. POR. SPA. GRE	0	0	0	0	0
IRE	1	0	1	0	1

Step 3: Resolving Contradictory Configurations

Contradictory configurations are a normal part of QCA and are found in cases of reiterative dialogue between data and theory. The evaluators need to resolve the contradictions through their knowledge of the cases and reconsider their theoretical perspective to obtain more coherent data. There are several ways to resolve contradictions: a simple method is to add conditions that may explain why a configuration then leads to a [1] outcome or [0] outcome. Other strategies include re-calibrating how some conditions are operationalised; for example, moving the threshold of dichotomisation for a given contradiction may possibly resolve the contradiction. This process or recalibration is a normal part of QCA and supports the refinement of the theory underpinning the programme. Other strategies are also available, see Rihoux and De Meur (2009).

Venn diagrams (see Figure 11) provide a visualisation of the membership and non-membership of cases and are particularly useful when considering whether conditions are necessary or sufficient. They can also indicate which variable may need recoding to obtain a more parsimonious configuration.

Figure 11: Example of Venn Diagram produced by Tosmana software and corresponding to conditions in Figure 1 (Rihoux & De Meur, 2009, p. 46)



Step 4: Boolean Minimisation

This step is usually completed by the software and synthesises the truth table. It identifies conditions that are either present or absent in configurations leading to the same outcome. As their presence or absence does not change the outcome, they are considered ‘trivial’: they do not help the evaluators to discriminate between success and failure. They are therefore removed to identify the shortest configuration possible.

Boolean minimisation is first applied to configurations with a [1] outcome, then to those with a [0] outcome, then to those without either, including the logical remainders.

Step 5: Bringing in the ‘Logical Remainders’ Cases

Logical remainders constitute a pool of potential cases that can be used by the software to produce a shorter (i.e. more parsimonious) minimal formula. By including hypothetical cases, the software can develop broader categories of memberships and make ‘simplifying assumptions’.

Similar to the previous step, minimisation is applied to configurations with a [1] outcome, then to those with a [0] outcome.

Consistency and coverage

When running a Boolean minimisation, specialist software (e.g. Tosmana or fsQCA) calculates consistency and coverage for each configuration and the solution as a whole. Consistency measures the degree to which the configuration is a subset of the outcome. Coverage measures how far the outcome is explained by each configuration (Rihoux, 2017).

3.7.3 MULTI-METHOD APPROACHES

Increasingly, evaluators are considering QCA as an exploratory method best used in combination with other methods. It can be integrated with statistical analysis to strengthen the theoretical contribution of the research (Meuer & Rupietta, 2017). According to Befani (2006), evaluation approaches based on generative causation such as QCA can be combined with:

- Systems-based evaluation: a holistic view of the factors affecting the outcome, including feedback loops, may lend itself to the simulation of complex dynamics.
- Realist evaluation: to provide a magnifying lens on specific interactions in the causal chain or system (e.g. a specific ‘arrow’) or as a means to explore CMO configurations.
- Contribution analysis: to explore causal chains, with risks and assumptions for each pathway.
- Process tracing: used to generalise process tracing mechanisms. The combination of these two methods increases inferential value. QCA can inform the choice of cases in process tracing and process-tracing findings can support better calibration (Schneider & Rohlfing, 2013).

As QCA and process tracing are increasingly used together in multi-method studies, further guidance is emerging on their combination (see Álamos-Concha et al., 2021). As noted by Rihoux et al. (2021), the application of QCA is rapidly evolving:

The various innovations currently under way in the field of QCA, both at the conceptual and technical levels, will undoubtedly provide multiple opportunities for further refinements, both for fundamental and applied research. (Rihoux et al., 2021, p. 194)

3.7.4 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 an understanding of the 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 do not 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 by the absence of a space: [A*B] can also be written as: [AB].
- Logical 'OR', represented by the [+] (addition) symbol.

Beyond competencies, QCA 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 for those not familiar with the method. While books and online resources are available, formal training is 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 usually offer a good introduction, focused on evaluation work, in a one-day workshop.

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

- fsQCA freeware, including CRISP and FUZZY QCA.
- TOSMANA freeware, used for crisp-set QCA and multi-valued outcome QCA.

Resource implications

QCA can be run with a small number of cases, typically between 10 and 50. However, the depth and quality of the data required to run the analysis make the method challenging. The iterative process between data and theory is time-consuming. In some cases, such as where the evaluator does not have an in-depth knowledge of an 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, p. 182; Schneider & Wagemann, 2010), but collaborative work will provide opportunities to raise the quality of the evaluation. An evaluation using QCA can take between three months (if data is QCA-ready) and six months (if data needs cleaning and clarifying). QCA relies on available data on the outcome of choice. Therefore, the evaluation can take longer if this data is not available from the outset.

3.7.5 CASE STUDY

Bingham et al. (2019) used QCA to evaluate a university programme that employed an early alert intervention in which students enrolled on courses could be alerted by their instructors or university staff of concerns. Instructors could alert students for various reasons, including attendance concerns, missing assignments, low quiz or test scores, or because they were in danger of failing or could not pass. The analysis focused on students enrolled in mathematics courses. Bingham et al. used crisp-set QCA to explore the relationship between several input variables and one of two outcome variables: whether students passed their autumn term maths course and whether they enrolled in courses for the following term. The dichotomous input variables included in the analysis were whether students were first time freshmen, full-time students, used the university's maths advice centre, had an early alert meeting with the Programme Coordinator and enrolled in other maths courses. Observations were collected from the termly sample of students who received alerts in their respective maths courses. Truth tables were constructed of the cases associated with each outcome and the number of occurrences observed in the sample. Boolean equations were generated directly from the observed cases, using the truth table. The use of csQCA allowed Bingham et al. to identify combinations of characteristics and behaviours that were associated with student success and retention, as well as the factors to take into account in considering the university policy on early alerts. The findings were holistic and provided a context for the interaction of the characteristics and behaviours of observed variables.

Reference

Bingham, A. J., Dean, S. & Castillo, J. (2019) Qualitative comparative analysis in educational policy research: Procedures, processes, and possibilities. *Methodological Innovations*, 12(2) https://doi.org/10.1177/2059799119840982

3.7.6 RESOURCES

Web resources

Compasss is a website that specialises in QCA, listing events and providing extensive resources including a very comprehensive bibliography: https://compasss.org/

Ragin's *User's Guide to fzQCA* (2017) is a good starting point for fzQCA: https://www.socsci.uci.edu/~cragin/fsQCA/download/fsQCAManual.pdf

Key reading

A good introduction to QCA:

Befani, B. (2016) *Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis* (QCA). Sztokholm: Expertgruppen för biståndsanalys (the Expert Group for Development Analysis). Available at: https://eba.se/wp-content/uploads/2016/07/QCA_BarbaraBefani-201605.pdf

More in-depth books on QCA:

Rihoux, B. & De Meur, G. (2009) Crisp-Set Qualitative Comparative Analysis. In Rihoux, B. & Ragin, C. C. (Eds.) *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*. Sage Publications.

Schneider, C. Q. & Wagemann, C. (2012) *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Using R for QCA:

Oana, I. E., Schneider, C. Q. & Thomann, E. (2021) *Qualitative Comparative Analysis Using R: A Beginner's Guide*. Cambridge University Press.

Further references

Álamos-Concha, P., Pattyn, V., Rihoux, B., Schalembier, B., Beach, D. & Cambré, B. (2021) Conservative solutions for progress: on solution types when combining QCA with in-depth process-tracing. *Quality & Quantity*, 1–33.

Befani, B. (2013) Between complexity and generalization: Addressing evaluation challenges with QCA. *Evaluation*, 19(3), 269–283.

Berg-Schlosser, D., Meur, G., Rihoux, B. & Ragin, C. (2009) Qualitative comparative analysis (qca) as an approach. In Rihoux, B. & Ragin, C. C. (eds.) *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, 1-18. Thousand Oaks and London: Sage.

Bingham, A. J., Dean, S. & Castillo, J. (2019) Qualitative comparative analysis in educational policy research: Procedures, processes, and possibilities. *Methodological Innovations*, 12(2), 1–13. https://doi.org/10.1177/2059799119840982

Cress, D.M. & Snow, D.A. (2000) 'The outcomes of homeless mobilization: The influence of organization, disruption, political mediation, and framing', *American Journal of Sociology*, 105(4), 1063–1104.

Meuer, J. & Rupietta, C. (2017) A review of integrated QCA and statistical analyses', *Qual Quant*. 51, 2063–2083.

Ott, U., Sinkovics, R. R. & Hoque, S. F. (2018) Advances in qualitative comparative analysis (QCA): Application of fuzzy set in business and management research. In Cassell, C., Cunliffe, A. L. & Grandy, G. (Eds.) *The SAGE Handbook of Qualitative Business and Management Research Methods*. Newcastle Upon Tyne: Sage Publications Ltd.

Pattyn, V., Molenveld, A. & Befani, B. (2019) Qualitative comparative analysis as an evaluation tool: Lessons from an application in development cooperation, *American Journal of Evaluation*, 40(1), 55–74.

Ragin, C. (1987) *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ragin, C. (2008) *Redesigning social inquiry: Set relations in social research*. Chicago: Chicago University Press.

Ragin, C. (2014) *The comparative method*. Berkeley: University of California Press.

Rihoux, B. (2006) Qualitative comparative analysis (QCA) and related systematic comparative methods: Recent advances and remaining challenges for social science research. *International Sociology*, 21(5), 679–706.

Rihoux, B., Álamos-Concha, P. & Lobe, B. (2021) Qualitative Comparative Analysis (QCA) An Integrative Approach Suited for Diverse Mixed Methods and Multimethod Research Strategies. In Onwuegbuzie A. J. & Johnson R. B. (eds.) *The Routledge Reviewer's Guide to Mixed Methods Analysis*, 185–195.

Rihoux, B. & Lobe, B. (2009) The case for qualitative comparative analysis (QCA): Adding leverage for thick cross-case comparison. *The Sage handbook of case-based methods*, 222–242.

Rihoux, B. & De Meur, G. (2009) Crisp-Set Qualitative Comparative Analysis (csQCA). In Rihoux, B. & Ragin C. C. (eds), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks and London: Sage.

Schneider, C. Q. & Rohlfing, I. (2013) Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research. *Sociological Methods & Research*. 42(4), 559–597.

Schneider, C. Q. & Wagemann, C. (2010) Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. *Comparative Sociology*, 9(3), 397–418.

3.8 COMPARATIVE CASE STUDY

3.8.1 OVERVIEW

A comparative case study is defined as 'the systematic comparison of two or more data points ('cases') obtained through the use of the case study method' (Kaarbo & Beasley, 1999, p. 372). A case may represent a participant, an intervention site, a programme or a policy.

Case studies have a long history in the social sciences, yet case-based methods were long treated with scepticism (Harrison et al., 2017). However, the advent of grounded theory in the 1960s led to a revival in the use of case-based approaches. From the early 1980s, the uptake in case study research in the field of political sciences led to the integration of formal, statistical and narrative methods as well as the use of empirical case selection and causal inference (George & Bennett, 2005), which largely contributed to its methodological advancement. Now, a comparative case study:

> *has grown in sophistication and is viewed as a valid form of inquiry to explore a broad scope of complex issues, particularly when human behavior and social interactions are central to understanding topics of interest.*

(Harrison et al., 2017)

It is claimed that comparative case studies can be applied to detect causal attribution and contribution when the use of a comparison or control group is not feasible (or not preferred). Advocates maintain that comparative case studies can produce generalisable knowledge about why and how an intervention (programme or policy) is successful or fails to work. They are particularly effective in exploring the role of context in influencing intended outcomes and are helpful in navigating multifaceted, multimodal programme components that generate different causal outcomes.

Comparative case studies can make use of quantitative as well as qualitative methods and employ similar data collection techniques to single case studies. Yet, as they move beyond merely supporting claims that propose the success or failure of an intervention towards examining causality, they require comprehensive effort in setting up propositions, conducting analytic work and synthesising findings.

3.8.2 KEY ELEMENTS OF METHODOLOGY

Comparing cases enables evaluators to tackle causal inference through assessing regularity (patterns) and/or excluding other plausible explanations. The approach to causality that underpins the comparative case study (CCS) approach is described by Byrne and Ragin (2009). The essence of causality for these authors is 'complexity' and 'multiplicity'. Causality is understood as *complex*, as it cannot be attributed to a single variable, given that the system from which it derives is also a complex system and the influence of variables are closely intertwined, such that causal impact resists quantification. While complex systems are not to be understood as holistic or chaotic (given that 'their parameters do not change in *any* fundamental, qualitative fashion' Ragin, 2009, p. 103), they do have the potential of non-linear change as a response to changes in internal parameters or the external environment. Furthermore, *multiple* causality means that the same outcomes may be generated by different causal configurations. It follows that – as Ragin asserts – the outcome is '... not the product of any single cause, but rather of the interaction of multiple causes, which causes are not 'variables' external to the cases but rather embodied aspects of the cases' (Ragin, 2009, p. 101).

Thus, CCS – as a configurational approach – differs from linear modelling techniques in allowing multiple causation and seeing cases as complex systems that cannot be reduced to variables (Ragin, 2009).

Moreover, in the social realm, these generative mechanisms (systems of complex causes) are always contingent and emerge through interaction with context. Therefore, understanding how context influences results is a crucial aspect of CCS research.

CCS applications may include scenarios when (1) a programme model is implemented in multiple sites (or contexts) and – contrary to expectations – different outcomes are observed, or (2) when a range of different interventions (programmes) lead to similar outcomes.

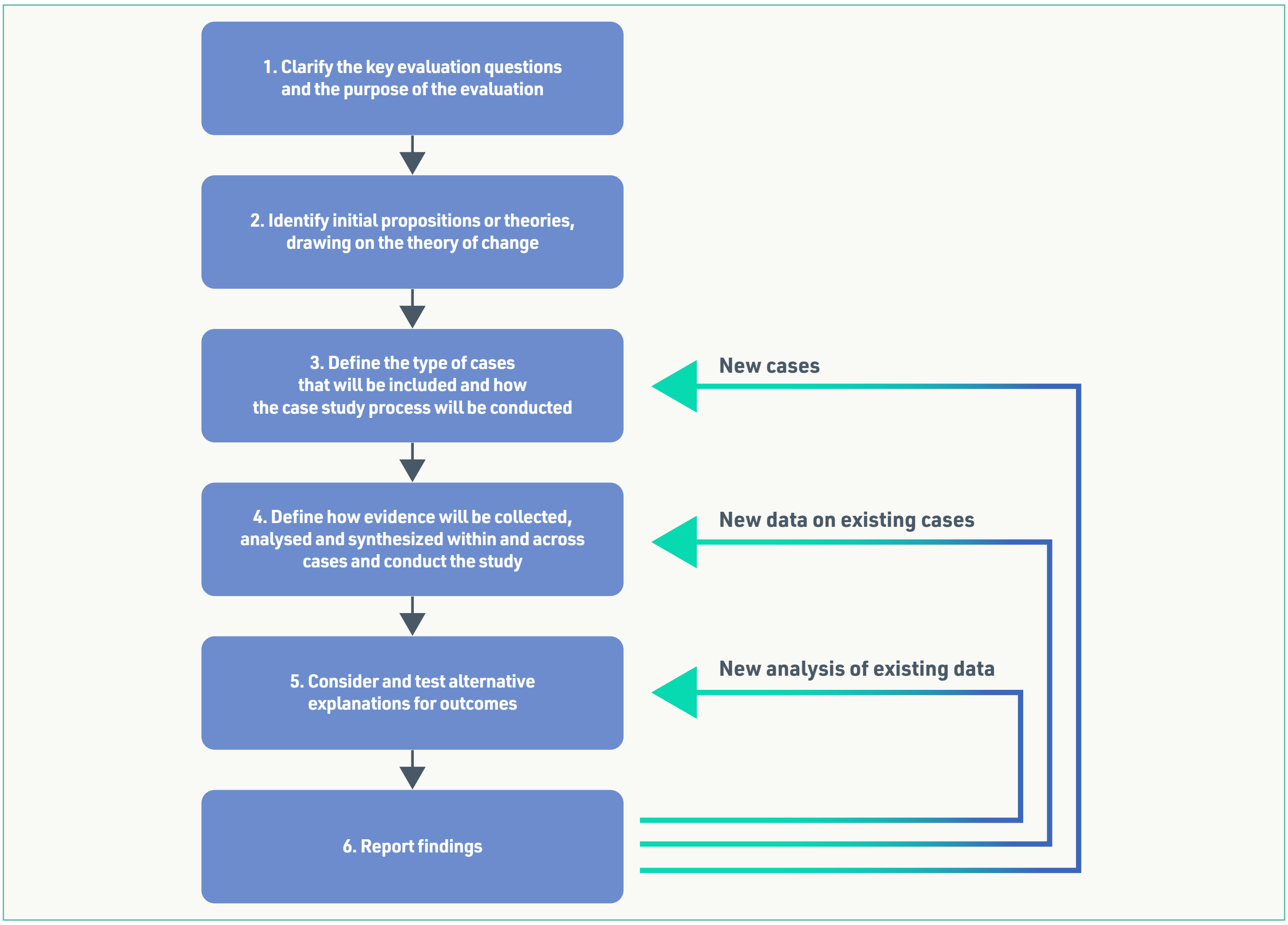
Depending on the quality of pre-existing theoretical explanations in the field, CCS can be used for the analytic purposes of:

- Building a theory – if available theories addressing a research question are scarce
- Refining existing theories – provided that theories are existent but underdeveloped, or
- Testing theories – to assess the applicability of multiple elaborated but competing theories (Vogt et al., 2011).

In practical terms, CCSs involve proposing, analysing and synthesising patterns (similarities and differences) across cases that share common objectives.

Goodrick (2014), on the assumption that detailed evaluation hypotheses and/or a comprehensive theory of change have been developed, discusses the subsequent steps to be taken in undertaking a CCS study (for mere theory-building purposes, single case studies are more often used). These steps include the following six essential stages: the first three are concerned with establishing the dimensions of the study and the selection of cases, while the last three allow the evidence to be retrieved and tested (see Figure 12).

Figure 12. The logic of comparative case studies (Goodrick, 2014, p. 2)


1. Clarify the key evaluation questions and the purpose of the evaluation
2. Identify initial propositions or theories, drawing on the theory of change
3. Define the type of cases that will be included and how the case study process will be conducted
4. Define how evidence will be collected, analysed and synthesized within and across cases and conduct the study
5. Consider and test alternative explanations for outcomes
6. Report findings
New cases
New data on existing cases
New analysis of existing data

Key evaluation questions and the purpose of the evaluation

The evaluator should explicitly articulate the adequacy and purpose of using CCS (this should be guided by the evaluation questions) and define the primary interests. The formulation of key evaluation questions promotes the selection of the most appropriate cases to be used in the analysis. The evaluator may be interested in:

- Describing the similarities and differences between cases
- Interpreting the implications of those similarities/differences
- Identifying and testing explanatory theories about how and why a programme worked (or did not work) in the given context.

Propositions based on the theory of change

The theories and hypotheses to be explored should be derived from the theory of change (or, alternatively, informed by previous research around the initiative, existing policy or programme documentation).

Case selection

Advocates for CCS approaches claim that an important distinction between the case-oriented small *n* studies and the (most typically large-n) statistical/variable-focused approaches lies in the process of selecting cases: in case-based methods, selection is iterative and cannot rely on convenience or accessibility. Hence, purely mechanical procedures, such as opportunity or random sampling, are not suitable in CCS, nor is it feasible to fix the number of cases a priori. Case selection should be an ongoing tentative process, in which – similar to the procedure of initial selection – the addition or removal of cases is justified on theoretical grounds (Mahoney & Goertz, 2004; Ragin, 1994).

'Initial' cases should be identified in advance, but case selection may continue as evidence is gathered. Cases can represent micro, meso or macro-level social phenomena. The selection of the unit of analysis – i.e. individuals, programmes, groups or implementation processes – should again be theoretically grounded. Goodrick (2014) suggests reflecting on key research questions when defining this unit of analysis. The evaluator may ask questions such as:

- 'Which group's response to the intervention is of interest?'
- How will an understanding of this group's perspective assist in helping to understand what worked, how and why?
- Which dimensions of the cases require study to answer the [key evaluation questions]?' (Goodrick, 2014, p. 7)

Various case selection criteria may be used depending on the analytic purpose (Vogt et al., 2011). These may include:

- Very Similar Cases: Except for one outcome that is different: Selected cases are very similar yet experience different outcomes. Used to assess the reasons for the difference among otherwise similar cases (see also Mill's most similar method). Can be used in to explore or confirm.
- Very Different Cases: Except for one outcome that is the same: Very different cases have the same outcome. Causal variables that are different can be ruled out, and other explanations can be explored or postulated.
- Typical or Representative: A case may be selected because it is representative of a large number of cases.
- Extreme or Unusual: A case is chosen precisely because it is an extreme case of a phenomenon. For instance, to test the effect of an environmental regulation, the evaluator may want to look at those states or nations with the strongest regulations, assuming that the impact – if there is any – is most likely to be identified and explained here.
- Deviant or Unexpected: To be used when strong prior evidence suggests an expectation for a particular (set of) case(s), yet the expectation proves incorrect.
- Influential or Emblematic: Used to identify influential cases that are typically dropped from a data set (e.g. when using regression analysis) as their presence would significantly alter the results. In case studies, however, they are useful to examine the robustness of hypotheses or theories.

Identify how evidence will be collected, analysed and synthesised

Comparative case studies often apply mixed methods. While other methodologies may mix qualitative and quantitative methods to explore different research questions, CCS analyses data together to gain in-depth knowledge of the cases and causal propositions. The dimensions of the comparisons should be defined by the theory of change and may include:

- Comparing how different programmes operate across cases and contexts.
- Comparing anticipated outcomes against actual outcomes.
- Comparing the responses of different stakeholders to a programme over time (within the same, or different contexts).

Test alternative explanations for the outcomes

Following the identification of patterns and relationships, the evaluator may want to test the established propositions in a follow-up exploratory phase. Approaches applied here may include:

- Triangulation: using multiple data sources to verify and substantiate an assessment
- Selecting contradicting cases: to help critically test the proposition
- Analytic approach: using qualitative comparative analysis (QCA) to examine attributes of cases that may be associated with the outcome, or variable-oriented approach to assess specific variables and their average effects.

Report findings

Reporting formats should always be framed around the key evaluation question, but the structure should find the right balance between description, interpretation and explanation. In CCS, findings are often complex and multifaceted; therefore, evaluators can use summary tables and diagrams, and structure findings around themes and main theoretical insights.

3.8.3 MULTI-METHOD APPROACHES

CCS can be used as a single component or is suitable for nesting within other designs. It is often used to complement (quasi-) experimental design to elaborate on findings and explain why or how an intervention was or was not successful. Nesting may also be helpful in understanding the specific contexts in which change occurred (e.g. Luecking et al., 2020).

When evaluating complex interventions, CCS can be useful in addressing evidence gaps, such as questions around the mechanisms that make an intervention successful, or in explaining relevant differences across different socio-cultural contexts (e.g. Pfadenhauer et al., 2021).

As a stand-alone design, CCSs are most often used to explore similarities and differences across contexts. In this case, they are likely to use mixed methods: for instance, they may combine measures, surveys or QCA (to identify causal relationships) with process tracing (to explore possible explanations and test whether alternative hypotheses can be excluded).

3.8.4 RESOURCES REQUIRED

Evaluator skills and experience

To apply CCS effectively, a range of skills and expertise is required (Goodrick, 2014), including expertise in relevant qualitative and/or quantitative methods (depending on the type of methodology used). Evaluators must be able to construct propositions (theories), embrace the complexities of the case and possess strong synthesising skills to integrate divergent (or convergent) evidence. The ability to formulate coherent arguments around often complex findings is also important.

Resource implications

CCS evaluation can be resource-intensive, especially if the study involves extensive fieldwork. If resources are scarce, it may be preferable to select a small number of cases (see discussion around typical or representative cases), or to rely entirely on secondary data, provided that the quality of evidence is strong enough.

3.8.5 CASE STUDY

Makerspaces are informal sites for creative production in art, science and engineering, where people of all ages blend digital and physical technologies to explore ideas, learn technical skills and create new products. Sheridan et al. (2014) used a CCS to explore how maker spaces may function as learning environments. The authors chose a case study approach because it allows the integration of diverse sources of evidence to build a deep within-case understanding of each maker space, and a comparative case approach because they judged that this was particularly suited to analysing commonalities and differences across sites given the diversity of maker spaces and the trend towards designing youth and family spaces after adult maker spaces. They used purposive sampling to select three makerspaces that reflected some of the diversity in types of participants in maker spaces and the nature of the participation. Drawing on field observations, interviews and analysis of artefacts, videos and other documents, the authors describe features of three maker spaces and how participants learn and develop through complex design and making practices. They describe how the maker spaces help individuals identify problems, build models, learn and apply skills, revise ideas and share new knowledge with others.

Reference

Sheridan, K., Halverson, E., Litts, B., Brahms, L., Jacobs-Priebe, L. and Owens, T. (2014) Learning in the Making: A Comparative Case Study of Three Makerspaces, *Harvard Educational Review* 84(4) 505–531. doi: https://doi.org/10.17763/haer.84.4.brr34733723j648u

3.8.6 RESOURCES

Web resources

A webinar shared by Better Evaluation with an overview of using CCS for evaluation:

Better Evaluation. (2016) Impact Evaluation Webinar 6 Comparative Case Studies. [video] Available at: <https://www.youtube.com/watch?v=SgLSR55BxHg> [Accessed 7 September 2021].

Key reading

A short overview, describing how to apply Comparative Case Studies for evaluation:

Goodrick, D. (2014). *Comparative Case Studies, Methodological Briefs: Impact Evaluation 9*, UNICEF Office of Research, Florence.

An extensively used book that provides a comprehensive critical examination of case-based methods:

Byrne, D. & Ragin, C. C. (2009). *The Sage handbook of case-based methods*. Sage Publications.

This book focuses on how case study is applied in practice using exemplary case studies drawn from a wide variety of academic and applied fields:

Yin, R. K. (2014) *Case study research: Design and methods (5th ed.)*. Sage: Los Angeles.

Further references

Byrne, D. (2009). Complex realist and configurational approaches to cases: A radical synthesis. In Byrne, D. & Ragin, C. C. (Eds.). *The Sage handbook of case-based methods*. Sage Publications. 101–112.

George, A. L. & Bennett, A. (2005). *Case studies and theory development in the social sciences* (4th ed.). Cambridge, MA: MIT Press.

Harrison, H., Birks, M., Franklin, R. & Mills, J. (2017). Case Study Research: Foundations and Methodological Orientations. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 18(1). https://doi.org/10.17169/fqs-18.1.2655

Kaarbo, J. & Beasley, R. K. (1999). A practical guide to the comparative case study method in political psychology. *Political Psychology*, 20(2), 369–391

Luecking, C. T., Dobson, P. & Ward, D. S. (2020). Barriers and Facilitators of Parent Engagement with Health Promotion in Child Care: A Mixed-Methods Evaluation. *Health Education & Behavior*, 47(6), 914–926.

Mahoney, J. & Goertz, G. (2004). The possibility principle: Choosing negative cases in comparative research. *American political science review*, 98(4), 653–669.

Pfadenhauer, L. M., Grath, T., Delobelle, P., Jessani, N., Meerpohl, J. J., Rohwer, A. & Rehfuess, E. A. (2021). Mixed-method evaluation of the CEBHA+ integrated knowledge translation approach: a protocol. *Health Research Policy and Systems*, 19(1), 1–12.

Ragin, C. C. (1994). *Constructing social research. The unity and diversity of method*. Newbury Park, CA: Pine Forge Press.

Vogt, W., Gardner, D., Haeffele, L. & Baker, P. (2011). Innovations in program evaluation: comparative case studies as an alternative to rcts. In *The SAGE handbook of innovation in social research methods* (pp. 293–324). SAGE Publications Ltd. https://www.doi.org/10.4135/9781446268261

3.9 AGENT-BASED MODELLING

3.9.1 OVERVIEW

Social sciences often aspire to move beyond exploring individual behaviours and seek to understand how interaction between individuals leads to large-scale outcomes. In this process, understanding complex systems requires more than understanding their parts. Agent-based modelling (ABM) is a bottom-up modelling approach in which macro-level system behaviour is modelled through the behaviours of micro-level autonomous, interacting agents. ABM can generate deep quantitative and qualitative insights into complex socioeconomic, natural and man-made systems through simulating interacting processes, diversity and behaviours at different scales. Social interventions are generally designed to influence micro-level behaviour (e.g. the behaviour of an individual or household). Similarly, evaluations are usually interested in explaining micro-level behavioural changes. ABM, in contrast, allows the representation of macro-level mechanisms, making it well-suited to evaluate complex programmes and policies, for example, evaluating multi-intervention outreach programmes.

The use of ABM expanded in the social sciences during the early 1990s to simulate large-scale social phenomena such as pollution, migration and disease (Epstein & Axtell, 1996). Later, many ABM studies turned their focus to designing effective teams and exploring the behaviours of social networks. According to a more recent survey of ABM practices (Heath et al., 2009), the method became widely used in a range of fields, including traffic, public policy and the military, yet the most popular areas using ABM are still economics, social science and biology.

Recent applications of ABM have been made possible by advances in the development of specialised agent-based modelling software, more granular and larger data sets and advancements in computer performance (Macal & North, 2010). However, ABM is a complex and resource-intensive method: the implementation process requires expert modellers, the results can be difficult to understand and communicate, and its application can be costly.

3.9.2 KEY ELEMENTS OF METHODOLOGY

The underlying premise of ABM is its 'complex system thinking'. That is, the complex world comprises numerous interrelated individuals, whose interactions bring about higher-level features. In ABM, such emergent phenomena are generated from the bottom up (Bonabeau, 2002) by seeking to reveal the underlying rules that govern the behaviours of whole systems. ABM presumes that simple rules behind individual actions can lead to coherent group behaviour, and that even a small change in these rules can radically change group behaviour. ABM characterises individual behaviour as non-linear and defined by thresholds and if-then logic. It can capture discontinuities in individual behaviour that are difficult to describe through structural statistical models (Bonabeau, 2002). Furthermore, ABM – unlike other modelling approaches – does not assume equilibrium within the social realm. Instead, systems are understood as dynamic and adaptable, consisting of the complex interactions of autonomous, decision-making agents, who influence each other, learn from their experiences and adapt their behaviour to better fit the environment (Macal & North, 2010).

A typical ABM contains three main elements:

1. Agents (with their attitudes and behaviour)
2. The relationships between agents
3. Environments (with which agents also interact) (Macal & North, 2010)

In ABM, the most essential property of *agents* is their autonomy. They are 'active, initiating their actions to achieve their internal goals, rather than merely passive, reactively responding to other agents and the environment' (Macal & North, 2010, p. 153). From a practical standpoint, agents have the following characteristics:

- They are *self-contained* (they have a distinct boundary that differentiates them from other agents).
- They have a *state* that varies over time (an agent's behaviour is conditioned on its state, which is conditioned by the agent's attributes as well as the state of other interacting agents and the state of the environment).
- They are social, meaning their behaviour is influenced by their dynamic interactions with their social environment.
- They may also be *adaptive* (i.e. they may modify their behaviour as a result of learning), and *heterogeneous* (their behaviours and characteristics are diverse and may vary in terms of their extent and sophistication).

Agent relationships and interactions are just as important for ABM as their behaviours. ABM is concerned with two principal questions: who interacts with whom (as agents only connect to a subset of agents – termed *neighbours*), and how these neighbours are connected (referred to as the *topology* of connectedness). Regarding the former, ABM understands (social) environments as decentralised systems, in which information is gained from an agent's interactions with its neighbours and the local environment – both can change rapidly as the agent moves. The topology of connectedness refers to the spatial (or social) network of agents and their relationships. Topology also models the direction of information. The neighbourhood in which agents interact can be specified spatially as well as socially. For instance, in the event of a pandemic, infection may be transmitted through physical actions that emerge in daily activities (special network) but also through interactions with friends and family (social network).

Information about the *environment* in which agents interact may be needed beyond spatial location: while interacting with the environment, agents are constrained in their actions by the infrastructure, resources, capacities or links that the environment can provide.

These elements need to be identified, modelled and programmed to create an ABM. The model developer must then use a computational engine to simulate the behaviours and interactions of agents. This step is executed using general software or programming language, or specifically designed toolkits. The underlying mainspring of ABM is that these behaviours and interactions are repeatedly executed by agents, and the resulting processes can be modelled to run in different structures: they can be activity-based, time-speed or discrete events, or they may operate over a timeline (Macal & North, 2010).

ABM is not purely inductive, nor is it deductive; it is often referred to as a third way of doing science (Axelrod & Tesfatsion, 2021). Deduction is used to specify a set of assumptions (about agents and the environment) and derive theorems about the system of interest, while induction is applied to identify patterns in the empirical data. Furthermore, as Axelrod and Tesfatsion (2021: unnumbered) explain,

> *... simulation does not prove theorems with generality. Rather, simulation generates data suitable for analysis by induction. In contrast to typical induction, however, the simulated data comes from controlled experiments rather than from direct measurements of the real world.*

Given these strengths, ABM is typically applied for three purposes:

- Theory development: theory is implemented in a model and tested to assess whether it can generate observed outcomes (Castellini et al., 2019).
- Analysis of real-world issues: ABM can draw on empirical research results to simulate future scenarios, potential interventions and counterfactuals to inform decision-making.
- Engagement of stakeholders in discussions and thinking: modellers and stakeholders work in collaboration to design models by editing and changing parameters. Such experimentation in the virtual world can trigger discussions about agent behaviours or the potential effects of different interventions (Gilbert et al., 2018).

ABM has been used in a variety of fields including the physical, biological, social and management sciences. According to Axelrod (1997), its application can be especially beneficial in systems that rely on competition and collaboration among its agents. Thus, it may be useful in higher education settings, where complex interactions may be difficult to understand using conventional analytic techniques (Triulzi et al., 2011).

3.9.3 MULTI-METHOD APPROACHES

ABM is most typically combined with other modelling approaches, such as social network analysis (for a recent review, see Will et al., 2020) and activity-based modelling (Müller et al., 2021).

ABM is commonly seen as methodologically incompatible with case-based methods (CBM) because (1) there is a conceptual difference between 'agent' and 'case', (2) ABM focuses on simulating processes for theory testing or scenario analysis while CBM seeks patterns in real data, and (3) the ingrained distance between quantitative and qualitative methodologies. However, recent studies have discussed the possibility of combining ABM with other small *n* case-based methods, such as using it in tandem with QCA (Castellini et al., 2019).

3.9.4 RESOURCES REQUIRED

Evaluator skills and experience

ABM requires expert modellers and facilitators; if the system is not captured in sufficient detail, the findings may not be meaningful. Computer scientists will be involved in building the complex computational models that underpin ABM. The findings are likely to be complex, hence results are often difficult to comprehend and communicate. However, it is argued that the complexity captured reflects that found in human behaviours and interactions.

Resource implications

In addition to assembling a team with the necessary skills to build complex computational models based on a detailed understanding of a social phenomenon, ABM is likely also to require specialist software and high-performance computers able to run simulations.

3.9.5 CASE STUDY

A study by Reardon et al. (2013) implemented ABM to simulate how socioeconomic background influences university application choices and enrolment. The goal of the simulation was to 'build intuition about the relative strength of some of the resource-based mechanisms that shape the distribution of students among more- and less-selective colleges and universities' (Reardon et al., 2013). The related mechanisms included: (1) prior educational performance (students with greater resources have a disproportionate ability to enhance their apparent academic preparation for university); (2) the quality of information used when selecting a university; (3) the number of applications submitted; (4) the perceived utility of university enrolment, and (5) the differences between students with high and low levels of resources in valuing higher/lower profile universities.

The model shows how the link between socioeconomic inequality and student performance drives observed patterns of resource stratification in university enrolment.

The study builds on a very basic framework and over-simplifies the issues surrounding university enrolment; it is not, therefore, a substitute for policy evaluation. Nevertheless, it is a useful method to explore the dynamic interdependent processes underlying the apparent patterns of stratification, and it can 'develop intuition about how student characteristics and behavior influence the sorting of students into colleges of varying quality' (Reardon et al., 2013).

Reference

Reardon, S., Kasman, M., Klasik, D., & Baker, R. (2016). Agent-based simulation models of the college sorting process. *Journal of Artificial Societies and Social Simulation*, 19(1), 8. doi: 10.18564/jasss.2993. Download at: https://www.jasss.org/19/1/8.html

3.9.6 RESOURCES

Web resources

The following website provides a good introduction to ABM:

Axelrod, R. & Tesfatsion, L. (2021) *Online Guide for Newcomers to ABM*. [online] Available at: <http://www2.econ.iastate.edu/tesfatsi/abmread.htm> [Accessed 24 August 2021].

Key reading

An introductory tutorial into the background and application of ABM:

Macal, C. & North, M. (2014) *Introductory tutorial: Agent-based modeling and simulation*. In Proceedings of the Winter Simulation Conference 2014, pp. –20. IEEE.

An early book that is regarded as launching the field of social ABM in a sustained way:

Epstein, J. M. and Axtell, R. (1996) *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: MIT Press.

A widely read book that provides a simple overview of the methodology including how to construct simple ABM:

Gilbert, N. and Troitzsch, K. (2005) *Simulation for the Social Scientist*. McGraw-Hill.

Exemplary ABM applications are scattered across disciplines. There is no single publication outlet for studies applying ABM, but the following journal has been for many years considered a high-quality source:

Jasss.org. 2021. JASSS. [online] Available at: <https://www.jasss.org/JASSS.html> [Accessed 24 August 2021].

Further references

Axelrod, R. (1997). *The complexity of cooperation*. Princeton University Press.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(3), 7280–7287.

Castellani, B., Barbrook-Johnson, P. & Schimpf, C. (2019). Case-based methods and agent-based modelling: bridging the divide to leverage their combined strengths. *International Journal of Social Research Methodology*, 22(4), 403–416.

Gilbert, N., Ahrweiler, P., Barbrook-Johnson, P., Narasimhan, K. P. & Wilkinson, H. (2018). Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation*, 21(1).

Heath, B., Hill, R. & Ciarallo, F. (2009). A survey of agent-based modeling practices (January 1998 to July 2008). *Journal of Artificial Societies and Social Simulation*, 12(4), 9.

Macal, C. M. & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3), 151–162. doi: 10.1057/jos.2010.3

Müller, S. A., Balmer, M., Charlton, W., Ewert, R., Neumann, A., Rakow, C., ... & Nagel, K. (2021). Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. *medRxiv*. doi: https://doi.org/10.1101/2021.02.27.21252583

Triulzi, G., Pyka, A. & Scholz, R. (2014). R&D and knowledge dynamics in university-industry relationships in biotech and pharmaceuticals: an agent-based model. *International Journal of Biotechnology 6*, 13(1–3), 137–179.

Will, M., Groeneveld, J., Frank, K., & Müller, B. (2020). Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review. *Socio-Environmental Systems Modelling*, 2, 16325-16325.

4. DISCUSSION: MOVING FORWARD WITH METHODOLOGIES FOR SMALL COHORTS

4.1 THE CASE FOR SMALL *N* METHODOLOGIES

One important point distinguishing small *n* from counterfactual approaches is how the respective strategies account for uncertainty. In counterfactual approaches, the two concepts of bias and precision are central. Control over both sources of uncertainty is exercised through research design (randomisation where possible and sample size/statistical power) and, in terms of precision, standard methods of statistical inference (e.g. computation of confidence intervals, *p*-values, effect sizes and/or Bayes factors). Small *n* approaches, in contrast, account for uncertainty through invoking concepts centring on ideas of complexity. In small *n* research and evaluation, the challenges of uncertainty and ignorance are in some senses greater due to the 'causes of effects' type strategies deployed, in which research attempts to account for all the possible relevant factors, besides the intervention, that give rise to the effect observed.

Small *n* methodologies can be used to address causal inference in situations where either comparison or counterfactual groups are not available and/or there are a small number of cases. However, they should not be seen as a 'second rate' alternative to a traditional, counterfactual impact evaluation. They draw on different understandings of causality, embrace the concept of complexity and answer different types of impact question; they provide different ways of 'unlocking the black box' associated with traditional, counterfactual impact evaluations. Nor are they a 'cheap' alternative to traditional, counterfactual impact evaluation. The focus on a small number of cases and the lack of a counterfactual does not necessarily reduce evaluation costs. Small *n* methods tend to rely on collecting large amounts of qualitative (and sometimes quantitative) data on a small number of cases and this can be time-consuming. While they do not require complex statistical analysis, they do involve complex and often iterative analysis of large, qualitative or mixed-method data sets.

4.2 SOME COMMON THEMES

While all the methodologies discussed in this guide are distinct, they share common elements.

Theory of change: All the small *n* methodologies discussed in this guide either explicitly or implicitly start with elaborating the theory of change that underpins the intervention being evaluated. Theory of change has become ubiquitous, to the extent that there is a risk that many exercises and outputs badged as 'theory of change' are superficial and will not offer the depth of analysis needed to provide an effective starting point for implementing a small *n* impact evaluation.

Hypothesis testing: In one form or another, all the small *n* impact evaluations discussed in this guide involve developing and testing hypotheses, and generally require the evaluator to specify what should be observed if the hypothesis is true or false (White & Phillips, 2012). Hypotheses must, therefore, be very specific. This is in part why well-developed theories of change are so important: good mid-level theories are needed from which to develop hypotheses that are sufficiently precise as to be testable using a small *n* methodology.

In-depth data collection: All the small *n* methodologies described in this guide involve detailed data collection at the case level. This will usually involve the collection of qualitative data and sometimes also quantitative data. It is difficult to specify the 'amount' of data that will be required in the abstract, but it seems likely that, in some small *n* impact evaluations, gathering in-depth qualitative data from one or more cases could be just as time-consuming and resource-intensive as data collection in a traditional, counterfactual impact evaluation.

Burden on participants: Except for ABM, all the small *n* impact methodologies described in this guide place some requirements on those participating in the service and on front-line staff delivering services. The need for in-depth qualitative data collection will make it hard to minimise this. Whereas a traditional counterfactual impact evaluation may rely primarily on administrative data or a survey, these data sources are unlikely to be sufficient for most small *n* methodologies.

Ownership: Most of the methodologies discussed in this guide require greater and more sustained involvement of the various stakeholders, be they those people who access a service or those who manage and deliver it. These more participatory approaches to evaluation will create different balances of power and ownership from traditional, counterfactual evaluation designs. Such shifts can be beneficial but also increase certain risks in the delivery of evaluation. Co-production can take more time, require additional resources and lead to evaluations proceeding in directions not originally envisaged.

Programme and sector knowledge: All evaluation requires that evaluators have the necessary skills, including in evaluation design, data collection, analysis and report writing. In a traditional, counterfactual impact evaluation, the burden of skills and experience required will be on the methods and evaluation approaches used. However, in a small *n* impact evaluation, the evaluator additionally needs a deeper knowledge of the programme, and the context within which it is being implemented, than might typically be required in a traditional, counterfactual impact evaluation. This is because most of the methodologies described in this report require the evaluator to develop a deep understanding of the case(s), the important context within which the cases operate and to develop detailed theories of change that draw on existing evidence and mid-level theory about the intervention and how it functions.

4.3 BUILDING INSTITUTIONAL CAPACITY

Before embarking upon a small *n* impact evaluation, evaluators will need to build up their knowledge and experience of small *n* methodologies. There are many ways to do this, from participating in formal training courses to self-directed learning.

Theory of change underpins all the methodologies in this guide so, for a team new to small *n* impact methodology, building skills and experience around theory of change will be a useful starting point. Theory of change is also one of the most frequently used methodologies outlined in this guidance and a range of resources, including guides and training courses, is available.

Realist evaluation is a widely used approach in evaluation, and many resources and courses are available to support an evaluator new to this approach. This may, therefore, be a methodology that is easier to take up and apply.

Some of the methodologies lend themselves to retrospective use. Process tracing, for example, has been used to analyse historical political events to better understand the causal process that preceded them. One option for 'practising' some of the small *n* methodologies covered here may, therefore, be to undertake ex-post or retrospective evaluations on interventions that are complete, but for which abundant data is available.

Some of the methodologies can be understood as extensions of one another or as broadly complimentary. For example, process tracing can be seen as a more complex version of GEM. As such, GEM may be a useful introduction to some of the more complex small *n* methodologies.

GLOSSARY

Attribution: Attribution involves a causal claim about the intervention as the cause of the impact.

Bias: Bias is a systematic source of error.

Case: The case is a complex entity in which multiple causes interact.

Case-based method: There are many different case-based methods. Their proponents agree that case-based methods can be used to make generalisations (i.e. develop explanations of causality that go beyond the specific instance being studied) and that case-based methods can be qualitative or quantitative (Byrne, 2009).

Case study: A case study is the detailed and intensive analysis of a single research case. More specifically, when applied to projects/programmes/interventions, case studies allow us to understand better the sequence of events, the different perspectives of key actors and, ultimately, the interplay between cause and effect.

Causality: Causality is the relationship between two events where one event is a source for the effect in the other event.

Contribution: Contribution makes a causal claim about whether and how the intervention has contributed to the impact (Stern et al., 2012). When the intervention is tightly defined and evaluation focuses on a single outcome, attribution may be possible. Where interventions are complex and multiple outcomes are being considered, focusing on contribution may be more appropriate (Stern et al., 2012).

Comparison group: A group of participants that share similar characteristics to those in the treatment group but do not receive the intervention. Comparison groups are constructed using statistical matching techniques rather than random allocation.

Complexity, complex systems: Complexity characterises the behaviour of a system whose components interact with each other and their environment in multiple ways, following local rules. These interactions give rise to emergent properties. The relationship and interactions of the differing parts of a complex system are dynamic and non-linear; thus, small changes can have disproportionate outcomes and vice versa.

Confidence interval: A study can only report an estimate of the effect size for the whole population because it is obtained from a sample. The true effect (i.e. the population value) will lie somewhere between a range of values known as the confidence interval. For example, a study may report an estimated effect size of 1.68 with 95% confidence that the true effect size is likely to fall between 1.20 and 2.36.

Confounding factors, confounding variables: A confounding factor or variable is a third variable that influences both the **dependent variable** and the **independent variable**. It is a causal concept.

Contamination: Contamination occurs when members of the control group are exposed to the intervention.

Control group: A group of participants equivalent to the treatment group as a result of random allocation which minimises any observed or unobserved differences.

Counterfactual: The counterfactual is what would have happened in the absence of an intervention. A control group is constructed to mimic this condition so that evaluators can compare the outcomes of participants who received an intervention with a group so similar that they could be the same participants who had not received the intervention.

Deductive: A deductive approach is primarily concerned with the testing of a hypothesis that has been derived from existing theory. It is an approach that goes from the general to the specific and tests a theory.

Dependent variable: The dependent variable refers to the situation researchers wish to explain with the **independent variable**. The dependent variable is thus the outcome (effect) of the treatment or intervention.

Dichotomous : A dichotomous variable contains two distinct values.

Effect size: The effect size is a standardised measure of the magnitude and direction of the difference in outcome between treatment and control groups after an intervention.

Epistemology: The philosophy of the justifications for human knowledge claims.

Ethics: Ethics is about how we behave or should behave both as individuals and as part of the society in which we live in interaction with others. In evaluation, ethical principles can be applied to participants' and stakeholders' rights, and govern evaluators' inherent ethics and ethical behaviour.

Ex-post evaluation: An evaluation is undertaken after programme implementation and when programme outcomes are known. Can be contrasted with an ex-ante evaluation undertaken prior to programme's implementation.

External validity: External validity describes how well the results of a study translate to other contexts.

Hypothesis: In its more 'technical' sense, a hypothesis is a predicted or expected answer to a research question. When used in this way, reference will often be made to a 'null hypothesis', which is the hypothesis that there is no relationship between two variables. The term 'hypothesis' is also used more generally as an idea to organise understanding of the evaluand and thereby guide observation (Robson, 2011).

Impact evaluation: Impact is the portion of an outcome change that can be attributed to the programme rather than other factors or influences. An impact evaluation is designed to evaluate the impact of a programme and implies a broader scope than an outcome evaluation.

Independent variable: The independent variable is the source of effect that causes changes in the **dependent variable**. We may more commonly refer to this as the intervention.

Inductive: An inductive approach extracts a likely premise from specific and limited observations to develop a hypothesis. It is an approach that goes from the specific to the general and can be used to build a theory.

Internal validity: Internal validity refers to the extent to which the design and conduct of a trial eliminate bias.

Longitudinal: Longitudinal research involves repeated observations of the same measurements among the same population. If data is collected periodically during the course of the intervention, the evaluators are better placed to construct an account of how the intervention works over time. Longitudinal designs can last many years.

Mechanism: A mechanism explains what it is about a programme that makes it successful. Mechanisms are not simply variables but accounts that encompass agency and structure. They demonstrate how programme outputs follow from stakeholders' choices and their capacity to put these into practice.

Mediator: A mediator (mediating variable) explains how two variables are related.

Meta-analysis: A meta-analysis is an analysis of analyses. It is a statistical method of combining the results of separate primary studies on a particular intervention to estimate the size of an effect on the population. Meta-analyses are typically conducted as part of a **systematic review**.

Mid-level theory: Mid-level or middle-range theory is a sociological concept that attempts to combine high-level abstract concepts and concrete empirical examples. As such, it makes hypothesis development possible by bringing together data and evidence with theoretical constructs to help make sense of the concept. Mid-level theories can also support a focus on causal mechanisms operating at a more general level than the context of specific individual interventions and, thus, encourage the identification of patterns and 'regularities' across a range of interventions or contexts. This process can support the development of generalisable conclusions and, therefore, the transfer of hypotheses between related interventions or contexts.

Moderator: A moderator (moderating variable) explains the strength and direction of a relationship between two variables.

Outcomes: Outcomes are the eventual benefits to society that a programme is intended to achieve.

Outputs: Outputs are a quantitative summary of an activity (The SROI Network 2012) that relate in some way to the outcomes the programme is designed to achieve.

Propensity Score Matching: Propensity Score Matching is a statistical technique for constructing comparison groups.

Process evaluation: This monitors whether an intervention was carried out as intended.

Quasi-experimental design: Evaluations that adopt a quasi-experimental design construct comparison groups using statistical techniques rather than random assignment to ensure that the characteristics of the group are similar to those exposed to the intervention.

Random assignment: Random assignment ensures that treatment and control groups are similar in both observable and unobservable characteristics by allocating participants on a chance basis.

Randomised Controlled Trial: In a randomised control trial, participants are randomly assigned to treatment (intervention) and control (placebo or 'business as usual') groups. Random assignment ensures a high degree of confidence that there are no systematic differences between treatment and control groups except that the treatment group participated in the intervention. By eliminating **selection bias**, we can be confident that any difference in group outcomes is the result of the intervention, not a characteristic of the groups. Equivalence between the treatment and control group is necessary for any causal claims about the effect of the intervention on the outcome of interest.

Rapid Evidence Assessment: A Rapid Evidence Assessment is an accelerated **systematic review**. It adopts a transparent and rigorous method but has a less exhaustive search strategy.

Sample size: The number of participants involved in a trial.

Selection Bias: Selection bias occurs when observable and unobservable characteristics are correlated with treatment assignment. If selection bias is present, it is impossible to disentangle the effect of an intervention from pre-existing differences between those exposed to it and those not.

Statistical power: The ability to detect whether an intervention has had an effect.

Systematic review: A review of primary studies that adopts explicit and reproducible methods to minimise bias.

REFERENCES

Armitage, P. (2003). Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology*, 32(6), 925–928.

Bamberger, M. (2015). Innovations in the use of mixed methods in real-world evaluation. *Journal of Development Effectiveness*, 7(3), 317–326.

Bamberger, M., Tarsilla, M. & Hesse-Biber, S. (2016). Why so many 'rigorous' evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning*, 55, 155–162.

Baron, R. M. & Kenny, D. A. (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

Befani, B. (2020) *Choosing Appropriate Evaluation Methods: A Tool for Assessment and Selection*. London: CECAN.

Befani, B. & Stedman-Bryce, G. (2016) Process Tracing and Bayesian updating for impact evaluation. *Evaluation*. http://doi.org/10.1177/1356389016654584

BIS (Department for Business, Innovation and Skills) (2014) *National Strategy for Access and Student Success in Higher Education*. BIS: London.

Boeije, H. R., Drabble, S. J. & O'Cathain, A. (2015) Methodological challenges of mixed methods intervention evaluations. *Methodology*, 11(4), 119–125.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. & Moore, L. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75, 2299–2306.

Byrne, D. (2009) Case-based methods: why we need them; what they are; how to do them. In Byrne, D. & Ragin, C. C. (eds.) *The SAGE Handbook of Case-Based Methods*. London: Sage.

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and quasi-experimental designs for research on teaching*. Houghton Mifflin and Company.

Cartwright, N. & Hardie, J. (2012) *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Crawford, C., Dytham, S. & Naylor, R. (2017) *Improving the evaluation of outreach: Interview report*. Bristol: OFFA.

Dawid, C. A. P. (2007) *Fundamentals of statistical causality*.

Fisher, R. A. (1925) *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Formby, A., Woodhouse, A. & Basham, J. (2020a) Reframing widening participation towards the community: a realist evaluation, *Widening participation and lifelong learning*, 22(2), pp. 184–204. doi:10.5456/WPLL.22.2.184.

Formby, A., Woodhouse, A., Basham, J. & Roe, F. (2020b) 'A presence in the community': developing innovative practice through realist evaluation of widening participation in West Yorkshire, *Widening participation and lifelong learning*, 22(3), pp. 173–186. doi:10.5456/WPLL.22.3.173.

Frazier, P. A., Tix, A. P. & Barron, K. E. (2004) Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1), 115.

Gerber, A.S. & Green, D.P. (2012) *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.

Gibson Smith, K., Alexander, K. and Cleland, J. (2021) 'Opening up the black box of a Gateway to Medicine programme: a realist evaluation', *BMJ Open*, 11(7), pp. e049993–e049993. doi:10.1136/bmjopen-2021-049993.

Gorard, S. & Smith, E. (2006) Beyond the 'learning society': What have we learnt from widening participation research? *International Journal of Lifelong Education*.

Hansen, A. B. G. & Jones, A. (2017) Advancing 'real-world' trials that take account of social context and human volition. *Trials*, 18(1), 531.

Harrison, N., Waller, R. & Last, K. (2015) *The evaluation of widening participation activities in higher education: A survey of institutional leaders in England*. [Online]. Bristol: University of West England. Available from file:///Users/user/Downloads/report-1-institutional-survey-June-2015.pdf

Harrison, N. & Waller, R. (2017a) Success and impact in widening participation policy: What works and how do we know? *Higher Education Policy*, 30(2), pp.141–160.

Harrison, N. and Waller, R. (2017b) 'Evaluating outreach activities: overcoming challenges through a realist 'small steps' approach', Perspectives: Policy and Practice in Higher Education: Themed Issue: *Access to Higher Education*, 21(2–3), pp. 81–87. doi:10.1080/13603108.2016.1256353.

Harrison, N., Vigurs, K., Crockford, J., McCaig, C., Squire, R. & Clark, L. (2019) *Evaluation of outreach interventions for under 16 year olds: Tools and guidance for higher education providers* (pp. 1–14).

Hill, J. & Stuart, E. (2015) 'Causal Inference: Overview' in Wright, J. (Ed.) *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier pp. 251–254.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396), 945–960.

Horton, M. & Hilton, G. (2020) *Evaluation Report: Does engagement in Aimhigher interventions increase the likelihood of disadvantaged learners progressing to HE? A mixed-methods approach employing a quasi-experimental design and case studies*. [Online]. Birmingham: Aimhigher West Midlands. Available from https://aimhigherwm.ac.uk/wp-content/uploads/2020/10/Aimhigher-Mixed-Methods-Impact-Evaluation-Study-2020.pdf

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. & Kerr, K. (2016) *Implementation and process evaluation (IPE) for interventions in educational settings: An introductory handbook*. Education Endowment Foundation.

Imai, K., Keele, L., Tingley, D. & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4), 765–789.

Jamal, F., Fletcher, A., Shackleton, A., Elbourne, D., Viner, R. & Bonell, C. (2015) The three stages of building and testing mid-level theories in a realist RCT: a case-example. *Trials, 16*. https://doi.org/10.1186/s13063-015-0980-y

Johnson, R. B. & Schoonenboom, J. (2016) Adding qualitative and mixed methods research to health intervention studies: Interacting with differences. *Qualitative Health Research*, 26(5), 587–602.

Keele, L. (2015) Causal Mediation Analysis: Warning! Assumptions Ahead. *American Journal of Evaluation*, 36(4), 500–513.

Lawson T. (1997) *Economics and Reality. Routledge*: London.

Lendrum, A. & Humphrey, N. (2012) The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, 38(5), 635–652.

Marchal, B., Belle, S., Olmen, J., Hoerée, T. & Kegels, G. (2012) Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research, *Evaluation* 18(2) pp.192–212.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., & Hardeman, W. (2015) Process evaluation of complex interventions: Medical Research Council guidance. *BMJ, 350*. https://doi.org/10.1136/bmj.h1258

Oakley, A., Strange, V., Bonell, C., Allen, E. & Stephenson, J. (2006) Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal*, 332, 413–415.

OFFA (Office for Fair Access) (2004) *Producing Access Agreements: OFFA Guidance to Institutions*. Bristol: OFFA.

OFFA (Office for Fair Access) (2013) *How to produce an access agreement for 2014–15*. Bristol: OFFA.

OfS (Office for Students) (2019) *Using standards of evidence to evaluation impact of outreach*. Bristol: OfS.

Passy, R., Morris, M. & Waldman, J. (2009) *Evaluation of the Impact of Aimigher and widening participation outreach programmes on learner attainment and progression – Interim Report*. Slough: NFER (National Foundation for Educational Research).

Passy, R. & Morris, M. (2010) *Evaluation of Aimhigher: Learner Attainment and Progression – Final Report*, Slough: NfER.

Pawson, R. (2008) Causality for beginners. In *NCRM Research Methods Festival 2008* (Unpublished).

Pawson, R. & Tilley, N. (1994) What works in evaluation research? *British Journal of Criminology* 34 pp.291–306.

Pawson, R. & Tilley, N. (1997) *Realistic evaluation*. London: Sage Publications.

Pickering, N. (2021) Enabling equality of access in higher education for underrepresented groups: a realist 'small step' approach to evaluating widening participation, *Research in post-compulsory education*, 26(1), pp. 111–130. doi:10.1080/13596748.2021.1873410.

Puttick, R. and Ludlow, J. (2012) *Standards of Evidence: An Approach that Balances the Need for Evidence with Innovation*, London: NESTA.

Robinson, D. & Salvestrini, V. (2020) *The impact of interventions for widening access to higher education: A review of the evidence*. TASO: London.

Rubin, D. B. (1977) Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1–26.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin and Company.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. & Befani, B. (2012) *Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development*. DFID: Department for International Development.

Suzuki, E. & VanderWeele, T. J. (2018) Mechanisms and uncertainty in randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*.

Torgerson, C., Gascoine, L., Heaps, C., Menzies, V. & Younger, K. (2014) *Higher Education access: Evidence of Effectiveness of university access strategies and approaches*: A report to the Sutton Trust. London: Sutton Trust.

White, H. (2009) Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, 1(3), 271–284.

White, H. (2013) The use of mixed methods in randomized control trials. *New Directions for Evaluation* (138), 61–73.

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

Wong, G., Westhorp, G., Pawson, R. & Greenhalgh, T. (2013) Realist Synthesis RAMESES Training Materials.

Younger, K., Gascoine, L., Menzies, V. & Torgerson, C. (2019) A systematic review of evidence on the effectiveness of interventions and strategies for widening participation in higher education. *Journal of Further and Higher Education*, 43, 742–773.

APPENDIX ONE: EVALUATION PARADIGMS

INTRODUCTION

An evaluation paradigm is a set of beliefs about the nature of the world and how to enquire into it (Guba & Lincoln, 1989). It contains various assumptions about the nature of the social world and the people in it and the role of social research in understanding and shaping that world. Each paradigm contains different ontological, epistemological and methodological positions:

- Questions about the form and nature of the social world are *ontological questions* and may take the form 'What is there that can be known?' or 'What is the nature of reality?' (Guba & Lincoln, 1989).
- Questions about the nature of knowledge and the relationship between the knower and what can be known are *epistemological questions* and may take the form 'How can we be sure that we know what we know?'
- Answers to ontological and epistemological questions raise questions about research strategy and practice known as *methodological questions* and may take the form 'How can we go about finding out things?' (Guba & Lincoln, 1989).

As Fox et al. (2018) explain, the different assumptions within different evaluation paradigms raise fundamental philosophical debates about the social world and the nature of social enquiry. Whether consciously or unconsciously, evaluators must take positions on the answers to these questions.

The positions scientists take in answering philosophical questions determine the questions they consider answerable by science and choose to address, as well as the methods they employ to answer them (Rosenberg, 2012, p. 3).

These issues are more acute for social scientists than in the field of natural sciences, where better-established frameworks exist for addressing these philosophical questions (Rosenberg, 2012).

Thus, it is important to be clear about the paradigm underpinning our approach to evaluation to ensure that the assumptions we make about the social world, the most appropriate way to evaluate it and how evaluation findings should be used are consistent. Guba and Lincoln warn of the dangers of mixing and matching paradigms in evaluation:

> *It is not appropriate to 'mix and match' paradigms in conducting an evaluation, for example, utilizing both scientific (positivist) and constructivist propositions within the same study. This is not a call for 'purity' nor is it intended to be exclusionary. It is simply a caveat that mixing paradigms may well result in nonsense approaches and conclusions.*

(Guba & Lincoln, 2001, p. 1)

The different impact evaluation methodologies discussed in this guide do not all belong in the same evaluation paradigm. This appendix describes three evaluation paradigms: the post-positivist paradigm, the scientific realist paradigm and the constructivist (fourth-generation evaluation) paradigm. These have been selected to represent a range of paradigms that suggest different understandings of and approaches to evaluation. RCTs and quasi-experiments are commonly understood to sit within the experimental paradigm. Realist evaluation and process tracing are commonly understood to sit in the scientific realist paradigm. MSC, it may be argued, sits within the constructivist paradigm. However, the philosophies underlying different methodologies are debated. For example, process tracing hypothesises causal mechanisms in order to arrive at causal explanations. Its mechanistic approach, and its assumption that explanation should combine social and institutional structure and context with individual agency and decision-making, means that – epistemologically and ontologically – it is closely related to scientific realism (Bennett & Checkel, 2015) and could be seen as a specific analytical process that fits within the broader scientific realist framework. However, process tracing could also align with approaches to evaluation grounded in other epistemologies and ontologies, such as pragmatism or constructivism (Bennett & Checkel, 2015).

POST-POSITIVISM

Post-positivism has a long history. While few, if any, evaluators would class themselves as positivists, as the key tenets of positivism have long since been discredited, many evaluators and much-published evaluation guidance can still broadly be placed in the post-positivist tradition. The term positivist is still loosely used to (inaccurately) describe evaluators who favour quantitative methods and experimental designs (Shadish et al., 2002).

The (quasi-) experimental approach to evaluation described in Section 2.3.1 is situated within the post-positivist paradigm. It is a scientific approach to evaluation that assumes there is an objective 'truth' external to the evaluator. However, although post-positivists pursue a degree of objectivity, they believe that knowledge is conjectural. If a scientific law inferred from observable facts is to be justified, then the number of observations forming the basis of the generalisation must be large, the observations must be repeated under a wide variety of conditions, and no observation should conflict with the derived law (Chalmers, 2013). An experiment or quasi-experiment provides the conditions for such a procedure. Interventions are captured as a series of variables and controls are applied, until all explanatory factors apart from the influence of the programme being evaluated are controlled for. However, even if these conditions are met, it is difficult to be sure that a general conclusion is correct when it is based on a limited number of observations. There are two main problems (Chalmers, 2013). The first is the issue of specifying what an adequate inductive argument is, e.g. how many observations are sufficient and what constitutes a wide variety of conditions. The second problem is the circularity involved in justifying the concept of induction – the so-called 'riddle of induction', whereby to justify the concept of induction we must attribute it to an inductive argument (the test of scientific knowledge).

Post-positivists often draw on the work of Popper who, in order to distinguish genuinely scientific theories from theories such as those developed by Marxists and Freudians, argued that scientific theories are falsifiable. Whereas the confirmation is logically flawed (e.g. all the swans I have seen are white, therefore all swans are white) the falsification is logically sound (e.g. one observation of a black swan is sufficient to falsify the statement 'all swans are white'). For Popper, scientific theories are speculative and tentative conjectures that should be rigorously and ruthlessly tested with the aim of refuting them (Chalmers, 2013). Hypotheses must be falsifiable. Theories that fail to stand up to testing must be eliminated and replaced with further conjectures. The principle of induction is not involved. Thus, 'science progresses by trial and error, by conjectures and refutations' (Chalmers, 2013, p. 56, echoing Popper).

(Quasi-) experimental evaluation is falsifiable because it requires experimenters to identify a causal claim and then generate and examine plausible alternative explanations that may falsify it (Shadish et al., 2002). More formally, evaluators attempt to falsify the null hypothesis that there will be no treatment effect. Thus, Cook and Campbell (1979, p. 94) note that 'casual inferences will never be proved with certainty since the inferences we make depend upon many assumptions that cannot be directly verified'.

Fox et al. (2018) describe the key beliefs of post-positivists. Post-positivists answer the ontological question 'What is the nature of reality?' by asserting that an objective reality exists that 'goes about its business irrespective of the interest that an enquirer may have in it' (Guba & Lincoln, 1989, p. 85). This is a realist ontology. Positivists believe that science uncovers reality (naïve realism); post-positivists recognise that this is not possible and adopt a critical realist position in which causes have a real nature, albeit one that can only be 'imperfectly grasped' (Cook & Campbell, 1979, p. 30).

The epistemological question is answered by maintaining that an evaluator can take an objective position with respect to the subject being evaluated (Guba & Lincoln, 1989). Evaluation is thus value-free.

For positivists, scientific knowledge is derived from facts. More specifically, scientific knowledge proceeds from statements about some events to statements about all events of a particular kind (Chalmers, 2013).

The realist perspective assumes that the world operates according to immutable, natural laws, many of which take the form of cause-effect relationships. This is a deterministic view of the social world and assumes that the approaches used in the natural sciences can be applied to the social world. However, post-positivists also recognise that studying the social world poses additional challenges. Thus, they advocate using methods (experiments and quantitative science) that 'are shared in part with the physical sciences' (Cook & Campbell 1979, p. 92).

Post-positivist evaluators who rely on the (quasi-) experimental model of evaluation are tied to a successionist theory of causation. This means that although 'causation' cannot be observed, it can be inferred from the repeated succession of one event by another (Pawson & Tilley, 1997). The existence of natural laws and a methodology (experimentation) that allows science to identify them enables scientists to predict and control social phenomena. For Campbell (1969), this leads to a vision of an experimenting society, where repeated experiments allow for the development of better social policy.

Methodology is structured to discover or test causal mechanisms. We cannot observe causality, but we can demonstrate regularity between a particular intervention and a particular outcome (Marchal et al., 2012). Because causation cannot be observed, RCTs attempt to attribute the observed outcome to the intervention (Marchal et al., 2012). This implies removing from the context any possible contaminating influences (confounding variables) either using physical controls, as in a laboratory or a randomised field experiment, or through statistical controls (Guba & Lincoln, 1989). This is an interventionist methodology.

SOME CRITICISMS OF POST-POSITIVISM

Debate about the merits and flaws of positivist – and, subsequently, post-positivist – positions has continued for over a hundred years. The main critiques are outlined below (Fox et al., 2018).

An interventionist methodology places an emphasis on control, but the result of this is to strip out context (Guba & Lincoln, 1989). The problem is that (quasi-) experimental evaluation designs reduce initiatives to a series of variables and apply controls until we can be sure that all explanatory factors, save for the influence of the programme itself, are 'squeezed out' (Pawson & Tilley 1994, p. 294). For Pawson and Tilley, this seemingly 'scientific' and 'objective' approach to evaluation hides a particular set of overly deterministic (mis) understandings about what interventions and programmes are and how they function, with the result that:

> *For us, the experimental paradigm constitutes a heroic failure, promising so much and yet ending up in ironic anticlimax. The underlying logic ... seems meticulous, clear-headed and militarily precise, and yet findings seem to emerge in a typically non-cumulative, low impact, prone-to-equivocation sort of way.*

(Pawson & Tilley 1997, p. 8)

For Guba and Lincoln, the result of context-stripping is evaluations that are 'often found to be irrelevant at the local level' (Guba & Lincoln, 1989, p. 37).

A post-positivist approach assumes that the evaluator is objective and independent of the subject of the evaluation; however, the subject of evaluations is people, and they are not inert, passive 'objects' of the kind studied in the natural sciences. Their realities are socially constructed and they also have values. Moreover, evaluators are also human beings and, whether it is acknowledged or not, evaluation involves reactivity between evaluators and their respondents. 'Facts' are socially constructed and, thus, evaluation is not value-neutral (Guba & Lincoln, 1989):

> *If science is not value-free, then it is the case not only that findings are subject to different interpretations but that the 'facts' themselves are determined in interaction with the value system the evaluator (probably unknowingly) brings to bear. Then every act of evaluation becomes a political act.*

(Guba & Lincoln, 1989, p. 35)

A commitment to objectivity and control leads to an over-reliance on quantitative methods and:

> *after a time these measuring instruments take on a life of their own; while initially intended as 'operationalizations' of scientific variables, they become, in the end, the variables themselves. It follows that what cannot be measured cannot be real.*

(Guba & Lincoln, 1989, p. 37)

CONSTRUCTIVISM (FOURTH-GENERATION EVALUATION)

Constructivists believe that multiple, socially constructed realities are ungoverned by natural laws (Guba & Lincoln, 1989). This is a relativist ontology. Constructions are devised by individuals and can be –and usually are – shared, and, over time, may become more sophisticated. This does not, however, make them more 'real', but simply more commonly used (Guba & Lincoln, 1989). Some constructions may also have law-like attributions, but 'there is a world of difference between believing that there is some law that one has "discovered" in nature versus believing that it may be useful for a variety of purposes to think in law-like terms' (Guba & Lincoln, 1989, p. 86).

Most Significant Change can be framed within the constructivist paradigm. Thus, accounts of significant change are constructed through unique meaning-making processes (that is, the storyteller's interaction with the world and the interpretation drawn from such interaction) and then reconstructed by reviewers who engage with those accounts and draw further new meanings. However, Davies and Dart (2005) note that MSC involves verification stages in which the credibility of accounts are checked; therefore, it may best be described as employing a constructivist epistemology and a realist ontology.

If the world is socially constructed, from an epistemological perspective, objectivity makes no sense, because it is impossible to separate the evaluator from that which is evaluated (Guba & Lincoln, 1989). Evaluation is thus value-laden, and knowledge emerges from the interaction between evaluators and the evaluated: 'Inquirers are human and cannot escape their humanness' (Guba & Lincoln, 1989, p. 88).

Methodology is constructed to 'expose the constructions of the variety of concerned parties, open each to critique in the terms of other constructions, and provide the opportunity for revised or entirely new constructions to emerge' (Guba & Lincoln, 1989, p. 89). This is a dialectic hermeneutic methodology (Guba & Lincoln, 1989). 'Dialectic' implies a process of debate used to examine and discuss opposing ideas in order to reach an agreed position, while 'hermeneutic' implies an interpretive process. It requires that evaluations be undertaken in a natural setting (as opposed to a laboratory) and implies that the methodology will be non-linear (Guba & Lincoln, 1989). In a constructivist methodology, evaluation questions cannot be predetermined, but must arise out of the dialectic hermeneutic process. As a result, a highly adaptable research instrument is required that can determine what is salient to the evaluation and what is not and, for Guba and Lincoln (1989), this means that a human is the instrument of choice for the constructivist. This, in turn, has implications for the choice of research methods:

> *... given that the human instrument is to be employed, the question of which methods to use is easily answered: those that come most readily to hand for a human. Such methods are, clearly, qualitative methods.*

(Guba & Lincoln, 1989, p. 175)

However, there is no reason why quantitative methods cannot also feature in a constructivist evaluation (Guba & Lincoln, 1989):

> *There is nothing in this formulation that militates against the use of quantitative methods; the constructivist is obviously free to use such methods without prejudice when it is appropriate to do so.*

(Guba & Lincoln, 1989, p. 176)

SOME CRITICISMS OF RESPONSIVE CONSTRUCTIVISM

The dialectic hermeneutic approach represents a radical departure for evaluation, and not everyone finds it convincing, because it is difficult to envisage in practice, particularly in scenarios with power differences between stakeholders. Pawson and Tilley argue that there is a:

> *... deep-seated air of unreality about the evaluations-as-negotiations perspective, namely its failure to appreciate the asymmetries of power which are assumed in and left untouched by the vast majority of policy initiatives.*

(Pawson & Tilley, 1997, p. 20)

One of the main criticisms of the constructivist approach is its relativism. If the world is socially constructed, then evaluations are also socially constructed and so have no claim to truth:

> *Evaluation data derived from constructivist inquiry have neither special status nor legitimation; they represent simply another construction to be taken into account in the move toward consensus.*

(Guba & Lincoln, 1989, p. 45)

The implication of this is spelled out by Guba and Lincoln, who recognise the challenges that this position poses for many evaluators:

> *... there is no objective truth on which inquiries can converge. One cannot find out how things really are or how they really work. That level of ambiguity is almost too much to tolerate. If evaluations cannot ferret out the truth, what use can there be in doing them?*

(Guba & Lincoln 1989, p. 46)

Pawson and Tilley (1997) argue that this extreme relativism is flawed in its understanding of the social world. For them, the social world, including policies and programmes, consists of more than the sum of people's beliefs, hopes and expectations. Echoing Giddens' (1984) structuration theory (a duality of structure and agency), they argue that the social world features structures and institutions which are in some respects independent of people's reasoning and desires.

SCIENTIFIC REALISM

Scientific realist evaluation sits somewhere between (post) positivism – characterised by Wong et al. as 'there is a real world which we can apprehend directly through observation' – and constructivism, characterised as 'given that all we can know has been interpreted through human senses and the human brain, we cannot know for sure what the nature of reality is' (Wong et al., 2013, p. 2). Scientific realists believe that 'there is a [social] reality that cannot be measured directly (because it is processed through our brains, language, culture and so on), but can be known indirectly' (Wong et al., 2013, p.2). Thus, scientific realism retains post-positivism's commitment to scientific rigour and experimentation but rejects its reliance on experimental evaluation designs based on an intervention and control group, arguing that this form of control 'squeezes out' what is important to our understanding of why social programmes work. From constructivism, scientific realism retains an emphasis on understanding the social context of interventions but rejects relativism.

Pawson and Tilley's starting point in their elaboration of scientific realism is to argue that the post-positivist 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 & Tilley, 1997, p. 11). They assume a different, 'realist' model of explanation in which 'causal outcomes follow from mechanisms acting in contexts' (Pawson & Tilley 1997, p. 58). Thus, the basic task of social enquiry:

> *...is to explain interesting, puzzling, socially significant regularities... Explanation takes the form of positing some underlying mechanism... which generates the regularity and thus consists of propositions about how the interplay between structure and agency has constituted the regularity. Within realist investigation there is also investigation of how the workings of such mechanisms are contingent and conditional, and thus only fired in particular local, historical or institutional contexts...*

(Pawson & Tilley, 1997, p. 71)

For Pawson and Tilley (1994, 1997), post-positivists misunderstand what makes programmes work: 'Programmes cannot be considered as some kind of external, impinging "force" to which subjects "respond"' (Pawson & Tilley, 1994, p. 294). Rather, social programmes are social systems involving an interplay between individual and institution or, in the language of Giddens (1984), agency and structure. Therefore, it is not that programmes work but, rather, that people co-operate and choose to make them work. Scientific realists do not, however, adopt the same formulation as constructivists. Instead, they see people's choices as constrained by social structures:

> *[P]rogrammes 'work', if subjects choose to make them work and are placed in the right conditions to enable them to do so. This process of 'constrained choice' is at the heart of social and individual change to which all programmes aspire ...*

(Pawson & Tilley 1994, p. 294)

A substantial part of Pawson and Tilley's key texts (1994, 1997), in which they set out the case for scientific realist evaluation, is given over to a discussion of causation. For Pawson and Tilley (1997), the model of causation adopted in the (quasi-) experimental evaluation design favoured by post-posivitists is external, successionist causation. This is the idea that causation itself is unobservable but that it can be inferred on the basis of observation. Scientific realists prefer a model of generative causation that sees causation acting internally as well as externally. Pawson and Tilley argue that, in reality, most experiments in the natural sciences do assume generative causation:

> *Do we understand the action of gravity on a falling body by observing the motion of a cannon ball dropped from a leaning tower and comparing it with the motion of one that remains atop?*

(Pawson & Tilley, 1997, p. 57)

They argue that the main exception to this is medical research, where experimental and control group evaluation based on successionist causation is the dominant model.

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 (Marchal et al., 2012). The best they can offer is plausible explanations.

Scientific realist evaluation is theory-led. However, this does not mean 'grand' sociological theories that explain society as a whole (Pawson & Tilley, 1997). Rather, it is what we might call mid-level theory, which explains in detail how a particular intervention works in a particular context. In other words, the theory that drives scientific realism is CMO configurations.

To be clear, Pawson and Tilley argue in favour of an experimental method. However, they reject the model of an experiment based on similar intervention and control groups. They argue instead, following philosophers such as Bhaskar, that the two essential elements of an experiment are triggering the mechanism being studied to ensure 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.

CRITIQUE OF SCIENTIFIC REALISM

Several questions have been asked of scientific realist evaluation.

Some commentators question whether scientific realist evaluation really represents a new paradigm for evaluation. Is the greater emphasis on theory testing a new paradigm or a refinement of existing post-positivist positions? Many advocates of (quasi-)experiments have argued that theory-based evaluation should be used together with experimental designs rather than in opposition to them (see, for example, Bonell et al., 2012). Pawson and Tilley argue to the contrary:

> *It is not a matter of the more 'sensitive experimentalist' being able to incorporate such features [as context-mechanism-outcome configurations] within yet more complex designs.*

(Pawson & Tilley 1997, p. 54)

Conversely, is scientific realism really a reworking of constructivism? The challenge for Pawson and Tilley is that they attempt to argue that greater attention should be paid to actors' purposes and understandings (realism), while also arguing that evaluation can retain a degree of objectivity and certainty (scientific). However, the 'ontological' shift that this requires towards more 'interpretive' approaches to social science makes it hard to see how objectivity can be maintained. For example, Farrington, a psychologist and proponent of the greater use of RCTs in the social sciences, particularly criminology, argues that:

> *Pawson and Tilley suggest that mechanisms essentially provide reasons and resources (the will?) to change behavior… [I]t is not clear how reasons could be investigated. Many psychologists are reluctant to ask people to give reasons for their behavior, because of the widespread belief that people have little or no introspective access to their complex mental processes… Hence, it is not clear that reasons in particular and verbal reports in general have any validity, which is why psychologists emphasize observation, experiments, validity checks, causes and the scientific study of behavior.*

(Farrington, 1998, p. 207)

IMPLICATIONS FOR EVALUATION

Fox et al. (2018) demonstrate how the different philosophies underpinning the three paradigms we have examined help to demonstrate Rosenberg's argument (see above) that the positions we take on philosophical questions determine the questions that we, as evaluators, can answer and the methods that we employ to answer them. It is also worth reiterating that this is not a debate about the relative merits of quantitative and qualitative research. Each of the paradigms we have reviewed favours either quantitative or qualitative research but also allows a role for the other.

The importance of and differences between evaluation paradigms are particularly clear when we consider their implications both for the role of the evaluator and for evidence-based policy.

Whereas in the post-positivist paradigm, evaluators are 'the communication channels through which literally true data are passed to the audience of evaluation reports' (Guba & Lincoln, 1989, p. 110), the role of the evaluator is more complicated in the two other paradigms we have considered. In the constructivist paradigm, evaluators are 'orchestrators of a negotiation process that aims to culminate in consensus on better informed and more sophisticated constructions [of the social world]' (Guba & Lincoln, 1989, p. 175). In the scientific realist paradigm, the evaluator is a researcher and theorist with a detailed understanding of social programmes, able to construct mid-level theories (groups of CMO configurations) for subsequent testing.

The post-positivist evaluator is an objective 'scientist' whose data and findings have a special status. However, at the other end of the spectrum in the relativistic constructivist paradigm, neither evaluation data and analysis nor evaluators have any special status. Evaluation data is simply another social construction and evaluators themselves are simply one group of social actors amongst many stakeholders (Guba & Lincoln, 1989).

The different paradigms suggest different expectations around the status and use of evaluation findings. A post-positivist approach sees great potential for social science research to inform policy. So, for example, Campbell (1969) – a key proponent of the experimental and quasi-experimental approach to evaluation – envisaged an 'experimenting society' in which policy and programmes emerge from continual social experimentation.

By contrast, scientific realists are more circumspect. Drawing on the ideas of Karl Popper, scientific realists argue that:

> *Realist evaluation begins with theory and ends with further theory... The grand evaluation payoff is thus nothing other than improved theory, which can then be subjected to further testing and refinement, through implementation in the next program. And so the cycle continues.*

(Pawson & Tilley, 1998, pp. 89–90)

This is a more incremental approach to improving the evidence underpinning policy, and the emphasis that scientific realists place on context suggests natural limits to the potential for generalisation and replication.

A responsive constructivist approach offers policymakers even less certainty. In the form of relativism envisaged by Guba and Lincoln, there is no basis from which to generalise evaluation findings from one context to another. However, for Guba and Lincoln, this is not problematic. They argue that this fear of the loss of absolutes is itself only a construction within which 'scientific' evaluators are trapped. Instead, they argue that 'it is precisely because of our preoccupation with finding universal solutions that we fail to see how to devise solutions with local meaning and utility' (Guba & Lincoln, 1989, p. 47).

REFERENCES

Bennett, A. & Checkel, J. (2015) Process tracing: from philosophical roots to best practice. In Bennett, A. & Checkel, J. (Eds.) *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. & Moore, L. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine* 75, 2299–2306.

Campbell, D. T. (1969) Reforms as Experiments. *American Psychologist* 24, 409–29.

Chalmers, A. (2013) *What is This Thing Called Science?* Maidenhead: OUP.

Cook, T. & Campbell, D. (1979) *Quasi-experimental Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.

Davies, R. & Dart, J. (2005) *The 'most significant change' (MSC) technique. A guide to its use*. PDF available at: https://www.researchgate.net/publication/275409002_The_'Most_Significant_Change'_MSC_Technique_A_Guide_to_Its_Use

Farrington, D. (1998) Evaluating 'communities that care': Realistic scientific considerations. *Evaluation* 4 pp. 204–10.

Fox, C., Grimm, R. & Caldeira, R. (2016) *An Introduction to Evaluation*, London: Sage.

Giddens, A. (1984) *The Constitution of Society. Cambridge*: Polity Press.

Guba, E. G. & Lincoln Y. S. (1989) *Fourth Generation Evaluation*, London: Sage.

Guba, E. G. & Lincoln, Y. S. (2001) *Guidelines and Checklist for Constructivist (a.k.a Fourth Generation)* evaluation.

Marchal, B., Belle, S., Olmen, J., Hoerée, T. and Kegels, G. (2012) 'Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research' *Evaluation* 18(2) pp.192–212.

Pawson, R. & Tilley N. (1994) What works in evaluation research? *British Journal of Criminology* 34, 291–306.

Pawson, R. & Tilley N. (1997) *Realistic evaluation*. London: Sage.

Pawson, R. & Tilley, N. (1998) Caring communities, paradigm polemics, design debates. *Evaluation* 4 73–90.

Rosenberg, A. (2012) Philosophy of Social Science. Boulder: Westview.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Wong, G., Westhorp, G., Pawson, R. & Greenhalgh, T. (2013) Realist Synthesis RAMESES Training Materials.

TASO

Transforming Access
and Student Outcomes
in Higher Education

Evidence Quarter
Floor 4, Albany House
94-96 Petty France
London SW1H 9EA

info@taso.org.uk
taso.org.uk

TASO is an independent charity that aims to improve lives through evidence-based practice in higher education (HE). We support HE professionals through research, toolkits and evaluation guidance on what works best to eliminate equality gaps. We inform practitioners of the best available evidence and produce new evidence on the most effective approaches.
TASO is an affiliate 'What Works' centre and is part of the UK Government's What Works Movement.