

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework

Howard White and Daniel Phillips

June 2012



About 3ie

The International Initiative for Impact Evaluation (3ie) works to improve the lives of people in the developing world by supporting the production and use of evidence on what works, when, why and for how much. 3ie is a new initiative which responds to demands for better evidence, and will enhance development effectiveness by promoting better-informed policies. 3ie finances high-quality impact evaluations and campaigns to inform better programme and policy design in developing countries.

The **3ie Working Paper series** covers both conceptual issues related to impact evaluation and findings from specific studies or systematic reviews. The views in the papers are those of the authors, and cannot be taken to represent the views of 3ie, its members, or any of its funders.

This Working Paper was written by Howard White and Daniel Phillips, 3ie, and edited by Catherine Robinson. Production and cover design by Radhika Menon and Mukul Soni, 3ie.

Acknowledgements

The authors would like to thank members of the cross-college working group on impact evaluation, especially Clare Chandler, who offered advice and useful information which helped in the paper's development. We would like to thank Fred Carden, Rick Davies, Katherine Hay, Lina Payne, Wendy Olsen, John Mayne, Bruno Marchal, Robert Brinkerhoff, Julia Betts, Karl Hughes, Kimberly Bowman and Eric Djimeu for constructive criticism, ideas and information that have led to revisions of the paper. We are also grateful to colleagues from 3ie whose ideas and information have contributed to our work.

Photograph: Dana Smillie / World Bank

© 3ie, 2012

Contacts

International Initiative for Impact Evaluation
c/o Global Development Network
Post Box No. 7510
Vasant Kunj P.O. New Delhi – 110070, India
Tel: +91-11-2613-9494/6885
www.3ieimpact.org

Contents

Abstract	3
1. Introduction	4
2. Defining small <i>n</i>	5
3. Evaluation approaches for small <i>n</i> impact analysis	7
3.1 Group I approaches	7
3.2 Group II approaches	13
3.3 Small <i>n</i> attribution.....	17
4. Tackling bias	19
4.1 Respondent biases.....	20
4.2 Evaluator biases	22
4.3 Addressing bias.....	23
5. Towards an integrated framework: common steps for causal inference in small <i>n</i> cases	24
6. Conclusion	28
7. Bibliography	29
Annex 1: Approaches for small <i>n</i> analysis: extended summaries	35
1.1 Realist Evaluation	35
1.2 General Elimination Methodology (GEM)	38
1.3 Process Tracing	40
1.4 Contribution Analysis	42
1.5 Most Significant Change (MSC)	45
1.6 Success Case Method (SCM)	49
1.7 Outcome Mapping (OM)	51
1.8 Method for impact Assessment of Programs and Projects (MAPP)	54
1.9 Qualitative Comparative Analysis (QCA)	56
Annex 2: Small <i>n</i> approaches: comparison tables	61

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework

Howard White

International Initiative for Impact Evaluation (3ie)

Daniel Phillips

International Initiative for Impact Evaluation (3ie)

Email: dphillips@3ieimpact.org

Abstract

With the results agenda in the ascendancy in the development community, there is an increasing need to demonstrate that development spending makes a difference, that it has an impact. This requirement to demonstrate results has fuelled an increase in the demand for, and production of, impact evaluations. There exists considerable consensus among impact evaluators conducting large n impact evaluations involving tests of statistical difference in outcomes between the treatment group and a properly constructed comparison group. However, no such consensus exists when it comes to assessing attribution in small n cases, i.e. when there are too few units of assignment to permit tests of statistical difference in outcomes between the treatment group and a properly constructed comparison group.

We examine various evaluation approaches that could potentially be suitable for small n analysis and find that a number of them share a methodological core which could provide a basis for consensus. This common core involves the specification of a theory of change together with a number of further alternative causal hypotheses. Causation is established beyond reasonable doubt by collecting evidence to validate, invalidate, or revise the hypothesised explanations, with the goal of rigorously evidencing the links in the actual causal chain.

We argue that, properly applied, approaches which undertake these steps can be used to address attribution of cause and effect. However, we also find that more needs to be done to ensure that small n evaluations minimise the biases which are likely to arise from the collection, analysis and reporting of qualitative data. Drawing on insights from the field of cognitive psychology, we argue that there is scope for considerable bias, both in the way in which respondents report causal relationships, and in the way in which evaluators gather and present data; this points to the need to incorporate explicit and systematic approaches to qualitative data collection and analysis as part of any small n evaluation.

1. Introduction

The results agenda is in the ascendancy in the development community. Multilateral agencies such as the World Bank are placing increasing emphasis on the importance of addressing development effectiveness. Many countries in Latin America are following the example of Mexico and Colombia (Briceño and Gaarder, 2010) by setting up institutions and developing legislation which require social programmes to demonstrate impact in order to ensure continued public funding. South Africa and Uganda have adopted nationwide policies for monitoring and evaluation as part of a focus on results; in South Africa's case, the policies are organised around 12 key development outcomes adopted by the Cabinet, which also approved a new Evaluation Framework in late 2011. India has established an Independent Evaluation Office. The agenda for the Fourth High Level Forum on Aid Effectiveness in Busan in November 2011 focused on results and development effectiveness.

Effective development means that development spending makes a difference, that it has an impact. The need to demonstrate results has fuelled an increase in the demand for, and production of, impact evaluations, defined here as studies that address the question of attribution. In other words, we are concerned with studies which lie along the central “research” branch of the “Evaluation Tree” (Alkin and Christie, 2004: 12–14).

This increased demand for results has resulted in a growth in large n impact evaluations involving tests of statistical significance between outcomes for treatment and comparison groups. However, this type of statistical analysis is not suitable for all evaluations. When there are insufficient units of assignment to conduct tests of statistical significance, then small n approaches are the best available option. However, while there is considerable consensus among large n researchers concerning what constitutes a valid approach, the same is not true for small n approaches. In fact, there is a plethora of possible approaches available to an evaluator, but little consensus concerning what approach, methodology, or methods are most suitable for an impact evaluation, and what constitutes valid causal evidence. This paper seeks to advance this discussion.

We begin Section 2 by sketching out some terminological issues, defining small n . Section 3 reviews a number of evaluation approaches suitable for small n analysis, assessing how they go about addressing attribution, and outlining the underlying conceptual frameworks and methods which provide a basis for making causal inferences. Throughout, we use the term “approach” to indicate the underlying logic by which an evaluation addresses attribution and causation, “methods” to refer to the tools or techniques which can be utilised in support of an evaluation approach, and “methodology” to indicate a procedure or system by which evaluation methods are organised.

For the evaluation approaches that we examine, we find that conceptual and methodological steps are clearly mapped out, and that for several of the approaches there is a methodological core in common. However, we also conclude that while the logic underlying the methodologies is usually well developed, less has been done to set out how evaluation methods could be systematically applied to promote the validity of conclusions. With this in mind, Section 4 assesses some of the potential biases that need

to be considered when carrying out this type of analysis. Section 5 sets out a suggested methodological framework for small n analysis, drawing both on the approaches examined earlier and on the wider literature, and goes on to discuss some ways of overcoming biases. We set out our conclusions in Section 6.

2. Defining small n

The need for small n approaches arises when data are available for only one or a few units of assignment, with the result that experiments or quasi-experiments in which tests of statistical differences in outcomes between treatment and comparison groups are not possible. A large n evaluation such as Miguel and Kremer's examination of de-worming for school children in Kenya (2004) was able to compare groups of students receiving treatment with a control group of students who did not. However, for a small n evaluation which, for example, looks at national policy change or a capacity-building intervention in a single organisation, it may not be possible to construct a comparison group. Here, the number of units of assignment, or " n ", refers to the sample size available for an evaluation. This is distinct from the " N ", or total population from which that sample is drawn.

Statistical tests of significance between treatment and comparison groups may not be possible in a number of circumstances. Small n approaches are the best available option in the following situations:

(1) When there is a small N ; a small N will mean that there is also a small n , as sample size can never be larger than population size. One example where $N=1$ would be capacity building in a single organisation, or advocacy campaigns at the policy level.

(2) When there is sufficient heterogeneity affecting at least one of the treatment population, the wider context of the intervention, or the treatment itself. Such heterogeneity would mean that the average treatment effect using a large n approach is of no possible value, as the different sub-groups in which we have an interest will be too small for statistical analysis. Examples of such heterogeneity may result from the way in which an evaluation is designed, as in the case of the Bangladesh Integrated Nutrition Project. For this project, data were collected from six different project sub-districts. In one of these areas, nutritional counselling was provided to couples, while in the other five it was provided only to women. This meant that it was not possible to test if the "couples" approach was more effective, as it was implemented in only one sub-district, with the result that area effects could not be separated from any differential programme effects. In principle, a small n approach might have been used to assess whether the "couples" approach helped to overcome the intra-household constraints that women faced in adopting the nutritional knowledge offered by the project. For a more extensive discussion of this project, see World Bank (2005), White and Masset (2006), and White (2005).

(3) If a treatment is national in scope, then N could be considered to be equal to one. Thus, a small n approach may be needed in order to tackle macro issues such as exchange-rate or tariff reform. However, for many national policies (for example, with a national health-insurance scheme) an "encouragement" design can be used to generate a control group, so that a large n approach is possible (Gertler et al., 2010: 69–79).

(4) When budgetary or political constraints prevent a sufficient sample size or use of a comparison group. In such cases, a small n approach may be adopted, but it is also worth exploring if additional resources can be raised for a large n approach.

In addition to those outlined above, a number of other situations have been put forward as being unsuitable for statistical tests of significance and therefore necessarily requiring a small n approach. However, this is not in fact the case. For example, if no untreated comparison group can be used for ethical or political reasons, either an encouragement design may again be used or a “comparison” can be made, using the current standard treatment or some variation thereof. In either case there is no need for an “untreated comparison group”, which helps to overcome ethical objections. Large n designs are also sometimes deemed inadequate for tackling “complex” interventions involving multiple agencies or simultaneous causal strands or where causality is, for example, recursive or outcomes emergent. This argument, however, fails to take into account the strength of a good theory-based impact evaluation combining both quantitative and qualitative methods to unpack the causal chain and uncover how complicated or complex social processes mediate impact (White, 2009). Furthermore, where interventions are complex it can also be an advantage to adopt a “black box” design which has the advantage of cutting through these problems and delivering a verdict on impact. For example, rural electrification is claimed to have extensive benefits for learning outcomes, fertility, income levels and so on (Barnes, 2007; World Bank, 2008). The causal chain behind some of these outcomes is indeed quite complex, but it can still be useful to know if there is an overall impact, even if we do not fully understand why that is so.

Large n should not be taken to simply equate to a quantitative approach and small n to equal a qualitative one. There are well-established small n quantitative approaches, especially in the field of economics. For example, national-level policies have often been analysed using a variety of modelling approaches, notably computable general equilibrium models (Bourguignon and Pereira da Silva, 2003; Haddad et al., 2011). Conversely, small n approaches may well be applicable to circumstances more normally associated with large n evaluations. For example, the work of one of this paper’s authors examining the macroeconomic impact of development aid rejected a large n approach involving cross-country regressions in favour of a small n case-study oriented approach which emphasised the need to set out the causal chain and construct counterfactual national and government accounts (White, 1996 and 1998, and White and Dijkstra, 2002).

This paper does not consider small n modelling-based approaches, focusing exclusively on mostly qualitative approaches for small n impact evaluation. This focus is justified by the lack of clarity about what constitutes best practice when tackling attribution. Evaluators undertaking large n studies and small n modelling-based approaches are broadly agreed on what a well-designed study looks like. The same is not as true for the qualitative analysis which is a major part of many small n approaches. For the most part, the small n approaches examined here make use of both qualitative and quantitative methods, although they rely far more extensively on the former. As Miles (1979) pointed out some years ago, “For quantitative data, there are clear conventions the researcher can use. But the analyst faced with a bank of qualitative data has very few guidelines to guard against self-delusion, let alone the presentation of unreliable or invalid conclusions to scientific or policy-making audiences.”

If development practice is to be guided by evidence of what works, it is important for development organisations to have a number of different tools capable of addressing questions of causal inference in a variety of settings. The initial motivation for this paper was to investigate whether the variety of small n approaches currently in use share a common core which could form the basis for an integrated framework for small n analysis. A number of such approaches are laid out in the next section.

3. Evaluation approaches for small n impact analysis

There are a wide variety of evaluation approaches suitable for small n analysis currently in use. Those examined here were selected purely on the basis that they are well known and have been used or proposed for use in the field of development evaluation. The list is by no means intended to be exhaustive. Our focus is also on approaches suitable for carrying out impact evaluations for just one or a few case studies, and for this reason we do not include approaches such as Qualitative Comparative Analysis (QCA) in our discussion.¹

The approaches that we do assess sometimes stem from quite different philosophical paradigms and are designed to fulfil a number of different monitoring and evaluation functions. The discussion here is not intended to be an assessment of their overall value for monitoring and evaluation. Instead, the focus in the following section is on the utility of these approaches as tools for carrying out *impact evaluations* geared towards making causal inferences. By making this comparison, the intention is to show that there are some important steps that are common to many of these approaches.

In the following section the approaches are divided into two groups, with each group examined in turn, and a brief summary of each approach provided together with case-study examples to illustrate how they have been applied. More extensive summaries of each approach are set out in *Annex 1*, together with relevant further reading. *Annex 2* contains tables which compare the general characteristics and methodological steps of each of the approaches examined here. After each group of approaches has been outlined, we discuss some of the parallels in the methodological steps that they set out. We conclude the section by discussing the causal logic which the small n approaches outlined here employ to tackle attribution.

3.1 Group I approaches

These approaches explicitly set out to discover the causes of observed effects with the goal of establishing beyond reasonable doubt how an outcome or set of outcomes occurred. Most of these approaches emphasise the need to draw on the implicit theory of change lying behind an intervention and to map out steps by which an evaluator can assess whether the theoretically predicted changes occurred as expected, or whether the

¹ Qualitative Comparative Analysis (QCA) is an evaluation approach widely adapted for causal analysis, usually for a small or medium number of cases. The consensus position is that QCA requires more than just a few cases if it is to reach causal conclusions (see, for example, Rihoux and Ragin, 2009). A brief summary of QCA is included in *Annex 1* along with a list of suggested further reading.

causes and assumptions set out in the theory of change varied or whether the observed outcomes were a result, in part or whole, of external factors.

***Realist Evaluation*²**

Realist evaluation (Pawson and Tilley, 1997) differs from the other approaches examined here in its fundamental identification with the realist paradigm. Realism is a school of philosophical thought which holds that there is a “real” world which exists independent of (or interdependent with) our interpretations of it, with the result that it is possible to work towards a closer understanding of how programmes can cause change. According to a realist perspective, programmes can be seen as theories incarnate; when a programme is implemented, it is testing a theory about what actions can help to bring about change (Westhorp et al., 2011). In practice, this means that a realist evaluation sets out to test a Middle Range Theory (MRT) detailing how the mechanisms initiated by a programme should cause desired outcomes. Programmes themselves are viewed as being akin to open systems in which there are always multiple and competing mechanisms which interact with the surrounding context to produce outcomes. Pawson and Tilley (1997) sum this up as “mechanisms + context = outcomes”.

As all mechanisms interact with context, when replicated in new environments programmes cannot be expected to achieve the same outcomes. Realist evaluation is designed to address the question of causation and find out “what works, how, in which conditions and for whom” (Pawson and Tilley, 1997). In order to do this, realist evaluators are asked to consider how underlying mechanisms are likely to interact with historical and cultural context, location, economic and political structures, participants, and so on, to produce varying outcomes. The construction of the Middle Range Theory will already have drawn on background research to establish prevailing theory and relevant previous experiences. Evaluators now consider the nature of a planned programme, the target population, and the contexts and settings in which the programme will operate to map out a series of conjectural mini-theories called Context Mechanism Outcome (CMO) configurations which relate the various contexts of a programme to the multiple mechanisms by which it might function to produce various outcomes.

One realist evaluation (Marchal et al., 2010a) examined how the introduction of a “bundle” of combined management practices might contribute to improving the performance of an urban district hospital in Ho, Volta Region, in Ghana. Drawing on a literature review of health-care performance and past case studies in Ghana, the team outlined a Middle Range Theory (MRT). A number of conjectural Context Mechanism Outcome theories were then drawn up, setting out how mechanisms designed to promote High Commitment Management practices such as leadership style, hospital social structure, and organisational climate and culture might operate within the programme context to produce a range of possible outcomes.

² Realist Evaluation is more a logic of evaluation than an approach that has explicit methodological steps. Although Pawson and Tilley set out their core ideas in considerable detail, the absence of a step-by-step framework means that their ideas have been interpreted quite widely. This summary reflects our interpretation of Pawson and Tilley’s ideas and takes into account those of other authors who have drawn on Realist Evaluation, in particular those of Marchal et al. (2010a; 2010b, forthcoming).

Realist evaluation draws on data collection utilising both quantitative and qualitative data sources to build up a picture of the programme in action and identify how mechanisms are operating in reality in order to revise, substantiate, or invalidate conjectural CMOs. In the case of the Ho Hospital evaluation, data collection drew on a document review, in-depth interviews, group discussions, observations and a review of routine health information. NVIVO 7 software was used for qualitative data management and analysis, with initial coding organised around the MRT and the research questions. A second round of coding then built on this in order to organise the emerging themes and patterns produced. Evidence from the interviews and observations was used to differentiate between the vision (what the management team wanted to achieve), the discourse (what they were saying had been achieved) and the actual practices (what they were doing).

Realist evaluation draws on a generative notion of causation which involves an iterative process of theory building, testing and refinement which allows causal statements about attribution to be made. Evaluation findings should demonstrate what worked, for whom, how, and in what circumstances. These findings can then be used to further refine existing Middle Range Theory, which can then itself be tested by future evaluations. The results of the Ho Hospital case study confirmed many of the main elements of the initial MRT, but also found that some additional factors had been key to improving the organisational climate and the perception of organisational support. The Ho Hospital study was one of several case studies in Ghana and Tanzania examining the links between hospital management and performance. In keeping with this, the final stage of the study involved updating the initial MRT to include the new findings, with the updated MRT to be used as the starting hypothesis for future studies.

General Elimination Methodology (GEM) – aka The Modus Operandi Method

Scriven's GEM (2008) builds upon his earlier Modus Operandi Method (1976) to provide an approach specifically geared towards substantiating causal claims. The methodology entails systematically identifying and then ruling out alternative causal explanations of observed results. It is based on the idea that for any event it is possible to draw up *Lists of Possible Causes* (LOPCs) or alternative hypothetical explanations for an outcome of interest. One evaluation which employed some of the logic of GEM examined the extent to which a judicial advocacy campaign called the "Final Push Campaign" had influenced a Supreme Court decision (Patton, 2008). Alternative competing explanations put forward included the possibility that the Supreme Court would not take into account external influences in making its decisions, and the idea that external influences other than the Final Push campaign would have more impact.

With GEM, each putative cause will have its own set of "footprints" or *Modus Operandi* (MO) – "a sequence of intermediate or concurrent events, a set of conditions or a chain of events that has to be present when the cause is effective" (Scriven, 2008). For example, a criminal investigation might be able to identify a criminal from a list of suspects by examining the means, motives and opportunity pertaining to each of them.

GEM sets out to identify potential causes of effects by examining the facts of a case and establishing which MOs are present and which are not. Any cause for which the *Modus Operandi* is not present can be dismissed, leaving only causal explanations that have a

genuine causal link. GEM is intended to provide a framework for evaluation which can establish causal claims beyond reasonable doubt. In this way, the Final Push evaluation aimed to build up evidence which could demonstrate how much influence the campaign appeared to have had, through triangulation and by aiming to seek out cumulative evidence to allow conclusions to be drawn.

Scriven provides a list of suggested ways or methods that can be used within this framework, including critical observation, interviews, theoretical inference, quasi-experiments and cross-sectional data, among others (see Scriven, 2008 for the full list). The Final Push evaluation drew on a detailed examination of campaign documents, legal briefs, court documents, media reports and a literature review on the Supreme Court, related cases and policy issues. Interviews and group discussions with people directly involved in and knowledgeable about the campaign also provided key information.

Process Tracing

Process Tracing sets out to “unwrap” the causal links between putative causes and outcomes by identifying the intervening causal processes or mechanisms at work (George and Bennett, 2005; Reilly, 2010). The approach has been adopted by Oxfam to evaluate policy influencing and citizen-engagement interventions such as its “Fair Play for Africa” campaign (Hughes and Hutchings, 2011a).

The evaluator starts by carrying out “process induction”, drawing on evidence to generate a number of (preferably competing) hypotheses about how an intervention may connect to an outcome. Evaluators should set out a series of hypothetical causal mechanisms which might be initiated by the programme, together with what should be observed if each hypothesis is true or false. For example, one goal of Oxfam’s “Fair Play for Africa” campaign was to facilitate greater collaboration and joint action among Civil Society Organisations (CSOs), with the ultimate goal of strengthening campaigns to improve health care across the continent, particularly for the most vulnerable (Stedman-Bryce, 2011). Oxfam’s evaluation set out to assess whether any improved CSO collaboration resulted from its own campaign, or from those of other civil-society actors, or from a combination of the two. This involves identifying a series of “diagnostic” pieces of evidence which will be present for each of the theoretically predicted links for each hypothetical explanation (the causal chain) if they are observed in practice.

Data collection should be designed so as to match the research questions being asked. To date, Oxfam’s Process Tracing evaluations have drawn on individual and group stakeholder interviews, analysis of programme documentation and policy documents, and media analysis designed to find references to the Oxfam campaign. In-depth case-study analysis, drawing largely on qualitative data including historical reports, interviews and other documents, but possibly also using quantitative data, is used to develop an explicit chronology of events, setting out the causal links between each stage. This evidence is then used to overturn or substantiate rival hypothetical explanations (“process verification”), the ultimate goal being to establish whether the actual mechanisms at work fit with those predicted (Bennett, 2010). Oxfam’s evaluation found that the “Fair Play for Africa” campaign had improved collaboration among African CSOs, but also concluded that in some cases significant work had already been done in this area. For example, the “*Medicines for All*” campaign in Malawi had already successfully

brought together CSOs in a national advocacy campaign to improve the medicine-supply chain in the country. The Fair Play for Africa campaign was able to build on that success.

By assessing each of the hypotheses drawn from theory, the aim is to be able to make strong causal claims about what mechanism(s) caused a given set of outcomes in any given case. All findings should be triangulated using multiple methods, and a key goal of data collection and analysis should be to search for forensic evidence, or “smoking guns”, which could be used as evidence of the influence of Oxfam’s campaign, such as the high-profile usage of an Oxfam campaign slogan by a key political figure. In Oxfam’s use of the approach, evaluators are asked to provide subjective contribution scores for all programme outcomes, intended and unintended, based on the evidence produced by the evaluation. These assessments are not expected to provide a precise measure of contribution, but rather a qualitative sense of how much the campaign was responsible for the observed change(s). Scores are awarded both according to the degree to which the targeted outcome in question materialised and according to the campaign’s contribution to this change.

Contribution Analysis

Contribution Analysis is an approach to evaluation developed by Mayne (2001, 2008, and 2011) which aims to compare an intervention’s postulated theory of change against the evidence, in order to come to robust conclusions about the contribution that it has made to observed outcomes. Clear methodological steps for carrying out contribution analysis are set out by Mayne (2001), with a recent revision (2011) updating them to explicitly adapt Contribution Analysis for more complex interventions. The aim of Contribution Analysis is to critically construct a “contribution story” which builds up evidence to demonstrate the contribution made by an intervention, while also establishing the relative importance of other influences on outcomes. The approach draws on the idea that an intervention’s theory of change can be used to infer causation by assessing whether the mechanisms or processes that it aims to initiate have in fact occurred.

The evaluation of the 2005 Paris Declaration (Wood et al., 2011) drew on Contribution Analysis among other approaches, with evaluation teams asked to evaluate the extent to which the implementation and consequent effects of the Paris Declaration contributed to development results, how, and why. For the Paris Declaration Evaluation, teams drew on a programme theory which set out the declaration’s underlying logic, including desired outcomes, both intermediate (in improved aid effectiveness) and longer-term (in contributions to improved development results), and a description of how programmatic actions were intended to achieve them. The evaluation’s theory-based approach was designed to seek out key actors and causal mechanisms which might drive or inhibit change. Various “complex pathways to change” were set out, with consideration given to the important potential contextual factors which might influence the causal chain between Declaration objectives and development results. Evaluators were also asked to assess Core Evaluation Questions about whether the Paris Declaration had been successfully implemented, the extent to which it had improved the efficiency of aid delivery, and the contribution that it had made to development results.

Contribution Analysis sets out to demonstrate a plausible association between a programme and observed outcomes, using weight of evidence – by building a credible contribution story in which each step lying between programme inputs and outcomes is

clearly evidenced, with the result that a "reasonable person, knowing what has occurred in the programme and that the intended outcomes actually occurred, agrees that the program contributed to these outcomes" (Mayne, 2001). A plausible association can be said to have been made if the following criteria are met: (1) a reasoned theory of change is set out; (2) the activities of an intervention are shown to have been implemented as set out in the theory of change; (3) the chain of expected results can be shown to have occurred; and (4) other influencing factors have been shown not to have made a difference, or else their relative contribution has been recognised. Where interventions are more complex, involving numerous sub-elements or simultaneous causal strands or where, for example, causality is recursive or outcomes emergent, Mayne suggests that multiple causal strands may be developed, each of these incorporated within a general theory of change.

For the evaluation of the Paris Declaration, teams were asked to assess the "plausible contributions" made by the Declaration to development results, and they were asked to provide clear evidence of any changes and connections observed and to state as explicitly as possible any other plausible explanations. Emphasis was placed on the relevance of the evidence found to core evaluation questions, the extent to which it could be triangulated and therefore be considered reliable, and the degree to which data were from recent, credible sources, and the extent to which data-collection methods and analysis provided a reasonable basis for the findings and conclusions drawn.

A matrix was designed to provide teams with a structured way to assess the plausible contributions of the Paris Declaration to development results in each context and to provide them with a comprehensive framework for evaluation, setting out the sorts of indicator and evidence that teams should be looking for and the methods or forms of analysis that teams could apply, and providing a ratings system for key questions. Evaluation teams were expected to adopt a multi-method approach designed to promote triangulation of results and the validity of conclusions made. Methods employed by evaluation teams included literature and document reviews, quantitative/statistical analysis of the most relevant available data, survey instruments, interviews, and focus groups and stakeholder analysis.

Summary: Group 1 approaches

All four of the approaches outlined above aim to address attribution by examining the facts of a case to gain an in-depth understanding of the causal chain connecting observed outcomes to an intervention. Their goal is to explain what has occurred and how it has occurred. They either seek out evidence to substantiate whether a programme's specified theory of change occurred in practice or they do the same for a number of alternative causal hypotheses which outline what might have occurred if causes or assumptions set out in the theory of change had varied. Evidence is gathered to assess each of the hypothesised explanations and account for any external factors which may have played a role. Causation is established beyond reasonable doubt by collecting evidence to validate, invalidate, or revise the hypothesised explanations, with the goal of documenting the links in the actual causal chain.

These approaches provide a set of logical steps to guide an evaluation so that researchers systematically consider how an outcome or set of outcomes might have occurred, consider what evidence will be needed to rule out or substantiate alternative

causal explanations, and target data collection to gather the most useful data. However, for the most part they do not specify what methods or tools to use, or how to use them. Instead they advocate a mixed-methods approach, making use of whatever quantitative and qualitative tools are available to build credible evidence that can demonstrate how observed outcomes came about.

3.2 Group II approaches

The next group of approaches are distinguished from the first group here by the fact that they do not set out to address attribution of cause and effect as explicitly as the Group I approaches. In general, the Group II approaches place stakeholder participation at the heart of data collection and analysis. They target programme beneficiaries, implementers and other key stakeholders in order to establish what factors are perceived to have been important in producing change; in so doing, they aim to gain an insight into how a programme is performing and the part that it is playing in driving change. However, in terms of Alkin's Evaluation Tree, they lie mostly along the "Valuing Branch".

Most Significant Change (MSC)

Most Significant Change (Davies and Dart, 2005) is a form of participatory monitoring and evaluation that involves the collection and selection of stories of *significant changes* which have occurred in the field. MSC was first developed to help NGOs monitor the impacts of participatory development projects by providing a technique flexible enough to identify development outcomes across a variety of locations; it emphasised the need to respect participants' own judgement regarding the changes that an initiative has made to their lives (Davies, 1998). Davies and Dart set out clear steps for using the approach in their MSC guide (2005).

The central element of MSC involves the systematic collection and selection of a purposive sample of significant-change stories. The stories themselves are elicited from programme participants by asking them to relate what significant changes (positive or negative) have occurred in their lives in the recent past, and enquiring why they think that these changes occurred and why they regard them as being significant. The Adventist Development and Relief Agency (ADRA) piloted the Most Significant Change (MSC) approach in two water and sanitation projects in the north and south of Laos between 2003 and 2004, with a view to implementing it more widely across other ADRA projects in Laos (Keriger, 2004; Willetts, 2004; Willets and Crawford, 2007). In this case, the stories were collected by field staff on a monthly basis through one-to-one interviews with beneficiaries and through their own direct observation. Four broad categories, or "domains of change", were chosen to manage analysis and guide those collecting the stories: "Changes in people's health; Changes in people's behaviour; Changes that reflect an area to improve (negative); Changes in any other area (open)" (Keriger, 2004). Stories were elicited by asking beneficiaries two questions about the domains of change: "What is the best change that has happened in the villages?" and "What is the biggest problem because of ADRA's project?" A template for recording the stories was designed to structure story reports and outline information regarding the "what, who, where, why, and how?" of the stories.

A key step in MSC is the process by which the most significant of the “significant change stories” are selected. After stories of significant change have been collected, they are then passed upwards through the organisational hierarchy and systematically reviewed by panels of designated stakeholders. The intention is for stakeholders to engage in in-depth discussion at each stage of the selection process regarding the significance of each story, the wider implications of the changes which they relate, and the quality of evidence which they contain. The use of multiple levels of selection enables large numbers of significant-change stories to be reduced to a smaller number of stories viewed as being most significant by a majority of stakeholders. Selected stories should be verified for authenticity by drawing on participants’ own experience or through direct observation. The stories of significant change collected by ADRA were evaluated and filtered for their relative significance by groups of stakeholders at increasingly high levels of organisational hierarchy, until only the most salient stories emerged.

MSC was originally developed as an approach for impact monitoring, rather than as an evaluation approach designed to generate summative statements about aggregate change. As an approach to impact monitoring, it is designed to report on the diverse impacts which can result from a development programme, and the significance of these changes as perceived by participants. It is intended to be an ongoing process occurring at regular intervals during the programme cycle, with the information gathered fed back into the programme to improve its management and running. The ADRA pilot process was reiterated monthly and included the provision of feedback to staff, setting out which stories had been selected, reasons for their selection, and managerial reaction to the programme-response measures suggested by local staff.

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

Success Case Method (SCM)

The Success Case Method (Brinkerhoff, 2003 and 2008) is a narrative technique based upon naturalistic inquiry and in-depth case-study analysis. It is intended to be a quick and simple evaluation process geared towards understanding whether an initiative (such as a training or educational programme) is actually working. SCM sets out to discover whether an intervention is working or not by searching for particularly successful or unsuccessful instances (“success” and “non-success” cases). The Success Case Method does not set out to find out about the “average” participant, but instead intentionally seeks out the very best (and worst) that a programme has produced, in order to understand the contribution that the programme has made to results, the role that contextual factors have played in influencing the different outcomes, and the way in which this information can be used to improve programme performance.

Either using already available data or, more often, by means of a survey, the SCM identifies individual examples that have been particularly successful (and unsuccessful) and uses case-study analytical methods to develop credible arguments about the contribution that the intervention made to them. The first step involves setting out the

purpose of the study and outlining practical and time constraints for the evaluation. An *impact model* should then be drawn up, detailing the behaviours and positive outcomes that should ideally be achieved if the programme functions as intended. Coryn et al. (2009) adapted the SCM as part of a larger evaluation of a programme aimed at reducing chronic homelessness and unemployment. They set the criteria for success or non-success in terms of the long-term retention of economically self-sufficient employment, stability of housing, and any reduction in the need for government or other social support systems.

The next stage of the SCM involves administering a brief survey to participants with the intention of eliciting best and worst cases. One-to-one interviews with success and non-success cases then determine specifics of intervention usage and results, the perceived value of the results, and environmental factors that enabled or prevented success. Coryn et al. classified participants according to the three criteria outlined above and then interviewed a random sample of success and non-success cases. Interviews were intended to elicit perceived reasons for success or failure and investigate possible alternative explanations for success or failure not directly attributable to the programme. The evaluation further adapted the SCM by adding a time-series component whereby interviews were undertaken at three separate points in time, and by drawing on some of the logic of GEM to identify alternative explanations and search for patterns which might eliminate some, if not all, of them and therefore strengthen any causal findings.

Outcome Mapping (OM)

Outcome Mapping (Earl et al., 2001) is a participatory approach for planning, monitoring and evaluating development programmes which was developed by the International Development Research Centre (IDRC). It aims to help programmes to report and learn from outcomes in a realistic manner by tracking the intervening steps between programme activities and programme outcomes. Development results are measured not in terms of “impact” but instead in terms of the programme’s contribution to changes in the behaviour of, and relationships between, the actors with whom the programme interacts directly.

Outcome Mapping can be split into three stages.

(1) In stage one, an “Intentional Design” outlines programme priorities, setting out programme goals, programme activities and potential boundary partners (all actors with which the programme interacts directly) and sets out how to promote progress towards anticipated results. The “Quality Education and Vulnerability Programme in Zimbabwe” has combined Outcome Mapping with elements from a Logical Framework Approach to monitor and evaluate its work (Van Ongevalle et al., 2009a, 2009b and n.d). The team held an initial series of workshops alongside relevant stakeholders to draw up a programme vision and mission, set out key result areas, and identify boundary partners. The workshops were also used to plan how the programme could measure its contribution towards achieving key results, desired changes in boundary-partner behaviour (known as outcome challenges) and the strategies that the programme should employ to support its partners in achieving these.

(2) Stage two, “Outcome and Performance Monitoring”, provides a framework for the ongoing assessment of a programme’s actions, and boundary partners’ progress towards

desired goals/outcome. Mixed methods are used to gather data for an outcome journal which monitors boundary-partner actions and relationships, a strategy journal which records programme strategies and activities, and a performance journal which monitors organisational practices designed to keep the programme relevant and viable. Under the “Quality Education and Vulnerability Programme”, data collection and analysis takes place over six-monthly cycles on an ongoing basis, with data collected through formal and informal processes. Boundary partners fill out the outcome journals themselves, and strategy journals are filled out by the team in a one-day meeting designed to provide information on the programme’s support activities. Further data are also gathered from informal conversations, direct observations and the systematic collection of Most Significant Change (MSC) stories from lecturers and students. The extra data are used to provide a means to triangulate findings and gain a deeper understanding of how the programme is functioning.

The completion of the outcome and strategy journals is itself intended to provide a first level of analysis through which progress towards desired outcomes can be gauged and potential future actions planned. The programme team then carry out a second level of analysis, drawing on the principles of grounded theory to identify lessons from important and unexpected outcomes, and using the key result areas specified by the programme’s log frame as a useful framework for considering boundary-partner progress, with qualitative data from the journals used as the basis for a quantification of programme results (such as the number of colleges that have introduced a particular measure). Results from the analysis are compiled and then used to provide feedback on the programme to all stakeholders.

(3) Step three, “Evaluation Planning”, prioritises the elements of the programme which are to be evaluated and works out the logistics for the evaluation to be carried out. The Zimbabwe Quality Education and Vulnerability Programme team plan to carry out evaluations every three years; they have prepared a preliminary framework for evaluations which emphasises assessing the impact on beneficiaries and the logic of the overall programme.

Method for Impact Assessment of Programs and Projects (MAPP)

MAPP (Neubert, 2000 and 2010) combines a qualitative approach utilising participatory assessment tools with a quantification step. It is essentially a “before and after” design, intended to compare the outcomes produced by a programme, as recalled by target groups. MAPP sets out to select only enough relevant data for conclusions to be reached, and is intended to be a cost- and time-efficient evaluation process. Originally developed for social analysis of the outcomes produced by resource-management projects in Sub-Saharan Africa, MAPP can also be applied to other types of programme and in other regions. It was used to evaluate the “Projet de Gestion des Ressources Naturelles” (PGRN) in Mali, financed by the World Bank (Neubert, 2010), which was intended to improve village development by introducing a sustainable land-use system.

MAPP uses a series of tools to evaluate the perceived influence of the development project on the daily life of the population, taking into account the project’s context. It then sets out to establish the reasons for changes, whether they be the result of project-specific measures, initiatives organised independently by the local population, the result of intervention on the part of other donors, or determined by further external factors.

One of the tools which the PGRN evaluation employed was an “Activity List”, which outlined all activities contributing to the life of the village, whether part of the PGRN or organised by villagers themselves or by other agencies. The list also compared the relative importance of each of these activities to the lives of the villagers.

MAPP sets out to be parsimonious in its information gathering as far as possible, selecting only relevant data in order to avoid information overkill. The various MAPP tools are intended to triangulate findings to improve validity. Instead of simply monitoring programme results or performance using a log frame, MAPP aims to capture a mix of effects, positive and negative, intended and unintended. The PGRN evaluation drew up a “Development and Impact Profile” which set out all changes occurring in the village over time and assessed what perceived factors had contributed to observed outcomes. For example, villagers reported that agricultural yields still depended primarily on rainfall, but that fluctuations had been reduced by PGRN work to reduce soil erosion. Villagers also reported that PGRN’s tree plantation had provided a useful resource for the village but had also led to conflict with local herders.

Summary: Group II approaches

As with the Group I methodologies above, the Group II approaches gather information which can help to reconstruct the intermediate steps between cause and effect. However, they do not make causal explanation their primary goal, instead focusing on reporting on participant-behaviour changes and assessing how existing programmes may be improved. Where they do make causal inferences, these approaches rely on the actions and reported experiences and perceptions of stakeholders in order to do so. They tend to draw on largely qualitative methods, although some also use quantitative data or contain a quantification step for qualitative information. In practice these approaches are not necessarily intended to be stand-alone tools but can instead be usefully employed as one element within a wider evaluation framework. However, as argued in Section 4, stakeholder perception of causal relationships is subject to various biases.

3.3 Small n attribution

For large n analyses, experiments provide a powerful tool for attributing cause and effect. The basis for experimental causal inference stems from the manipulation of one (or more) putative causal variables and the subsequent comparison of observed outcomes for a group receiving the intervention (the treatment group) with those for a control group which is similar in all respects to the group receiving the intervention, except in that it has not received the intervention (Duflo et al., 2008, White, 2011). The creation of an empirical “closest possible world” comparison group facilitates a comparison between an intervention’s observed outcomes and the counterfactual scenario of what would have happened had the intervention not taken place.

The small n approaches outlined above draw on a different basis for causal inference. They set out to explain social phenomena by examining the underlying processes or mechanisms which lie between cause and effect. References to processes or mechanisms are common within philosophy-of-science explanations of causation (for example, Brady, 2008; Hedström and Ylikoski, 2010, etc.). However, there is also increasing recognition of the utility of mechanistic or process-based causal inference for social-science

evaluation (examples include Pawson and Tilley 1997; Maxwell, 2004a and 2004b; Mohr, 1999; Pawson, 2008; Astbury and Leeuw, 2010).

At the heart of the process-based or mechanism-based³ approach to causal inference is the idea that outcomes can be explained by seeking to discover the cogs and wheels which drive causal processes. 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 (Hedström and Ylikoski, 2010). There are plenty of observable mechanisms in the biological sciences, such as DNA replication, but there are also many examples in the social sciences, including the bringing together of buyers and sellers in markets or the electoral process by which voters choose candidates (Hedström and Swedberg, 1998).

Whereas experimental approaches infer causality by identifying the outcomes resulting from manipulated causes, a mechanism-based approach searches for the causes of observed outcomes. The small *n* approaches examined here, especially the Group I approaches outlined above, utilise in-depth examination of the facts of a given case or cases to identify the mechanism(s) connecting cause and effect. Hypothesised causal relations (such as those lying between a programme and observed outcomes) can be investigated by searching for mechanisms and recognising the relevance of different parts of the mechanism, their properties, their interactions and their relationship with context.

Mechanism-based explanations are not merely historical narratives; detailed evidence must be provided in order to understand how different parts are organised and inter-related, so as to critically reconstruct each link in the causal chain. Good mechanism-based explanations should also seek not only to ascertain whether the evidence supports a theorised explanation of cause and effect, but also to assess whether the observed effects might not actually be produced by other (known or unknown) mechanisms. Thus, mechanism-based explanations include implicit counterfactuals; they set out not only to find rigorous empirical evidence that supports the assumptions of one explanation, but also to plausibly demonstrate that it is absent for alternative counterfactual hypotheses.

Of the two groups discussed above, only the Group I set of approaches explicitly set out to make a causal link. Their overarching aim is to build a credible case which will demonstrate that there is a causal relationship between an intervention and observed outcomes. Mohr (1999) suggests the analogy of a medical diagnosis or a detective investigating a case as being a good one to describe the process of elimination and accumulation of evidence by which causal conclusions can be reached. Multiple causal hypotheses are investigated and critically assessed. Evidence is built up to demonstrate the different connections in the causal chain, with the ultimate goal of providing sufficient proof to demonstrate a plausible association, as in the case of Contribution Analysis, or to substantiate a causal claim “beyond reasonable doubt”. This latter phrase

³ Hereafter we refer to this as “mechanism-based” causal explanation. The use of the word *mechanism* should not be interpreted in a purely mechanical sense, but instead is used here to signify the inner workings of a social process.

is one which Scriven uses in his discussion of General Elimination Methodology (2008) and one commonly used as the standard of proof required to validate a criminal conviction in many countries. What the phrases imply is that there should be sufficient weight of evidence to convince a reasonable and independent person that the evidenced version of events occurred.

In specifying and assessing a number of causal hypotheses and reconstructing the actual causal chain lying between intervention and outcomes, these approaches should also be able to make a judgement about the extent to which outcomes may be attributable to a particular programme, and the extent to which they are attributable to other relevant factors. If it is possible to reach a robust conclusion as to whether the programme caused the outcomes rather than the outcomes being attributable in part, or in whole, to some other factor(s), it should also be possible to estimate the extent to which outcomes should be attributed to that programme or not. This may not involve placing an exact numerical estimate of the effect produced by an intervention, but will instead mean interpreting it in terms of a scale of magnitude (for example, from no discernible impact through to high impact) and necessity and sufficiency (was the intervention sufficient by itself to produce the desired impact, or was it a necessary but individually insufficient factor in producing the observed outcome?).

In contrast with the Group I approaches, the Group II approaches are designed for a variety of monitoring and evaluation purposes, rather than for the explicit purpose of causal attribution. They do not set out to assess causation in the systematic and thorough way that the Group I approaches do. However, if employed as part of a wider evaluation which more explicitly sets out to assess rival explanations for observed outcomes, they can offer a systematic method for gathering the necessary information to rigorously attribute outcomes to a programme. For example, Van Ongevalle et al. (2009, n.d.) employed Outcome Mapping in conjunction with a Logical Framework Approach, and Coryn et al. applied the Success Case Method as part of a wider evaluation framework (2009).

4. Tackling bias

The Group I approaches discussed in Section 3 above provide evaluators with a clear logic or set of methodological steps to follow. However, with some exceptions, they do not prescribe specific methods, preferring instead to recommend that evaluators choose the most appropriate mix of tools and techniques to fit the particular circumstances of any given evaluation. Furthermore, for the most part neither do these approaches tackle the subject of data collection and analysis, or the issue of how evaluators can ensure that an approach is followed in a systematic way which minimises bias.

All methods are subject to potential biases. In experiments and quasi-experiments, analysis bias can arise when there is a systematic deviation between a statistical estimate of a given parameter and its true value in the population of interest. For large n studies, systematic error can arise from, for example, researcher or selection bias. For randomised control trials, long lists of possible biases have been outlined (for examples, see Scriven, 2008; Jadad and Enkin, 2007). But in the field of development evaluation, far less attention has been paid to potential biases in other approaches.

Bias affects all evaluation approaches and methods; it affects the choices that evaluators make, the questions that they ask, the data they collect and the way that they carry out analysis. Much has been written about the biases that can affect our thought processes and our ability to make judgements. For example, Daniel Kahneman (2011) outlines cognitive limitations which can affect us all. Elsewhere, Robert Chambers has written at length about some of the biases especially relevant to those working in development (1983, 1997, 2006). He sets out six biases in particular which may prejudice findings: spatial biases focus our attention on urban areas or more accessible locations with better transport links; there are biases in the projects on which we choose to focus; person biases determine whom we are likely to talk to; bias arises from the fact that projects are often visited in the dry season; diplomatic and professional biases such as politeness and timidity may also prevent us from collecting the information that we need.

For the type of largely qualitative analysis pursued by the small n approaches for impact evaluation which are examined here, bias may result from the systematic tendency to either under- or over-estimate the strength of a causal relationship. Indeed there is a well-established tendency for people to see a pattern where there is none, seeing “runs” in random numbers. This belief is the view that “things must happen for a reason”, or the “teleological fallacy” (Kay, 2011).

These biases may arise either on the part of the evaluator or from the nature of data collection itself. Following the work of Heider (1944 and 1958), attribution theory in cognitive psychology addresses biases in people's ability to assess causal relationships. The discussion here draws partly on this literature, but also elaborates more generally on other biases.⁴

4.1 Respondent biases

The problem of “courtesy bias”, whereby a respondent tells you what they think you want to hear, is well established. In structured surveys, courtesy bias can affect both people's reported behaviour and self-reported outcomes relating to matters such as health.

For example, in Madagascar, user satisfaction regarding their treatment in a health facility was found to be higher in exit interviews – when the interviewer was seen to be associated with the clinic – than in the results of representative household surveys carried out in the same areas (Glick, 2009). This discrepancy was not replicated for more objective measures of facility conditions and the availability of supplies. A study of cook-stoves in Ghana found no impact on exposure to harmful fumes, but there was a significant impact on self-reported health. This finding is most likely an example of courtesy bias, as the training for the cook-stoves had emphasised health benefits, but it may also be a placebo effect.

Courtesy bias has a clear relevance for qualitative interviews, for example when interviewing respondents about how influential a particular agency or programme has

⁴ A recent book, Mark et al. (2011), explores the links between social psychology and evaluation. However, it is mainly concerned with outlining the insights offered by social psychology for programme design and implementation, rather than with the implications for evaluation methods.

been in affecting a change. Courtesy bias has been found to be stronger in Asian and Latin American countries.

A related form of bias is that of “social acceptability” or “political correctness” bias, where people provide responses which reflect what they regard as being the socially acceptable thing to say. This kind of bias is likely to appear if there is a prevailing consensus concerning how things have occurred, or concerning the role that a particular agency has played in achieving outcomes.

The above biases are known and well recognised. However, development evaluations have paid less attention to biases arising from respondents' perception of events. People see things from a different perspective, depending on where they are standing. This self-evident statement is of vital importance when assessing qualitative data. The task facing an evaluator is one of constructing an analytical narrative which is consistent with all the data and which reconciles conflicting interpretations through an understanding of context and of how different perspectives inter-relate. We might label the tendency to see things only from where one is standing as “positional bias”, although rather than interpreting it as a bias it may be better represented as an inherent challenge which faces any evaluator, that of interpreting different people's conflicting constructions of reality. However, there are specific instances of positional bias which certainly do deserve the term bias and which are well documented in social and cognitive psychology. Very appropriately these biases go under the name of “attribution biases”, some of which are especially relevant when thinking about small *n* evaluation methods.

The first is the “fundamental error of attribution” whereby people are more likely to attribute changes to individuals than to contextual factors. Hence the appeal of the “Great Men” approach to history, according to which, for example, Hitler was responsible for the Second World War, Garibaldi unified Italy, and Henry VII put an end to the English Middle Ages at about 10 am on 22nd August 1485 by slaying Richard III on Bosworth Field. These examples illustrate the tendency to under-estimate general social forces which may contribute to a policy change and to over-estimate the role of particular individuals.

Experiments have shown that, when presented data in which the outcome varies according to individual characteristics, respondents are more likely to correctly identify this causal relationship than when outcome varies by contextual factors (McArthur, 1972, cited in Fösterling, 2001: 82).

Hence there is a significant danger that respondents will emphasise the role of individual actors, while ignoring the more general social or political context and its effects on the timing and course of events. A particular individual or agency generally has an effect usually because they are in the right place at the right time. To quote rather different authorities, “Men make their own history, but they do not make it as they please; they do not make it under self-selected circumstances, but under circumstances existing already” (Marx, 1852) and “eighty percent of success is showing up” (Woody Allen, n.d.).

A second important attribution bias arises when respondents have a biased view of their own contribution to changes. This bias, called “self-serving bias”, occurs where an individual takes credit when things go well, but blames other factors (other people, the

situation, fate...) when things go badly. Teachers identify their teaching methods as the source of students' success, but blame the students when they do badly: a tendency borne out in experiments in which there has in fact been no association between teacher and student outcomes (quoted in Fösterling, 2001: 88–89).

Self-serving bias can be present in groups as well as individuals. For example, in assessing the outcomes of its structural adjustment programmes, the World Bank was ready to trumpet success when a programme was followed by improved macroeconomic performance; but when performance did not improve, or even worsened, the blame was attributed to others, usually the government in question for not having implemented the Bank's recommendations for adjustment completely (see White and Dijkstra, 2002, for a discussion of this phenomenon).

A final example is "self-importance bias". People are indeed the centre of their own universe; the problem comes when they think that they are also the centre of everyone else's universe, and as a result they overstate their role in events. An example of self-importance bias can also be drawn from the age of adjustment programmes. The World Bank typically dated the start of reform from the date of the first adjustment loans, overlooking the preceding years of reform efforts which laid the basis for the subsequent Bank-supported change (White and Dijkstra, 2002). More generally it has been found that it is domestic political processes that actually drive policy change (Killick et al., 1998), so any account which ascribes a central role to external actors is likely to be overstating the importance of those actors.

These biases present a real challenge for evaluators. Outcome monitoring will show that change took place. Analysis of the causal chain will show that programme activities took place, and many of those taking part will say that these activities played a role in the changes – and bias means that they are systematically more likely to say there was such a role than is "objectively" the case. The causal chain will appear to have been confirmed, and therefore causation will be inferred. As stressed by many of the approaches outlined earlier, it is important to ensure that all potentially relevant external factors have also been accounted for and consideration has been given to the possibility that assumptions and theorised causes may have varied. However, evaluators also need to ensure that they systematically apply evaluation methods in such a way as to overcome biases. Unfortunately, however, evaluators are themselves also prone to bias.

4.2 Evaluator biases

It is commonly held that there may also be biases pushing evaluators towards positive findings, the main one being "contract renewal bias". Another bias which may actually be even more influential is "friendship bias". If an evaluator has spent time with programme staff and has developed a good relationship with them, it becomes difficult to upset them with a critical report.

Related to friendship bias is the issue of independence in evaluation. Although the need for independence is normally couched in terms of constraints or incentives which may inhibit critical judgement, there is also a cognitive basis to the arguments in favour of independence: that is, a greater likelihood of finding a positive influence from "ingroup" individuals than "outgroup" individuals (Fösterling, 2001: 105).

Reinforcing these effects are further cognitive biases associated with giving greater weight to the accounts of people with whom we have had direct contact, and those with whom we most strongly identify. The former bias can be demonstrated by an experiment in which subjects are shown a video of a discussion between two people, A and B. Half the subjects shown the video see only images of person A, and the other half of the subjects see only person B. Most subjects consider that the person whose image they were shown played a more important role in the discussion. The implications for evaluation methodology are clear. We will give more weight to the narratives of people whom we interview, reinforcing their already over-inflated role arising from self-importance and self-serving biases. Evaluators typically speak mostly to staff of the agency that they are evaluating, staff of other relevant external agencies, and selected government officials. They often conduct very limited work in the field, and rarely speak to parliamentarians, traditional authorities (such as village chiefs), trade unions, journalists and other key actors.

The biases which affect the choice of whom evaluators speak to are a manifestation of “similar-person bias”. This affects not only the choice of the people we speak to, but also whose evidence we are most likely to trust. Evaluators will most readily make sense of, and therefore trust, data from people “like them”, who have a similar world view. It requires a mental effort and considerable empathy to understand the quite different perspective from which other people may see things.

4.3 Addressing bias

In summary, there are potential biases in the data themselves, the way they are collected, and their interpretation and reporting by the evaluator. Many of these biases might well be of significance for any impact evaluation, but they are of particular relevance for small n evaluations, which usually rely largely on qualitative research. If small n approaches are to guard against possible biases, clear strategies for data collection and analysis are needed to identify the risks and counter them. Without such systematic analysis, evaluators may either cherry-pick the cases and data which support their argument, or sub-consciously note only evidence fitting with their prior expectations, omitting any contrary evidence.

There are implications of the above discussion for the use of participatory data. Such data can yield valuable insights into addressing key evaluation questions. But asking programme beneficiaries direct questions about attribution is of limited, if any, value.

If a small n evaluation is to adequately tackle the question of attribution, it needs to plan for and address these potential biases. For bias to be minimised, an evaluation must have a clear plan which sets out the nature and sources of data to be collected and includes instrument design and protocols for qualitative field work and data analysis. As well laid out in *Qualitative Data Analysis: an Expanded Sourcebook* (Miles and Huberman, 1994), categorising, coding and matrices are key to systematic analysis of qualitative data. Checklists for qualitative evaluation (for example, Patton 2003) can also provide evaluators with guidelines that can help them to prepare for and address possible biases.

Stakeholder mapping and careful consideration of sampling can ensure that views from all key stakeholders are taken into consideration. Interviews should be carefully planned, semi-structured and targeted to explore key parts of the causal chain. They should also be recorded, documented and independently carried out. Results can be triangulated, with information collected from a diverse range of individuals and settings or using a variety of methods, so as to reduce the risk of systematic biases due to a specific source or method. All this may be basic good practice, but the point is that if evaluators are aware of the possible biases that they face, they can build systematic data collection and analysis processes designed to address them and minimise their potential impact. Furthermore, in any evaluation, these data collection and analysis processes should be clearly documented.

5. Towards an integrated framework: common steps for causal inference in small n cases

The following section sets out a framework for small n causal inference. The framework steps are derived to a large extent from the logic of the Group I approaches discussed earlier, as well as from examples of their application (see especially Pawson and Tilley, 1997; Scriven, 2008; George and Bennett, 2005; Hughes and Hutchings, 2011b; Mayne, 2011) and also more widely from the evaluation literature. By seeking to find a common methodological core in these approaches, the intention is not to outline a new approach for impact evaluation, but instead to produce a list of general steps that a small n evaluation should ideally go through in order to rigorously examine the causal relationship between an intervention and observed outcomes.

The key steps that we set out involve establishing what causal questions are being asked, clearly setting out a programme's theory of change, outlining a plan for data collection and analysis, and generating a series of alternative causal explanations, before finally examining each of these hypotheses against the facts of the case and systematically providing evidence for each link in the actual causal chain. The emphasis on theory advocated here does mean that considerable systematic effort will be required to carry out a thorough evaluation. This is likely to mean that a small n evaluation which tackles issues of causation in a rigorous manner will require a reasonable commitment of time and resources. The goal is to assess the plausibility of causal claims – that is, to demonstrate beyond reasonable doubt that there is a causal relationship between an intervention and outcomes, and to estimate to what extent outcomes can be attributed to that intervention.

The framework is not intended to necessarily be a single, linear, step-by-step procedure: evaluation design and implementation is an iterative process. For example, data collection and analysis should be an ongoing process and should inform several of the steps in the framework. For strong conclusions to be reached, reasoning at each step should be submitted to systematic criticism, with new evidence sought out and, where necessary, previous steps revisited.

The attribution question(s)

An important step is to set out what attribution questions must be answered and to state how the evaluation will address them. This entails explicitly stating what causal questions are to be tested. Useful questions to consider include the following.

- What is the observed change in outcomes of interest?
- To what extent can observed outcomes be attributed to the intervention in question?
- What contextual factors or external influences, such as the social or cultural setting, political or economic trends, and parallel interventions or other stakeholder actions, are likely to have influenced outcomes?
- If causes or assumptions vary, what alternative causal hypotheses might there be for observed outcomes?

Setting out a programme's Theory of Change

A theory of change⁵ seeks to establish the links between an intervention, its wider context, and its outcomes (Weiss, 1995, Mason and Barnes, 2007, White, 2009). The aim should be to clearly set out the theory behind an intervention, outlining how it will be implemented, why it is being implemented, what its aims are, how it is supposed to work, and in what conditions. This should go far beyond simply describing a programme, listing inputs, activities, outputs and outcomes. It should explain how a programme is intended to function in a particular context and how it will bring about the planned outcomes. It should provide an explanation of what causal mechanisms an intervention is intended to trigger, and what the underlying assumptions for those causal mechanisms to operate are.

The procedure of drawing up a theory of change should be a consultative process between evaluators, programme implementers and other stakeholders. The whole evaluation team should be involved in the process, identifying what sort of evidence they can bring to bear to answer the different evaluation questions and different stages of the causal chain. It should make use of any relevant past case studies and incorporate stakeholder inputs as well as wider theory. By setting this out in detail, an evaluator provides a blueprint for how a programme should function. It also ensures that not only the intervention goals but also the evaluation aims are clearly outlined. The theory of change can also then be used to inform later steps in the evaluation, including data collection and analysis.

The theory of change should do the following:

- Set out the underlying logic of the intervention, specifying each link in the theoretically predicted causal chain.
- Outline the planned programme inputs, activities, expected outputs and desired intermediate and final outcomes.
- Include possible spill-over effects, both positive and negative.
- List the potential programme participants and all other project-affected persons, along with the timelines involved and any indicators being used to monitor change.

⁵ The Theory of Change concept was initially developed at the Roundtable on Community Change at the Aspen Institute (Weiss, 1995).

- Outline assumptions and risks for each link in the theoretically predicted causal chain; the evaluation questions should “interrogate” the causal chain to see if the assumptions necessary for the intervention to function hold as expected.
- Set out contextual factors and external influences which might influence the causal chain or otherwise affect the outcomes of interest.

Evaluation plan for data collection and analysis

The attribution questions and the theory of change set out the assumptions to be tested and the evaluation questions to be answered. However, the evaluator should also make an evaluation plan which outlines the nature and sources of data to be collected in order to answer these questions, as well as protocols for qualitative field work and categories for data analysis. Some consideration should also be given to the methods likely to be best able to produce causal statements. These judgements need to take account of both respondent and evaluator biases in the collection and analysis of qualitative data, with the goal of putting in place mechanisms to overcome them. For example, criteria for interviews should require that they are recorded, that notes are taken by multiple team members and then compared, and that questions are carefully drawn up to avoid leading interviewees.

In studies of multiple case studies (such as countries, agencies or communities) it is useful to have a template which sets out the information that should be collected for each case, to ensure that the same range of data is available for each. This is especially important if different researchers are working on different cases. The researchers should discuss their understanding of the template to ensure a common understanding, so that the template is consistently applied.

The development of a good chronology of events is also usually invaluable. It will help to avoid claims that “X caused Y” when in fact Y preceded X. It may also be possible that the anticipation of X alone may also have affected Y – so did people know that X was going to happen? Evaluators should be aware of this type of problem and explore it where feasible. Looking for footprints is a useful analogy: evidence that the intervention has “been there”. Sometimes an intervention will not be associated with the name of the supporting agency, for example if it worked through partners and did not insist on heavy branding of the activities that it supported. However, the theory of change should have identified the activities which were supported by the intervention, for example workshops or media campaigns, so that the influence of those activities can still be attributable to the intervention.

Data analysis itself should be an iterative process between theory and data. Initial themes, categories and coding of data should be structured around the theory of change. Once alternative causal hypotheses have been developed, subsequent rounds of analysis can be built around the different causal explanations.

Identifying alternative causal hypotheses

In addition to the programme’s theory of change, an evaluation should also identify a series of plausible hypothetical causal explanations for outcomes.

Scriven's General Elimination Method (2008) requires evaluators to produce a list of possible causes and a modus operandi for each of them. Similarly, realist evaluators theorise about the various ways in which different mechanisms might interact with context to produce a series of conjectural Context Mechanism Outcome theories (Pawson and Tilley, 1997). When carrying out Process Tracing, evaluators conduct "process induction" to generate a number of (preferably competing) hypotheses about how an intervention may connect to an outcome.

The rival hypotheses should take the form of carefully constructed counterfactuals which detail what might occur as a result of variation in causes, contexts or the assumptions of the theory of change. The alternative hypotheses should draw on the emerging substantive evidence gathered thus far, especially that relating to intended and unintended programme outcomes and external factors, but they should also be informed by stakeholder consultation, theory and past case studies.

As with the original theory of change, these alternative explanations should take the form of causal hypotheses in their own right, with each detailing the precise mechanisms thought to have triggered observed outcomes. For each hypothesis, careful consideration should also be given as to what further evidence would be observed were it to be proved correct, and what might be observed if it were false. This can help to guide later data collection and analysis – although it should not limit it – as many explanatory factors cannot be predicted in advance.

Verifying the causal chain

Existing and new data should be used to critically examine each of the hypothesised causal explanations, including the programme's theory of change. Analysis should examine the degree of congruence between predicted and actual observations for each causal hypothesis, with the ultimate goal of building up rigorous empirical evidence to demonstrate each link in the actual causal chain.

Evaluators should set out to examine whether programme actions were implemented as planned, what outcomes have occurred, and whether programme assumptions or those predicted by alternative causal explanations have been proved right or wrong. The analogy of a detective is a useful one to bear in mind when thinking about how evaluations can reach causal conclusions by investigating multiple hypothetical causal explanations and critically assessing evidence in order to substantiate, modify or invalidate each of them, ultimately leaving standing only those that have had a causal influence.

Where possible, evaluators should seek to identify "diagnostic" evidence which by itself might serve to reject or confirm part or all of a hypothesised explanation. Bennett's (2010) four tests to assess the strength of alternative hypotheses are a useful logic to apply to this process: "Straw in the Wind" tests provide evidence for or against a hypothesis, but by themselves cannot confirm or deny it; "Hoop" tests, if passed, can affirm the relevance of a hypothesis but cannot fully confirm it and, if failed, can eliminate a hypothesis; "Smoking Gun" tests can confirm a hypothesis if passed, or weaken it if failed; "Doubly Decisive" tests confirm a given hypothesis and in doing so eliminate any others.

In critically assessing the different causal hypotheses, the goal is to reconstruct the actual causal chain. The process of constructing a sufficiently strong case should emphasise the triangulation of key findings, with a judgement regarding the number of sources thought necessary, made on the basis of how important they are to the evaluation's ultimate conclusions and the evaluator's confidence in their validity. Ultimately, as set out in Section 3.3 above, the goal is to establish a causal link by accumulating sufficient weight of evidence to demonstrate the different links in the causal chain.

6. Conclusion

While there is a plethora of possible approaches suitable for small n impact evaluation, there is also an implicit consensus concerning what the approach should look like. It involves the specification of a theory of change (or similar) and usually of a number of further alternative hypothetical causal hypotheses. It also involves the subsequent assessment of the different hypothesised explanations and consideration of any other factors which may have influenced observed outcomes. Causation is established by collecting evidence to validate, invalidate or revise the hypothesised explanations, with the ultimate goal of rigorously documenting the links in the actual causal chain.

The Group I approaches outlined above offer good conceptual steps for carrying out small n impact evaluation. The Group II approaches do not address cause and effect in the same explicit way but, if they are combined with an approach that does do so, they can provide an effective tool for gathering some of the evidence needed to make a causal claim.

If properly applied, the Group I approaches outlined can tackle attribution; they can demonstrate the extent to which an intervention alters the state of the world. However, if we want to adequately tackle attribution in a small n evaluation, it is not enough for the evaluation to cover the conceptual steps from the framework above, or for the logic of one of the Group I approaches outlined to have been applied. There is a need for systematic data collection and analysis in order to be able to make causal statements. There are possible biases in how respondents perceive and report causality. There are also possible biases in the people to whom evaluators choose to speak and in whose information they assume credibility, and in the way these data are analysed and reported. Unless these biases are explicitly confronted, largely qualitative studies are likely to systematically over-estimate the impact of development interventions. However, there are ways of addressing and minimising the impact of these biases; ultimately if a small n impact evaluation is to adequately tackle attribution, it should build systematic data collection and analysis processes into a wider approach explicitly designed to tackle causation.

7. Bibliography

Allen, Woody (n.d.) http://www.quotationspage.com/quotes/Woody_Allen

Alvin Alkin, Marvin and Christina Christie (2004) 'An Evaluation Theory Tree' in Marvin Alkin, *Evaluation Roots: Tracing Theorists' Views and Influence*. London: Sage.

Astbury, B. and Leeuw, F. (2010) Unpacking black boxes: mechanisms and theory building in evaluation. *American Journal of Evaluation*, 31 (3): 363–381.

Barnes, D. F. (2007) *The Challenge of Rural Electrification: Strategies for Developing Countries*. Washington D.C.: Rff Press.

Brady, H. E. (2008) Causation and Explanation in Social Science, in Janet M. Box-Steffensmeier, Henry E. Brady and David Collier (eds.) *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Bennett, A. (2010) Process Tracing and Causal Inference, in Henry Brady and David Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield.

Bourguignon F. and Pereira da Silva, L. A. (eds.) (2003) [*The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*](#). New York: World Bank and Oxford University Press.

Briceño, B. and Gaarder, M. M. (2011) Institutionalisation of Government Evaluation: Balancing Trade-Offs. 3ie Working Paper No. 8. Online at http://www.3ieimpact.org/admin/pdfs_papers/Working_paper_8%20MG%20final.pdf.

Brinkerhoff, R.O. (2003) *The Success Case Method: Find Out Quickly What's Working and What's Not*. San Francisco, CA: Berrett-Koehler.

Brinkerhoff, R.O. (2008) *Telling Training's Story: Evaluation Made Simple, Credible, and Effective*. San Francisco, CA: Berrett-Koehler.

Chambers, R. (1983) *Rural Development: Putting the First Last*. London: Prentice Hall.

Chambers, R. (1997) *Whose Reality Counts? Putting the First Last*. London: Intermediate Technology Publications.

Chambers, R. (2006) *Poverty Unperceived: Traps, Biases and Agenda*. Institute of Development Studies, Working Paper 270, University of Sussex, UK. Online at <http://www.ids.ac.uk/files/Wp270.pdf>.

Coryn, C., Schröter, D. C. and Hanssen, C. E. (2009) Adding a time-series design element to the Success Case Method to improve methodological rigor: an application for non-profit programme evaluation. *American Journal of Evaluation*, 30 (1): 80-92.

Davies, R. (1998) An evolutionary approach to facilitating organisational learning: an experiment by the Christian Commission for Development in Bangladesh. *Impact Assessment and Project Appraisal*, 16 (3): 243–50.

Davies, R. and Dart, J. (2005) *The “Most Significant Change” (MSC) Technique: A Guide to Its Use*, April 2005. Online at <http://mande.co.uk/docs/MSCGuide.pdf>

Duflo, E., Glennerster, R. and Kremer, M. (2008) Using Randomization in Development Economics Research: a Toolkit, in Paul Schultz and John Strauss (eds.) *Handbook of Development Economics*, Vol. 4. Amsterdam: Elsevier.

Earl, S., Carden, F. and Smutylo T. (2001) *Outcome Mapping: Building Learning and Reflection into Development Programs*. International Development Research Centre (IDRC). Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>

Engela, R. and Ajam, T. (2010) *Implementing a Government-wide Monitoring and Evaluation System in South Africa*. Independent Evaluation Group (IEG), World Bank. ECD Working Paper Series, No. 21. Online at http://siteresources.worldbank.org/EXTEVACAPDEV/Resources/ecd_wp_21_south_africa.pdf

Fösterling, Friedrich (2001) *Attribution: An Introduction to Theories, Research and Applications*. Hove: Psychology Press.

George A. L. and Bennett, A. (2005) *Case Studies and Theory Development in the Social Sciences* (Cambridge, MS: MIT Press).

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B. and Vermeersch, C. M. J. (2010) *Impact Evaluation in Practice*. Washington D.C., World Bank. Online at http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf

Glick, P. (2009) How reliable are surveys of client satisfaction with healthcare services? Evidence from matched facility and household data in Madagascar. [*Social Science and Medicine*](#), 68 (2): 368–79.

Haddad, E. A., Perobelli, S., Domingues, E. P. and Aguiar, M. (2011) Assessing the ex ante economic impacts of transportation infrastructure policies in Brazil. *The Journal of Development Effectiveness*, 3 (1): 44–61.

Hedström, P. and Swedberg, R. (1998) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hedström, P. and Ylikoski, P. (2010) Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36: 49–67.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, 51, 358–374.

- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Hughes, K. and Hutchings, C. (2011a) Can We Obtain the Required Rigour Without Randomisation? Oxfam GB's Non-experimental Global Performance Framework. 3ie Working Paper No. 13. Online at http://www.3ieimpact.org/admin/pdfs_papers/WP%2013_Final.pdf
- Hughes, K. and Hutchings, C. (2011b) *Oxfam GB Guidance on Evaluating Policy Influencing Work Using Process Tracing*. Oxford: Oxfam GB.
- [Jadad](#), R. A and Enkin, M .R. (2007) *Randomized Controlled Trials: Questions, Answers and Musings*. UK: Wiley-Blackwell.
- Kahneman, D. (2011) *Thinking, Fast and Slow*. UK: Allan Lane.
- Kay, J. (2010) *Obliquity: Why our goals are best achieved indirectly*. UK: Profile Books.
- Keriger, R. (2004) MSC Guide: Based on the Experience of ADRA Laos: A Guide to Implementing the Most Significant Changes (MSC) Monitoring System in ADRA Country Offices. Adventist Development and Relief Agency (ADRA). Available from the MSC mailing list file repository at <http://groups.yahoo.com/group/MostSignificantChanges/>
- Killick, T., Gunatilaka, R. and Marr, A. (1998) *Aid and the Political Economy of Policy Change*. London and New York: Routledge.
- Marchal, B., Dedzo. M. and Kegels, G. (2010a) Turning around an ailing district hospital: a realist evaluation of strategic changes at Ho Municipal Hospital (Ghana). *BMC Public Health* 2010, 10: 787.
- Marchal, B., Dedzo. M. and Kegels, G. (2010b) A realist evaluation of the management of a well performing regional hospital in Ghana. *BMC Health Services Research*, 10(24).
- Marchal, B. Van Belle, S., van Olmen, J. Hoérée, T. and Kegels, G. (forthcoming) *A Review of Realist Evaluation Studies in the Field of Health Systems Research*.
- Mark, M. M., Donaldson, S. I. and Campbell B. (2011) *Social Psychology and Evaluation*. New York: Guilford.
- Marx, Karl (1852) 'The 18th Brumaire of Louis Bonaparte'. *Die Revolution*.
- Mason, P. and Barnes, M. (2007) Constructing theories of change: methods and sources. *Evaluation*, 13: 151–170.
- Maxwell, J. A. (2004a) Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, 33 (2): 3–11.
- Maxwell, J. A. (2004b) Using qualitative methods for causal explanation. *Field Methods*, 16 (3): 243–264.

- Mayne, J. (2001) Addressing attribution through contribution analysis: using performance measures sensibly. *Canadian Journal of Program Evaluation*, 16: 1–24.
- Mayne, J. (2008) Contribution Analysis: An Approach to Exploring Cause and Effect. Institutional Learning and Change Brief No. 7. Online at http://www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf
- Mayne, J. (2011) Contribution Analysis: Addressing Cause and Effect, in K. Forss, M. Marra and R. Schwartz (eds.) *Evaluating the Complex*. New Brunswick, NJ: Transaction Publishers.
- Miguel, E. and Kremer, M. (2004) Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, [72 \(1\)](#): 159–217.
- Miles, M. B. (1979) Qualitative data as an attractive nuisance: the problem of analysis. *Administrative Science Quarterly*, 24: 590–601.
- Miles, M.B. and Huberman, A. M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. UK: SAGE.
- Mohr, L. B. (1999) The qualitative method of impact analysis. *American Journal of Evaluation*, 30 (1): 69–84.
- Neubert, S. (2000) *Social Impact Analysis of Poverty Alleviation Programmes and Projects: A Contribution to the Debate on the Methodology of Evaluation in Development Cooperation*. GDI Book Series No. 14. London: Routledge.
- Neubert, S. (2010) Description and Examples of MAPP Method for Impact Assessment of Programmes and Projects. German Development Institute (GDI). Online at www.ngo-ideas.net/mediaCache/MAPP/.
- Patton, M. Q. (2003) Qualitative Evaluation Checklist. Online at http://www.wmich.edu/evalctr/archive_checklists/qec.pdf.
- Patton, M. Q. (2008) Advocacy impact evaluation. *Journal of MultiDisciplinary Evaluation*, 5 (9), March 2008.
- Pawson, R. (2008) Causality for Beginners. NCRM Research Methods Festival 2008 (unpublished).
- Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: SAGE Publications.
- Reilly, R. (2010) Process Tracing, in A. J. Mills, G. Durepos and E. Wiebe (eds.) *Encyclopedia of Case Study Research*. London, SAGE Publications.
- Scriven, M. (1976) Maximising the power of causal investigations: the Modus Operandi Method, in Gene V. Glass (ed.) *Evaluation Studies Review Annual, Vol. 1*. London: SAGE Publications.

Scriven, M. (2008) A summative evaluation of RCT methodology: and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, 5(9): 11–24.

Smutylo, T. (2005) Outcome Mapping: A Method for Tracking Behavioural Changes in Development Programs. Institutional Learning and Change (ILAC) Brief 7, August 2005. Online at <http://web.idrc.ca/uploads/user-S/11235064481Brief-FINAL.pdf>.

Stedman-Bryce, G. (2011) Fair Play for Africa Impact Evaluation: Report Prepared for Oxfam GB. Pamoja Consulting.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, New Series, 185 (4157): 1124–1131.

Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (n.d.) *Analysing Outcome Mapping Monitoring Data: The Case of the Quality Education and Vulnerability Programme in Zimbabwe 2008–2013*. Working paper. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>

Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (2009a) Monitoring for Impact in a Programme's Sphere of Influence – a Case Study of the Quality Education and Vulnerability Programme in Zimbabwe. Presented at the Afrea (African Evaluation Association) Conference, 29 March–3 April 2009, Cairo, Egypt. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>

Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (2009b) Monitoring for Impact in a Programme's Sphere of Influence – a Case Study of the Quality Education and Vulnerability Programme in Zimbabwe. Presented at the Afrea (African Evaluation Association) Conference, 29 March–3 April 2009, Cairo, Egypt. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>

Weiss, C. (1995) Nothing as Practical as Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives, in Connell, J. Kubisch, A., Schorr, L. and Weiss, C. (eds.) *New Approaches to Evaluating Comprehensive Community Initiatives: Concepts, Methods and Contexts*. Washington, DC: The Aspen Institute.

Westhorp, G., Prins, E., Kusters, C., Hultink, M., Guijt, I. and Brouwer, J. (2011) Realist Evaluation: An Overview. Report from an Expert Seminar with Dr Gill Westhorp. Online at http://www.cdi.wur.nl/NR/rdonlyres/66764817-54E3-4DCB-BD4F-806B403F892C/143307/2011_WP_RealistEvaluationSeminar_CecileKusters.pdf

White, H. (1996) Evaluating programme aid. *IDS Bulletin* 27 (4).

White, H. (1998) *The Macroeconomics of Aid: Case Studies of Four Countries*. Basingstoke, MacMillan.

White, H. (2005) Comment on contributions regarding the impact of the Bangladesh Integrated Nutrition Project. *Health Policy and Planning*, 20 (6): 408–411.

White, H. (2009) *Some Reflections on Current Debates in Impact Evaluation*. 3ie Working Paper No. 1.

White, H. and Dijkstra, G. (eds.) (2002) *Programme Aid and Development: Beyond Conditionality*. London: Routledge.

White, H. and Masset, E. (2006) The Bangladesh Integrated Nutrition Program: findings from an impact evaluation. *Journal of International Development*, 19: 627–652.

White, H. (2011) *An introduction to the use of randomized control trials to evaluate development interventions*. 3ie Working paper No. 9.

Willettts, J. (2004) *Most Significant Change Pilot Project: Evaluation Report*, prepared by the Institute for Sustainable Futures (Sydney, ADRA Laos). Available from the MSC mailing list file repository at <http://groups.yahoo.com/group/MostSignificantChanges/>

Willettts, J. and Crawford, P. (2011) The most significant lessons about the Most Significant Change technique. *Development in Practice*, 17 (3), June 2007.
The Most Significant Change approach egroup/online forum:
<http://groups.yahoo.com/group/MostSignificantChanges/>

Wood, B., J. Betts, F. Etta, J. Gayfer, D. Kabell, N. Ngwira, F. Sagasti and M. Samaranayake (2011) *The Evaluation of the Paris Declaration: Final Report*. Copenhagen: Danish Institute for African Studies.

World Bank (2005) *Maintaining Momentum to 2015? An Impact Evaluation of Interventions to Improve Maternal and Child Health and Nutrition in Bangladesh*. Washington D.C.: OED and World Bank.

World Bank (2008) *The Welfare Impact of Rural Electrification: A Re-assessment of the Costs and Benefits*. Washington D.C.: Independent Evaluation Group (IEG), World Bank. Online at http://siteresources.worldbank.org/EXTRURELECT/Resources/full_doc.pdf.

Annex 1: Approaches for small *n* analysis: extended summaries

1.1 Realist Evaluation

Overview

Realist evaluation (Pawson and Tilley, 1997) differs from the other approaches examined here in its fundamental identification with the realist paradigm, a school of philosophical thought which holds that there is a “real” world which exists independent of (or interdependent with) our interpretations of it, and therefore it is possible to work towards a closer understanding of how programmes can cause change. According to a realist perspective, programmes can be seen as theories incarnate; when a programme is implemented it is testing a theory about what actions can help to bring about change (Westhorp et al., 2011). In practice, this means that a realist evaluation sets out to test a Middle Range Theory (MRT), detailing how the mechanisms initiated by a programme should cause desired outcomes. Programmes themselves are viewed as being akin to open systems in which there are always multiple and competing mechanisms which interact with the surrounding context to produce outcomes. Pawson and Tilley (1997) sum this up as “mechanisms + context = outcomes”.

As all mechanisms interact with context, when replicated in new environments programmes cannot be expected to achieve the same outcomes. Realist evaluation is designed to address the question of causation and find out “what works, how, in which conditions and for whom” (Pawson and Tilley, 1997). In order to do this, realist evaluators are asked to consider how underlying mechanisms are likely to interact with historical and cultural context, location, economic and political structures, participants, and so on, to produce varying outcomes. The construction of the Middle Range Theory will already have drawn on background research to establish prevailing theory and relevant previous experiences. Evaluators now consider the nature of a planned programme, the target population, and the contexts and settings in which the programme will operate in order to map out a series of conjectural mini-theories called Context Mechanism Outcome (CMO) configurations which relate the various contexts of a programme to the multiple mechanisms by which it might function to produce various outcomes.

Data collection utilising both quantitative and qualitative data sources is then used to build up a picture of the programme in action, identifying how mechanisms are operating in reality in order to revise, substantiate or invalidate conjectural CMOs. Realist evaluation draws on a generative notion of causation which involves an iterative process of theory building, testing and refinement which allows causal statements about attribution to be made. Evaluation findings should demonstrate what worked, for whom, how, and in what circumstances. These findings can then be used to further refine existing Middle Range Theories, which can then themselves be tested by future evaluations.

Methodology

Realist Evaluation is more of a logic of evaluation than a methodology. Although Pawson and Tilley set out their core ideas in considerable detail, the absence of a step-by-step framework means that Realist Evaluation has been interpreted quite widely. The brief steps outlined below reflect our interpretation of Pawson and Tilley's ideas and take into account those of other authors who have drawn on Realist Evaluation, in particular those of Marchal et al. (2010a; 2010b, forthcoming).

- 1) Theory and hypothesis formulation.** Research is carried out to establish the prevailing Middle Range Theory (MRT), drawing on relevant previous programme experiences and evaluations, documentation, social- science literature, and consultation with programme implementers. This research should also establish programme goals, the contexts in which it will operate.

Realist evaluation assumes that the workings of a particular programme or intervention can be explained by an underlying theory, known as the Middle Range Theory (MRT), which explains how a particular programme should function. Background research drawing on relevant previous programme experiences and evaluations, documentation, social-science literature and consultation with programme implementers should help to establish the nature of a social programme, the target population and the contexts and settings in which it will operate. This information should inform the data collection and analytical tools to be used and enables an evaluator to map out a series of conjectural mini-theories called CMOs. A CMO has three constituent parts: a Context, a Mechanism and an Outcome. The *context* signifies the precise circumstances into which a particular intervention is introduced, and the *mechanism* is the precise way in which this measure works within a given context to produce a particular observable *outcome*. The CMO configurations for a given intervention bring together the different programme contexts with the multiple potential mechanisms which together might produce various outcomes. In Pawson and Tilley's (1997) discussion of how CCTV installation in car parks could produce various outcomes, depending on context, they set out various potential CMO configurations, for example:

- (1) Thieves are deterred (*context*) by the presence of a camera (*mechanism*); therefore crime goes down (*outcome*).
- (2) In a car park which is far away from response teams (*context*), the presence of a camera (*mechanism*) may not deter thieves, as they will be able to escape before response teams can arrive and therefore crime will not go down (*outcome*).

- 2) Data collection.** Quantitative and qualitative research builds up a picture of the programme in action. Documentary evidence, direct observation, surveys, interviews, focus groups, quantitative data, etc. may all be used. The aim should be to increase understanding of the different CMO configurations which have been triggered by the programme. Data collection should be designed in such a way as to collect information which can refine, refute or demonstrate how conjectural CMO configurations have operated in practice.

- 3) Data analysis and conclusions.** Realist theory expects there to be different outcome patterns for different groups of actors or contexts within a given

programme. Analysis should aim to understand how mechanisms have operated in programme contexts to generate a set of outcomes – which CMO configurations were substantiated, which were invalidated, and which need to be revised. This does not represent “final” knowledge as such, but the conclusions reached should demonstrate for the programme in question what worked, for whom and how. These findings can then be incorporated into the existing MRT and the evaluation cycle restarted so that theory can be further refined.

Useful Resources

Blamey, A. and Mackenzie, M. (2007) Theories of change and realistic evaluation: peas in a pod or apples and oranges? *Evaluation*, October 2007, 13 (4): 439–455.

Gill, M. and Spriggs, A. (2002) The development of Realistic Evaluation Theory through the evaluation of national crime prevention programmes. *Crime Prevention Studies*, 10: 179–199.

Holma, K. and Kontinen, T. (2011) Realistic evaluation as an avenue to learning for development NGOs, *Evaluation*, 17:181–192.

Marchal, B., Dedzo, M. and Kegels, G. (2010) Turning around an ailing district hospital: a realist evaluation of strategic changes at Ho Municipal Hospital (Ghana). *BMC Public Health*, 10: 787.

Marchal, B., Dedzo, M. and Kegels, G. (2010) A realist evaluation of the management of a well performing regional hospital in Ghana. *BMC Health Services Research*, 10: 24.

Marchal, B., Van Belle, S., van Olmen, J. Hoerée, T. and Kegels, G. (forthcoming) *A Review of Realist Evaluation Studies in the Field of Health Systems Research*.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: SAGE Publications.

Pawson, R. and Tilley, N. (2004) *Realist Evaluation*. London: SAGE Publications.

Online at <http://www.communitymatters.com.au/gpage1.html>

Pawson, R. (2008) Causality for beginners, in ‘NCRM Research Methods Festival 2008’ (unpublished).

Tilley, N. (2000) *Realistic Evaluation: An Overview*.

Westhorp, G., Prins, E., Kusters, C., Hultink, M., Guijt, I. and Brouwer, J. (2011) Realist Evaluation: An Overview. Report from an Expert Seminar with Dr Gill Westhorp.

Online at http://www.cdi.wur.nl/NR/rdonlyres/66764817-54E3-4DCB-BD4F-806B403F892C/143307/2011_WP_RealistEvaluationSeminar_CecileKusters.pdf

1.2 General Elimination Methodology (GEM)

Overview

Scriven's GEM (2008) builds upon his earlier Modus Operandi Method (1976) to provide an approach specifically geared towards substantiating causal claims. The methodology entails systematically identifying and then ruling out alternative causal explanations of observed results. It is based on the idea that for any event it is possible to draw up *Lists of Possible Causes* (LOPCs) or alternative hypothetical explanations for an outcome of interest. Each putative cause will have its own set of "footprints", or *Modus Operandi* (MO) – "a sequence of intermediate or concurrent events, a set of conditions or a chain of events that has to be present when the cause is effective" (Scriven, 2008). For example, a criminal investigation might be able to identify a criminal from a list of suspects by examining the means, motives and opportunity pertaining to each of them.

GEM sets out to identify potential causes of effects by examining the facts of a case and establishing which MOs are present and which are not. Any cause for which the Modus Operandi is not present can be dismissed, leaving only causal explanations that have a genuine causal link. General Elimination Methodology is intended to provide a framework for evaluation which can establish causal claims beyond reasonable doubt. A list of suggested ways of doing this is provided, and these include critical observation, interviews, theoretical inference, quasi-experiments and cross-sectional data, among others (see Scriven, 2008 for the full list).

Methodology

- 1) A List of Possible Causes or competing explanations for an event, outcome or set of outcomes which are consistent with the circumstances of the evaluation should be drawn up.
- 2) For each Possible Cause, outline the Modus Operandi which will be present if that cause is found to have had a causal influence. Each Possible Cause should have a distinct set of footprints which would allow an evaluator to identify it.
- 3) Systematically establish the "facts of the case" in order to demonstrate whether the MO for each Possible Cause is present or not. Key evidence likely to prove or disprove each Possible Cause should be sought out. The logic here is two-fold: identifying elements of a Modus Operandi which are present provides evidence that a Possible Cause might have been an actual cause; identifying elements of Modus Operandi which 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.

Useful Resources

Patton, M. Q. (2008) Advocacy impact evaluation. *Journal of Multi-Disciplinary Evaluation*, 5 (9), March 2008.

Scriven, M. (1976) Maximising the power of causal investigations: the Modus Operandi method, in Gene V. Glass (ed.) *Evaluation Studies Review Annual, Vol. 1*. London: Sage Publications.

Scriven, M. (2005) Can We Infer Causation from Cross-sectional Data? Paper presented at a Symposium on the Use of School-Level Data for Evaluating Federal Education Programmes, Washington, DC. Online at http://www7.national-academies.org/BOTA/School-Level%20Data_Michael%20Scriven-Paper.pdf

Scriven, M. (2008) A summative evaluation of RCT methodology and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, 5 (9): 11–24.

1.3 Process Tracing

Overview

Process Tracing sets out to “unwrap” the causal links between putative causes and outcomes by identifying the intervening causal processes or mechanisms at work (George and Bennett, 2005; Reilly, 2010). The evaluator starts by carrying out “process induction”, drawing on evidence to generate a number of (preferably competing) hypotheses about how an intervention may connect to an outcome. Evaluators should set out a series of hypothetical causal mechanisms which might be initiated by the programme, together with what should be observed if each hypothesis is true or false. This involves identifying a series of “diagnostic” pieces of evidence which will be present for each of the theoretically predicted links for each hypothetical explanation (the causal chain) if they are observed in practice. Research (usually in the form of in-depth case-study analysis, drawing largely on qualitative data including historical reports, interviews and other documents, but also able to make use of any evidence available including quantitative data) is then used to develop an explicit chronology of actual events, setting out the causal links between each stage. This evidence is then used to overturn or substantiate rival hypothetical explanations (“process verification”), the ultimate goal being to establish whether the actual mechanisms at work fit with those predicted (Bennett, 2010).

Bennett (2010) sets out four alternative tests that can be used to assess the strength of alternative hypotheses: “Straw in the wind” tests, which can provide evidence for or against a hypothesis but by themselves cannot confirm or deny it; “Hoop tests”, which, if passed, can affirm the relevance of a hypothesis but cannot fully confirm it and, if failed, eliminate a hypothesis; “Smoking Gun” tests, which can confirm a hypothesis if passed, or weaken it if failed; “Doubly Decisive Tests”, which in confirming a given hypothesis eliminate any others. By assessing each of the hypotheses drawn from theory, the aim is to be able to make strong causal claims about what mechanism(s) caused a given set of outcomes in any given case.

Methodology

- 1) Process induction.** There should be a body of evidence available to an evaluator concerning how a given programme should bring about change; social-science theory, documentation, and past evaluations and programme experiences should all indicate how a programme may function in practice. Process induction involves drawing on evidence to generate a set of (preferably competing) hypotheses or putative explanations linking cause and effect. Each hypothetical explanation should detail the processes or causal mechanisms that should be observed if they function as expected. This also involves setting out what diagnostic patterns of evidence will be observed if each causal mechanism is to be shown to be valid, and also what might be observed for it to be proved false. Consideration should be given to the types of evidence that might be most helpful for this process.
- 2) Data collection / Chronology.** Process Tracing generally relies largely on qualitative data, although quantitative data may also be brought to bear. The aim should be to understand what actual processes or mechanisms have been generated by the

programme. Often this involves the construction of an explicit chronology of the process under investigation. This should take the form of a narrative, detailing the events that have occurred and placing them within an analytical framework which sets out their causal significance.

- 3) Process verification.** The final step involves examining the evidence for congruence or incongruence between actual observations and those predicted by hypothetical explanations. The aim should be to systematically analyse each link in the causal chain to see how it was reached. Evaluators can employ Bennett's four tests (see above) in order to substantiate or invalidate the causal claims of rival hypotheses. This may involve seeking out additional data in order to triangulate findings and increase certainty regarding conclusions.

Useful Resources

Bennett, A. (2008) Process-tracing: a Bayesian perspective, in Janet M. Box-Steffensmeier, Henry E. Brady and David Collier (eds.) *The Oxford Handbook of Political Methodology*, pp. 702–721. Oxford: Oxford University Press.

Bennett, A. (2010) Process tracing and causal inference, in Henry Brady and David Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield.

Checkel, J. T. (2006) *It's the Process Stupid! Process Tracing in the Study of European and International Politics*, Arena Centre for European Studies, University of Oslo. Online at http://www.sv.uio.no/arena/english/research/publications/arena-publications/workingpapers/working-papers2005/wp05_26.pdf

Checkel, Jeffrey T. (2008) Tracing causal mechanisms. *International Studies Review*, 8 (2): 362–370.

Collier, D. (2011) Understanding process tracing. *Political Science and Politics*, 44 (4): 823–830.

George A. L. and Bennett, A. (2005) *Case Studies and Theory Development in the Social Sciences*, MIT Press.

Hughes, K. and Hutchings, C. (2011) Can We Obtain the Required Rigour Without Randomisation? Oxfam GB's Non-experimental Global Performance Framework. 3ie Working Paper No. 13. Online at http://www.3ieimpact.org/admin/pdfs_papers/WP%2013_Final.pdf

Reilly, R. (2010) Process tracing, in A. J. Mills, G. Durepos and E. Wiebe (eds.) *Encyclopedia of Case Study Research*. London, SAGE Publications.

1.4 Contribution Analysis

Overview

Contribution Analysis is an approach to evaluation developed by Mayne (Mayne, 2001, 2008, 2011) which aims to compare an intervention's postulated theory of change against the evidence in order to come to robust conclusions about the contribution that it has made to observed outcomes. The aim is to critically construct a "contribution story" which builds up evidence to demonstrate the contribution made by an intervention, while also establishing the relative importance of other influences on outcomes. Contribution Analysis draws on the idea that an intervention's theory of change can be used to infer causation by assessing whether the mechanisms or processes that it aims to initiate have in fact occurred. It sets out to demonstrate a plausible association between a programme and observed outcomes using weight of evidence – by building a credible contribution story in which each step lying between programme inputs and outcomes is clearly evidenced, with the result that a "reasonable person, knowing what has occurred in the programme and that the intended outcomes actually occurred, agrees that the programme contributed to these outcomes" (Mayne, 2001). A plausible association can be said to have made if the following criteria are met: (1) a reasoned theory of change is set out; (2) the activities of an intervention are shown to have been implemented as set out in the theory of change; (3) the chain of expected results can be shown to have occurred; and (4) other influencing factors have either been shown not to have made a difference, or their relative contribution has been recognised. Where interventions are more complex, involving numerous sub-elements or simultaneous causal strands or where, for example, causality is recursive or outcomes emergent, Mayne suggests that multiple causal strands may be developed and each of these summarised within a general theory of change which incorporates them all.

Mayne also distinguishes between minimalist Contribution Analysis, Contribution Analysis of Direct Influence, and Contribution Analysis of Indirect Influence. Minimalist analysis bases causal claims on the fact that a theory of change was clearly developed and that evidence was found to show that the expected outputs were delivered. Analysis of Direct Influence builds on minimalist analysis by setting out to confirm that direct results predicted by a programme's theory of change, such as changes in knowledge, skills or attitude, were observed and that the programme was influential in causing those changes, taking account of other factors. Analysis of Indirect Influence goes further by examining whether the theoretically predicted indirect influences of a programme were observed, and the extent to which they can be attributed to that programme in the light of other factors.

Mayne does not set out specific methods to be used within the framework of Contribution Analysis which he sets out, instead advocating that an evaluator should use a mix of available qualitative and quantitative methods in order to provide comprehensive evidence. Analysis should be carried out iteratively, with new evidence sought out to strengthen the contribution story and increase understanding of how outcomes occurred. Where settings are complex and there are multiple "arms" or elements to interventions, separate contribution stories should be developed for each arm, as well as for the intervention as a whole.

Methodology

Clear methodological steps for carrying out contribution analysis are set out by Mayne (2001), with a recent revision (2011) updating them to explicitly adapt Contribution Analysis for more complex interventions. In this latest work Mayne sets out seven iterative steps, each of which should build the contribution story and address any weaknesses identified by the previous stages:

- 1) Set out the cause–effect question(s) which must be addressed.
- 2) Draw up a carefully reasoned theory of change, identifying potential influencing factors and outlining the different links in the theory of change and the risks and assumptions associated with them.
- 3) Gather existing evidence on the theory of change (i) for observed results, (ii) for each of the links in the results chain, and (iii) for the other influencing factors.
- 4) Assemble and assess the contribution story, outlining whether an intervention was implemented as planned, what the role of external factors was, and whether the predicted theory of change and expected results occurred.
- 5) Seek out additional evidence to reinforce the credibility of the contribution story.
- 6) Revise and strengthen the contribution story.
- 7) In complex settings, assemble and assess the complex contribution story.

Useful Resources

Kotvojs, F. (2006) Contribution Analysis: A New Approach to Evaluation in International Development. Paper presented at the Australian Evaluation Society 2006 International Conference, Darwin.

Larbi, G., Christensen, J., Jackson, P., Ura, K. (2006) *Capacity Development in Bhutan: Capacity Development Outcome Evaluation of Danish Supported Organisations in Bhutan*. Copenhagen: Danida. Online at <http://www.oecd.org/dataoecd/45/9/42218416.pdf>

Lemire, S. (2010) Contribution Analysis: The Promising New Approach to Causal Claims. Paper presented at European Evaluation Society International Conference, Prague 2010. Online at http://www.europeanevaluation.org/images/file/Conference/Past_Conference/2010_Prague/FullPapers/3_Lemire_Sebastian.pdf

Mayne, J. (2001). Addressing attribution through contribution analysis: using performance measures sensibly. *Canadian Journal of Programme Evaluation*, 16: 1–24.

Mayne, J. (2003) *Reporting on Outcomes: Setting Performance Expectations and Telling Performance Stories*. Office of the Auditor General of Canada.

Mayne, J. (2008) Contribution Analysis: An Approach to Exploring Cause and Effect. Institutional Learning and Change Brief No. 7. Online at http://www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief16_Contribution_Analysis.pdf

Mayne, J. (2011) Contribution analysis: addressing cause and effect, in K. Forss, M. Marra and R. Schwartz (eds.) *Evaluating the Complex*. New Brunswick, NJ: Transaction Publishers.

1.5 Most Significant Change (MSC)

Overview

Most Significant Change (Davies and Dart, 2005) is a form of participatory monitoring and evaluation that involves the collection and selection of stories of *significant changes* which have occurred in the field. “It is participatory because project stakeholders are involved in deciding the sorts of changes or stories of significant change to be recorded and in analysing the data collected. It is a form of monitoring because it occurs throughout the programme cycle and provides information to help people manage the programme. It contributes to evaluation by providing data on short-term and long-term outcomes that can be used to help assess and improve the performance of the programme as a whole” (Davies and Dart, 2005). MSC was first developed to help NGOs to monitor the impacts of participatory development projects by providing a technique flexible enough to identify development outcomes across a variety of locations and which emphasised the need to respect participants’ own judgement regarding the changes that an initiative has made to their lives (Davies, 1998). Davies and Dart set out clear steps for using the approach in their MSC guide (2005).

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

A key step in MSC is the process by which the most significant of the “significant change stories” are selected. After stories of significant change have been collected, they are then passed upwards through the organisational hierarchy and systematically reviewed by panels of designated stakeholders. The intention is for stakeholders to engage in in-depth discussion at each stage of the selection process regarding the significance of each story, the wider implications of the changes that they relate, and the quality of evidence which they contain. The use of multiple levels of selection enables large numbers of significant change stories to be reduced to a smaller number of stories viewed as being most significant by a majority of stakeholders. Selected stories should be verified for authenticity by drawing on participants’ own experiences or through direct observation.

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

MSC has since been adapted for use in impact evaluations, by expanding the scale of story collection, extending the range of stakeholders engaged in story selection, and using it alongside other evaluation methods, such as in tandem with a log-frame/theory-

of-change approach (for example, Van Ongevalle et al., n.d.). The stories of significant change which MSC generates can provide useful sources of information for the specification and subsequent assessment of a theory of change.

Methodology

Davies and Dart's *Guide to MSC Use* (2005) provides an in-depth manual for MSC. We briefly summarise below the steps that they set out.

- 1) Getting started.** "Champions" are appointed to raise organisational interest and identify how MSC can be implemented throughout an organisation. Past programme examples are used to show how the technique can be effective and to show the type of results that it can produce.
- 2) Establishing "domains of change".** Reported Significant Changes (SCs) should be categorised within broad categories to manage analysis and guide those collecting stories without being too prescriptive. Examples of domains include changes in the quality of people's lives and the nature of people's participation in development activities, and changes in the sustainability of organisations and activities.
- 3) Defining the reporting period.** Organisations often start by reporting more frequently (every three months, for example) and then, as any backlog of stories is cleared, settling into a regular rhythm (for example, bi-annually).
- 4) Collecting stories of change.** The central MSC question is: *"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 question can be adapted to time periods, domain of change, etc.:
 - Fieldworkers can record unsolicited stories and carry out interviews or group discussions, or beneficiaries can be asked to write up or record the stories themselves.
 - Each story should clearly set out why it is significant from the storyteller's point of view. Information about who collected/provided the story and when events occurred should be provided. Stories should be short but contain enough detail to make them comprehensible to all readers. Many organisations develop their own standardised form.
- 5) Reviewing the stories.** A group process selects the most significant stories at successive hierarchical levels until the most salient stories emerge. Reviewers read the stories in groups or separately, and then hold in-depth discussions about which are most significant. Selection can be through majority, iterative voting, scoring, pre-scoring followed by a group vote, or through secret ballot.
 - The reasons for choices made should be documented, as transparency is crucial.
 - Stories that are filtered out should be kept on file for later review and content analysis.
- 6) Feedback regarding the review process.** The results of the selection process should be fed back to those who provided the stories. This can aid the future search for stories by expanding or challenging views of what is significant or of value.

- 7) **Verify the stories if necessary.** A verification step should seek out confirmation of reported stories, checking that they have been reported accurately and honestly, and seeking out more information about experiences which are considered to be especially significant.
- 8) **Quantification.** Although MSC places a strong emphasis on qualitative research, it also sets out steps for the quantification of changes. Quantitative information can include the number of people affected by a change, the activities that took place, and their effects. Checks can ascertain whether reported changes from stories have affected others too. Secondary analysis of stories can assess the number of times a type of change occurred.
- 9) **Conducting secondary analysis.** Secondary analysis involving the examination, classification and analysis of the content and themes in the stories can be useful for management of programmes. For example, meta-monitoring of trends in reporting, examination of whether recommendations have been acted upon, and coding to see what topics come up most often and to compare actual with expected outcomes.
- 10) **Revising the MSC process.** Revising MSC implementation can make it more efficient and inclusive. Examples of revisions include the domains of change, the structure of meetings for selecting most significant stories, frequency of reporting, etc.

Useful Resources

Davies, R. (1998) An evolutionary approach to facilitating organisational learning: an experiment by the Christian Commission for Development in Bangladesh. *Impact Assessment and Project Appraisal*, 16 (3): 243–50.

Davies, R. and Dart, J. (2005) *The “Most Significant Change” (MSC) Technique: A Guide to Its Use*, April 2005. Online at <http://mande.co.uk/docs/MSCGuide.pdf>

Davies, R. and Delaney, T. (2011) *Connecting Communities? A Review of World Vision’s Use of MSC*. Cambridge and Melbourne, March 2011. Online at <http://mande.co.uk/blog/wp-content/uploads/2011/06/Connecting-communities-Main-Report-April-2011-RD-TD.pdf>

India HIV/AIDS Alliance (2007) *Stories of Significance: Redefining Change. A Report Based on an Evaluation of a Programme on “Community Driven Approaches to Address the Feminisation of HIV/AIDS in India” by Means of the “Most Significant Change” Technique*. The India HIV/AIDS Alliance (Alliance India), July 2007.

Keriger, R. (2004) *MSC Guide: Based on the Experience of ADRA Laos: A Guide to Implementing the Most Significant Changes (MSC) Monitoring System in ADRA Country Offices*. Adventist Development and Relief Agency (ADRA). Available from the MSC mailing list file repository at <http://groups.yahoo.com/group/MostSignificantChanges/>

Kotvojs, F. and Lasambouw, C. (2009) *MSC: Misconceptions, Strengths and Challenges*. Presented at the Australasian Evaluation Conference, September 2009. Online at <http://mande.co.uk/blog/wp-content/uploads/2010/08/2009-MSC-misconceptions-Kotvojs.pdf>

Sigsgaard, P. (2002) Monitoring without indicators: an ongoing testing of the MSC approach. *Evaluation Journal of Australasia*, 2 (1): 8–15.

Willetts, J. (2004) *Most Significant Change Pilot Project: Evaluation Report*, prepared by the Institute for Sustainable Futures, Sydney, ADRA Laos. Available from the MSC mailing list file repository at <http://groups.yahoo.com/group/MostSignificantChanges/>

Willetts, J. and Crawford, P. (2011) The most significant lessons about the Most Significant Change technique. *Development in Practice*, 17 (3), June 2007.

The Most Significant Change approach egroup/online forum:
<http://groups.yahoo.com/group/MostSignificantChanges/>

1.6 Success Case Method (SCM)

Overview

The Success Case Method (Brinkerhoff, 2003) is a narrative technique based upon naturalistic inquiry and in-depth case-study analysis. It is intended to be a quick and simple evaluation process geared towards understanding whether an initiative (such as a training or educational programme) is actually working. SCM sets out to discover whether an intervention is working or not by searching for particularly successful or unsuccessful instances (“success” and “non-success” cases). The Success Case Method does not set out to find out about the “average” participant, but instead intentionally seeks out the very best (and worst) that a programme has produced, in order to understand the contribution that the programme has made to results, the role that contextual factors have played in influencing the different outcomes, and the way in which this information can be used to improve programme performance. The SCM “searches out and surfaces successes, bringing them to light in persuasive and compelling stories so that they can be weighed... provided as motivating and concrete examples to others, and learned from so that we have a better understanding of why things worked and why they did not”.

Either using already available data or, more often, by means of a survey, the SCM identifies individual examples that have been particularly successful (and unsuccessful) and uses case-study analytical methods to develop credible arguments about the contribution that the intervention made to them. The first step involves setting out the purpose of the study and outlining practical and time constraints for the evaluation. An *impact model* should then be drawn up, detailing the behaviours and positive outcomes that should ideally be achieved if the programme functions as intended. A brief survey is then administered to participants with the intention of eliciting best and worst cases. One-to-one interviews with success and non-success cases then determine specifics of intervention usage and results, the perceived value of the results, and environmental factors that enabled or prevented success.

Methodology

- 1) **Evaluation goals.** The first step is to identify key stakeholders and their interests and then propose an evaluation plan which sets out a realistic study timeframe and clearly defines its purpose.
- 2) **“Impact model”.** The next step involves outlining what desired programme performance should be. The impact model should set out what behaviours and positive outcomes should be achieved if the programme being evaluated is functioning well.
- 3) **Survey.** The next step involves sending a survey to participants, with the intention of eliciting best and worst cases. The survey should set out to find out whether the programme has achieved its intended outcomes. Questions should be tailored to the individual purpose of a given programme but broadly should aim to find out whether the intervention was correctly implemented, whether participants understood it, and whether it produced the intended outcomes. Cases should be coded as high (H), that

is, success cases; moderate (M), or average cases; or low (L), that is, failure or non-success cases.

- 4) Interviewing and documenting success cases.** Analysis of survey results should allow an evaluator to select a purposive random sample of success and non-success cases. One-to-one interviews with success cases should then determine specifics of intervention uptake and results, perceived value of the results, and the environmental supports that enabled success. One-to-one interviews with non-success cases are employed to determine why they were not able to apply or benefit from the intervention, and to identify barriers to success. Interviews should be open-ended inasmuch as they should be flexible enough that interviewees should be able to “lead” the interviewer to unexpected but useful information. At the same time, interviews should also be carefully structured to ensure that the success/non-success story is credible and all stages or links in the stories are understood. Where necessary, the evaluator should seek corroborating information from third parties and other data sources in order to triangulate findings.
- 5) Findings.** Conclusions and recommendations should be presented in the form of in-depth stories, setting out the details of the most compelling cases so that findings can be confronted, the implications discussed, and a plan of response put together.

Useful Resources

Bell, C. and McDonald, D. (2007) The Success Case Method: A Simple Evaluation Tool to Identify Critical Success Factors and Programme Impact. Evaluation in Emerging Areas: Australasian Evaluation Society International Conference, Darwin, 4–7 September 2006. Online at

<http://www.aes.asn.au/conferences/2006/papers/063%20Catherine%20Bell.pdf>

Brinkerhoff, R. O. (2003) *The Success Case Method: Find Out Quickly What's Working and What's Not*. San Francisco, CA: Berrett-Koehler.

Brinkerhoff, R.O. (2008) *Telling Training's Story: Evaluation Made Simple, Credible, and Effective*. San Francisco, CA: Berrett-Koehler.

Coryn, C., Schröter, D. C. and Hanssen, C. E. (2009) Adding a time-series design element to the success case method to improve methodological rigor: an application for non-profit programme evaluation. *American Journal of Evaluation*, 30 (1): 80-92.

1.7 Outcome Mapping (OM)

Overview

Outcome Mapping (Earl, Carden and Smutylo, 2001) is a participatory approach for planning, monitoring and evaluating development programmes which was developed by the International Development Research Centre (IDRC). “The process of outcome mapping is intended to help a project team or programme to be specific about the actors it targets, the changes it expects to see and the strategies employed to achieve this” (Smutylo, 2005). It aims to help programmes to report and learn from outcomes in a realistic manner by tracking the intervening steps between programme activities and programme outcomes. Development results are measured not in terms of “impact” but instead in terms of the programme’s contribution to changes in the behaviour of, and relationships between, the actors with whom the programme interacts directly.

Outcome Mapping can be split into three stages: (1) an “Intentional Design” outlines programme priorities, setting out programme goals, programme activities and potential partners, and sets out how to promote progress towards anticipated results; (2) “outcome and performance monitoring” provides a framework for ongoing assessment of a programme’s actions and the progress of boundary partners (all actors with which the programme interacts directly) towards desired goals/outcomes; mixed methods are used to gather data for an outcome journal which monitors boundary-partner actions and relationships, a strategy journal which records programme strategies and activities, and a performance journal which monitors organisational practices designed to keep the programme relevant and viable; (3) “evaluation planning” prioritises what elements of the programme are to be evaluated and works out the logistics for the evaluation to be carried out.

Methodology

Outcome mapping is designed to introduce monitoring and evaluation at an early stage, with the goal of linking them with the programme’s ongoing management. The planning process should be a participatory process involving boundary partners. Clear methodological steps for Outcome Mapping are set out by Earl, Carden and Smutylo (2001) and are briefly summarised here.

Outcome Mapping is split into three stages, each of which has a number of smaller steps.

1) Intentional Design sets out the programme priorities. These goals provide reference points to guide strategy formulation and action plans (rather than acting as performance indicators).

- **1: Programme vision:** presenting the broad and long-term economic, political, social or environmental changes that the programme is intended to bring about, as well as the desired behaviour changes in key boundary partners.
- **2: Mission statement:** outlining how the programme hopes to achieve the desired outcomes (without listing all activities).
- **3: Identification of boundary partners:** including all groups with whom the programme works directly and who may contribute to the achievement of

programme goals. They can include local communities/beneficiaries, government officials and policy makers, NGOs, private-sector, academic, and research institutions, international institutions, the private sector, etc.

- **4: Outcome challenges:** setting out how the behaviour, relationships, activities, or actions of each boundary partner should change if a programme is successful.
- **5: Progress markers:** developing for each boundary partner graduated indicators for tracking changed behaviours: indicators which focus on the depth or quality of change that identify the incremental changes that the programme can realistically hope to influence.
- **6: Strategy map** for each outcome challenge: setting out strategies used by the programme to contribute to the achievement of an outcome. The map should be divided into *causal, persuasive and supportive* behaviours. It should help to clarify the mix of programme strategies employed, indicate the relative effect that the programme is likely to have on different boundary partners, pinpoint strategic gaps, and help to provide key clues as to how to evaluate the success of the programme.
- **7: Organisational practices:** setting out the steps needed to deliver each programme activity and ensure future continuity and ongoing programme maintenance. There should be ongoing consideration of how programmes can innovate, seek feedback and maintain support for their work.

2) Outcome and performance monitoring: this framework builds on the progress markers, strategy maps and organisational practices developed in the “intentional design” stage.

- **8: Monitoring priorities,** set to focus data collection. Both qualitative and quantitative methods are used to gather data, which may include a document review, interviews, focus groups, workshops, etc. and should encompass boundary-partner as well as staff inputs. Three tools help the team to reflect on the data they have collected and how they can be used to improve performance:
- **9: Outcome journal** monitors boundary-partner actions and relationships.
- **10: Strategy journal** monitors strategies and activities.
- **11: Performance journal** monitors the organisational practices that keep the programme relevant and viable.

3) Evaluation planning: this should help the team to set priorities so that they can target evaluation resources and activities where they will be most useful.

- **12: Evaluation plan:** describing key evaluation questions, resources, methods, the team, time-line, how findings will be used and by whom.

Useful Resources

Ambrose, K. and Roduner, D. (2009) *A Conceptual Fusion of the Logical Framework Approach and Outcome Mapping*. Outcome Mapping Ideas Paper No. 1, May 2009.

Armstrong, J., Carden F., Coe, A. and Earl, S. (2000) *International Model Forest Network Secretariat (IMFNS): Outcomes Assessment*, July 2000.

- Earl, S., Carden, F. and Smutylo T. (2001) *Outcome Mapping: Building Learning and Reflection into Development Programmes*. International Development Research Centre (IDRC).
- Earl, S. and Carden, F. (2002) Learning from complexity: the International Development Research Centre's experience with outcome mapping. [Development in Practice](#), 12 (3/4), August 2002.
- Howard, G., Jeger, M. and Wilson-Grau, R. (2011) *BioNET 2007–2010: An Outcomes Evaluation and Assessment of the Prospects for BioNET to Increase its Impact on Food Security, In Particular through Greater Support to Plant Health Systems*, February 2011.
- Jones, H. and Hearn, S. (2009) *Outcome Mapping: a Realistic Alternative for Planning, Monitoring and Evaluation*. Overseas Development Institute (ODI) Background Note, October 2009.
- Majot, J., Richert, W. and Wilson-Grau, R. (2010) *Evaluation of Oxfam Novib's Global Programme 2005–2008*, March 2010.
- Outcome Mapping Learning Community website: <http://www.outcomemapping.ca/>
- Roduner, D. and Schläppi, A. (2008) *Logical Framework Approach and Outcome Mapping: A Constructive Attempt of Synthesis*. Zurich, April 2008.
- Smutylo, T. (2005) Outcome Mapping: A Method for Tracking Behavioural Changes in Development Programmes. Institutional Learning and Change (ILAC) Brief 7, August 2005. Online at <http://web.idrc.ca/uploads/user-S/11235064481Brief-FINAL.pdf>
- Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (n.d.) *Analysing Outcome Mapping Monitoring Data: The Case of the Quality Education and Vulnerability Programme in Zimbabwe 2008–2013*. Working paper. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>
- Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (2009a) Monitoring for Impact in a Programme's Sphere of Influence – a Case Study of the Quality Education and Vulnerability Programme in Zimbabwe. Presented at the Afrea (African Evaluation Association) Conference, 29 March–3 April 2009, Cairo, Egypt. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>
- Van Ongevalle, J., Chipimbi, R. and Sibanda, M. (2009b) Monitoring for Impact in a Programme's Sphere of Influence – a Case Study of the Quality Education and Vulnerability Programme in Zimbabwe. Presented at the Afrea (African Evaluation Association) Conference, 29 March–3 April 2009, Cairo, Egypt. Available from the Outcome Mapping Learning Community file repository at <http://www.outcomemapping.ca/>

1.8 Method for impact Assessment of Programs and Projects (MAPP)

Overview

MAPP (Neubert, 2000 and 2010) combines a qualitative approach utilising participatory assessment tools with a quantification step. Originally developed for social analysis of the outcomes produced by resource-management projects in Sub-Saharan Africa, MAPP can also be applied to other types of programme and in other regions. It is essentially a “before and after” design, intended to compare the outcomes produced by a programme, as recalled by target groups. MAPP sets out to select only enough relevant data for conclusions to be reached and is intended to be a cost- and time-efficient evaluation process.

MAPP uses a series of tools to evaluate the perceived influence of the development project on the daily life of the population, taking into account the project's context. It then sets out to establish the reasons for changes, whether they be the result of project-specific measures, initiatives organised independently by the local population, the result of intervention on the part of other donors, or determined by further external factors. Instead of monitoring programme results or performance using causal chains, MAPP aims to capture a mix of effects, positive and negative, intended and unintended. It sets out to be parsimonious in its information gathering as far as possible, selecting only relevant data, in order to avoid information overkill. The various MAPP tools are intended to triangulate findings to improve validity.

Methodology

A major part of the methodology involves conducting structured group discussions with relevant stakeholders. Applications of MAPP to agricultural programmes have involved groups of benefiting and non-benefiting farmers, together with a range of other stakeholders including programme managers. As representatives of stakeholder groups are encouraged to take part in the workshops, mutual control or validation of individual statements and assessments is an inbuilt part of the evaluation process. Usually workshops should be conducted in various project regions – where results appear to be good, bad, or mediocre and where projects have been running for a range of time periods.

These group discussions provide the basis for a number of interlinked tools. Most of these tools use a rating system, which makes it possible to quantify and aggregate the originally qualitative results.

1. **Life line:** outlines development trends, with a five-point scale for quality of life, setting out how their community has developed over the past 20 years or so. Drawing up the lifeline and discussing key changes will give an indication of potentially important external influencing factors.
2. **Trend analysis:** a matrix sets out detailed development trends for the following suggested key dimensions: changes in living standards, access to resources, expansion of knowledge, participation on rights and power. Changes are tracked

using a five-point scale from *very negative* to *very positive* for the period covered by the programme (or longer), and the reasons for variations in trend lines are discussed and recorded.

3. **Transect cross-checking:** all visible programme measures are inspected for their effectiveness, for example magnitude, soundness. Practical tools, such as transect walks (which note the location and distribution of resources, features, landscape, main land uses, etc.), can provide useful information about the scale, relevance and maintenance of interventions as well as poverty dimensions in the community, and can provide a means of comparison for feedback gained from group discussion.
4. **Programme activity list:** all programme activities, as well as all donor and partner organisations active in the community, are listed. The group(s) benefiting from the activity are identified and rated (again, on a five-point scale) according to their relevance to the day-to-day lives of beneficiaries. The activities are also evaluated according to the financial inputs and work required to implement and maintain them. Thus, a programme can be compared with other projects for perceived importance. The list also provides a cost–benefit analysis of activities and examines how they are distributed between beneficiary groups.
5. **Influence matrix:** evaluates the influence of each programme activity (on a five-point scale) on a range of social criteria such as the improvement of living standards, or access to resources.
6. **Development and impact profile:** a chart is drawn up to summarise the findings of the different MAPP tools, with the intention of providing a rapid comparison of the different trends. The summary can be used to analyse whether programme activities have provided stable progress. In addition, the chart can be used to help to draw up a development profile, setting out the relative influence of different stakeholders or other external influences for reported changes.
7. **Participatory development planning:** any indicators that did not show positive results are discussed with stakeholder groups. Then a plan to remove any bottleneck is made, and steps are agreed for its implementation. Members of the community should be assigned to concrete activities such as writing an application for credit.

Useful Resources

Neubert, S. (2000) *Social Impact Analysis of Poverty Alleviation Programmes and Projects: A Contribution to the Debate on the Methodology of Evaluation in Development Cooperation*. GDI Book Series No. 14. London: Routledge.

Neubert, S. (2010) Description and Examples of MAPP Method for Impact Assessment of Programmes and Projects. German Development Institute (GDI)

Online at www.ngo-ideas.net/mediaCache/MAPP/

1.9 Qualitative Comparative Analysis (QCA)

Overview

Qualitative Comparative Analysis (QCA) is a case-oriented comparative approach intended to bridge qualitative and quantitative research by combining in-depth case studies with the identification and interpretation of causal patterns in the cases under examination. Each of the cases examined is conceived of as consisting of a configuration of relevant characteristics or conditions, and it is the combination of these different conditions which is seen as producing a particular outcome. QCA sets out to compare the different combinations of conditions and outcomes pertaining to each case, with the goal of discovering what configurations of conditions lead to what outcomes, and which of those conditions are key in producing certain outcomes.

Typically, QCA looks at a small or intermediate number of cases, although the approach has also been applied to large n research designs as well (Berg-Schlosser, De Meur, Rihoux, and Ragin, 2009). Analysis begins with the selection of a number of different cases for which a specified outcome has or has not occurred. Case selection itself should be based upon in-depth substantive knowledge of cases and theoretical knowledge about the causal conditions relevant to outcomes. To illustrate, we borrow from an example set out by Rihoux and De Meur (2009) in which the authors analyse the survival or breakdown of democratic systems in Europe during the inter-war period. Drawing on theory as well as substantive knowledge, they assess 18 European countries for “Lipset conditions” (wealth, industrialisation, education and urbanisation) along with a further condition, governmental stability.

QCA can actually be thought of both as an underlying logic or approach for research, and as a family of analytical tools or techniques. The term QCA itself is often used as an umbrella term to describe three main variants of analysis. Using conventional Boolean or “Crisp Set” QCA (also known as csQCA), cases are coded using the dichotomous values “0” or “1” to indicate either the absence or presence of a particular condition. “Multi-value” QCA (mvQCA) permits multiple-category nominal-scale conditions such as religion to be coded, but also allows for more refined coding of ordinal data. Finally, “Fuzzy Set” QCA (fsQCA) permits a researcher to code for the presence or absence of a particular causal condition on a continuous scale anywhere between 0 and 1, inclusive.

To begin with, the selected cases are summarised in a data table, with each case reduced to a combination of conditions, along with an outcome. In our extended example from Rihoux and De Meur (2009), “Crisp Set” QCA is used to code each of the conditions and the final outcome dichotomously for each of the cases. In order to do this, thresholds for dichotomisation were drawn up with reference to empirical and theoretical knowledge. For example, *Literacy* was coded 0 if below 75%, 1 if above. This information is then summarised, as in Table 1.

Table 1: Data Matrix

Case ID	GNP per capita	Urbanisation	Literacy	Industrial labour force	Government stability	Survival
AUS	1	0	1	1	0	0
BEL	1	1	1	1	1	1
CZE	1	1	1	1	1	1
EST	0	0	1	0	1	0
FIN	1	0	1	0	1	1
FRA	1	0	1	1	1	1
GER	1	1	1	1	0	0
GRE	0	0	0	0	0	0
HUN	0	0	1	0	0	0
IRE	1	0	1	0	1	1
ITA	0	0	0	0	1	0
NET	1	1	1	1	1	1
POL	0	0	1	0	0	0
POR	0	0	0	0	0	0
ROM	0	0	0	0	1	0
SPA	0	0	0	0	0	0
SWE	1	0	1	1	1	1
UK	1	1	1	1	1	1

Source: Adapted from Rihoux and De Meur (2009: 54)

The next step reduces these data to a “truth table” which summarises the different configurations of conditions associated with a given outcome.

Table 2: Truth Table

Case ID	GNP per capita	Urbanisation	Literacy	Industrial labour force	Government stability	Survival
AUS	1	0	1	1	0	0
BEL, CZE, NET, UK	1	1	1	1	1	1
EST	0	0	1	0	1	0
FRA, SWE	1	0	1	1	1	1
GER	1	1	1	1	0	0
GRE, POR, SPA	0	0	0	0	0	0
HUN, POL	0	0	1	0	0	0
FIN, IRE	1	0	1	0	1	1
ITA, ROM	0	0	0	0	1	0

Source: Adapted from Rihoux and De Meur (2009: 55)

QCA draws on an equifinal conception of causation which holds that it is possible for the same outcome to result from different configurations of conditions, or for a given condition to have a different impact on an outcome, depending on the context (Berg-Schlosser, De Meur, Rihoux, and Ragin, 2009). Ultimately, causal analysis is based upon the examination of the different cases and their different permutations of conditions and outcomes. QCA draws on the logic of “necessity” and “sufficiency” to draw conclusions about which of the conditions are necessary parts of a “causal recipe” – itself sufficient to bring about a given outcome (Ragin, 2008).

In this way, QCA software can be used to strip away elements not causally related to the outcome and reduce the information held in the truth table into minimised expressions of the conditions which lead to either the presence or absence of an outcome. In the example from Rihoux and De Meur (2009), the authors use csQCA software to reduce the information held in the truth tables down to a few parsimonious statements about the causal conditions necessary in determining whether democracy survived or not. For example, they are able to conclude that for countries in which democracy did survive, a combination of high GDP per capita and governmental stability was the key factor in determining this outcome.

QCA will usually produce multiple “causal recipes”, relating to the different conjunctions of causal conditions which produce a given outcome for a certain group of cases. The “causal recipes” produced from the analysis should be examined with reference, once again, to theory and to the case studies themselves. The way in which cases are grouped may also offer interesting conclusions. Findings should be evaluated to see whether they support, challenge or offer a revised version of existing theory. Ultimately, the findings can also be used as the basis for further case-based research into the mechanisms implied by each causal recipe.

Methodology

As briefly outlined above, QCA can be thought of as a series of tools or techniques for analysis which share a common underlying logic. The steps set out below represent a very brief overview of the logical steps which comprise QCA. For a more in-depth explanation of how to carry out QCA variants, csQCA, mvQCA and fsQCA, see especially Rihoux and Ragin (2009), chapters 3–5.

- 1) Identifying relevant cases and causal conditions.** Typically, this first step involves the identification of cases (both positive and negative) that exemplify the outcome of interest. Relevant theoretical and substantive knowledge is then used to identify the major causal conditions relevant to each case.
- 2) Construction of a truth table.** The second step involves summarising the required information on the selected cases into a data matrix based on the causal conditions outlined in step 1. The cases are then subsequently sorted into a “truth table” which outlines the various combinations of conditions and outcomes present in the cases being examined.
- 3) Truth-table analysis.** The next step involves the comparison of the different configurations of conditions linked to the different outcomes. The goal is to identify decisive differences between positive and negative outcomes and outline which

conditions are necessary parts of “causal recipes” which themselves will be sufficient to cause an outcome.

- 4) Evaluation of the results.** The final step involves the critical analysis and interpretation of the resulting causal recipes. Consideration should be given to whether the conclusions fit in with existing theory or provide new leads which may form the basis for further case-study analysis investigating the possible mechanisms implied by each causal recipe.

Useful resources

Berg-Schlosser, D., De Meur, G., Rihoux, B. and Ragin, C.C. (2009) Qualitative Comparative Analysis (QCA) As an Approach, in Rihoux, B. and Ragin, C. (eds.) *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA and London: Sage.

Byrne, D. and Ragin, C. (eds.) (2009) *Handbook of Case-Centred Research Methods*. London: Sage.

Cronqvist, L. and Berg-Schlosser, D. (2009) Multi-Value QCA (mvQCA), in Benoit Rihoux and Charles Ragin (eds.) *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related techniques*. Thousand Oaks, CA and London: Sage Publications.

Huntjens, P., Pahl-Wostl, C., Rihoux, B., Flachner, Z., Neto, S., Koskova, R., Schlueter, M., Nabide Kiti, I. and Dickens, C. (2008) *The Role of Adaptive and Integrated Water Management (AIWM) in Developing Climate Change Adaptation Strategies for Dealing with Floods and Droughts – A Formal Comparative Analysis of Eight Water Management Regimes in Europe, Asia, and Africa*. University of Osnabruck: NeWater project, Institute of Environmental Systems Research.

Ottitsch, A. and Weiss, G. (1998) Mountain Forest Policies in European Countries – a Comparison Using Ragin's Qualitative Comparative Analysis Method, in Glück, P. and Weber, M. (eds.) *Mountain Forestry in Europe. Evaluation of Silvicultural and Policy Means*. Wien: Institute for Forest Sector Policy and Economics.

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

Ragin, C. (1994) *Constructing Social Research*. Thousand Oaks: Pine Forge Press.

Ragin, C. (2008) *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Ragin, C. (2009) Qualitative Comparative Analysis Using Fuzzy Sets (fsQCA), in Benoit Rihoux and Charles Ragin (eds.) *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA and London: Sage Publications.

Rihoux, B. and De Meur, G. (2009) Crisp-Set Qualitative Comparative Analysis (csQCA), in Benoit Rihoux and Charles Ragin (eds.) *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA and London: Sage Publications.

Rihoux, B. and Grimm, H. (2006) Introduction: Beyond the “Qualitative–Quantitative” Divide: Innovative Comparative Methods for Policy Analysis, in Rihoux, B. and Grimm, H. (eds.) *Innovative Comparative Methods for Policy Analysis*. New York: Springer.

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

Rihoux, B., Rezsöházy I. and Bol, D. (2011) Qualitative Comparative Analysis (QCA) in Public Policy Analysis: an Extensive Review. *German Policy Studies*, 7 (3): 9–82. Center for Political Science and Comparative Politics (CESPOL) and COMPASSS, Université catholique de Louvain (Belgium).

Scouvard, M., Adams, R. T., Caldas, M., Dale, V., Mertens, B., Nedelec, V., Pacheco, P. et al. (2007) Causes of deforestation in the Brazilian Amazon: a qualitative comparative analysis. *Journal of Land Use Science*, 2 (4): 257–282.

Varone, F., Rihoux, B. and Marx, A. (2006) A new method for policy evaluation? Longstanding challenges and the possibilities of Qualitative Comparative Analysis (QCA), in Rihoux, B. and Grimm, H. (eds.) *Innovative Comparative Methods for Policy Analysis*. New York: Springer.

See also the resources available on the following websites:

- Qualitative Comparative Analysis: www.fsqca.com
- Comparative methods: www.compassss.org
- QCA applications: www.compassss.org/pages/resources/international.html

Annex 2: Small *n* approaches: comparison tables

Group I approaches: general characteristics			
Approach	What question(s) does the approach explicitly set out to answer?	Philosophy / paradigm	How does the approach tackle attribution?
Realist Evaluation	• Sets out to establish: - an “inequivocal causal relationship between a programme and outcome(s)”; - that it was, beyond doubt, a programme that caused a given measurable change, and not some other factor(s). • Designed to explain how, and in what circumstances, programmes generate outcomes, by asking “what works for whom, in what contexts, in what respects and how”.	• Realism holds that there is a “real” world, which exists independent of (or interdependent with) our interpretations of it. Thus, programmes can have real effects – positive, negative, intended and unintended. While there is no “final truth”, improvement in knowledge and understanding is possible.	• Programmes regarded as “theories incarnate” and their implementation as a test of an implicit theory or “Middle Range Theory” about what causes change. • This theory, together with contextual research, is used to map out a series of conjectural mini-theories called CMOs (Context, Mechanism and Outcome) which relate the various contexts of a programme to the multiple mechanisms by which it might operate to produce various outcomes. • These “causal mechanisms” are not directly observable, but must be hypothesised and tested. Whether mechanisms function as intended or not is dependent on context. • Quantitative and qualitative research builds up a picture of a programme in action. This is used to revise, substantiate or invalidate conjectural CMOs, leaving only those that explain how the programme led to observed outcome(s).
General Elimination Methodology (GEM)	• Intended to provide an overarching logic geared towards investigating	• Though not explicitly stated, has much in common with realist	• Draws on the principle that all macro events have a cause. For a given event it is possible to construct a List of Possible Causes (an LOPC). Each possible cause will have its

	and substantiating causal claims.	approaches which assert that there is an identifiable “reality” and that it is possible to understand what causes changes.	own empirically verifiable modus operandi (MO), or footprints. • By gathering the “facts of the case”, we can examine which modus operandi are present and which are not, and thereby establish which possible cause(s) had a causal influence.
Process Tracing	• Aims to “unwrap” the causal links between programmes and outcomes by identifying the intervening causal mechanisms at work. (the causal chain).	• Its focus on causal dynamics draws on the positivist tradition, but it also employs much qualitative data, themselves associated with constructivist epistemological assumptions (i.e. subjectivism).	• “Process induction” draws on evidence to generate a set of (preferably competing) hypothetical explanations about how an intervention may connect to an outcome. Researchers set out what should be observed if each theory is true or false and develop an explicit chronology of events, outlining the causal chain. • “Process verification” examines the observable implications for each step of the hypothesised explanations in order to establish whether the actual events or processes within the case fit those predicted.
Contribution Analysis	• Verifies a hypothesised theory of change by building up a “contribution story”. • Demonstrates whether intervention X contributed to outcome Y; it aims to conclude that a “reasonable person, knowing what has occurred in the programme and	• Similar to some theory- based evaluation approaches such as Realist Evaluation, in that it assumes a verifiable reality.	• Does not set out to show that a specific intervention “caused” an outcome but instead develops a theoretical “results chain” that sets out the intended interplay between policy activities, contexts and outcomes. • Background research assembles a plausible “contribution story”, with available evidence used to verify the different stages in the results chain, set out any weaknesses and put forward alternative theories.

	that intended outcomes actually occurred, agrees that the programme contributed to outcomes".		
--	---	--	--

Group I approaches: methodology				
Approach	Sets out a hypothesised theory of change	Outlines rival causal hypotheses / assesses confounding or contextual factors	Cross-checks hypotheses against the evidence	Contribution to theory / external validity
Realist Evaluation (RE)	Sets out to test a Middle Range Theory (MRT) detailing how the mechanisms initiated by a programme should cause desired outcomes.	Rival hypotheses in the form of CMOs (Context, Mechanism, Outcome) configurations explore how different potential combinations of contexts and mechanisms could deliver outcomes.	Systematic analysis should eliminate/revise CMO configurations so that only those that have had a causal influence remain.	- Generates evidence that "this programme theory works in these respects, for these subjects, in these kinds of situations". - Should start with a "Middle Range Theory" (MRT); any conclusions can be used to update the initial MRT.
General Elimination Methodology (GEM)	Does not explicitly require an evaluator to draw up a theory of change, but research should aim to draw up a Modus Operandi (a	A list of possible causes (LOPC) is constructed, with a Modus Operandi or likely set of "footprints" outlined for each of the possible causes.	Analysis is geared towards gathering evidence selectively as to the presence or absence of factors listed in the Modus Operandi of	Scriven (2008) states that GEM is able to produce both specific and generalisable results: "GEM works equally well with both general and particular causal claims, i.e. with

	series of events which will be present when a cause is effective) for a number of possible causes.		each possible cause.	both causes of classes of effects (or with the effects in a large class of subjects), or with the causes of a particular effect on one occasion, or with one subject."
Process Tracing	• If not already set out, "Process Induction" should identify a programme's implicit theory of change as one of several possible potential causal explanations.	• "Process Induction" should identify all plausible causal explanations for observed outcomes.	• "Process Verification" should systematically identify "diagnostic" evidence for each hypothetical causal explanation's theoretically predicted causal chain to see whether it fits that evidence.	• Most useful as a method for identifying, testing and validating hypothesised causal mechanisms within case studies. As a result, is not intended to produce generalisable theories. However, examination of multiple cases may be used to build up a body of evidence.
Contribution Analysis	• A reasoned theory of change should set out potential influencing factors and outline the different links in the theory of change and the risks and assumptions associated with them.	• Evidence should be gathered for potential external influencing factors so that they may be discounted (if possible) or assessed for their relative importance in contributing to observed outcomes.	• Contribution Analysis emphasises the need to construct a plausible "contribution story" which sets out and links the different stages in a "contribution story". The process should be	• The key goal for Contribution Analysis is one of analysing a programme's contribution to observed outcomes. The conclusions that can be drawn are therefore largely programme-specific, although, as with the other

		• In complex settings where there are multiple elements or “arms” of an intervention, a theory of change for each arm should be developed.	iterative; further research should be carried out to address weaknesses or gaps in the story.	Group I approaches discussed, conclusions can still contribute to wider theory without being directly generalisable.
--	--	--	--	---

Group II approaches: general characteristics			
Approach	What question(s) does the approach explicitly set out to answer?	Philosophy / paradigm	How does the approach tackle attribution?
Most Significant Change (MSC)	• Intended as a monitoring and evaluation tool, it sets out to discover what significant changes have occurred in the lives of beneficiaries since its inception. • By seeking out stories of change, MSC provides data on short- and long-term outcomes that can be used to help assess and improve programme implementation and identify and address negative or unexpected outcomes.	• Draws on a constructivist epistemology, but realist ontology. Meaning is constructed in the telling and interpretation of stories, but a verifiable reality is also assumed; verification stages check whether stories occurred as reported.	• Searches for (the most) interesting stories of change; these may, or may not, relate to outcomes that have resulted from a programme. Where they do, the stories themselves provide the causal link between outcome and programme. (A verification stage is included to check their accuracy.) • Does not explicitly present causality as working in any single way. Storytellers, or those interpreting the stories, interpret the causal pattern themselves. • Provides for the analysis of individual stories and secondary analysis of multiple stories as a way of examining general patterns in changes.
Success Case Method (SCM)	• Sets out to discover whether an intervention has been successful or not. By finding corroborating evidence and documentation for “success cases” (and non-success cases), the method aims to establish the extent to which the programme was implemented as intended, whether participants understood it, and whether it produced the intended results.	• A narrative technique based upon naturalistic/ constructivist inquiry and in-depth case-study analysis. However, as with other approaches here, SCM draws on mutual validation to discover a verifiable reality.	• Seeks corroborating evidence for an intervention's theoretical “impact model”. • Surveys and key interviews help to develop an understanding of the relationship between intervention and outcomes. • Intentionally seeks out cases which represent the best (and worst) that a programme is producing. (A purposive survey followed by random sampling of success cases focuses the inquiry on relatively few.) The success and non-success cases are then used to evaluate the programme

			as a whole.
Outcome Mapping (OM)	• Intended for planning, monitoring and evaluating development initiatives. • Aims to help programmes be specific about the actors that it targets, the changes that it expects and the strategies that should be employed. • Sets out to realistically reports on achievements by tracking connections between programme activities and outcomes.	• Has its roots in systems thinking and organisational learning theories, both of which draw on constructivism .	• Examines the relationship between a programme and observed changes by documenting a programme's planning and execution and by gathering information on how far changes influenced outcomes. • Seeks evidence of how programme initiatives influenced "boundary partner" behaviours and how these contributed to outcomes. • Aims to prove that changes in behaviour can be linked to programme activities, though the causal link may be indirect rather than direct.
Method for Impact Assessment of Programs and Projects (MAPP)	• An evaluation tool designed to give an insight into an intervention's outcomes and impact. • Aims to establish both the intended and unintended outcomes resulting from a programme and, in so doing, establish how an intervention can be adapted and improved.	• Draws on beneficiary understanding and interpretation of programme activities, their value and significance, and employs mutual validation to discover a verifiable reality.	• Utilises the collective recall of a representative group of beneficiaries and programme operatives to attribute changes to programme-specific measures or to other activities or factors. • The presence of a number of different groups in workshops is intended to ensure that mutual validation is an inbuilt part of the evaluation process. • MAPP's multiple tools are designed to triangulate findings.

Group II approaches: methodology				
Approach	Sets out a hypothesised theory of change	Outlines rival causal hypotheses / assesses confounding or contextual factors	Cross-checks hypotheses against the evidence	Contribution to theory / external validity
Most Significant Change (MSC)	Not over-concerned with the underlying programme logic. It assumes that the programme is already in place and is instead concerned with how and to what extent it works and how it may be improved.	- Collects stories of changes occurring in participants' lives, whether directly or indirectly linked to a programme, intended or unintended. - The intention is not to systematically evidence all changes occurring at a given location, but to find the most salient stories of change and extract learning from them.	MSC stories are not designed to provide comprehensive feedback about a programme's performance but instead to offer key learning by providing insight into participant perceptions of that programme. However, there is a verification stage designed to ensure that reports are accurate and honest.	MSC stories should drive organisational learning and ability to adapt and improve a programme on an ongoing basis. Learning from the stories should be incorporated back into a programme to improve its effectiveness. Though stories are particularly valid on a programme-specific level, there may still be transferable lessons of relevance to how other similar programmes are implemented.
Success Case Method (SCM)	An "impact model" outlines programme activities and the behaviours and positive	Outcomes are examined to assess whether success (or non-success) cases adhere	Only exceptional success (or non-success) cases are examined in detail to	Geared towards finding out which aspects of an intervention are working and which are

	outcomes that should result from the proper implementation of that programme.	to the impact model. • Only looks at exceptional cases and therefore does not set out to exhaustively examine the relationship between a programme and outcomes.	discover whether desired goals have been achieved or not, and what key factors led to success or failure. • Where there is a mismatch between Impact Model and observed outcomes, further evidence can be sought to explain what really occurred.	not. • Conclusions relate only to a small sample of cases relating to a given intervention and are not intended to be generalisable, either to the average programme participant or more widely.
Outcome Mapping (OM)	• An “Intentional Design” sets out intended outcomes, the relationship between them, the impacts the programme is designed to achieve, and the strategies needed to support this work.	• External factors such as local organisations or institutions are factored in from the start, when they are set out in the planning stage, right through to the assessment of outcomes, strategies and activities.	• Aims to assess a programme’s contribution to desired behavioural changes or “outcomes” by drawing on the perceptions of boundary partners and stakeholders with the goal of triangulating findings.	• The process of OM is intended to be ongoing, with learning fed back into the programme to facilitate learning and adaptation to better achieve goals. • Any conclusions will be context-specific, although some learning may still have wider relevance.
Method for Impact Assessment of Programmes and Projects (MAPP)	• The focus is not on programme design, but instead on establishing a programme’s context as a	• Does not identify external factors in the explicit and thorough way that “Group I” approaches	• MAPP’s tools set out to list project activities and outline and compare their relative importance to	• The primary goal is to establish development trends and demonstrate how a programme

	starting point and then evaluating development trends affecting the daily life of the population, as they perceive it.	do, though some MAPP tools are designed to identify key recent changes, perceived contributing factors and perceptions as to whether outcomes resulted from a programme or not.	participants' daily lives. These tools should aid an evaluator to establish which activities have resulted from which programme measures / external interventions / beneficiaries own work.	can be adapted and improved. • It is also intended to facilitate the comparison of results for different locations where the same programme has been implemented, but where length of implementation and results produced have been different.
--	---	--	--	---

Publications in the 3ie Working Paper series

Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework by Howard White and Daniel Phillips, Working Paper 15, July 2012

Behind the scenes: managing and conducting large scale impact evaluations in Colombia by Bertha Briceño, Laura Cuesta and Orazio Attanasio, Working Paper 14, December 2011

Can we obtain the required rigour without randomisation? by Karl Hughes and Claire Hutchings, Working Paper 13, August 2011

Sound expectations: from impact evaluations to policy change by Vanessa Weyrauch and Gala Díaz Langou, 3ie Working Paper 12, April 2011

A can of worms? Implications of rigorous impact evaluations for development agencies by Eric Roetman, 3ie Working Paper 11, March 2011

Conducting influential impact evaluations in China: the experience of the Rural Education Action Project by Mathew Boswell, Scott Rozelle, Linxiu Zhang, Chengfang Liu, Renfu Luo, Yaojiang Shi, 3ie Working Paper 10, February 2011

An introduction to the use of randomized control trials to evaluate development interventions by Howard White, 3ie Working Paper 9, February 2011

Institutionalisation of government evaluation: balancing trade-Offs by Marie Gaarder and Bertha Briceño, 3ie Working Paper 8, July 2010

Impact Evaluation and interventions to address climate change: a scoping study by Martin Prowse and Birte Snilstveit, 3ie Working Paper 7, March 2010

A checklist for the reporting of randomized control trials of social and economic policy interventions in developing countries by Ron Bose, 3ie working paper 6, January 2010

Impact evaluation in the post-disaster setting by Alison Buttenheim, 3ie Working Paper 5, December 2009

Designing impact evaluations: different perspectives, contributions from Robert Chambers, Dean Karlan, Martin Ravallion, and Patricia Rogers, 3ie Working Paper 4, July 2009. Also available in Spanish, French and Chinese

Theory-based impact evaluation by Howard White, 3ie Working Paper 3, June 2009. Also available in French and Chinese.

Better evidence for a better world edited by Mark W. Lipsey University and Eamonn Noonan, 3ie & The Campbell Collaboration, 3ie Working Paper 2, April 2009

Some reflections on current debates in impact evaluation by Howard White, 3ie Working Paper 1, April 2009

For the latest 3ie working papers visit:

http://www.3ieimpact.org/3ie_working_papers.html