report

DFID Department for International Development

WORKING PAPER 38
April 2012

BROADENING THE RANGE OF DESIGNS AND METHODS FOR IMPACT EVALUATIONS

Report of a study commissioned by the Department for International Development

DEPARTMENT FOR INTERNATIONAL DEVELOPMENT

Working Paper 38

BROADENING THE RANGE OF DESIGNS AND METHODS FOR IMPACT EVALUATIONS

Report of a study commissioned by the

Department for International Development

APRIL 2012

Elliot Stern (Team Leader), Nicoletta Stame, John Mayne,

Kim Forss, Rick Davies, Barbara Befani

Table of Contents

Reading hints

This is a study report dealing with difficult methodological and theoretical challenges faced by those who wish to evaluate the impacts of international development policies. It is in parts difficult to read as it was also difficult to write! Knowing that not everybody interested in this study will have the time to work through the entire report when it crosses their desk for the first time, there are some short-cuts built-in to the report. First the Executive Summary provides an overview in fairly general terms. Second each of the core chapters has a 'Main Messages' box at the beginning that highlights the key points in that chapter. Third the final chapter – 'Conclusions and Next Steps' draws together 10 study conclusions and briefly outlines some proposed follow ups to the study.

Hopefully those who read such a cut-down version will be encouraged to dig deeper second time around!

Acknowledgements

This report would not have been possible without extensive support, guidance and advice. The management of the study by Richard Martini, Lina Payne and Mark Clayton of the Department for International Development (DFID) has throughout helped solve a host of practical problems whilst their insightful questions encouraged us to clarify and focus the report. We are also grateful to Nick York for his ongoing support and strategic perspective. The study benefitted from an advisory group whose members included Nancy Cartwright (London School of Economics), Penny Hawkins (Rockefeller Foundation), Michael Patton (Independent Consultant, USA), Ray Pawson (Leeds University), Mark Petticrew (London School of Hygiene and Tropical Medicine), Charles Ragin (University of Arizona), Patricia Rogers (RMIT University), Oumoul Khayri Ba Tall (Independent Consultant, Mauretania) and Jos Vaessen (UNESCO). We are also grateful to practitioners from a number of development agencies who attended a workshop that DFID hosted in London in October 2011; and to those who commented on an early draft of this report. The inputs received from all of the above were invaluable even though it is the study team that is entirely responsible for the judgements and contents of the report.

Elliot Stern
April 2012

Executive Summary

S1. This report covers a study commissioned by DFID entitled 'Broadening the Range of Designs and Methods for Impact Evaluations'.

S2. Impact Evaluation (IE) aims to demonstrate that development programmes lead to development results, that the intervention as *cause* has an *effect*. Accountability for expenditure and development results is central to IE, but at the same time as policy makers often wish to replicate, generalise and scale up, they also need to accumulate lessons for the future. Explanatory analysis, by answering the 'hows' and 'whys' of programme effectiveness, is central to policy learning.

S3. IE must also fit with contemporary development architecture that post Paris Declaration is decentralised, works through partnership and where developing countries are expected to be in the lead. These normative principles have practical implications for IE. For example, working through partners leads to multi-stage, indirect causal chains that IE has to analyse; and using a country's own systems can limit access to certain kinds of data.

S4. Up to now most investment in IE has gone into a narrow range of mainly experimental and statistical methods and designs that according to the study's Terms of Reference, DFID has found are only applicable to a small proportion of their current programme portfolio. This study is intended to broaden that range and open up complex and difficult to evaluate programmes to the possibility of IE.

S5. The study has considered existing IE practice, reviewed methodological literatures and assessed how state-of-the art evaluation designs and methods might be applied given contemporary development programmes.

S6. Three elements - evaluation questions; appropriate designs and methods; and programme attributes - have to be reconciled when designing IE. Reviews of existing evaluations suggests that sometimes methods chosen are unable to answer the evaluation questions posed; or the characteristics of development programmes are not taken into account when choosing designs or deciding on evaluation questions.

S7. Demonstrating that interventions cause development effects depends on theories and rules of causal inference that can support causal claims. Some of the most potentially useful approaches to causal inference are not generally known or applied in the evaluation of international development and aid. Multiple causality and configurations; and theory-based evaluation that can analyse causal mechanisms are particularly weak. There is greater understanding of counterfactual logics, the approach to causal inference that underpins experimental approaches to IE.

S8. Designs need to build on causal inference approaches each of which have their strengths and weaknesses, one of the reasons that combining designs and methods – so called 'mixed methods' – are valuable. Combining methods has also become easier because the clear distinctions between quantitative and qualitative methods have become blurred, with quantitative methods that are non-statistical and new forms of within-case analysis made easier by computer aided tools.

S9. On the basis of literature and practice, a basic classification of potential designs is outlined. Of the five design approaches identified - Experimental, Statistical, Theory-based, Case-based and Participatory – the study has in line with its ToR concentrated on the potential of the latter three.

S10. The study has concluded that most development interventions are 'contributory causes'. They 'work' as part of a causal package in combination with other 'helping factors' such as stakeholder behaviour, related programmes and policies, institutional capacities, cultural factors or socio-economic trends. Designs and methods for IE need to be able to unpick these causal packages.

S11. This also has implications for the kind of evaluation questions that can usefully be asked. It is often more informative to ask: 'Did the intervention make a difference?' which allows space for combinations of causes rather than 'Did the intervention work?' which expects an intervention to be cause acting on its own.

S12. Development programmes that are difficult to evaluate such as those concerned with governance, democracy strengthening or climate change mitigation, are often described as 'complex'. In the case of humanitarian relief or state-building the term complexity can also be applied to the setting in which the programme is located.

S13. Instead of classifying programmes into how complicated they are, attributes of programmes were identified on the basis of literature reviews and analysing a portfolio of current DFID programmes. These attributes included duration and time scale; non-linearity and unpredictability; local customisation of programmes; indirect delivery through intermediate agents such as funds; multiple interventions that influence each other – whether as sub-programmes within the same programme or in separate but overlapping programmes.

S14. Tailored evaluation strategies are needed to respond to these attributes. For example careful analysis is needed to decide under what circumstances large multi-dimensional programmes, such as those characteristic of the governance area, can be broken down into sub-parts or have to be evaluated holistically.

S15. A reality that often has to be faced in IE is that there is a trade off between the scope of a programme and strength of causal inference. It is easier to make strong causal claims for narrowly defined interventions and more difficult to do so for broadly defined programmes. The temptation to break programmes down into sub-parts is therefore strong, however this risks failing to evaluate synergies between programme parts and basing claims of success or failure on incomplete analysis.

S16. Similarly when the effects of programmes are long-term and have unpredictable trajectories, designs for IE will need to take this into account. Results monitoring may need to be prioritised alongside a staged evaluation strategy able to respond to changes in implementation trajectories not anticipated at programme launch.

S17. Quality Assurance (QA) for IE is needed both to reassure policy makers that evaluation conclusions are defensible and to reinforce good practice within the evaluation community. Existing QA systems already in use by DFID cover many generic aspects of evaluation quality. We have therefore focussed on what else is needed to assure quality in relation to IE.

S18. Having reviewed the literature on quality in research and evaluation, the study concluded that a common framework could be applied across different designs and methods. Standards such as validity, reliability, rigour and transparency have therefore been incorporated into a three part QA framework covering: the conduct of the evaluation; the technical quality of methods used and normative aspects appropriate to IE in an international development setting.

S19. This has been a 'proof of concept' study and many of the newer designs and methods identified have not previously been applied in international development IE. Making these IE approaches usable in the development setting will require field-testing in a number of targeted programmes in cooperation with DFID's decentralised offices and evaluation specialists.

S20. IEs should not be regarded as an everyday commission. Any IE that is thorough and rigorous will be costly in terms of time and money and will have to be justified. Criteria are therefore needed to decide when such an investment should be made. More generally IE raises questions of capacity both within development agencies such as DFID and among evaluators. For this reason 'Next Steps' are outlined that would enhance capacities needed to support the take-up of a broader range of designs and methods for IE in international development.

S21. It is important to recognise that even when IE is inappropriate, enhancing results and impacts can still be addressed through evaluation strategies other than IE. Results monitoring can make major contributions to accountability driven evaluations. There will also be occasions, when real-time, operational, action-research oriented and formative evaluations can all make serious contributions to filling gaps in evidence and understanding.

Abbreviations

3ie	International Initiative for Impact Evaluation
CDF	Comprehensive Development Framework
CSGF	Civil Society Governance Fund (Malawi)
DDP	District Development Programme
DFID	Department for International Development
DIME	Development Impact Evaluation Initiative
EBP	Evidence Based Policy
EQ	Evaluation Question
GEF	Global Environment Fund
GTZ(GIZ)	Deutsche Gesellschaft fur Internationale Zusammenarbeit
HTN	How to Note
IE	Impact Evaluation
J-PAL	Abdul Latif Jameel Poverty Action Lab

MDG	Millennium Development Goals
NGOs	Non Governmental Organisations
NONIE	Network of Networks on Impact Evaluation
NORAD	Norwegian Agency for Development Cooperation
OECD-DAC	Organisation for Economic Co-operation and Development – Development Assistance Committee
PD	Paris Declaration
PPA	Programme Partnership Agreements
PRA	Participatory Rural Appraisal
QA	Quality Assurance
QCA	Qualitative Comparative Analysis
RBA	Results Based Aid
RCT	Randomised Control Trials
SCIP	Strategic Climate Institutions Programme (Ethiopia)

SIDA	Swedish International Development Cooperation Agency
TBE	'Theory Based' Evaluations
ToC	Theory of Change
ToR	Terms of Reference
VAW	Violence Against Women (Bihar, India)
VFM	Value for Money

Chapter 1: Introducing the study

1.1 Terms of Reference

1.1 This report brings together the findings and conclusions of a study on 'Impact Evaluation' (IE) commissioned by the UK's Department for International Development. The full title of the study in the Terms of Reference is:

'*Developing a broader range of rigorous designs and methods for impact evaluations'*

1.2 The assumption underlying the study is that whilst IE has made progress in international development in recent years, this has been on a narrow front, mainly deploying quantitative and experimental methods. More was now needed. As the ToR explains:

'*The aim of this study is to test the case that alternatives to experimental and quasi-experimental designs can have equivalent robustness and credibility – they are different but equal'*

1.3 Whilst the ToR emphasises the potential for 'qualitative methods' it also acknowledges the continued importance of established IE traditions and the importance of 'mixed methods'. The rationale for this interest in 'a broader range' of designs and methods centres on the characteristics of development itself. First the challenges posed by complex programmes:

'The primary purpose of this research is to establish and promote a credible and robust expanded set of designs and methods that are suitable for assessing the impact of complex development programmes.'

1.4 The importance of this is also set into a wider context:

'Some have suggested that only 5% of development programmes of donors such as DFID are suitable for randomised controlled trials. Even if this is a conservative estimate, it is clear that this will be the case for a significant proportion of the profile of development spending.'

1.5 The ToR takes the view that there are risks attached to any 'assumed superiority' of a limited set of methods. Programmes evaluated via other methods may well be dismissed as 'less effective' or even 'distorted' to meet the needs of evaluation designs. Finally the ToR emphasises the importance of IE being able to:

'.....withstand criticism or scepticism over the robustness of [the] findings. The present study is intended to identify these approaches and explain how they can be rigorously quality assured.'

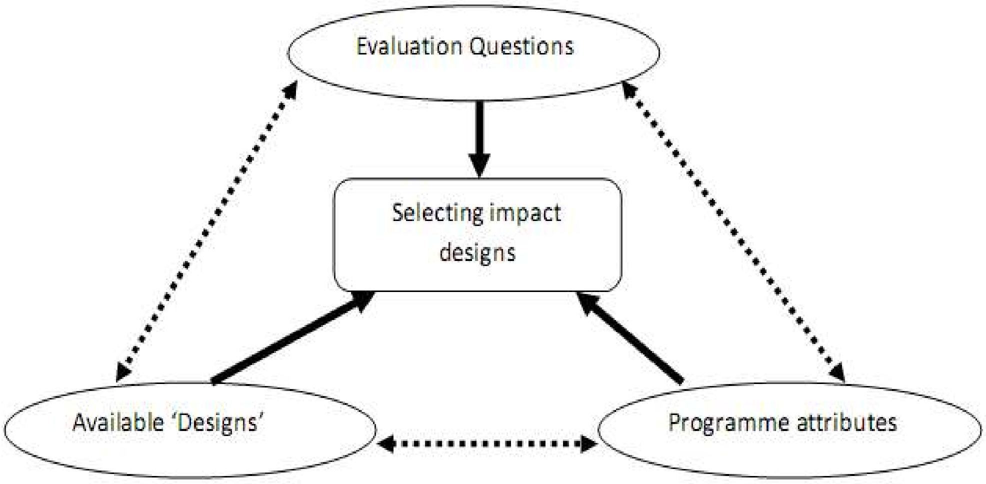
1.2 Structure of this Report

1.6 The body of this report follows the logic of the ToR:

- It starts by defining IE in a way that reflects the current 'profile of development spending' and the current aid environment post Paris Declaration. This process of definition also situates IE within a broader debate about Evidence Based Policy (EBP), the importance of demonstrating a link between 'causes' and 'effects'; and the kinds of evidence that should inform policy-making.
- It goes on to consider a 'broader range' of designs and methods in relation to different understandings of causal inference – being able to justify links between cause and effect. This analysis draws both on literature reviews and an analysis of actual IEs

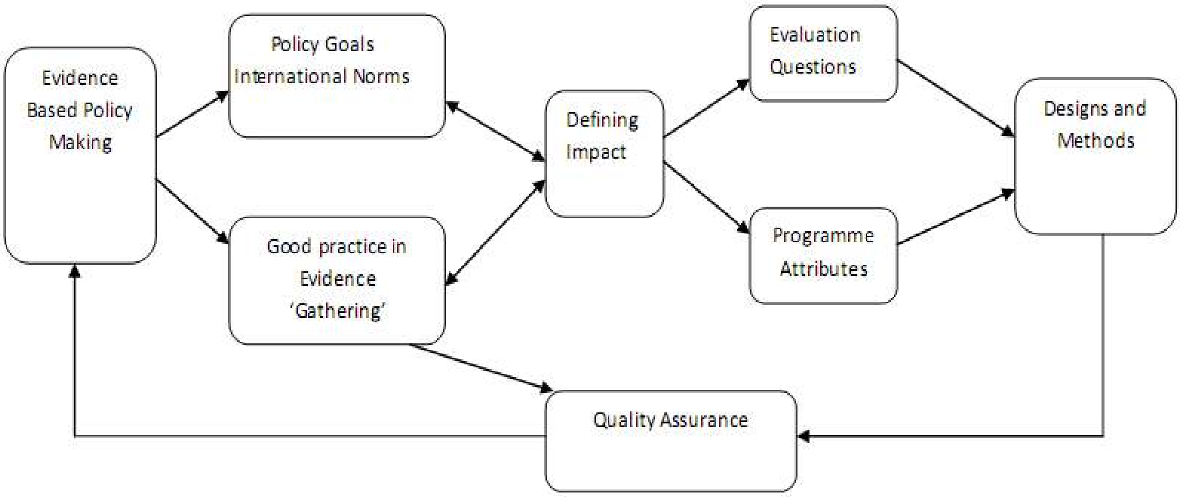
sourced through a number of development agencies. It is argued that demonstrating causal links and explaining how these links work is at the heart of IE.

- The report identifies the kinds of evaluation questions that are of interest to policy makers. It does so from the perspective that certain questions can only be answered by certain designs.
- It then addresses the challenges of 'complex development programmes'. It does this by analysing programme 'attributes' and the implications they have for evaluation designs and methods drawing both on literatures about complexity and an analysis of actual development programmes.
- The linkage between these sets of ideas is central to the way the study team has conceived of its work: selecting 'designs' for IE should be seen as a process of aligning evaluation questions – which must be at the heart of any evaluation – with an available repertoire of designs *and* the attributes of development programmes. This is depicted in **Figure 1**.



- The report then discusses issues of how to quality assure designs and associated methods. It does so in the knowledge that there are many existing QA systems in evaluation, and that they have already seriously addressed a wide range of methods, both quantitative and qualitative.
- Finally conclusions and the next steps recommended to take forward the broadening of methods and designs for IE are outlined.

1.7 The overall approach taken in this study makes certain assumptions about how to identify high quality IE designs and methods. These necessarily rest on a number of key 'building blocks' as **Figure 2** suggests.


Evidence Based Policy Making
Policy Goals International Norms
Good practice in Evidence 'Gathering'
Defining Impact
Evaluation Questions
Programme Attributes
Designs and Methods
Quality Assurance

1.8 This framework advances the argument that in order to define 'impact' one needs first to have an understanding of the current international development policy environment; what are the current debates in EBP and what is considered 'good practice' when gathering evidence. Furthermore it is only through a consideration of evaluation questions and programme attributes that designs and methods can be selected.

1.9 A definition of IE has direct implications for designs and methods, albeit mediated by evaluation questions and programme attributes. For example by building on the current aid relationship between donors and partner countries as expressed in OECD-DAC guidelines and agreements, certain evaluation questions become more relevant just as certain types of programme attributes become more commonplace.

1.3 Managing expectations

1.10 Evaluators are at their best when links between 'interventions' and policy goal are relatively direct and short term; and when policy interventions are relatively self contained - parallel and overlapping interventions make it difficult to disentangle causes from effects. We find it easier to deal with standardised interventions that go about things in the same way whatever the circumstances. And we find programmes that are long term, embedded in a changing context with extended causal chains or pathways especially challenging.

1.11 This study focuses on the very settings and programmes that have been difficult to evaluate with the traditional evaluators' toolkit. It is for this reason that a health-warning is needed. An overall 'conclusion' that can be shared at the beginning of this report has both positive and cautionary elements.

1.12 First the positives: this study confirms that deploying a wider range of designs and methods in the evaluation of development aid and interventions has the potential to answer pressing IE questions that cannot otherwise be answered by the more limited range of designs and methods currently favoured by development agencies. This conclusion is supported by both academic and practitioner experience and by reviewing existing evaluations, literatures and exemplary programmes. However many of the designs and methods identified in this study are complicated and expensive. Whilst agency managers and policy makers may want the kinds of answers that only these kinds of designs can deliver, they may need persuading that the extra effort and resource is justified. One approach that we favour is to acknowledge that elaborate IEs (of whatever design) are likely to be rare and will need to be justified. Hence the suggestion (see Next Steps, Chapter 7) that a set of criteria should be developed to decide under what circumstances sophisticated IE designs are justified.

1.13 This study has been a 'proof of concept' exercise that remains to be tested and refined in practice. The ideas and approaches outlined in the report need to be tried out in relation to real evaluation assignments before the case can be regarded as 'proven'. We hope that this process of field-testing will take place.

1.14 There are two further warnings:

- First this is essentially a report prepared by researchers and evaluators based on desk-research and with limited involvement of development practitioners. It is for the most part conceptual analytic and methodological. Before the suggestions contained in this report could be put into practice they would need to be 'translated' into practitioner-friendly language of guidelines, checklists and toolkits. We believe this could be done but we have not had the time or resources to take this next step.
- Second there are limits to what should be expected of IE, supported by whatever designs and methods are used. The sheer complexity and ambition of international development programmes means that there will always be limits to what can be said about the links between development interventions as 'causes' and development results as 'effects'. Programmes about governance, empowerment and accountability, climate change and post-conflict stabilisation are exploring new areas of practice for which little evidence exists. Many of the problems being addressed are long-term and inter-generational; some of their impacts will only become clear to policy-makers still to be born!

1.15 This is not intended to suggest that impact evaluation has no role in evaluating these programmes, only to caution against expectations of precise estimation of the long-term future effects even when direction of travel can be plausibly predicted. However there will be times when impact evaluation will not be possible and it will be evaluation more generally rather than IE in particular that is best able to make a contribution to enhancing programme impact. For example, evaluation can contribute to developing theories of change, providing rapid feedback, improving programmes whilst they are ongoing and involving stakeholders and beneficiaries. For programmes for which IE is inappropriate, real-time, operational, action-research oriented and formative evaluations can help fill gaps in evidence and theory whilst also improving targeting and implementation.

Chapter 2: Defining Impact Evaluation

Introduction

2.1 This chapter defines impact evaluation in a way that is relevant to the tasks set out in the ToR and as outlined in the previous introductory chapter. It considers different understandings of impact and how these are linked to contemporary debates about aid and international development on the one hand, and about how evidence feeds-in to policy-making on the other. It is suggested that these debates support particular understandings of evidence, causal analysis and the content of what needs to be evaluated in IE. This Chapter also indicates the rationales for some of the main assumptions that were made in this study – about causality and explanation and about the characteristic of the contemporary aid and development environment thereby setting the scene for more in depth discussions in following chapters. The chapter ends by proposing a working definition of impact evaluation that has been used to guide the present exercise.

Main Messages[1]

- At the heart of IE is a requirement to link causes and effects and to explain 'how' and 'why'. This highlights the importance of theory in IE.
- If the purpose of an IE is accountability rather than policy learning explanation may be less important. In these circumstances causal analysis will be sufficient.
- The study has adopted a definition of Impact Evaluation consistent with the Paris Declaration and contemporary understandings of aid and the aid relationship.
- IEs therefore need to deal with contemporary interventions that are often complex, multi-dimensional, indirectly delivered, multi-partnered, long-term and sustainable.
- IE is not method specific – no single design or method can lay monopoly claim to the production of evidence for policy learning; and all established methods have difficulty with many contemporary interventions.
- IE should start with the kinds of evaluation questions to which policy-makers want answers.

2.1 The implications of how 'impact' and 'impact evaluation' are defined

2.2 From the outset of this study we need to define what is meant by 'impact' in international development evaluation. How impact is defined will necessarily determine the scope and content of the study because different definitions prioritize different aspects of 'impact'; imply different concepts of causality (what produces the impact); and how to estimate the impact (evaluation designs). The most widely shared definition

[1] The 'Main Messages' box summarise key-points in each chapter

is that of the OECD-DAC Glossary (2002), which starts from the content of international development and interventions funded through aid. It defines impact as:

'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'

2.3 This definition stresses:

- the search for *any* effect, not only those that are intended;
- recognizes that effects may be positive and negative;
- that effects of interest are 'produced' (somehow caused) by the intervention;
- suggests the possibility of different kinds of links between all kinds of development intervention (project, programme or policy) and effect; and
- focuses on the longer-term effects of development interventions.

2.4 The OECD-DAC definition is broad in scope and represents the consensus view of most international development actors. However in recent years other more focused definitions have been advocated. For example, the 'Poverty Action Lab' defines IE as:

The primary purpose of impact evaluation is to determine whether a program has an impact (on a few key outcomes), and more specifically, to quantify how large that impact is. What is impact?

2.5 The World Bank poverty/net website defines Impact Evaluation as:

'…assessing changes in the well-being of individuals, households, communities or firms that can be attributed *to a particular project, programme or policy'*

2.6 3ie defines IE in its 'foundation document' (2008) as:

'Rigorous impact evaluation studies are analyses that measure the net change in outcomes for a particular group of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question that is being investigated and to the specific context.'

2.7 And as:

'*the difference in the indicator of interest (Y) with the intervention (Y1) and without the intervention (Y0). That is, impact = Y1 – Y0. An impact evaluation is a study which tackles the issue of attribution by identifying the counterfactual value of Y (Y0) in a rigorous manner.' (White 2010)*

2.8 In these definitions:

- tend to be methods-led rather than content-led[2];
- the search is for a *given* effect (the indicator of interest) and often a particular cause;
- the emphasis on attribution suggests a direct link between a cause (often described in medical terms as a 'treatment') and an effect;
- there is a reliance on counterfactual logic - asking the question what would have happened without the intervention - in making causal inference;
- there is often an assumption that experimental and quasi-experimental methods are the best or default method; and

[2] Bamberger also distinguishes between a 'technical-statistical definition' (the counterfactual, which implies a 'rigorous' experimental methodology) and a 'substantive definition' (the DAC's, that does not require a particular methodology).

- there is no specific reference to the long term – indeed effects may as commonly be short-term.

2.9 Whilst as noted some proponents of IE (e.g. Poverty Action Lab and the Bill and Melinda Gates Foundation) only advocate experimental and quasi-experimental methods others argue that Randomised Control Trials (RCTs) are useful in some but not all circumstances. This is the rationale of this study.

2.2 Counterfactuals as one basis for causal inference

2.10 Counterfactuals, closely associated with RCTs, are currently one of the most widely discussed ways of thinking about causation in impact evaluation. Advocates of experimental methods have even succeeded to co-opt the word 'rigorous' into their own lexicon although we will argue (see Chapter 6) there are well-established processes and criteria for ensuring 'rigour' that are not necessarily method-specific. Counterfactual logics (Lewis 1973) seek to answer the question: 'what would have happened without the intervention?' by comparing an observable world with a theoretical one, where the latter is intended to be identical to the former except for the presence of the cause and effect. The latter is described as 'counterfactual' because it cannot be observed empirically.

2.11 In this study we have taken counterfactual thinking as the underpinning of experimental designs and as the basis for causal inference within these designs. However, as a strategy for causal inference, counterfactuals are based on Mill's Method of Difference, which does not necessarily imply adoption of experimental methods. It is simply a comparison of two quasi-identical entities: a situation where an effect and an event supposedly representing its cause have taken place, and the same situation without that effect and without that cause. If two such events can be shown to exist, the effect is attributed to the cause. But the assumptions of counterfactual thinking do not always hold (e.g. finding an identical match for the factual world, e.g. the world where the cause and the effect have been observed, may be difficult or impossible); and even when they do, counterfactuals associate a single cause with a given effect without providing information on what happens in-between: how the effect is produced. This information can be important for attribution purposes because knowing only the beginning and the end of the causal process might not be sufficiently fine-grained in some situations. This is especially so when pre-emption operates – that is some other causal process intervenes either before an effect ('early pre-emption') which could lead to the same outcome; or after the effect ('late pre-emption') which could have led to the effect even if the causal process of interest had not taken place. There are even more severe problems with counterfactuals when causation is 'chancy' or probabilistic (see Menzies 1989, Noordhof 1999 and Edgington 2004)[3].

[3] It is beyond the scope of this study to enter fully into the ongoing philosophical debate about the extent to which counterfactuals are a sufficient basis for understanding causation. (Interested readers can refer to the sources cited above and in addition to Brady 2002, Hall 2004, Dowe and Noordhof 2004).

2.12 When counterfactual inference is impossible or unsatisfactory, we have drawn on other ways to arrive at causal inferences: those based on Mill's Method of Agreement and others where the basic causal unit is a combination of causes or a causal chain, rather than a single cause (see for example the discussion in Chapters 3 and 4 below). This does not prevent evaluators and policy makers using counterfactual thinking as a mind-exercise or imaginary to think through different policy options. However this use of counterfactuals should not be regarded as the foundation for a methodology of causal inference or indicative of the best method. Even when counterfactual inference is possible and can be applied flawlessly, it can still be misleading. Counterfactuals answer contingent, setting-specific causal questions 'did it work there and then' and cannot be used for generalization to other settings and timeframes, unless they are accompanied by more fine-grained knowledge on the causal mechanisms actually operating within the process leading from potential cause to effect.

2.13 Another issue that is discussed in detail in Chapter 4 is the kinds of designs and methods suited to different kinds of evaluation questions. For example, Jonas et al (2009) observes that RCTs can answer 'does it work?' questions but not 'how could it be improved?'. Cartwright and Munro (2010: 265) note that experiments can answer the question 'did it work here?' but not 'will it work for us (elsewhere)?'. For the latter purpose 'we need theory to judge which factors have stable capacities (to promote the outcome, and how) and to hypothesize when they are worth implementing'. Ravallion (2009) makes a similar argument that RCTs are strong on internal validity (what works here and now) but are far weaker on external validity - whether an intervention would work in a similar or contrasting setting.

2.14 Whatever the limitations of experimental and quasi experimental approaches its advocates have foregrounded important questions about the strength of methods used by evaluators (see Chapter 6). The proponents of counterfactual and experimental ideas have succeeded in challenging all evaluators to address questions of causal claims and explaining the effectiveness of development interventions in ways that because they are reliable can allow lessons to be learnt.

2.15 It is these challenges that provide a backcloth for this study with its aim to identify a broader range of designs and methods that can potentially address causal analysis of 'impacts', contribution and attribution; and also answer a full range of evaluation questions that are important to policy makers. This is the focus of Chapters 3 and 4 of this report.

2.3 Impact and the 'Evidence Based Policy' movement

2.16 Different definitions of 'impact' and 'impact evaluation' are themselves embedded in wider policy-related debates. Thus the OECD-DAC definitions of impact reflects important shifts in understanding and values surrounding aid and development including shifts in the relationship between donors and partner countries – the beneficiaries of aid. On the other hand the methods-led definitions of impact evaluation have grown out of a quite different debate about 'evidence based policy'. This is concerned with how evidence from research and evaluation should inform policy making and policy decisions. Understanding these debates gives texture to our own definitions of IE by highlighting the context for policy-makers concerns.

2.17 The search for 'what works' - for the causal underpinning of interventions - is the basis for the Evidence Based Policy (EBP) movement. This movement argues that policy makers should take decisions based on evidence rather than on ideology or in response to special interests (Davies et al 2000, Nutley, Walter and Davies 2007, NAO 2003). The movement arose out of twin concerns. First that systematic knowledge and analysis was often not utilized by policy-makers; and second that policy makers had no way of judging the trustworthiness of the findings with which they were bombarded by academics, pressure groups and lobbyists.

2.18 Related ideas have inspired other rationalist governance 'reforms' of the late 20th century, including 'new public management', 'results based management' and of course evaluation more generally. However the origins of EBP were medical or more specifically based on protocols for quality assuring drug-trials. This goes some way to explaining the particular approach to judging causality through experiments that EBP and advocates of first generation IE favoured.

2.19 The EBP movement started from two main assumptions. First, it is assumed that there are only some specific methods that can 'rigorously' produce evidence of the efficacy of policy interventions. These superior methods (described by proponents as a 'gold standard') and coming from the bio-medical tradition favour RCTs both as a way of answering questions of attribution but also as a foundation for generalisation through the synthesis of the results of separate experiment. The second assumption was that once policy-makers are confronted with high-quality evidence of impact they will be more willing to use that evidence and take on-board the lessons of systematic research or evaluation. Understanding what works and what does not then feeds into a process of learning and the improvement of policy.

2.20 Both assumptions have been contested in more recent approaches to evidence-based policy (Pawson 2006, Cartwright and Hardie 2012). What constitutes evidence is often not clear-cut. Different forms of 'evidence' from practitioners, beneficiaries, partners and policy-makers themselves, vie with each other in real-world settings. When policy priorities and evidence collide, it is not always evidence that comes out ahead. But more routinely in all policy domains there are some priorities – in international development, humanitarian relief and concerns about 'failing' States - that are over-riding and cannot wait to be proved before being implemented. These are not arguments against EBP; rather they lead to a much more nuanced analysis of the circumstances within which policy is likely to be informed by evidence and what can be done to make this happen.

2.21 The idea that there is a single superior method of producing evidence has also been widely challenged (Ravallion 2009, Deaton 2010). Rather it is generally understood that methods and designs are fit for different purposes and when well-executed all have their strengths and weaknesses – one of the justifications of so called 'mixed-methods'. Furthermore the choice of methods (and overall designs) needs to follow from the kinds of questions that are being asked; whilst also taking into account the settings in which they are to be applied. This underlines the importance of identifying a range of designs; and understanding when each is appropriate.

2.22 EBP was steered by researchers who understandably chose questions amenable to answers that research could provide. This highlights a further weakness of EBP – questions posed by researchers can be distant from the more complex and even messy questions that policy makers need answers to. The importance of asking the right question in IE is discussed further in Chapter 4.

2.23 Finally, most social scientists now recognise that causation without explanation is insufficient for policy learning because policy-makers need to understand why as well as how if they are to use findings from research or evaluation for future policy-making. Explanation depends on the use of theory alongside appropriate methods, because without theory there is no basis for interpretation.

2.24 The evolution of the EBP debate has profound implications for IE. It suggests that:

- there are different kinds of evidence which often clash and have to be either reconciled or judgements have to be made between conflicting policy signposts;
- that evidence and the possibilities of causal inference can be found through many methods and designs;
- that even sound methods have their strengths and weaknesses and need to be combined in 'mixed' or 'hybrid' forms to ensure rigour; and
- there is a need for theory alongside methods to permit explanation, which is necessary for evidence to be useful outside of the single case.

2.25 These ideas are also central to this study and inform in particular the discussion in Chapter 3 of how different designs can contribute to understanding impact.

2.4 An evolving understanding of aid and the aid relationship

2.26 The role of aid in development and the aid relationship between donors and developing countries has undergone significant shifts since the late 1990s. For example:

- Aid is no longer seen as the sole or even main determinant of development. Since the mid-1990s (OECD-DAC 1996, World Bank 1998) other development drivers such as world trade, governments' own resources and the political economy of developing countries, are seen as working in combination with aid rather than alone. Aid works together with other factors and rarely leads to development results on its own.
- The focus on poverty reduction, MDGs and other 'Internationally Agreed Development Goals' (IADGs) such as gender equality, highlights the equity question: 'development for whom?'. It also places limits on how far aid can be managed by donors in traditional ways. This includes how the 'impacts' of aid can be evaluated in partner countries that take responsibility for their own development, thus placing a new emphasis on capacity development, decentralised evaluation and participatory and empowerment approaches to aid delivery and evaluation.
- Middle income countries and those coming out of the lowest quartiles of extreme poverty face particular problems of inequality and deprivation. These circumstances as in developed countries require multi-dimensional interventions at the policy, programme and

project level. Up to now IE has been less successful dealing with these multiple and overlapping interventions which require system-wide evaluation strategies and designs.

- Notions of partnership at the heart of the aid-relationship are most clearly spelt out in the Paris Declaration and the Accra Agenda for Action with their commitment to country ownership and leadership and to donor harmonisation. Development programmes now rarely involve the direct delivery of services by donors. Indirect modes of delivery and reliance on 'country systems' as in Sector Wide and General Budget Support are more common and create specific challenges for accountability and when assessing VFM. Long implementation chains; partnerships with governments and civil society; and bottom-up or devolved programmes are characteristic of today's 'complex' interventions.
- The frequency of humanitarian emergencies, post conflict stabilisation and the containment of terrorism have all focussed attention on governance, peace-building and the empowerment of civil society actors. Many development interventions are now longer-term and high risk for which there are few pre-tested policy models or Theories of Change. These programmes are 'emergent' and their impacts are often difficult to evaluate or measure using established tools.

2.27 The evolving nature of the aid relationship and shifts in aid priorities and modalities has many consequences for IE designs. Such designs have to be:

- Appropriate to the characteristics of programmes, which are often complex, delivered indirect through agents, multi-partnered and only a small part of a wider development portfolio.
- Able to answer evaluation questions, that go beyond 'did it work' to include explanatory questions such as 'how did it work', and equity questions 'for whom do interventions make a difference?'.
- Able to address multi-dimensional overlapping interventions that are often also long-term in their ambitions.
- Adapted to devolved, partnership and trust-based relationships with the recipients of aid that also limit what donors can evaluate on their own.

2.28 These issues are discussed in greater depth in chapter 4 below.

2.5 Defining impact evaluation

2.29 The above discussion has reinforced the relevance of the OECD-DAC definition of impact as an appropriate starting point for the way we define IE. It is broad enough to accommodate the complex and long-term nature of the interventions we have been asked to work with; it acknowledges unintended consequences and uncertainty; and does not place restrictions on the range of designs and methods that could be used.

2.30 At the same time, the concerns of those evaluators who have come out of the EBP tradition cannot be ignored. In particular their wish as far as possible to establish a clear link between causes (interventions) and subsequent effects has not been sufficiently emphasised by evaluators hitherto. However given the way development aid is understood nowadays we are assuming it more likely that 'causes' will contribute to 'impacts' rather than be more exclusively connected as the term 'attribution' implies.

2.31 On this basis our definition of IE is as follows:

- evaluating the positive and negative, primary and secondary long-term effects on final beneficiaries that result from a development intervention;
- assessing the direct and indirect causal contribution claims of these interventions to such effects especially for the poor whether intended or unintended; and
- explaining how policy interventions contribute to an effect so that lessons can be learnt

2.32 Although this definition follows the OECD-DAC definition in its focus on the 'long-term' it is assumed that even IE will always pay attention to intermediate effects as a step along the way. As subsequent Chapters will argue impact evaluators must be interested in 'implementation trajectories' and the 'shape' and speed of movement along a causal pathway (Woolcock 2009). This is also consistent with the logic of 'Theories of Change'.

2.33 This definition is also open as to whether IE will always be able to measure impact. Understandably policy makers will want to answer 'how much' and 'to what extent' questions wherever possible for accountability purposes. However as discussed further in Chapters 3 and 4 long-term interventions and more complex programmes around governance, accountability and democracy may not be amenable to measurement with interval scale data and estimates of net-effects. On the other hand as the following Chapters also argue, there are still other ways available to answer accountability questions. For example, monitoring of results and estimations of 'distance travelled' remains possible; and the measurement of sub-parts of programmes using 'hard-edged' methods (electoral interventions which are just one part of a wider 'Governance' or 'Democracy' programme) may also be appropriate even if a full large-scale programme cannot be evaluated in the same way.

Conclusions

2.34 This chapter has taken the OECD-DAC definition of impact with its emphasis on long-term effects as the basis for defining IE. The definition of IE looks to establishing cause and effect relationships between development interventions and development results, although not exclusively through counterfactual-based methods. To that extent our definition of IE emphasises the logic of contribution as much as attribution. Some argue that this distinction is weak insofar as both perspectives are interested in making causal claims and that the distinction is between quantitative measurement and qualitative judgements. Furthermore all those who see themselves as working with rigorous attribution analysis are also concerned with other causes that affect the 'variable of interest'. However as the discussion in Chapters 3 and 4 argues as we understand it, the distinctive character of a contribution perspective focuses less on a single cause and more on the way different causal and supporting or contextual factors combine together.

2.35 This emphasis on 'contributory' causes is a major theme in this study consistent with the broad consensus that development aid interventions work best in combination with other non-aid factors. A contribution-based logic is also consistent with the complex and multi-dimensional nature of many contemporary development interventions. The Paris Declaration has led to an increase in devolved and locally customised interventions which increasingly rely on devolved evaluations often jointly conducted by donors and beneficiary countries.

2.36 Finally we have introduced the argument that IE designs need to be able to address a full range of evaluation questions that policy makers are interested in. These include explanation and lesson-learning questions as well as the accountability question ‘has this intervention worked?’. This concern for evaluation questions is taken up in greater detail in Chapter 4.

2.37 Because policy makers often need to generalise, we have emphasised learning and hence explanation in the way IE is defined. Evaluators sometimes distinguish between learning and accountability purposes of evaluation. This distinction is not always clear-cut, for example potentially accountability can and should reinforce policy learning. However short-term accountability evaluations that do not address long-term development effects (even if only by anticipating longer-term causal chains) are on the margins of how we have defined IE.

Chapter 3: Choosing Designs and Methods

Introduction

3.1 This chapter focuses on 'choosing' designs and methods because it starts from the assumption that there are alternative approaches to conducting an IE; and the most suitable designs needs to be deliberately chosen in each instance. The chapter starts by clarifying the distinction between designs and methods – and between evaluation design and research design. In order to focus more narrowly, it then discusses possible ways that IE can deal with 'causes' and explanation. This was already noted as critical to IE because of the need to establish links between interventions (i.e. causes) and effects.

3.2 As the ToR for this study emphasises, with the growing importance of 'qualitative' methods the distinction between quantitative and qualitative approaches is both clarified and challenged. This is done explicitly in two contexts: new paradigms of causal analysis that reject a simple distinction between quantitative and qualitative methods; and the potential of combining different methods – quantitative and qualitative as well as different qualitative methods.

3.3 We arrived at a limited set of designs via two 'routes'. First a selection of existing evaluations that claimed to address 'impact' was reviewed. Second the literature on causal inference was reviewed. This review is summarised in the chapter and the full review is annexed.

3.4 This is introductory to a synthesis of 'design options': these are presented in a table that allows for a degree of comparison between approaches and is followed by a short narrative description of each of the major designs identified. It is important to emphasise that consistent with a commitment to mixed methods we are not suggesting that these designs are alternatives: most real-world evaluations will involve hybrid designs.

Main Messages

- Approaches to 'causal inference' widely understood in the social sciences are not well established in IE.
- 'Theory-based', 'Case-based studies' and 'Participatory' approaches can offer IE real potential to link causes and effects.
- Recent developments in methodology have blurred the distinctions between quantitative and qualitative methods making combinations of methods more feasible.
- Studies of 'cases' that combine within-case analysis and comparisons across cases are especially suited to IE in complex settings.
- Methods that strongly support causal inference work best for narrowly specified interventions while methods that work well for broadly specified interventions are less strong on causal inference even though they can support plausible judgements of effectiveness.

3.1 Designs and Methods

3.1.1 Distinguishing designs and methods

3.5 The ToR for this study is concerned with 'designs and methods'. In the social sciences a 'design' refers to the overarching logic of how research is conducted. For King, Keohane and Verba (1994) a design should consist of four elements: research questions, theory, data and the use of data. We have focussed more narrowly – on forms of theory and uses of data that support causal inference. In these terms 'experiments' or 'theory based' evaluations (TBE) or 'case studies' would be examples of design approaches. As is explained below, particular designs within these broad approaches might include RCTs under the general category of experiments; 'realist' (or mechanism-based designs) under the general category of TBE; and ethnography under the general category of case-based studies.

3.6 Different designs may share similar methods and techniques: both experiments and case studies may use interview data, draw on focus groups and analyse statistical data. What holds them together is the fundamental logic of a design not their methods. This distinction is important for a number of reasons. For example, it maintains attention on causality which in our understanding is at the heart of IE; and it opens up the possibility that many of the difficulties with mixed methods are at the level of mixed designs, i.e. underlying assumptions rather than at a more technical level.

3.7 This tidy classification is not so tidy in practice: some designs are packaged in such a way that they can include elements that are both designs and methods – this could be true for interrupted time-series in relation to quasi-experiments; or comparative case analysis in 'configurational' designs. Real-world designs are also rarely pure types: as we discuss below 'hybrid' designs combining, for example, statistical models with in-depth case-studies or participatory designs with statistical surveys, have become common in development as in other areas of evaluation.

3.8 It would also be possible to impose a much more general classification, distinguishing only between true experimental designs that are based on control or 'manipulation' and other 'observational' studies. This distinction accounts for important differences in how far it is possible to draw causal inference and underpins the distinctive claims of RCTs. In those circumstances where 'manipulation' is possible causal claims are indeed strong. However this risks blurring the distinction between the different contributions of many different types of qualitative and quantitative design which in the right circumstances and when properly implemented are able to support causal and explanatory analysis.

3.9 In following sections of this Chapter major design types that can be considered when planning an evaluation are identified. This discussion is intended to support informed choice when commissioning an evaluation: about what designs are suitable given the type of intervention, the availability of data, the questions being asked and the wider setting.

3.1.2 Designs for research and evaluation

3.10 The logic of 'designs' is based on research – which raises the important question: how far is research design applicable to IE?.

3.11 There are for example 'real world' constraints in evaluation that are not the same as for research. These include constraints such as time, budget, data availability and skills (Bamberger, Rugh and Mabry 2006). Questions for most evaluations are set by policy makers or commissioners not scientists as they are for research. In evaluation there are also other 'design' choices – about the purpose of an evaluation (which can include lesson-learning as well as demonstrating impacts); about how to ensure utilisation; how to involve stakeholders and, in development settings in particular, to acknowledge country ownership; and how to address ethical dilemmas when working with very poor and marginalised people.

3.12 However we would argue following a definition of IE that specifically emphasises the causal links between interventions and 'effects', evaluation is inevitably pushed towards the norms and principles of scientific research. Without designs and methods, theories and quality standards there is no firm foundation for identifying causes and offering explanation.

3.2 Causal attribution and explanation

3.13 Causation is at the heart of traditional science – and social science. This topic has been central to methodological and philosophical debates since Compte and Hume, and was systematised in particular by J S Mill.4 These debates are very much alive today: there continue to be different schools each with respectable foundations that take fundamentally different position about causations. At the same time there have been moves towards breaking down these divisions in the early 21st century. First there have been serious attempts at 'bridge-building' based on the appropriateness of some approaches to some circumstances and the particular strengths of different traditions in causal analysis. Second there have been challenges to sharp distinctions between quantitative and qualitative methods. This is discussed further in relation to mixed methods (see section 3.5 below) and in relation to new ideas in case-based methods that emphasise the importance of configurations that locate variables in relation to each other and in their context.

3.14 Causation seeks to connect cause with effect and different approaches to 'causal inference' do this in different ways. There are various well-rehearsed classifications of these different approaches, for example in a review conducted for this study (see Appendix) we distinguished between four approaches each of which draws on particular theories of causation.

3.15 These were:

- Regularity frameworks that depend on the frequency of association between cause and effect - the inference basis for statistical approaches to IE.
- Counterfactual frameworks that depend on the difference between two otherwise identical cases – the inference basis for experimental and quasi experimental approaches to IE.

[4] This section draws on the paper prepared by Barbara Befani that is appended to the report.

- Multiple causation that depends on combinations of causes that lead to an effect - the inference basis for 'configurational' approaches to IE.
- Generative causation that depends on identifying the 'mechanisms' that explain effects – the inference basis for 'theory based' and 'realist' approaches to IE.

3.16 Each of these causal approaches has:

- Requirements, i.e. conditions under which they do and do not apply.
- Potential strengths.
- Potential weaknesses.

3.17 For example:

- 'Regularity' requires high numbers of diverse cases; without this it is not possible to capture sufficient diversity (or difference).
- Generative causation is strong on explanation but weak on estimating quantities or extent of impact.
- Experiments are good at answering the question; 'has this particular intervention made a difference here?' but weak on generalisation (external validity).
- Multiple causalities are good at dealing with limited complexity and interdependence but not at unpicking highly complex combinations.
- Both experiments and regularity/statistical association approaches work best when causal factors are independent.
- Neither experiments or statistical models are good at dealing with contextualisation.

Table 3.1: Requirements, Strengths and Weaknesses of Four Approaches to causal inference

	Regularity	**Experiments/ Counterfactuals**	**Multiple causation**	**Generative/mechanisms**
Requirements	Many (or highly diverse) cases. Independent causes.	Two identical cases for comparison. Ability to 'control' the intervention.	Sufficient number of cases. Availability of cases with comparable characteristics.	One case with good access to multiple data sources. Theory to identify 'supporting factors'.
Potential Strengths	Uncovering "laws".	Avoiding bias.	Discovery of typologies. Dealing with limited complexity.	In-depth understanding. Focus on the role of contexts. 'Fine-grained' explanation.
Potential weaknesses	Difficulties explaining 'how' and 'why'. Construct validity.	Generalisation. Role of contexts.	Difficulties interpreting highly complex combinations.	Estimating extent. Risks of bias/ loss of evaluators' independence.

3.18 Key design choices for IE therefore include:

- Matching designs to the settings to which they are best suited.
- Using more than one design to compensate for weaknesses in another, and
- Combining designs and methods – even within the same design 'approach' – so as to strengthen causal claims.

3.19 These choices are taken further in discussing 'mixed methods' below and Quality Assurance and quality standards in Chapter 6.

3.3 Learning from existing IEs

3.20 Approaches were made to 30 contact points: development agencies, NGOs, consulting firms and individual consultants, requesting that they provide examples of IEs that went beyond experimental and quantitative methods.[5] In the event having assessed some 20 eligible IEs 10 were selected for in-depth analysis. It is worth noting that many of those who were approached and who responded said that they did not undertake IEs.

[5] Several provided evaluations that fell outside the categories requested – e.g. econometric or experimental evaluations; and some respondents provided ToRs of IEs planned or just beginning. In addition DFID and OECD data bases were scanned and nominations were requested from DFID.

3.21 Reviewing existing evaluation examples that use a range of designs[6] highlighted some of the common challenges facing IE. These fall under several categories:

- How impact is defined;
- The extent to which the focus is on the short or long-term;
- The kinds of interventions that are considered or ignored;
- The kinds of designs and methods that are in use and their appropriateness; and
- The extent to which the needs of different groups, especially the poor are considered.

3.22 Overall these evaluations included serious attempts to address impacts at different levels – from basic services through to attitude and government policies; and including environmental and cultural change. However the results of this review confirmed that there were few examples of 'good practice' in relation to a broader range of methods that were not predominantly experimental or quantitative.

3.23 With one or two exceptions the cases were relatively weak in terms of being able to make causal claims. They were often stronger on descriptive inference (drawing lessons for a population from the accumulation of consistent information on a sample) than causal inference based on principals or rules to judge how far conclusions and generalisations on causes and effects were valid. Theories of change were not routinely articulated even when this would have helped draw causal inferences. There was a default position that comparisons and counterfactuals were the ideal form of IE. However there was little evidence of a good understanding of technical aspects of comparative and counterfactual thinking nor consequently of their strengths or limitations.

3.24 Measurement worked best when 'impacts' were specific and could be easily measured (access to services, transport, income levels etc.) and less when dealing with more abstract concepts such as governance, conflict and cultural change. However as qualitative methods were generally weak these kinds of outcomes were poorly covered. In relation to these more conceptual 'impacts' many evaluations faced 'construct validity' problems - doubts about how far what was described or measured was a good representation of the 'impact' of interest.

Construct validity

'Causal inferences are only reasonable if measurement is valid … this depends on careful attention to conceptualisation…… Issues of conceptualisation and measurement are more fundamental than the conventional problem of generalising from a sample to a population' (Collier, Brady and Seawright (2010;197)).

[6] Although given our ToR we deliberately did not review evaluations that used only experimental or quantitative methods where cases combined these methods with others they were considered.

3.25 Different designs and methods are best able to deal with different kinds of causal relations. For example in some programmes there is one main intervention and in others several are implemented as part of an overall programme. Some programmes aim for a single outcome or impact whilst others aim for several. Interventions may be multiple or single, interdependent rather than independent.

Table 3.2: Different Causal Patterns

Types of causal relation	Examples
One cause (the intervention) associated with one outcome.	A livelihood programme that aimed for an immediate reduction of income poverty.
One cause (the intervention) associated with multiple outcomes.	A programme to improve road infrastructure that aimed to thereby improve travel and transport, commerce, and access to basic services.
Multiple causes (or interventions) associated with multiple outcomes.	A 'deepening democracy' programme that combined supporting election with training members of parliament, and encouraging a culture of accountability in political parties.
Multiple causes (or interventions) associated with one main outcome.	Improving maternal health by a combination of improved neonatal services, health education, midwife training and targeting low income families.

3.26 We encountered examples of all of these 'causal patterns' in the evaluation cases reviewed however they did not appear to be reflected in the design of these evaluations. The review of existing evaluations was suggestive of various more specific conclusions, although they needed to be treated with some caution given the small numbers of cases considered:

- Many of the 'impacts' chosen would normally be classified as outcomes or intermediate effects and there were few attempts to links the shorter-term with the longer term. There were references to 'capacity development' which seems to be widely recognised as key to the longer term sustainability of the effects of interventions but capacity development was not then evaluated. There seemed to be limited understanding about how to evaluate capacity development from the cases we reviewed.
- These evaluations mostly focussed on short-term or intermediate outcomes and impacts. In some cases it would have been possible to have linked longer term effects with intermediates, by using a time-related Theory of Change. However this weakness

can be seen as much as a reflection of when these IEs were commissioned as on the quality of how the evaluations were implemented.

- Several IEs had relatively sophisticated ToRs or 'Approach Papers' that were not followed through. Weaknesses that were evident in the evaluation report were anticipated. This suggests that 'real-world' constraints (Bamberger et al 2006) such as budget, skill capacities and timescale may have had some role in how the evaluation was conducted. There are also raised questions about how IEs are structured. For example ensuring that enough time is allocated at the beginning of an evaluation to construct a sensible analytic framework and Theory of Change; or to proceed iteratively and revise an initial evaluation plan as the evaluation unfolds and new issues become clear.

- Common terms that are used in IE were used in inconsistent ways. For example 'rigour', 'Theory of Change' and even 'impact' were variously defined. In some cases there was a sense that the evaluators felt obliged to use certain terms they barely understood as often as possible.

- The analysis of qualitative data was especially weak. It was generally seen as 'supporting' quantitative data rather than a data source in its own terms. Even where potentially useful data appears to have been collected (e.g. via case studies or focus groups) this data was not always reported and where it was analysis did not use state-of-the-art designs and methods.

- Evaluation questions were poorly expressed – sometimes too narrow, at other times too broad and on occasions completely absent. Even when evaluation questions were clearly formulated they were only loosely connected with the methodologies adopted which may have been unable to answer the questions put forward.

- There is little assessment of contexts. The term is certainly used but inconsistently and without any typologies or theories. Context was often spatially defined (the country or district) but rarely linked to a causal analysis. This limited the potential for causal inference.

- Stakeholders are seldom involved in the IE process; participatory evaluation is usually a module within an overall design rather than an overarching principle. However the issue of "impact for whom" is usually addressed, in particular by dividing beneficiaries into target groups differing on gender, income, poverty level, risk-level and geographical area. The importance of addressing inequality and poverty seemed especially strong in DFID commissioned evaluations.

- There were some indications of bias in relation to both quantitative and qualitative aspects of the evaluations considered. For example there was some cherry-picking: successful cases were emphasised in preference to examples of failure; or failures were given less attention. The generally weak specification of designs included the absence of built in 'quality assurance' procedures and standards.

- Standards (when they are addressed) are mostly understood as statistical robustness and bias control. In some cases, validity, appropriateness for causal questions and relevance to goals are also recommended, as well as good practice for mixed methods research and ensuring the quality of the Theory of Change. However the dominant model of QA even for qualitative methods was derived from quantitative rather than qualitative logics.

- Evaluations take place either ex-post or ongoing; in the latter case, it is because programmes last several years. The time taken for impact to occur and the moment when the evaluation takes place are not always consistent or connected: when ultimate impacts have not yet developed, evaluations usually address intermediate outcomes or partial, "early" impact. However, while some narrative is devoted to it, the time 'trajectory' of the programme is seldom incorporated in the evaluation design.

3.27 Although this review mainly underlined weaknesses in current IE practice outside of experimental and statistical practice, it also pointed to important gaps that have been addressed in later stages of the study. For example the importance of evaluation questions, the subject of Chapter 4; the need for standards and quality assurance processes, the subject of Chapter 6; and the importance of capacity and skills in the evaluation and evaluation commissioning community that is addressed in Chapter 7, under 'next steps'.

3.28 As a complement to this work on cases IE guidance put out by a cross-section of development agencies was also scanned7. This was useful in explaining the approaches taken in the cases reviewed: these cases mainly followed the guidance that is currently available. For example:

- Experimental methods were described as the 'gold standard' to be applied wherever possible in most of these guides.
- Qualitative methods were generally regarded as complementary to quantitative methods.
- Theory based evaluations were also seen as mainly a way of interpreting quantitative results or deepening understanding of causal mechanisms derived from experiments (3ie, NONIE Guidance). In only two cases (GEF and NONIE Subgroup 2) was theory based evaluation seen as a core approach.
- It was almost always assumed that only experimental and statistical methods could be rigorous. In general this term was associated with counterfactual methods rather than to the underlying logic of an approach or the way IE was conducted.
- Participatory methods received relatively little attention. One guide suggested it was not appropriate; and in another that whilst providing insight it was not a rigorous or a sound design. NONIE Guidance - NONIE Guidance Subgroup 2 associated participatory methods with testing for validity and identifying alternative explanations.

3.29 An implication of this relatively cursory overview was that it would be important to instantiate the findings of this study within some form of guidance in order to

[7] Guidance was sourced from: NORAD, 3ie – various Working Papers, World Bank – IE in Practice, World Bank-DIME, GEF Outcomes to Impacts Handbook, NONIE Guidance (Leeuw and Vaessen 2009), NONIE Subgroup 2 (2008), German Federal Ministry for Economic Cooperation and Development (BMZ), ODI - Improving Impact Evaluation

disseminate methods and designs and support more diverse practice among those undertaking IE.

3.4 Design Approaches

3.30 On the basis of the review of methodological literature and of existing IE cases that used a 'broader range' of methods a limited shortlist of designs and methods was chosen. In the Table 3.3 that follows these are briefly summarised. The table starts from families of designs, labelled as 'Design Approaches', indicates some specific variants within each approach and then cross-refers to the previous discussion on causality and explanation[8].

[8] Although this table includes a full range of potential designs, most of the discussion in this and other chapters concentrates on qualitative and non-experimental designs and methods highlighted in the ToR for this study.

Table 3.3: Design Approaches, Variants and Causal Inference		
Design Approaches	**Specific Variants**	**Basis for Causal Inference**
Experimental	RCTs Quasi Experiments, Natural Experiments	Counterfactuals; the co-presence of cause and effects
Statistical	Statistical Modelling Longitudinal Studies Econometrics	Correlation between cause and effect or between variables, influence of (usually) isolatable multiple causes on a single effect Control for 'confounders'
Theory-based	*Causal process designs*: Theory of Change, Process tracing, Contribution Analysis, impact pathways *Causal mechanism designs*: Realist evaluation, Congruence analysis	Identification/confirmation of causal processes or 'chains' Supporting factors and mechanisms at work in context
'Case-based' approaches	*Interpretative:* Naturalistic, Grounded theory, Ethnography *Structured*: Configurations, QCA, Within-Case- Analysis, Simulations and network analysis	Comparison across and within cases of combinations of causal factors Analytic generalisation based on theory
Participatory	*Normative designs*: Participatory or democratic evaluation, Empowerment evaluation *Agency designs*: Learning by doing, Policy dialogue, Collaborative Action Research	Validation by participants that their actions and experienced effects are 'caused' by programme Adoption, customisation and commitment to a goal
Synthesis studies	Meta analysis, Narrative synthesis, Realist based synthesis	Accumulation and aggregation within a number of perspectives (statistical, theory based, ethnographic etc.)

3.31 We see three main design approaches that are not currently widely deployed in IE as offering considerable potential for linking interventions with outcomes and impacts, these are:

- Theory based approaches.
- 'Case-based' approaches.
- Participatory approaches[9].

3.4.1 Theory-based approaches

3.32 In order to explain we need theory to bridge the gap between data and interpretation of that data; and in the case of IE to bridge the gap between 'causes' and 'effects'. The term 'theory' is used in evaluation in various and inconsistent ways. However we can identify two main tendencies in theory-based approaches likely to have traction for IE purposes. In practice these two tendencies are inextricably interwoven and most practitioners would combine them (see for example Funnell and Rogers 2011).

3.33 The first tendency is process oriented. It regards the programme as a 'conjunction' of causes that follows a sequence. It follows the pathway of a programme from its initiation through various causal links in a chain of implementation, until intended outcomes are reached. The process is built around a 'theory' – a set of assumptions about how an intervention achieves its goals and under what conditions. This was the founding logic of how the Aspen Round Table and Kellogg Foundation were originally using 'Theories of Change' (ToCs) and continues to inform various flowchart-like representations of 'Causal' and 'Impact' pathways.

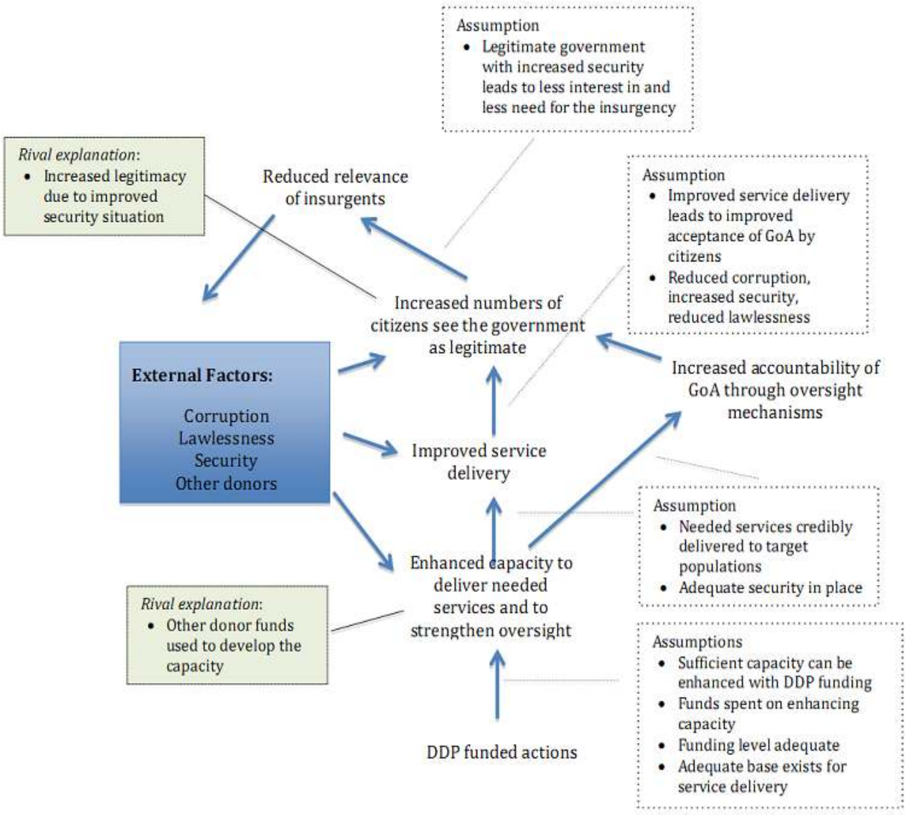
3.34 There are weak and strong depictions of a programme's theory. In weak forms it can be little more than a logic model that expresses the intentions of policy makers ignoring the actions and intentions of other stakeholders; and not making clear the assumptions that are made about the conditions under which the programme success is assumed. Stronger versions do take into account these other understandings. Evaluation and IE in particular, is an opportunity to test a programme's theory through the links in the causal chain. In terms of method, this tendency is close to 'process tracing' (George and McKeown, 1985, Collier 2011), defined by Aminzade (1993) as: 'theoretically explicit narratives that carefully trace and compare the sequences of events constituting the process…'. These causal chains are represented graphically as causal maps or neural networks.

3.35 The following figure (**Figure 3**) is an example of a causal map for a 'Theory of Change' that illustrates a theoretically explicit narrative for the UK's 'District Development Programme' (DDP) in Afghanistan.

[9] 'A distinction is made here between participation as a way of *conducting* an evaluation - well established in the evaluation of international development – and participation as a theory that can be used to link cause and effect.'

Figure 3

Theory of Change for DDP


Assumption
• Legitimate government with increased security leads to less interest in and less need for the insurgency
Rival explanation:
• Increased legitimacy due to improved security situation
Reduced relevance of insurgents
Assumption
• Improved service delivery leads to improved acceptance of GoA by citizens
• Reduced corruption, increased security, reduced lawlessness
Increased numbers of citizens see the government as legitimate
External Factors:
Corruption
Lawlessness
Security
Other donors
Increased accountability of GoA through oversight mechanisms
Improved service delivery
Assumption
• Needed services credibly delivered to target populations
• Adequate security in place
Enhanced capacity to deliver needed services and to strengthen oversight
Rival explanation:
• Other donor funds used to develop the capacity
Assumptions
• Sufficient capacity can be enhanced with DDP funding
• Funds spent on enhancing capacity
• Funding level adequate
• Adequate base exists for service delivery
DDP funded actions

3.36 ToCs and related descriptions of causal chains can also consider 'causal mechanisms' in greater depth although they more usually focus on a sequence of decisions or actions rather than on mechanisms. To clarify the distinction, designs that are explicitly focussed on mechanisms are discussed separately.

3.37 Mechanism focused understandings, build on the widespread assumption in the philosophy of science that in order to be able to make plausible causal claims there needs to be an identification of a mechanism (or what Cartwright calls 'capacities') that make things happen. Programme mechanisms 'take the step from asking whether a programme works to understanding what it is about the programme that makes it work' (Pawson and Tilley 1997:66). An education intervention ensures that children are able to remain at school after the age of eleven. The mechanism that ensures this leads to positive educational outcomes is the interaction between a child and a teacher in the school. However there are other conditions variously understood as 'contextual' or 'helping conditions' that need to be in place. These include: a) availability of a teacher, b) the training that the teacher has received, and c) to house both pupil and teacher the prior construction and equipping of a classroom. Mechanisms are not therefore 'context free' – they are contingent and depend on various starting conditions, supporting factors and predispositions to be sure that they will operate.

3.38 Designs that are mechanism-based look for the connection between cause and effect through in-depth theoretical analysis, rather than by demonstrating regularity or inevitability. Instead of comparing cases generative causality explains how causal factors interact. The causal explanation is not a matter of one element (X), or a combination of elements (X1.X2) asserting influence on another (Y); rather it is the association as a whole that is explained (Pawson 2007).

3.39 Most 'theory' and 'Case' oriented approaches that stem from a 'realist' understanding of the world also assume that even if similar mechanisms are in place, this does not guarantee a common outcome. If the context is different or if various 'helping' or 'support' factors are absent, there will not be common outcomes. The search for regularities and generalisations within theory-based evaluations depends on a combination of a mechanism in its context. It is noteworthy that 3ie has in recent years taken on board this aspect of theory-based approaches when looking to generalise beyond the site of specific interventions (White 2009). At the same time mechanism based understandings such as realist evaluation that analyse only a single case can be subject to external validity threats like RCTs. This can be overcome if theory development and testing in a particular case forms the basis for more general theory. This can then provide insights and even short-cuts that make it unnecessary to repeat all aspects of a mechanism based evaluation in all settings[10].

3.40 Causal mechanisms operate in specific contexts making it important to analyse contexts. This is often to be done only in an ad hoc way but typologies of context are a useful intermediate step towards generalisation in mechanisms based evaluations. Contexts may include related programmes affecting the same target population; socio-economic and cultural factors; and historical factors such as prior development initiatives. Developing typologies of context whilst not supporting universal generalization can support 'under these conditions' types generalizations.

3.4.2 Studies of the 'Case'

3.41 The choice of this title is intended to take us beyond traditional understandings of 'case-studies'. New methodologies for the systematic causal analysis of 'cases' emerged around the turn of the century. Cases may be policy interventions, institutions, individuals, events or even countries during a particular historical period[11]. This represents a shift from focusing causal analysis on variables taken out of their specific context. Locating variables in the context of the 'case' and conducting within-case analysis alongside comparisons across cases has opened up major new opportunities for causal analysis that are still largely ignored in evaluation practice.

3.42 There has been a tradition in evaluation of 'naturalistic', 'constructivist' and 'interpretative' case studies that generally focus on the unique characteristics of a single

[10] This is also the justification for 'realist' synthesis, one of the approaches to synthesis (see Table 2.1) that are now receiving much attention among evaluation practitioners.

[11] It is also important to note that 'cases' are generally understood as 'complex' systems. This is considered in greater depth in Chapter 4 in relation to programme 'attributes'.

case. These case studies eschew causal analysis even though they contribute to such analysis in several ways. For example interpretative case studies provide a rich understanding of contexts in theory based evaluations; help define construct validity in terms that make sense to stakeholders on the ground; and give voice to programme beneficiaries both at the stage that evaluation questions are formulated and when interpretations of findings are being made.

3.43 Newer approaches to studies of the 'case' are specifically interested in causal analysis (George and Bennett 2005; Ragin 2008; Byrne and Ragin 2009). They are also interested in generalising that goes beyond the single case - though not in 'universalising' (Byrne 2009). These approaches tend to generalise 'under certain conditions' and identify clusters or subsets of cases about which it is possible to make similar causal inferences. Case based causal inference may 'test' a theory (as in George and Bennett's notion of 'congruence') and also contribute to 'theory building'. However the role of theory is less pronounced within 'case' designs that aim for causal analysis. For example in QCA causal inference depends more on the conditions that are necessary and sufficient for an outcome basing this on comparisons of 'configurations' of cases and their attributes.

3.44 The single case also features in case based approaches especially through 'within-case' analysis – examining the attributes of a case so as to establish configurations. Notionally it should be possible to conduct single 'case' studies within this framework and locate the case in relation to a previously established cluster of other similar programmes or similar institutions.

3.45 One noticeably underdeveloped design that has been included in the table above is newer forms of simulation or modelling. 'Systems Dynamics' and 'Agent based modelling' in particular could be a useful design to conduct 'virtual' experiments with single-case complex interventions. These designs are beginning to be applied in some areas of policy research, but only being spoken of in international development (Porter 2011), and are entirely absent from IE in this sector.

3.4.3 Participatory Approaches

3.46 Development policy rests heavily on ideas – such as participation, ownership and democratic accountability just as major programme areas are also concerned with related ideas of 'Democracy', 'Transparency', 'Empowerment' and 'Accountability'. Paris Declaration principles such as 'ownership' and 'mutual responsibility' have a strong participatory flavour as does the overall philosophy of 'partnership' between donors and 'partner countries'. Development evaluation has expressed similar values through methods such as Participatory Rapid Assessment, Participatory Action Research, Most Significant Change (Dart and Davies 2003) etc. These approaches relate to IE, even if only indirectly. For example they can:

- Ensure that beneficiaries and the poor have a voice when programmes are planned thus improving targeting and relevance.
- Investigate local communities and circumstances, clarifying problems and constraints thus improving 'construct validity'.

- Add a beneficiary and stakeholder perspective to the conclusions and lessons learned from an IE.

3.47 This last point raises the question 'who participates?'. In development programmes those participating may include beneficiaries but may also include country-based officials and decision makers. There will be different implications from different patterns of participation. For example the participation of decision makers may have implications for implementation efficiency and sustainability.

3.48 To be seen as a design that contributes more directly to IE, participatory approaches need to support causal inference. One reading of the Paris Declaration (PD) and the wider 'consensus' on aid partnerships suggest that there is a basis for using participatory logics in this way. Specifically the proposition in the PD is that if there is ownership[12] by developing countries of their development plans – it is more likely that these plans will succeed. This perspective is also consistent with David Ellerman (2005), that it is only through self-directed actions that development is possible; and that it is the role of those supporting development to enhance that autonomy. Participatory approaches to causal inference do not see recipients of aid as passive recipients but rather as active 'agents'. Within this understanding, beneficiaries have 'agency' and can help 'cause' successful outcomes by their own actions and decisions. As suggested above this would be the case if country-based decision makers were actively involved in programmes and their evaluation.

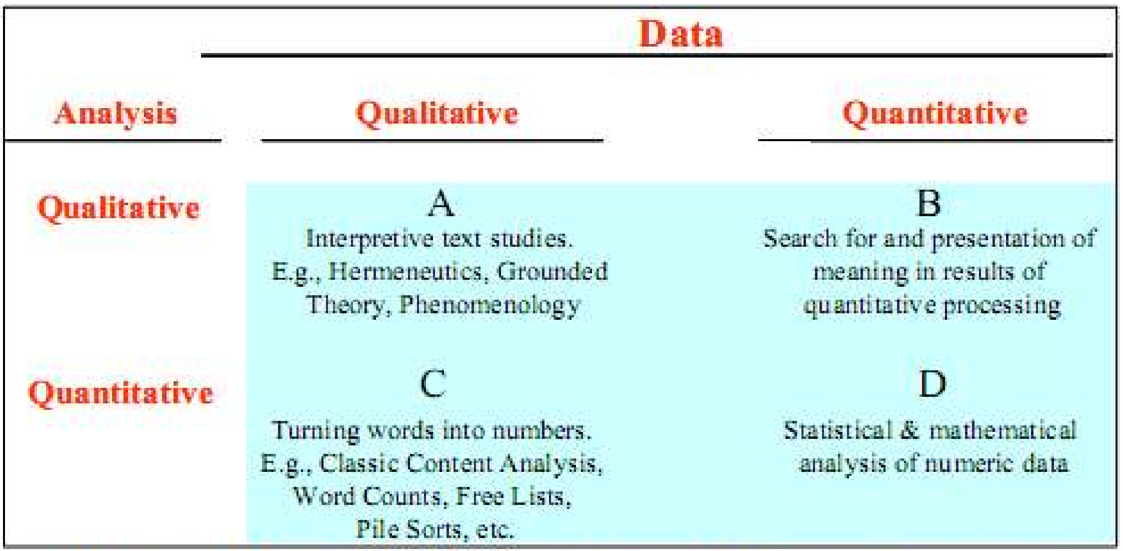
3.49 It should be possible to interpret impacts in terms of an interventions participatory content: for example the extent to which involvement, ownership and commitment improve development outcomes. These processes relate to programme design as well as evaluation design – which is also true of causal inferences using other designs. 'Theory based evaluations' for example look to causal aspects of programmes: participatory designs focus on the 'agency' of stakeholders and participatory methods to demonstrate that this has made a difference.

3.5 Qualitative, Quantitative and 'Mixed Methods'

3.50 The ToR for this study places an emphasis on qualitative methods but also notes the scope for the complementarity of quantitative and qualitative methods. The distinction between QUAL/QUANT is not straightforward. Collier, Brady and Seawright (2010) reject a 'sharp distinction' between quantitative and qualitative methods. For example, there are many gradations between 'small n' and 'large n'; and different judgements can be made as to 'levels of measurement' and the 'cut-off' point between nominal, ordinal and interval scales. At a basic level there is a distinction between data and analysis, thus Bamberger et al. argue (2010: 3) – 'since a lot of qualitative data can be coded, quantified and econometrically analyzed, this distinction should therefore be more appropriately between data collected from structured, closed-ended questions and non-structured, open-ended, modes of inquiry'. A similar perspective is summarised in the following Table (Ryan, ND).

[12] And post Accra 'leadership' as well

Table 3.4: Key qualitative and quantitative distinctions



	Data	
Analysis	**Qualitative**	**Quantitative**
Qualitative	A Interpretive text studies. E.g., Hermeneutics, Grounded Theory, Phenomenology	B Search for and presentation of meaning in results of quantitative processing
Quantitative	C Turning words into numbers. E.g., Classic Content Analysis, Word Counts, Free Lists, Pile Sorts, etc.	D Statistical & mathematical analysis of numeric data

Adapted from: Bernard, H. Russell. 1996. Qualitative data, quantitative analysis. *Cultural Anthropology Methods Journal* 8(1):9-11

3.51 Although this table highlights the blurring that can easily creep into QUAL/QUANT distinctions, it also reflects the mid 20th century view that associates qualitative methods with constructivism and 'interpretation' of meaning rather than causal analysis. As noted above there has been a shift towards using qualitative approaches as at least as one important part of causal analysis (Ragin and Becker 1992; Miles and Huberman 1994; Pawson and Tilley, 1997; Brady 2002; Yin 2003). However as Denzin (2010) points out most advocates of mixed methods have come from a quantitative (or post-positivist) not qualitative background. There remain questions for qualitative and interpretative researchers about uncritical advocacy of mixed methods. For example: can data derived from different theoretical perspectives be combined – might the data they produce be incommensurable, so the evaluator is left with adding up apples and pears?. Or what happens when different methods lead to different conclusions?.

3.52 There are two points of entry into the debate about the role of qualitative analysis in causal inference. The first is to question whether this is the right distinction to make given: a) a general shift away from 'variable' based to 'case' based understandings of data, and b) new non-statistical but still quantitative approaches to measurement. The second is to advocate the combination of 'multiple' or 'mixed' methods both as a compensatory principle and as a means of meeting criteria for causal inference. These topics are discussed further below.

3.5.1. 'Cases' and 'variables'

3.53 The nature of measurement based on variables that are intended to represent phenomena across settings has been countered by approaches that see the alternatives to a variable approach as 'case-based' or 'set-theoretic' (Abbott, 2001; Brady 2002; George and Bennett 2005; Rihoux and Ragin 2008; Byrne and Ragin (Eds) 2009). Classical quantitative research draws conclusions from variables that are de-contextualised and represent phenomena that may or may not be true to the real world (is the state rated as 'legitimate'; how many small-businesses are owned by women etc.?). Causality is down to 'covariance' among independent variables or causes, which are related to each other in what is assumed to be a predictable way.

3.54 The case-based view on the other hand, regards the case as a complex entity in which multiple 'causes' interact not always predictably. It is how these causes interact as a set that allows an understanding of causes (for example, the way that women owned businesses contributes to economic or income growth can only be understood in the context of sources of finance, stage in the economic cycle and family structure). This view does not ignore individual causes or variables but examines them as 'configurations' or 'sets' in their context. However what a case-based view does refocus on is 'within-case' analysis using process tracing and 'configurational' methods (George and Bennett 2005; Brady and Collier 2010). This contrasts with forms of analysis that solely rely on comparisons across cases.

3.55 Taking a case-based view has often been presented as necessarily 'qualitative' but this need not be so (see earlier description of 'Studies of Cases'). In fact it is argued that: 'Case centred methods are by no means limited to small-N research, but can also offer alternative for large datasets in situations where the limitations and assumptions of mainstream […] quantitative methods restrict conclusions that can be drawn from the analysis of quantitative data' (Ray Kent 2009). The application of Bayesian methods, configurational analysis and 'neural' network analysis can be applied to large data sets; however they do so from a perspective that situates variables in their case context. Furthermore such methods and techniques are often quantitative – even if they are not traditionally statistical.

3.5.2 Multiple or 'mixed' methods

3.56 There is widespread agreement (Greene, Caracelli and Graham 1997; Bennett 2002) on the benefits of combining quantitative and qualitative techniques, stemming from the practice of different disciplines - quantitative from economics, physical sciences, psychology, but also sociology; qualitative from history, anthropology, ethnography, sociology.

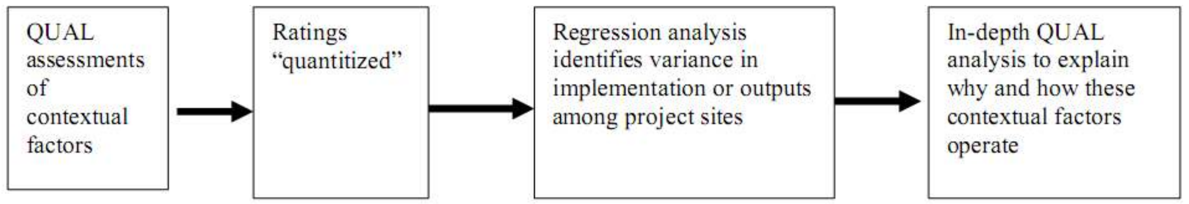
3.57 Different techniques meet specific purpose, from measurement and description of events and states to understanding of a situation or a process, bringing their own strengths and limitations. Combining methods is a way to overcome limitations and enhance strengths. Within this rationale, combining methods offers specific advantages in impact evaluation. For example, adding quantitative methods such as sampling or overcoming the risk of positive bias from field visits with administrative data; and adding qualitative methods to straightforward monitoring helps underpin indicator outputs with an understanding of process, which is central in the explanation of how impacts occur. Multiple methods can help reconstruct baseline data and overcome time and budget constraints.

3.58 Methods can be combined in different ways (Greene, Caracelli and Graham 1997[13]):

[13] This influential paper lists the following aims: triangulation, complementarity, development, new start and expansion. For the purposes of the present work, we consider the ones listed in the text as most relevant.

- '**Triangulation**': confirming and corroborating results reached by one method with other results reached by another method. For instance, when beneficiaries of a project's services state that they judge it good (or bad); this can be cross-checked by collecting quantitative data on coverage and accessibility of the service. Or when survey data indicated that the size of housing units had increased for those affected by a transport improvement programme; this was better understood when community interviews revealed that improved transport had led to a reduction in the costs of building materials. This may be 'parallel' if the aim is cross-checking or 'sequential' if the aim is to integrate methods. The following figure from Bamberger et al 2010 is an example of a 'sequential' design.

Figure 4: An iterative process for analyzing how contextual factors affect project outcomes


QUAL assessments of contextual factors
Ratings "quantitized"
Regression analysis identifies variance in implementation or outputs among project sites
In-depth QUAL analysis to explain why and how these contextual factors operate

- '**Complementarity**': results obtained by a method help better understand those obtained by another method. In-depth theory based approaches may help understand reasons why a project led to unexpected results; qualitative methods may help clarify concepts and define variables; and large-scale data sets may be analysed by multivariate and case-based methods to provide a context within which a small number of intensive case studies can be interpreted.

- '**New start**': diverging results, obtained by different methods may raise new questions and require an adjustment of a previous design. In such cases, new programme theories could be worked out and further tested. This seems particularly relevant with designs dealing with indirect and complex programs where there is little prior knowledge and that need flexible designs.

3.59 A different perspective on combining methods is the possibility of 'nested designs and methods' (Lieberman 2005). This follows from the view that programmes are complex multi-level entities (Cilliers 2001). For example, there might be a 'nested' design where a complex programme can be partitioned perhaps hierarchically or perhaps by analysing the interdependencies within a large scale programme[14]. This may allow different designs and methods to be used in different parts of an evaluation. For example a mainly quantitative experimental design may be applied at the level of a straightforward health intervention; this can then be nested into a mainly qualitative realist 'Theory Based' design to allow for a comparison across several such interventions; and a configurational approach with both qualitative and quantitative elements could be used at the macro-level to compare different programme elements and their effects.

[14] This relationship between complexity and IE design is further elaborated in Chapter 4 on programme.

3.60 Tarrow (2009) provides a useful table about 'bridging' the quantitative-qualitative divide. This table touches on several of the arguments for combining methods that have been raised in this chapter.

Table 3.5: Tools for Bridging the Qualitative-Quantitative Divide (Source Tarrow, 2009)

Tool	***Contribution to Bridging the Divide***
Process Tracing	Qualitative analysis focused on processes of change within cases may uncover the causal mechanisms that underlie quantitative findings.
Focus on Tipping Points	Qualitative analysis can explain turning points in quantitative time-series and changes over time in causal patterns established with quantitative data.
Typicality of Qualitative Inferences Established by Quantitative Comparison	Close qualitative analysis of a given set of cases provides leverage for causal inference and quantitative analysis then serves to establish the representativeness of these cases.
Quantitative Data as point of Departure for Qualitative Research	A quantitative data set serves as a starting point for framing a study that is primarily qualitative.
Sequencing of Qualitative and Quantitative Studies	Across multiple research projects in a given literature, researchers move between qualitative and quantitative analysis, retesting and expanding on previous findings.
Triangulation	Within a single research project, the combination of qualitative and quantitative data increases inferential leverage.

3.6 The 'strength' of causal inference

3.61 Throughout this chapter arguments have centred on 'causal inference': the ability of designs and methods to demonstrate that an intervention as cause leads to an effect. A commonly expressed concern is about the strength of the causal claims that can be made with other than experimental designs. This concern has been addressed by both philosophers of science and methodologists. For example, Shadish, Cook and Campbell leading proponents of experiments doubt the capacity of even correlational designs 'to support strong causal inferences in most cases' (2002:18).

3.62 Cartwright (2007) makes a distinction between two different categories of methods that support causal claims. 'There are those that clinch the conclusion but are narrow in their range of application; and those that merely vouch for the conclusion but are broad in their range of application'. The first type may be strong but 'narrow in scope'. The other type may be of much broader applicability but is only 'symptomatic of the conclusion but not sufficient'. Cartwright places RCTs and theory based approaches in the first category (strong but narrow) and QCA and those that use multiple forms of evidence in the second category (broad but weaker).

3.63 The implications of the above is that whilst designs and methods applied to narrowly specified interventions may lead to strong causal claims this will not be so for broadly specified interventions. At the same time various statistical evaluation approaches that seek to compensate for incomplete data or deploy quasi-experimental methods when control is not possible, also have less statistical power than straightforward experiments.

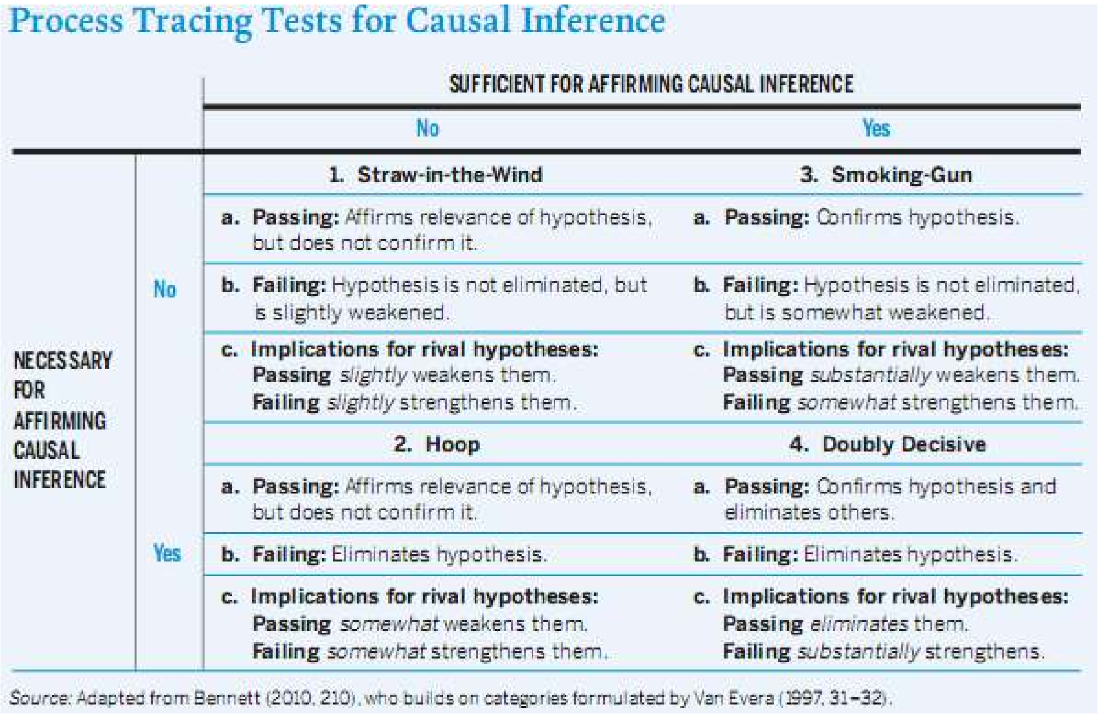
3.64 There have been various approaches to this problem:

- First to devise methods that compensate for the weaknesses of specific methods by combining them with others – an approach discussed in the section on combining methods.
- Second to apply 'tests' to the evidence produced by methods to see how far general causal indications stand up to scrutiny – whether for example the evidence both eliminates rival explanations and confirms a specific explanation[15].
- Third to see how far it is possible to narrow down the scope of an impact evaluation so as to be able to apply 'narrow' but 'strong' methods.

3.65 The extent to which this final strategy is feasible given the attributes of complex and large scale programmes is discussed in Chapter 6. However it should be acknowledged that for many IEs asking policy-relevant questions in relation to complex programmes the best that can be expected is plausible interpretation rather than firm 'proof'.

3.66 The ways in which the strength of causal inference using methods such as 'process tracing' can be assessed is exemplified in table 3.6 (Collier 2011).

Table 3.6:



Process Tracing Tests for Causal Inference

		SUFFICIENT FOR AFFIRMING CAUSAL INFERENCE	
		No	Yes
NECESSARY FOR AFFIRMING CAUSAL INFERENCE	No	**1. Straw-in-the-Wind**	**3. Smoking-Gun**
		a. Passing: Affirms relevance of hypothesis, but does not confirm it.	**a. Passing:** Confirms hypothesis.
		b. Failing: Hypothesis is not eliminated, but is slightly weakened.	**b. Failing:** Hypothesis is not eliminated, but is somewhat weakened.
		c. Implications for rival hypotheses: **Passing** *slightly* weakens them. **Failing** *slightly* strengthens them.	**c. Implications for rival hypotheses:** **Passing** *substantially* weakens them. **Failing** *somewhat* strengthens them.
	Yes	**2. Hoop**	**4. Doubly Decisive**
		a. Passing: Affirms relevance of hypothesis, but does not confirm it.	**a. Passing:** Confirms hypothesis and eliminates others.
		b. Failing: Eliminates hypothesis.	**b. Failing:** Eliminates hypothesis.
		c. Implications for rival hypotheses: **Passing** *somewhat* weakens them. **Failing** *somewhat* strengthens them.	**c. Implications for rival hypotheses:** **Passing** *eliminates* them. **Failing** *substantially* strengthens.

Source: Adapted from Bennett (2010, 210), who builds on categories formulated by Van Evera (1997, 31–32).

[15] The four tests proposed by Andrew Bennett for confirming and eliminating alternative explanations (Bennett 2010); and the application of a checklist of 'basic principles' and associated 'strategies' by Jane Davidson see http://realevaluation.com/pres/causation-anzea09.pdf

Conclusions

3.67 This has been a ground-clearing chapter: clarifying the distinction between designs and methods; different approaches to causal inference and the strengths and weaknesses of each. This literature-based judgement as to strengths and weaknesses is reinforced by reviews of existing examples of IE drawn from the experience of a cross-section of development agencies.

3.68 Three main designs that show promise to reinforce existing IE practice when dealing with complex programmes – theory-based; case-based and participatory – were identified and introduced. How these designs are used in practice is further elaborated in the next two chapters.

3.69 New understandings of case-based rather than variable-based methods; and challenges to a simple distinction between quantitative and qualitative methods have opened up new methodological possibilities for IE. In particular there are now new possibilities and frameworks to combine different methods and different designs in ways that allow causal inference to be systematically assessed.

Chapter 4: Evaluation Questions and Evaluation Designs

Introduction

4.1 This chapter starts from the questions that IE evaluators are faced with. These questions need to be clearly stated because they frame how cause-effect questions can be answered. The chapter revisits issues raised in Chapter 3 by showing how different impact evaluation questions are rooted in different notions of causal inference and the implications for IE design that follow.

4.2 The chapter suggests that some common causal questions such as 'Did the intervention cause the impact?' can be subject to widely differing interpretations. Also, that the question 'What would have happened without the intervention?' widely favoured by evaluators may not be a key question that decision makers need to have answered. As argued in Chapter 2, other non-aid factors are essential for development outcomes. This suggests that key causal questions should aim at better understanding how specific interventions contribute to bringing about intended impacts in combination with other causal factors. The chapter also discusses 'how' and 'why' questions because as argued in both Chapters 2 and 3, explanation is central to generalisation and lesson-learning.

Main Messages

- Evaluation questions matter: they affect which IE designs are feasible and appropriate.
- As foreshadowed in Chapter 3, some methods are stronger on internal than external validity. This may not matter if the evaluation purpose is more about accountability than policy learning.
- Most IE questions about the effects of an intervention are best answered through a 'contributory' lens: successful programme are part of a sufficient 'causal package' – they do not work in isolation.
- There are challenges for transferable learning through IE in today's development architecture built around indirect delivery of aid through partnerships. New forms of 'participation' through joint donor/beneficiary IEs may be one solution.
- To answer most evaluation questions 'hybrid' methods are needed: it is rare that a single design or method will be enough.

4.1 The Traditional Causal Question

4.3 The traditional causal question is: 'Did the development intervention cause the impact?'. This is a seemingly simple question but identifying general characteristics of both causes and effects can be problematic. For example:

- What does the impact look like?. A fixed situation observed at a given time, or a shifting process or trend, comparing states of the world at multiple times (e.g. an improvement from a previous situation)?.

- Even when we know what impact looks like, what can we say about the general form of the cause?. Is it the intervention as a whole?. Just some of its components?. Or particular ways some components have been implemented, perhaps in combination with other contextual factors?.

4.4 In development policy, the characteristics or general forms of causes and effects vary with the characteristics of interventions. These programme 'attributes' are the subject of Chapter 5. However at this point it can be noted that programme attributes affect what kinds of impact questions can be answered as well as design choices.

4.5 Causal questions in IE can take two basic forms. One can simply ask if the intervention resulted in a desired impact. The answer to whether we can make this type of causal claim is essentially either 'yes' or 'no'?. A second, usually more challenging question is to explain the causal link, to demonstrate how the intervention caused the impact and also to explain how it was that the intervention led to or produced the impact. On this basis, Table 4.1 lists four causal questions to be addressed in IE, the first two addressing the yes/no question and the second two the explanatory question.

Table: 4.1 Four Key Questions in Impact Evaluation

1. To what extent can a specific (net) impact be attributed to the intervention?
2. Did the intervention make a difference?
3. How has the intervention made a difference?
4. Will the intervention work elsewhere?

4.6 These questions are close but not identical to the questions asked by Nancy Cartwright who distinguishes between what has 'worked' in a particular place, what 'works' generally and everywhere; and what will work, in a particular setting as part of a future policy initiative[16].

4.2 Impact Evaluation Question One: To what extent can a specific (net) impact be attributed to the intervention?

4.7 There are various ways of framing this question, for example:

- Did the intervention work?
- How much of the impact can be attributed to the intervention?
- What is the 'net effect' of the intervention?

[16] Evidence-based Policy: Doing It Better: A Practical Guide to Predicting If a Policy Will Work for You. Nancy Cartwright and Jeremy Hardie. Oxford University Press. Forthcoming 2012.

- What would have happened without the intervention?

4.8 In classical approaches to causal inference, causality is established by seeking a strong association between a single cause and a single effect, either by observing a regular combined presence of cause and effect in a number of highly-diverse cases (Hume's regularity and Mill's Method of Agreement) or through the observation of quasi-identical cases whereby only the cause and the effect are different (Mill's Method of Difference, or the philosophical basis for experiments and other methods involving the construction of a counterfactual). The cause is mostly conceived as being both necessary and sufficient for the effect; and not usually able to untangle the complexities of causal relations when causes are interdependent and affect outcomes as 'causal packages' rather than independently.

4.9 A debated issue in impact evaluation is the focus on *attribution* and *contribution*. Are we trying to attribute an impact to the intervention or evaluate the contribution the intervention is making to the impact?. Attribution involves a causal claim about the intervention as the cause of the impact, and measurement of how much of the impact can be linked to the intervention. This contrasts with 'contribution', which makes a causal claim about whether and how an intervention has contributed to an observed impact. This usually involves verifying a theory of change in a way that takes account of other influencing factors and thus reduces uncertainty about the contribution the intervention is making[17].

4.10 The attribution question is the traditional question addressed by experimental designs. Experimental designs (RCTs, quasi-experimental and natural experiments) allow one to associate the intervention as a single cause to a measure of the net impact that can be attributed to the intervention. This is accomplished by answering the counterfactual question "What would have happened without the intervention?". Confirmation is sought of the assumption that without the intervention there would be no impact or a different impact - whilst still acknowledging that change is underway anyhow: the focus here is on *additional* change. Typically, this is approached by comparing like situations with and without the intervention, showing both that with the intervention there was an impact i.e. that the intervention was sufficient for a net change in the impact; and that without the intervention, there was no impact (necessity), i.e. that net change – beyond pre-existing trends or business as usual - in those conditions could only be achieved through the intervention (*ibid* Mill's Method of Difference).

4.11 True experimental designs (i.e. RCTs) are a powerful way to answer intervention specific questions about effectiveness. However as with all designs this works best under certain conditions; and in this instance the conditions are stringent. For example, the conditions include:

- Where there is likely to be one primary cause and one primary effect of interest although there are many possible causes – hence difficulties with complex interventions.

[17] This first IE question focuses on attribution. Contribution is discussed in greater detail in relation to subsequent IE questions.

- Where a 'control' group can be identified and contamination between treatment and comparison groups can be controlled.
- Where there is an interest in the success of a specific intervention in a particular setting rather than in wider generalisation.
- Where there are sufficient numbers to support statistical analysis and to provide mean differences or other statistical comparisons between the 'treated' and the control group.
- Where on balance there is more of a need for causal than explanatory analysis.

4.12 If all these conditions do not hold, some of the comparative, case-based designs described in Chapter 3 could also help answer these questions. For example both Rihoux and Ragin argue that QCA can address 'counterfactual' type questions; and participatory methods combined with comparative 'Case' based designs could further strengthen the claim of these alternative approaches. However it is important to recognise that the narrow IE question as to whether a particular focussed intervention works in a particular setting will mainly be useful to policy-makers who want answers to accountability questions about 'this programme, now and under these conditions'.

4.13 It is of course possible to replicate RCTs under many different conditions to begin to build up a 'theory' of appropriateness and effectiveness. As a recent World Bank Working Paper argues:

> ' ….. results may fail to carry over to different settings. Limited external validity of any study would not be a problem if we could replicate it easily. With enough replications, the sheer mass of evidence would provide the desired generality' (Cadot et al 2011).

4.14 However as the authors of this Working Paper recognise the logistical and cost implications of this can be prohibitive.

4.3 Impact Evaluation Question Two: Did the Intervention Make a Difference?

4.15 When we say that one event *causes* another we do not always imply a necessary and sufficient relation between cause (intervention) and effect (impact). Indeed, we may be saying that the first event is any one of the following:

- Both necessary and sufficient: The cause always leads to the intended effect and is the only way to get there.
- Necessary but not sufficient: The cause is a necessary precondition for intended effects but won't make them happen without what Nancy Cartwright calls other 'helping factors'.
- Sufficient but not necessary: The programme is one way to arrive at the effect but there are other ways.
- Neither necessary nor sufficient but a contributory cause: The programme is a vital part of a 'package' of causal factors that together are sufficient to produce the intended effect. However on its own the programme is neither sufficient nor always necessary – if for example other effective causal packages did not include the programme of interest.

4.3.1 The 'contributory' cause and causal 'packages'

4.16 Considering the types of programmes we are dealing with in this study, we would argue that seeing interventions as a 'contributory' cause describes many of them.

4.17 Simple *necessary* causation is a demanding condition. It suggests that the *only* way for the impact to occur is with the intervention. For many of the impacts of interest, such as quality education outcomes, empowerment, enhanced democracy and public accountability this would not appear to be the case. There are a variety of ways such impacts might be realized. For example, building a well does have a direct effect on water supply but water could also be bought-in or diverted from rivers.

4.18 Simple *sufficient* causation is perhaps more promising in that an intervention on its own may be sufficient to produce the impact. But again, the implication is that the intervention would bring about the impact whenever and wherever it occurs. It is rare that any intervention is effective in all and every context. It may have been sufficient for that one location and time – and for that population. But all these qualifying statements suggest that the intervention was in fact a 'contributory' cause and part of a 'causal package'.

4.19 The notion of a 'contributory' cause, recognizes that effects are produced by several causes at the same time, none of which might be necessary nor sufficient for impact. It is support for civil society, when combined with an effective poverty reduction strategy, suitable capacity development and policy coherence in partner government policies that legitimate governance and provide the pre-conditions for enhanced development results. It is unlikely to be support for civil society alone. Just as it is smoking along with other factors and conditions that result in lung cancer, not smoking on its own, so also it is development intervention *along with other factors* that produce an impact.

4.20 Statistical and econometric models can have difficulties with multiple causality and struggle to capture the interactions among variables or represent irregular, complex paths, particularly when irregularity is not well-modelled by probabilistic distributions. It can be difficult for these methods to capture the influence of combinations of causal factors rather than of each causal factor as a free-standing agent.

4.21 In complex settings causes interact unpredictably - in ways more typical of historical events or crises or chemical reactions. Historical events, such as wars and economic crises for example, do not have a single cause: they usually have a triggering cause but many other factors had to "prepare the ground" for the "final" cause to trigger the event. Thus the First World War was triggered by the Sarajevo assassination but this only had potency because of a preceding arms race, colonial competition among the great powers of the day and arguably the automaticity of the Schlieffen plan.

4.22 If a causal 'package', i.e. the intervention plus other factors, is the concept most likely to be relevant in the impact evaluation of complex development projects, this focuses attention on the role of the intervention in that package. Was it a necessary ground-preparing cause, a necessary triggering cause or something that did not make any difference and a similar effect would have occurred without the intervention?. If the intervention was indeed a trigger then a stronger claim becomes possible. If the intervention starts the causal chain and possibly supports change along the way it is possible to claim that it was the intervention that made the difference because it was an initiating contributory cause.

4.23 The idea of a 'contributory' cause[18], came to prominence in epidemiological studies of the role of tobacco as a cause of lung cancer. As part of a causal package of other lifestyle, environmental and genetic factors cigarettes can cause cancer; but they need not and sometimes cancer can be 'caused' by a quite different mix of causes in which tobacco plays no part. The causal package is sufficient but can also be unnecessary: i.e. there are other "paths" to impact, which may or may not include the intervention. The intervention is a *contributory cause* of the impact if:

- The causal package with the intervention was sufficient to bring about the impact, and
- The intervention was a necessary part of that causal package.

4.24 A common way of representing 'contribution' is via what are called causal 'pies' or 'cakes' (Rothman and Greenland 2005). The underlying idea is that the different ingredients of the cake together account for its taste, appearance, texture and shelf-life. However there are different ways of making cakes!. Two examples of such cakes follow (Figure 5 and Figure 6). The first takes an example of a civil-society strengthening programme in a fragile state setting, which it is suggested will only have an effect as part of a much larger 'causal package' of measures. The second puts forward two ways in which an intervention to counter domestic violence in India might be an effective 'cause' of the desired impact. It reinforces the idea that a particular intervention may indeed have potency when it combines with others. But it also underlines that even potent causes may not be the only way – there can also be more than one path to the same outcome.

4.25 The analysis of necessity and sufficiency of (combinations of) causal factors follows an approach to causal inference as put forward by J L Mackie has strong echoes in contemporary causal analysis and evaluation. It is described as multiple-conjunctural by Ragin, (1987) and as 'configurational' causation by Pawson (2007). 'Multiple' refers to the number of paths, each sufficient but possibly unnecessary, that can bring about an impact; while 'conjunctural' is related to the connection between the outcome and a package of multiple causes which, taken in isolation, are unlikely to be responsible for the outcome, or even for "bits" of it (often identified as the net effect).

[18] In the literature this is called an INUS cause: an Insufficient but Necessary part of a Condition that is itself Unnecessary but Sufficient for the occurrence of the effect. These ideas were developed by the philosopher JL Mackie (1974).

Figure 5

Supporting civil-society in a 'post-conflict' setting:

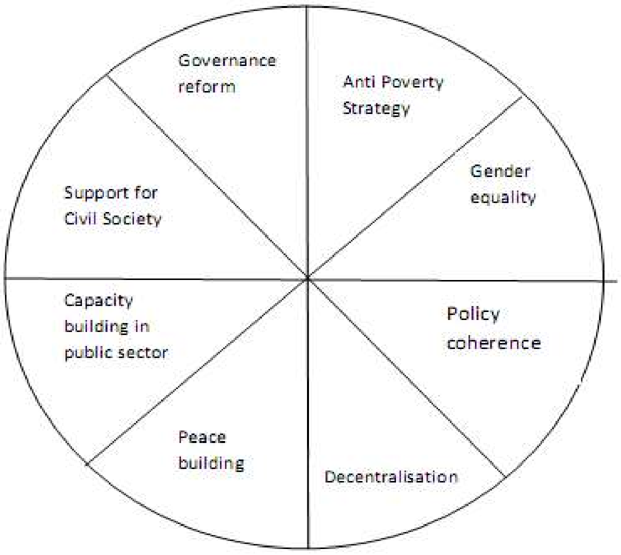
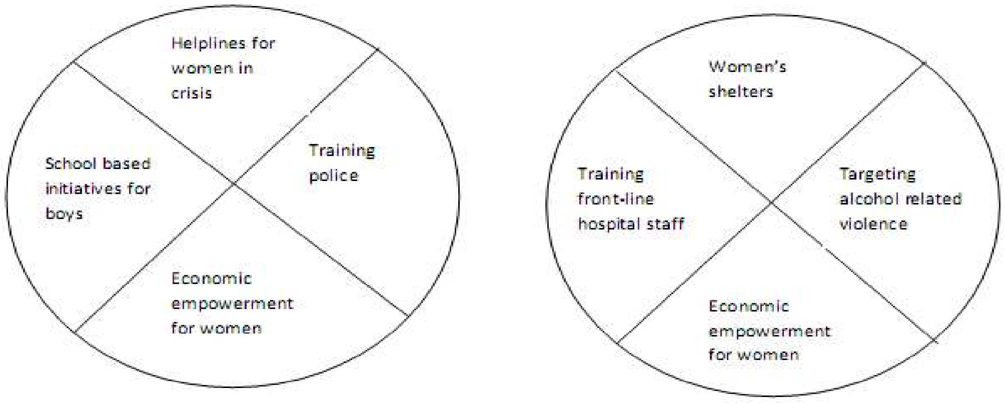


Figure 6

Countering domestic violence: Two 'causal packages':


Helplines for women in crisis
Training police
Economic empowerment for women
School based initiatives for boys
Women's shelters
Targeting alcohol related violence
Economic empowerment for women
Training front-line hospital staff

4.26 The text box below illustrates what a 'contribution' based answer to the 'make a difference' type of IE would look like.

Making 'Contribution' Claims about an Advocacy Campaign

This NGO funds advocacy campaigns that aim to impact on budget expenditure. These campaigns do not work in isolation - a number of other contextual factors may play a role in bringing about an observed budget impact. Key IE questions are:

- Has the 'package' of the campaign plus these other contextual factors resulted in the observed budget impact?
- Was this 'package' sufficient to bring about the budget impact?
- Was the combination of the other contextual factors without the campaign enough to bring about the budget impact?

If confirmed, the 'contribution claim' would be that the advocacy campaign 'worked' in the sense that it made a difference and contributed to bringing about the observed impact.

The Theory of Change for a campaign traces the logical sequence in the causal chain between the campaign's activities and the observed budget impacts, identifying the other contextual factors that are needed for the links in this chain to work. The causal claim will then be two stages:

- First the links in the causal chain must be shown to have happened and explained as to why they happened. This includes identification and discussion of contextual factors and other contributing causes that brought about each point in the sequence.
- Second, plausible rival explanations of why each link in the causal chain happened are identified and discussed in terms of their relative significance in bringing about impact.

Not all campaigns will work. Breakdowns in the expected sequence of impacts are also useful to help identify what could be done differently in the future.

4.27 Contributory causality is relevant when it is likely when there is more than one possible cause, i.e. the intervention is just one part of a causal package. If there are several different but similar programmes then comparative 'Case' designs (see Chapter 3) may also be relevant. If there is only one programme then unpacking the case through within-case analysis by constructing a 'Theory of Change' and using 'process tracing' (as in the above example) and finally eliminating alternative explanations may be most appropriate[19].

4.28 It is not possible to make the assumption that even when all causal factors combine together in favourable circumstances an effect will inevitably occur. For example, can a programme be said to have failed if other negative factors grew stronger at the same time as it was being implemented?. It may have despite apparent failure still raised the probability of success. In the previous example of smoking and lung-cancer individual

[19] Eliminating alternative explanations is a standard way to leverage causal inference in science. In evaluation this has been formalised by Michael Scriven (1976) in his General Eliminative Method (GEM) approach.

predispositions to cancer will vary making the outcomes of even an otherwise sufficient causal package less certain. In an indeterministic world even a causal package that can bring about an effect (in the right circumstances) will not always do so. A cause may only be 'chance-raising' and we may only be dealing with 'conditional probability' (Suppes 1970; Dowe and Noordhof 2004). Mahoney (2008) suggests that comparative researchers split into case-oriented researchers who are interested in a cause and population-oriented researchers who tend to adopt a probabilistic view. He advocates a 'unified theory of causality' for comparative researchers that combines both orientations.

4.29 For policy-makers who have high expectations of their programmes the probable and context-specific nature of impacts presents difficulties. IEs that offer 'only' probable and 'only' plausible conclusions are not always welcome. Backing such evaluations up with results monitoring may be especially important in these circumstances.

4.4 Impact Evaluation Question Three: How has the intervention made a difference?

4.30 The 'making a difference' causal question does not offer an explanation for *how* the intervention caused an impact or made a difference. In many situations when assessing a causal package we also want to understand and be able to explain.

4.31 There are various forms of causal question that get at explaining *how* an intervention has led to or contributed to an impact. For example:

- How and why have the observed impacts come about?
- What causal factors or mechanisms in what combination have resulted in the observed impacts?
- Has the intervention resulted in any unintended impacts, and if so, how?
- For whom has the intervention made a difference?

4.32 There is always a balance to be struck between explaining and simply demonstrating that there has been a causal relationship between a programme and certain effects. As noted above if the purpose of an evaluation is accountability, then demonstrating causality will be enough. This may take the form of a single ToC analysis or an experiment if the preconditions for these designs are in place. If the purpose is lesson-learning for whatever reason – to better plan future interventions or to scale-up and replicate - then the 'black-box' of the intervention needs to be opened up. In that case if there are already well-developed theories some kind of 'theory-testing' may be appropriate, e.g. what was described as 'congruence' analysis in Chapter 3. Where no well-developed theory exists and where there is little prior knowledge (as is often the case in interventions in newer areas such as governance and democracy strengthening) then a 'Causal mechanism' design such as 'realist' evaluation that allowed for theory building as well as theory testing would be appropriate.

4.33 Thinking about the purpose of an IE underlines a further rationale of hybrid approaches (see discussion on combined and 'mixed' methods in Chapter 3). There are many cases where there are demands to use IE for both accountability and explanation. There may also be expectations that IE will contribute to ongoing improvement during the short to medium part of the intervention cycle. In these circumstances there is a need to think not only about combining *methods* but also combining *designs*. An experiment; a participatory design to ensure relevance and targeting; and comparative studies of 'cases' may all be combined in response to the demands of different evaluation purposes.

4.34 Such hybrids may even go beyond IE, i.e. an evaluation that sets out to establish a causal connection between an intervention and an intended effect. For example, in order to improve the results orientation of an intervention, introducing a monitoring system may be the best way to provide timely results information; or in the early stages of a new intervention where little is known, 'real-time' evaluation that supported programme improvement may also be needed. IE is not completely separable from other forms of evaluation.

4.35 A particular strength of any evaluation design that explores causal processes in depth is that it should also be able to contribute to the 'impacts for whom' question. Realist designs try to provide an answer by asking the question "what works for whom, why and under what circumstances" (Pawson and Tilley, 1997). Many specific methods and designs can address this question, in particular 'distributional' analyses that use quantitative methods to identify beneficiaries[20]. However within the scope of theory oriented and qualitative designs, we would argue that 'realist' designs probably in combination with participatory designs can also contribute to understanding the 'impact for whom' question. These designs would however need to be combined with others to answer the more demanding 'how much impact for whom' question.

4.36 In general it is not argued that there is a one-to-one match between these kinds of IE questions and any particular design. However it is possible to indicate the conditions that would make the kinds of designs discussed above most appropriate. For example:

- If the content of a programme is difficult to represent quantitatively there are grounds to use qualitative methods.
- In depth theory-building approaches are suited to circumstances where little prior theory exists and building up new knowledge is a priority.
- Situations that can be judged a priori as 'complex' and 'multi-dimensional' and where there are many possible 'confounding' factors are suited to mechanism based designs.

4.5 Impact Evaluation Question Four: Will the intervention work elsewhere?

4.37 The fourth type of impact evaluation question aims at understanding the conditions under which the findings are generalizable. Variants of questions include:

- Can this 'pilot' be transferred elsewhere and scaled up?
- Is the intervention sustainable?
- What generalisable lessons have we learned about impact?

Or, if the hoped for impacts are not being realized, we want to know:

- Why have the impacts not been realized?
- What contribution has the intervention made?

[20] See for example benefits incidence and expenditures tracking in EU evaluations of General Budget Support.

- Were the impacts not realized because of programme failure or implementation failure?

4.38 These causal questions all aim at learning more about how and why development interventions make a difference. This 'what have we learned?' perspective can also be at the heart of accountability for development assistance – especially in the post-Paris paradigm of 'mutual accountability'. To that extent accountability and learning need not be seen as alternative evaluation purposes.

4.39 This question highlights a contradiction in many discussions about IE. On the one hand the expectation is for aid to be delivered through partnerships and in a decentralised way with partner countries in the 'lead', even though commitment to evaluation among beneficiary governments is limited, especially if it is seen as part of donor oversight and control. On the other hand donors want to ensure effectiveness and Value-for-Money – and to learn from specific decentralised experiences to inform their broader policies.

4.40 Resolving this contradiction requires participatory designs that engage the recipients of aid in an evaluation as part of mutual accountability and to support learning. Participatory designs will be especially useful where there are local instances of an intervention that is being implemented more widely and there is a wish to learn across cases.

4.41 At a macro level it is possible to conceive of 'joint evaluations' between donors and partner countries as a form of participatory design. Participatory approaches that involve stakeholders in both programme implementation and in learning through evaluation would match such programmes. These would span both 'normative' and 'agency' designs as described in Chapter 3. For example, normatively, they would operate within the assumptions of the OECD-DAC about the aid relationship as a partnership. At the same time beneficiary countries would be seen to have agency, to be actively involved in shaping and learning from the evaluation process – thereby improving the prospects for successful programme design and implementation.

4.42 There is also scope for hybrid designs that include experimental elements to answer the 'will this work elsewhere?' question. For example, if the intention is to expand or generalise from a particular set of interventions, there are arguments for the first generation interventions in a particular policy area to be planned as an experiment that intentionally tests out different alternatives in clusters of different settings. This would be on the lines of 'natural experiments' taking advantage of a diversified set of *programme* designs in order to learn through their implementation. Such a strategy would be well-suited to customised, decentralised initiatives that are attempting to achieve similar goals - the 'experiment' is about means rather than ends (this is an important element of the World Bank's DIME programme). Finally once a set of interventions have been implemented there will be opportunities for 'Synthesis' studies that try to accumulate lessons from across these interventions.

4.43 Underlying assumptions and requirements that follow from this understanding of the question 'Can this be expected to work elsewhere?' include:

- A joint commitment to learning and mutual accountability.
- Identification of different implementation strategies, that could be expressed as 'causal packages', identified in dialogue with partners.
- Generic understanding of contexts e.g. typologies of context so that extension/scaling-up can be systematically planned.
- Building in innovation and diffusion processes so that successful 'configurations' in different contexts could be extended.

Conclusions

4.44 This Chapter has taken a different 'cut' through the design options identified in Chapter 3. By starting with different types of evaluation questions that can be asked in IE it has set out to illustrate how different designs can be used and as important how different designs can be combined. It has concentrated on the three design approaches (Theory based, Studies of 'Cases' and Participatory) but has also identified complementarities with other designs and methods.

4.45 Table 4.2 summarises the logic of this chapter. It takes each of the four key evaluation question identified, indicates related questions, describes the underlying assumptions, some requirement for these designs to be a practical proposition; and provides examples of suitable designs.

Table 4.2: Summarising the Design Implications of Different Impact Evaluation Questions

Key Evaluation Question	*Related Evaluation Questions*	*Underlying Assumptions*	*Requirements*	*Suitable Designs*
To what extent can a specific (net) impact be attributed to the intervention?	What is the net effect of the intervention? How much of the impact can be attributed to the intervention? What would have happened without the intervention?	Expected outcomes and the intervention itself clearly understood and specifiable. Likelihood of primary cause and primary effect. Interest in particular intervention rather than generalisation.	Can manipulate interventions. Sufficient numbers (beneficiaries, households etc.) for statistical analysis.	Experiments. Statistical studies. Hybrids with 'Case' based and Participatory designs.
Has the intervention made a difference?	What causes are necessary or sufficient for the effect? Was the intervention needed to produce the effect? Would these impacts have happened anyhow?	There are several relevant causes that need to be disentangled. Interventions are just one part of a causal package.	Comparable cases where a common set of causes are present and evidence exists as to their potency.	Experiments. Theory based evaluation, e.g. contribution analysis. Case-based designs, e.g. QCA.
How has the intervention made a difference?	How and why have the impacts come about? What causal factors have resulted in the observed impacts? Has the intervention resulted in any unintended impacts? For whom has the intervention made a difference?	Interventions interact with other causal factors. It is possible to clearly represent the causal process through which the intervention made a difference – may require 'theory development'.	Understanding how supporting & contextual factors that connect intervention with effects. Theory that allows for the identification of supporting factors - proximate, contextual and historical.	Theory based evaluation especially 'realist' variants. Participatory approaches.
Can this be expected to work elsewhere?	Can this 'pilot' be transferred elsewhere and scaled up? Is the intervention sustainable? What generalisable lessons have we learned about impact?	What has worked in one place can work somewhere else. Stakeholders will cooperate in joint donor/ beneficiary evaluations.	Generic understanding of contexts e.g. typologies of context. Clusters of causal packages. Innovation diffusion mechanisms.	Participatory approaches. Natural experiments. Synthesis studies.

Chapter 5: Programme Attributes and Designs

Introduction

5.1 The kinds of programmes that we were given to consider by DFID were variously described as complex, challenging or difficult to evaluate. They included programmes concerned with empowerment, accountability, governance, state-building, climate change and humanitarian relief. This is consistent with the ToR which stated: 'The primary purpose of this research is to establish and promote a credible and robust expanded set of designs and methods that are suitable for assessing the impact of *complex* development programmes' (emphasis added). Some were 'complex' because of their own characteristics whilst others were embedded in complex and sometimes dangerous country environments. It follows that programme attributes are not only a consequence of their own goals and modalities; but have to be seen as embedded objects in which the constraints imposed by institutional, social, cultural and economic environment are an integral part of the 'object' being evaluated.

5.2 'Attributes' along with evaluation questions also determine critical evaluation design choices in IE. We therefore:

- Scanned relevant literatures in order to identify attributes that have design implications.
- Reviewed actual programmes sourced from DFID in terms of these attributes.

5.3 The chapter begins with 'complexity' literatures but broadens this to include related ways of classifying programmes. It then cross-checks what research and evaluation literatures suggest with a review of specific DFID programmes.

Main messages

- Many contemporary programmes that concern governance, accountability, empowerment, state-building and climate change mitigation are difficult to evaluate.
- Difficulties can follow from the complex nature of what they try to do or from difficult setting which may hinder implementation.
- Deciding whether the unit of analysis for IE should be an individual programme or a set of linked programmes depends on the degree of vertical linkage with larger 'hierarchical' programmes or overlap with other related programmes.
- Selecting for IE a single sub-programme which is highly interdependent within a larger programme is a risky basis for judging overall programme success or failure.
- The duration of programmes and when their effects should be expected affects when a programme should be evaluated and how pre-planned or responsive an evaluation should be.
- There is a scope to deploy newer methods such as QCA, Realist Synthesis, system dynamics & Agent Based Modelling which are not widely used in international development IE at present.

5.1 Complex systems in evaluation

5.4 ‘Complexity’ and related concepts from ‘systems theory’ have become popular words in the evaluators’ lexicon. Two books, (Forss et al 2011; Williams & Imam 2007) have recently been published dedicated to ‘complexity’ and ‘systems’ in evaluation and related articles are regularly submitted to evaluation journals. This preoccupation with complexity derives from recognition that programmes nowadays often address chronic problems, use new policy instruments, take place in multi-institutional settings and are located in unstable and sometimes dangerous settings.

Table 5.1: Complicated and Complex Aspects of Interventions (Rogers, 2008)

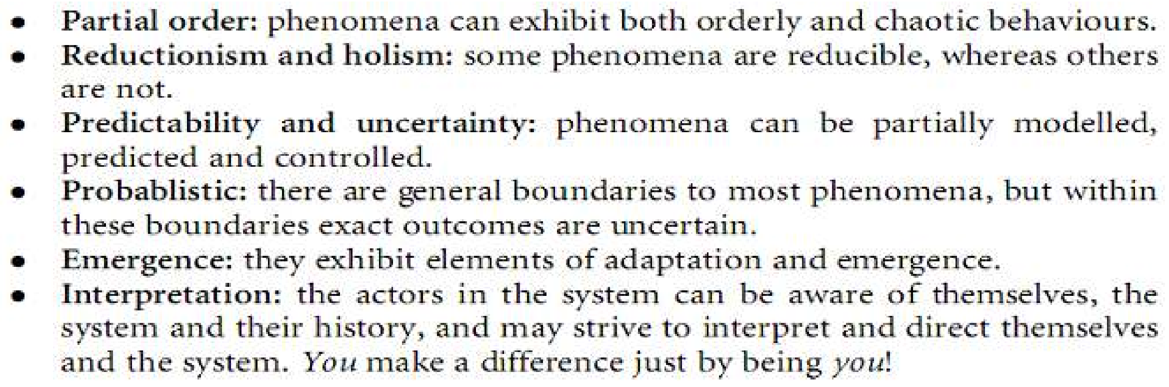
Aspect	**Simple version**	**Not simple version**	**Challenges for evaluation**	**Suggested label**
1. Governance and implementation	Single organisation	Multiple agencies, often interdisciplinary and cross-jurisdictional	More work required to negotiate agreement about evaluation parameters and to achieve effective data collection and analysis.	Complicated
2. Simultaneous causal strands	Single causal strand	Multiple simultaneous causal strands	Effective programs may need to optimise several causal paths, not just one; evaluation should both document and support this.	
3. Alternative causal strands	Universal mechanism	Different causal mechanisms operating in different contexts	Replication of an effective program may depend on understanding the context that supports it. The counter-factual argument may be inappropriate when there are alternative ways to achieve the outcome.	
4. Non-linearity and disproportionate outcomes	Linear causality, proportional impact	Recursive, with feedback loops	A small initial effect may lead to a large ultimate effect through a reinforcing loop or critical tipping point.	Complex
5. Emergent outcomes	Pre-identified outcomes	Emergent outcomes	Specific measures may not be able to be developed in advance, making pre- and post-comparisons difficult.	

5.5 One typology that has gained prominence in evaluation is Glouberman and Zimmerman’s distinction – building on Ralph Stacey’s work (Stacey 1992) – between the simple, complicated and complex. This has been taken up by Patricia Rogers in 2008 (see Table 5.1) and by Funnell and Rogers (2011).

5.6 Simple programmes are 'standardised interventions' that consistently follow a causal chain from inputs through to outputs, outcomes and impact. Typically they would involve a single causal strand; whereas complicated programmes involve multiple and often alternative causal strands. In complex programmes on the other hand causality is less ordered: causal links can be reversed, virtuous and vicious circles occur, new properties - new outcomes and means emerge whilst a programme is underway. The aspects of programmes that Rogers (2008) emphasises are summarised in the above table: 'Complex and Complicated Aspects of Interventions'.

5.7 Ideas about complexity preoccupy practitioners in many fields, for example in science policy, organisation theory and healthcare management. The following table from an analysis of diabetes care from a complexity perspective (Cooper and Geyer 2007) echoes many of the debates found in evaluation and indeed in international development policy.

Table 5.2: 'Golden Rules' in a Complexity Paradigm (source Cooper and Geyer)



- **Partial order:** phenomena can exhibit both orderly and chaotic behaviours.
- **Reductionism and holism:** some phenomena are reducible, whereas others are not.
- **Predictability and uncertainty:** phenomena can be partially modelled, predicted and controlled.
- **Probablistic:** there are general boundaries to most phenomena, but within these boundaries exact outcomes are uncertain.
- **Emergence:** they exhibit elements of adaptation and emergence.
- **Interpretation:** the actors in the system can be aware of themselves, the system and their history, and may strive to interpret and direct themselves and the system. *You* make a difference just by being *you*!

5.8 Although Rogers distinguishes between the big categories of the simple, complicated and complex she is mainly interested in 'aspects of them rather than the interventions themselves'. This is because, 'A complex intervention may have some simple aspects' as indeed relatively simple programmes may have isolated complex features – an insight akin to that of Cooper and Geyer's notion of 'partial order'. We have also concluded that it is better to look for design implications within interventions rather than to try and distinguish interventions themselves.

5.9 The next section focuses on specific attributes of programmes that are suggested by systems thinking but are also more widely debated within policy research.

5.2 Specific programme attributes

5.2.1 Linear causality and 'trajectories'

5.10 Few social systems display the mathematical properties of linear systems where outputs are proportionate to inputs; and where effects are understandable in terms of a preceding cause. Systems frameworks emphasise the iterative and disproportionate qualities of social systems. In development evaluation, challenges to *'linear causality'* have been elaborated by Woolcock (2009) who argues that the trajectory of change over time may vary considerably. For example the empowerment of marginalised groups may lead to negative results before they become positive i.e. there will be a J-shaped curve. Or there may be a step-function, where after a long period of no change there may be a dramatic improvement e.g. when a new group is elected to power. Elsewhere, anti-corruption efforts making use of public expenditure tracking surveys may have a high initial impact that then fade over time (as those in power find alternative ways to misuse funds).

5.11 Woolcock argues that few programme designs are explicit on the shape of the expected impact trajectory and that this omission can easily undermine the validity of evaluation findings based on impact data which has been collected at a particular point in time. Clarifying the expected trajectory of change is most relevant to ToCs. However, it is also relevant to other evaluation designs that are informed by programme theory e.g. statistical association and configurational analyses, as well as experimental designs.

5.2.2 Timing and long term effects

5.12 The OECD-DAC definition of 'impact' emphasises 'long term' effects which has implications for when an evaluative judgement about 'impact' can be made. Because of their goals some programmes are necessarily long-term and others short term. Reforming a national education system; supporting the development of a democratic culture; changing the traditional power relations between women and men are likely to take longer than introducing an educational subsidy or making available mosquito nets.

5.13 Programmes also transform or evolve over time. Partly this is because the policy focus changes as do a host of features of the country's political economy, world trade etc. (This was evident in past DFID Country Programme Evaluations; and in the evaluation we reviewed of a 25 year engagement by SIDA in Tanzanian forestry – see Katila et al 2003). It is also because development processes are inherently uncertain and change over time. This is close to systems notions of 'emergence'.

5.14 It is noteworthy that despite the long term assumption built into many definitions of IE, they rarely take a long-term perspective. Many experimental designs are ex ante (preceding scale-up or roll-out) and mixed method designs often take place at an early stage in the programme implementation cycle. Assessing the duration of programmes and how they evolve should be part of planning when to evaluate for impacts. It may for example be sensible to time-slice an evaluation, undertaking some activities in the short run and others over an extended period. It may also be useful to set up monitoring systems and panel studies in the short term that can track change over the longer term. Such approaches are likely to make more of a contribution to learning how to achieve improved impacts than to judge whether impacts have taken place.

5.2.3 Standardisation and customisation

5.15 Standardisation of delivery is an important attribute of many service oriented development programmes and is an attribute of interventions that make them amenable to experiment based designs. Funnel and Rogers discuss 'standardised interventions' which work 'Pretty much the same everywhere'. The "treatment" being tested has to be delivered as described, in multiple settings. Sometimes standardisation follows from a desire to use particular methods. This was the case with conditional cash transfers provided to poor families. These had to be provided according to the agreed rules (so that programme fidelity was ensured) in order that they could be evaluated using experimental methods. When there are many beneficiaries of an intervention and that intervention is standardised, an evaluation will be better able to use statistical methods than when the intervention is targeted at relatively few (people, firms or institutions). This is the approach to multi-site evaluations favoured by Herrel and Straw (2002) which they describe as being 'in the clinical trial tradition'.

5.16 Standardisation contrasts with diversity and customisation which have become common features of post Paris Declaration programmes 'owned' by partner countries. Customisation is often associated with devolved or indirectly delivered programmes. Doing away with 'parallel implementation structures' was an important justification for the Paris Declaration and indirect delivery is one of the aspects of programmes that also has implications for IE (see below on specific programmes). Diversified and locally customised programmes lend themselves to comparative, participatory and synthesis types of designs.

5.2.4 Hierarchies and 'nested' systems

5.17 Mitta Marra (in Forss et al 2011) drawing on complexity science, also discusses how 'mechanism based explanations' allow interventions to 'be decomposed into its constituent parts'. 'Complex settings can be seen as nested systems, where mechanisms can be analyzed at different levels' (op cit page 323). Marra also contrasts the logic of subsystems understood in terms of mechanisms with the logic of 'causal relationships'. From a complexity science viewpoint causes and effects are not independent; causes and effects influence each other. Marra argues for the need for *both* analyses of causal chains; and for a more dispersed mechanism-based understanding of nested systems (some of the large scale complex initiatives that we reviewed led to a similar conclusion from a 'bottom-up' perspective, see below).

5.18 The concept of 'nested' hierarchical systems is also discussed by Lieberman (2005) as a specific research strategy building on ideas of Frissell et al (1986) in relations to rivers and stream habitats. This approach has been used in various settings - by Kirsch et al (2008) in healthcare; and by Hummelbrunner (2010) in international development. The notion that complex systems can be broken down into a series of smaller and smaller subsystems – like Russian dolls – is an attractive one. However systems theorists (e.g. Cilliers 2001) are also concerned with 'intersections' – that even deeply embedded subsystems are 'interpenetrated' by the wider macro-system and cannot be viewed in isolation. In IE for example, is it legitimate to evaluate specific interventions within a governance or social policy programme in isolation from the wider programme?. Is it legitimate to infer that because a particular 'nested' sub-programme on election reform or women's empowerment 'works' that the overall programme is a success?. This has echoes of the 'micro-macro paradox' (Mosley 1987) – the observation that whilst most development projects are successful, the programmes and polices of which these projects are a part often fail. The analysis of specific cases (see below), suggests that whilst breaking down complex systems into smaller subsystems for evaluation purposes is sometimes justified, it is often not.

5.2.5 Comprehensive and targeted interventions

5.19 In both national and international policy design there is a tension between comprehensive and targeted policies. For example in urban regeneration in the UK the comprehensive strand is represented by 'Area Based Initiatives' where multiple dimensions of poverty and disadvantage such as poor education, limited work opportunities, bad standards of healthcare, inadequate housing and high levels of crime are addressed simultaneously[21]. On the other hand single-dimension, dedicated policies targeted at child protection, crime reduction etc. are also common. Whilst there is some evidence that simultaneously addressing multiple-deprivation is more effective – there are also questions about the targeting, cost and the administrative efficiency of comprehensive programmes. In international development similarly there are comprehensive poverty reduction programmes such as the CDF (Comprehensive Development Framework) alongside narrowly targeted health and education initiatives. In trade policy there has been a shift from the comprehensive - as in structural adjustment and tariff reforms - towards interventions targeted more narrowly at export promotion (Cadot et al 2011). Governance, democracy enhancing and empowerment programmes are often comprehensive of their nature – although they are often evaluated through their component parts.

5.20 The extent to which a programme is comprehensive has implications for IE. In more comprehensive programmes multiple measures and the synergies between them have to be evaluated. It can be argued that from policy coherence perspective, when programme are comprehensive single measures are the wrong unit of analysis. This would certainly be the case if there was a high level of interdependence between measures such as electoral reform, voter education, education of parliamentarians, empowerment of women etc. (see Moehler 2010 for fuller discussion of this issue). Another variant of interdependence is when separate interventions are targeted at the same group in the same area or region. In these circumstances looking at these interventions together is also justified, indeed given the risks of double counting of 'benefits' it would be risky not to.

5.21 'Diffusion of innovation' models (Rogers 2003; Denis et al 2002) would suggest that in some circumstances a single intervention can seed many other related measures. If such diffusion could be demonstrated, it might be justified to focus on a single measure for an IE directed at learning. But it is still necessary to analyse programme attributes to understand how one programme can reinforce or undermine others.

5.2.6 Uncertainty and risk

5.22 Jonathan Morell (2010) and Jay Forrest (2007) - like Cooper and Geyer - are also preoccupied with uncertainty. Morell accepts uncertainty as inevitable. Forrest who works with qualitative systems analysis and 'foresight' studies, suggests that drawing the boundary too narrowly around a 'system' can stand in the way of identifying sources of

[21] Among prominent examples of this in the UK would be the Community Development Programme of the 1960s and the 'New Deal for Communities' of the early 2000s.

feedback and influence that can lead to instability. Hence the need for whole systems analysis rather than a partial analysis of sub-systems when programmes have relatively open boundaries. Morell suggests that in the face of uncertainty, early detection is needed, an argument for monitoring and environmental scanning. High levels of uncertainty lead to a consequent need of managers and policy makers for speedy feedback. Similar points are made by Funnel and Rogers. This again highlights the value of formative and 'real-time' evaluation designs.

5.23 A systems perspective underlines the contextual as well as programmatic aspects of uncertainty and risk. Whilst programmes attempting something new and ambitious will be inherently risky other programmes following well-trodden paths may face risks because of their context. Uncertainty and risk that is contextual can take on a political or security form. For example, there are political risks in pursuing long term programmes with no guarantee of success; and many international development programmes take place in pre or post conflict settings where the personal safety of personnel cannot be guaranteed. This will for example affect the implementation of programmes: inoculation programmes may fail if healthcare workers are unable to work in safe conditions. From an evaluation standpoint unstable contexts can limit data collection (difficulties collecting data was one of the main constraints on the IE design used in the GTZ evaluation of programmes in NE Afghanistan).

5.3 Attributes of more complex programmes and implications for IE

5.24 The table that follows summarises the main attributes of complex programmes and programmes embedded in complex settings suggested by a number of bodies of literature and research. It identifies the possible implications of these attributes for IE design. Attributes are presented in this table along dimensions that run from the more to the less complex. Design implications are then described for the 'more complex' end of the spectrum. However no attempt has been made to calibrate these dimensions or suggest precisely how attributes should be positioned along them.

Table 5.3: IE Design Implications of Complex Programme Attributes

Least ← – – – – **Complex**	– – – – → **Most** **Complex**	**Possible Design Implications for 'more complex' programme and settings**
Bounded	Embedded	Systems analysis – based on an expanded 'unit of analysis' that includes wider system. Mapping of feedback loops and 'external influences' and their interaction. Comparisons of 'contexts' and mechanisms – realist logic.
Centrally specified	'Locally' Customised	Need to assess the relevance of local programme variant Input from programme stakeholders. Data collection more complex should be participative.
Standardised interventions	Diversified interventions	Process track and compare different ToCs. Meta evaluations and comparisons of different programmes – QCA /matrical designs.
Predictable and linear impacts	Difficult to predict and non-linear impacts	Assess the 'trajectory' of intended change. Ensure timing of evaluation matches this trajectory. Monitoring systems that provide rapid feedback. Real-time evaluation.
Single or few causal strands that are independent	Multiple causal strands that are interdependent	Process track different causal strands and how they interact (ToC). Identify alternative potential causal paths to same goal. Identify 'blockages', virtuous and vicious circles.
Mechanisms are universal	Different causal mechanisms operate in different contexts	Study mechanisms in context – realist/mechanism based designs. Develop typologies of contexts. Realist synthesis. Configurational designs/QCA.
Causes and effects are independent of each other	Causes and effects influence each other	Review different context/mechanism configurations. Matrix of effects. Conduct case studies of cause/effect interactions.
Homogeneous systems	Nested diversified systems	Systems mapping to identify 'zones' of greater or less complexity. Differentiate designs accordingly.
Pre-identified effects	Emergent effects	Real-time evaluation to track how goals are re-defined. Staged designs to adapt to evolving ToC. System analysis and monitoring.

5.4 The attributes of programmes reviewed

5.25 This section reviews actual programmes, their attributes and design implications. It takes a bottom-up approach to cut through the kinds of attributes identified in the literature discussed above. Some attributes are seen to follow from programme theory: the intentions, activities and assumed ways of working of a programme. Some are closer to technical constraints. They cover how often an intervention occurs, data availability etc.. Both types of attributes have implications for design. On the other hand the kinds of design implications identified in this chapter extend beyond choosing a design exclusively to link causes and effects; or to explain how and why effects occur. Programme attributes also have implications as to when an impact evaluation should take place, how stakeholders should be involved and what data can be relied on.

5.26 In order to decide on a small number of cases to analyse in detail a larger sample of programmes were suggested by DFID and reviewed before selections were made. The aim of this first review was to ensure that the 'attributes' that were characteristic of these programmes were likely to be representative of a larger population of programmes – bearing in mind that they would all fall into what was described as 'complex' and 'challenging' categories. The preliminary reviews involved several activities:

- The nominated programmes fell under broad categories such as climate change, public sector reform, governance, empowerment and accountability, humanitarian relief and post-conflict or security. In some of these categories several programmes were put forward. For example, under the broad heading of 'Deepening Democracy' four different programmes in Africa were considered before selecting one part of one such programme - the Civil Society Governance Fund in Malawi (CSGF).
- Although we eventually selected one programme in Afghanistan we read documentation for two others before making this selection and also reviewed another evaluation conducted in Afghanistan that we had previously assessed in terms of its methodology. Other evaluations in post conflict or fragile states were also reviewed.
- Several programmes shared key characteristics, even if they were not in the same broad policy heading. An example of this was setting up a 'Fund' that would then select and support a portfolio of projects in line with programme aims. This appears to be a common modality post Paris Declaration.
- We also held discussions with DFID staff to check how far specific projects – such as the Response to the Haiti Earthquake could be seen as typical of other disaster relief programmes.

5.27 On the basis of the above we are confident that the six programmes reviewed in depth were not a-typical and the attributes identified would be likely to be of wider relevance. The six programmes were:

- Strategic Climate Institutions Programme (SCIP) Ethiopia;
- District Delivery Programme Afghanistan;
- 'Violence Against Women' Bihar India;
- Civil Society Governance Fund, Malawi;

- Response to Haiti Earthquake – January 2010; and
- Results Based Aid (RBA) in the Education Sector in Ethiopia.

5.28 All were reviewed by pairs of team members using a common template and were then discussed in virtual team meetings and written exchanges. From these discussions a set of 'attributes' were identified that were likely to have design implications. These attributes were common across most of the cases reviewed and included:

- *Overlap with other interventions with similar aims.* These could be funded by partner governments or by other donors.
- *Multiple and diverse activities and projects* – implementing different projects or using different measures (projects, funds, policy dialogue, technical assistance) as part of the same programme.
- *Customised non standard projects* often in combination with interventions in diverse contexts. Even when they were implementing similar projects these projects were not standardised; and the settings in which they were implemented varied – e.g. in terms of communities, institutions, effected populations.
- *The likely impacts of programmes were long term.* However even though end goals for intended beneficiaries were long term the planned life-cycle of the programme itself was often relatively short. The most that could be expected were intermediate outcomes.
- *Working in areas of limited understanding/experience.* Such programmes could be innovative to the point of being unable to clearly articulate their Theory of Change or their assumptions. Alternatively a ToC was put forward but was not backed by evidence.
- *Working in areas of high risk or uncertainty.* This could include dangerous security situations or post disaster. It could also include ambitious social programmes where how progress would be achieved was unclear and set-backs could be expected.
- *Stated impacts are difficult to measure, possibly intangible* and often composite different goals into one heading. Several of the programmes reviewed set themselves ambitious almost rhetorical goals – 'to reduce violence and fear of violence among women in the State of Bihar'; or to achieve 'increasingly accountable, inclusive and responsive governance in Malawi'.
- *Programmes working 'indirectly' through 'agents' and often at different levels and stages.* They might for example set up a programme that would itself set up a fund or other intermediary structure which could be outsourced to contractors. This intermediary body would then fund projects, which for their part would undertake activities, deliver services etc.

5.29 These attributes were common across most of the selected programmes even though they were attempting to do very different things. The aim was to home in on a common but limited set of 'attributes' so that some general implications could be drawn for the future evaluation of many programmes. The table below summarises these attributes and their incidence across the selected programmes.

Table 5.4: Presence of Attributes in Selected Programmes

	VAW Bihar	SCIP Ethiopia	CSGF Malawi	DDP Afghanistan	Earthquake Response HAITI	Results based Education Ethiopia
Overlap with other interventions with similar aims	Y	Y	Y	Y	Y	Y
Multiple and diverse activities and projects	Y	Y	Y	Y	Y	
Customised non standard activities/ interventions in diverse contexts	Y	Y	Y	Y	Y	Y
Programmes working 'indirectly' through 'agents' and often at different levels and stages	Y	Y	Y	Y	Y	Y
The likely impacts of programmes were long term	Y	Y	Y	Y		Y
Working in areas of limited understanding/experi ence		Y	Y	Y		Y
Working in areas of risk or uncertainty	Y	Y		Y	Y	Y
Intended impacts are difficult to measure and in parts intangible	Y	Y	Y	Y		Y

5.5 IE Design implications

5.30 Because all programmes have more than one attribute IE design implications cannot simply be read across – which is another reason that 'combined' or 'mixed' designs are necessary. In practice any one of the selected programmes should have an evaluation design that takes cognisance of all its design-sensitive attributes. Table 5.5 provides an overview of the design implications of the common evaluation challenges posed by this set of 'complex' programmes that were reviewed.

Table 5.5: Programme Attributes and Design Implications

Programme attributes	**Evaluation challenge**	**Design Implications**
Overlap with other interventions with similar aims	Disentangling effects of this programme from others.	Consider multi-programme evaluations where purposes are related. Possibility of 'Contribution' analysis. Consider joint evaluations with other donors or with government partners.
Multiple and diverse activities and projects	How to assess programme impacts and distinguish them from component impacts.	Ensure that programme goals inform the criteria for 'project' evaluation. Provide a common evaluation framework for projects to guide their evaluation. Conduct case studies of linked policy decisions/implications/ 'spill-overs'. Provided they all aim for common goals and are similarly implemented use configurational/QCA type methods.
Customised non standard projects often implemented in diverse contexts	How to add up or synthesise 'apples and pears'.	Identify alternative Theories of Change. Focus on mechanisms rather than effects (e.g. realist synthesis). Develop typologies of 'contexts' or settings. Involve stakeholders and beneficiaries to obtain 'local knowledge' – participatory designs.
Programmes working 'indirectly' through 'agents' and often at different levels and stages	What should be the evaluation strategy at different stages? How to break the evaluation down into manageable parts.	Distinguish between different programme stages – e.g. setup, implementation and delivery. Devise 'nested' evaluation strategy. Consider need for technical assistance alongside intervention to support 'agents' & intermediaries.
The likely impacts of programmes were long term	When to conduct an evaluation and how to judge success.	Construct a time extended Theory of Change. Use monitoring systems and indicators to estimate 'distance travelled'. Assess the implementation trajectory in order to judge when evaluations are likely to register 'impacts'. Track critical events that could re-direct programme.
Working in areas of limited understanding/experience	Constructing a valid Theory of Change. Finding evidence to support a ToC that was put forward.	Regard Theory of Change as an 'evolving' object –options to revise it. Explore alternative causal pathways. Involve stakeholders in articulating their own Theories of Change.
Working in areas of risk or uncertainty	Likelihood of set-backs and uneven progress (step level	Conduct an evaluability assessment. Design an iterative or staged evaluation strategy that can respond to 'emergence'.

	change). No proportional relation between cause and effect.	Identify ‘trigger points’ and bottlenecks’. Real-time evaluation – with formative elements – rapid feedback capacities.
Intended impacts are difficult to measure, possibly intangible	How to know what impact is – it may mean different things to different stakeholders and beneficiaries.	Jointly develop a Theory of Change with inputs from stakeholders and beneficiaries. Agree on intermediate and long-term impacts along an implementation trajectory. Participatory inputs to better define impacts

5.31 Although many of the attributes of these complex programmes are suited to the kinds of designs and methods that were identified in Chapters 2 & 3, an analysis of ‘attributes’ also highlights the limitations of the methods being used. Complex programmes would be suited to Realist approaches and case-based analyses such as QCA; and in some circumstances simulation and modelling techniques even though there is little experience in applying these approaches to IE in general and to IE in international development in particular.

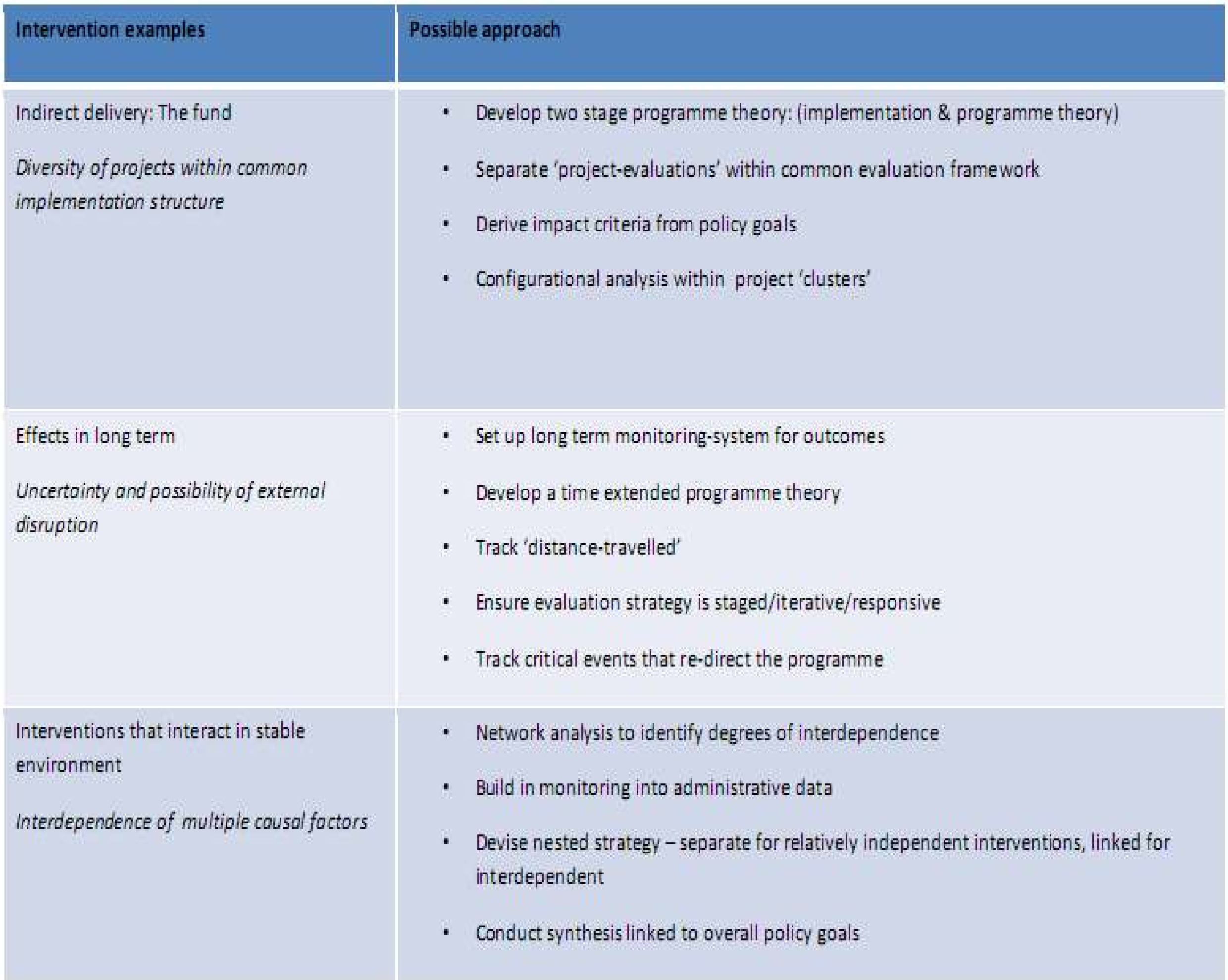
5.32 Individual programmes can be analysed in terms of these kinds of attributes as in table 5.6. This puts together a ‘typical’ programme composited from several of the programmes we reviewed and assesses them in terms of the frameworks developed in this chapter.

5.33 Two key IE design questions have come up at several points in this chapter:

- How to determine the appropriate ‘unit of analysis’ for evaluating complex programmes made up of multiple and interdependent sub-programmes or interventions.
- How to sequence the evaluation of long-term programmes (i.e. those that extend over a long time) or emerging programmes (i.e. those that are likely to evolve and change their character).

5.34 These are likely to be important questions to answer in many complex programmes which are often made up of related interventions; and which are also non-linear and prone to unplanned changes in the longer-term.

Table 5.6: Example of a 'composite' programme



Intervention examples	Possible approach
Indirect delivery: The fund *Diversity of projects within common implementation structure*	• Develop two stage programme theory: (implementation & programme theory) • Separate 'project-evaluations' within common evaluation framework • Derive impact criteria from policy goals • Configurational analysis within project 'clusters'
Effects in long term *Uncertainty and possibility of external disruption*	• Set up long term monitoring-system for outcomes • Develop a time extended programme theory • Track 'distance-travelled' • Ensure evaluation strategy is staged/iterative/responsive • Track critical events that re-direct the programme
Interventions that interact in stable environment *Interdependence of multiple causal factors*	• Network analysis to identify degrees of interdependence • Build in monitoring into administrative data • Devise nested strategy – separate for relatively independent interventions, linked for interdependent • Conduct synthesis linked to overall policy goals

5.35 From reviews of actual programmes it seems to be possible to structure design decisions related to interdependence by identifying potentially related programmes whether in a single large scale programmes or in other 'nearby' or overlapping programmes or in separate programmes that target the same population. This requires:

- Assessing the extent of 'embeddedness' of interdependent programmes and sub-programmes and distinguishing between those that are relatively free-standing and those that are interdependent or embedded or otherwise linked (e.g. overlapping target populations).
- Developing different evaluation strategies for those programmes or sub programmes that are more or less interdependent.

5.36 For example in the Bihar 'Violence Against Women' Programme there are some sub-programmes that are closely linked together. The effectiveness of referral of women victims to social work and health agencies is likely to be highly dependent on the training and professional development of front-line workers in these agencies. They should therefore be evaluated as a single unit of analysis. On the other hand school-based interventions to encourage boys to develop different attitudes and behaviours towards women and girls are more free-standing and can be evaluated on their own.

However it will not be possible to assess the success of the overall programme from evaluations of separate parts.

5.37 For long-term programmes a similar structured approach is needed. As suggested in Chapter 4, a Theory of Change offers one way to map likely future change. However emergent programmes are unlikely to follow a linear or predictable path especially if they are long-term. This suggests the need to:

- Assess the likely elapse time for the programme to have an impact. For those with an extended elapse time set up monitoring or longitudinal systems to track change over time.
- Use a Theory of Change to distinguish between programmes that are likely to be more or less predictable in their effects.
- For less predictable programmes institute a 'staged' or 'iterative' evaluation strategy. This should aim to capture short, medium and longer term effects with an expectation to review designs and methods at stages in the programme 'cycle'.

5.38 For example, in a 'Climate Change Mitigation' programme 'proofing' village agriculture against drought may be relatively short-run; building capacity to maintain these capacities may take longer; and related measures to make infrastructure (healthcare, utilities, transport) more resilient to the consequences of climate change may take longer still. At the same time various consequences of climate change are difficult to predict. Floods may become more common than floods in a particular region; and other changes in public policy may have big effects on the ways capacity can be strengthened. An IE strategy will therefore need to be responsive to real-time changes that are not anticipated when a programme is launched.

Conclusions

5.39 This chapter has argued that there are specific evaluation challenges thrown up by complex, multi-dimensional and 'difficult-to-evaluate' programmes. They include duration and time scale; non-linearity and unpredictability; local customisation of programmes; multiple interventions that influence each other – whether as sub-programmes within the same programme or in separate but overlapping programmes. Many of these attributes are highlighted by systems and complexity theory and similar attributes can be identified in specific programmes that have been reviewed as part of this study.

5.40 Programme attributes add a further dimension to the design of IE which can be addressed by the designs and methods already discussed in this report such as theory-based, comparative case-based and participatory designs, even though often these designs and associated methods need further development (for example, as noted in Chapter 4, case-based approaches in development evaluation rarely reflect newer methods and techniques now becoming mainstream in other policy domains). There are also other gaps in the evaluation toolkit that are thrown into sharp relief by considering the attributes of diversified, multi-dimensional and complex programmes. For example, evaluation still lacks some of the simulation and modelling approaches for complex systems that have been applied in other domains. System dynamics and Agent-

Based Modelling have been identified as a particular approach that could make a contribution.

5.41 A number of the key design decisions have been highlighted. These include how to decide on the unit of analysis when there are many related interventions taking place simultaneously and likely to influence each other; and when to conduct an evaluation of a programme that is long-term, 'emergent' and unpredictable. Suggested responses to these challenges such as identifying vertical and horizontal linkages between programmes; and 'time slicing' an evaluation so as to be able to respond to short and long-term effects, both new tasks for 'evaluability assessment'. A preliminary analysis of programmes is needed so as to assess programme attributes and their implications for what elements of a programme should be evaluated (the unit of analysis problem) and how an evaluation should be organised over time.

Chapter 6: Quality Assurance

Introduction

6.1 Many frameworks have been developed to judge the quality of research and evaluation. However there are three features of this study that are distinctive and have implications for quality assurance (QA) approaches:

- It is concerned with qualitative as well as quantitative evaluation approaches whilst the most prevalent QA approaches have grown out of quantitative research.
- It is interested specifically in IE which although sharing characteristics with other forms of evaluation also has distinctive methodological features.
- It is located in international development assistance which has its own normative assumptions and institutional arrangements.

6.2 The ToR for this study indicates why 'quality' and 'quality assurance' is important for impact evaluation designs and methods and emphasises in particular:

- That IE needs to 'withstand criticism or scepticism over the robustness of findings'.
- To this end, it must be possible to 'tell the difference between appropriate, high quality use of the approach and inappropriate/poor quality use'.

6.3 This chapter proposes an approach to QA that spans the evaluation 'life-cycle', including:

- The choice of appropriate designs and methods.
- The proper application of these designs and method.
- The drawing of legitimate conclusions based on these designs and methods.

6.4 The chapter aims to judge quality whilst at the same time contributing to quality improvement during the course of an evaluation.

Main messages

- Although followers of different methods and paradigms traditionally advocate different QA approaches there is equivalence at the level of overarching standards.
- Common criteria and standards for IE designs are possible across quantitative and qualitative methods even though criteria need to be adapted to different methodological approaches.
- One of the main contributions of QA to qualitative evaluation and research is a focus on the conduct or process of an investigation over the evaluation life-cycle – from planning to reporting.
- Existing frameworks are able to address most generic QA needs and should be drawn on rather than replicated however there is a need to focus on what is distinctive in the range of designs and contexts identified in this report.
- A framework for QA in three parts is proposed that includes criteria that operationalise standards such as Reliability, Robustness, Transparency, Validity and Rigour.

6.1 An 'inclusive' framework

6.5 The ToR compares the kinds of standards that are required for any set of designs and methods with those suitable for experimental and or quasi experimental designs: those for 'alternative designs' should be 'different but equal' whilst having 'equivalent robustness'. Consistent with the approach taken throughout this report we consider the question of quality 'inclusively', assuming from the outset that the rationale for quality assurance and the form it should take will need to be coherent across different designs, whether qualitative or quantitative, theory-based or experimental – and including mixed designs.

6.6 There is a longstanding debate between those who advocate a common approach to QA for all studies based on quantitative and particularly statistical norms (following on from Campbell and Stanley 1966; Shadish Cook and Campbell 2002); and those who argue for a distinctive approach for qualitative research and evaluation22. Quantitative concepts such as rigour building on notions of 'reliability' - understood as how to make an investigation replicable; and 'validity'- understood as accuracy or something that is true to the phenomenon – have been fought over for at least 50 years. Validity has been a particularly hard-fought battleground when the distinction is made between internal and external validity. External validity - whether findings are applicable to other populations or settings than where they were first originated – is the foundation for generalisability claims. And being able to generalise is central for any kind of research and evaluation that aims to be policy relevant.

6.7 Guba and Lincoln (1981, 1989), the most influential proponents of a distinctive qualitative set of criteria, argued that 'trustworthiness' is more appropriate for qualitative approaches rather than 'rigour'. Instead of validity and reliability that emerged from statistical or 'scientific' traditions these authors suggest that for 'naturalistic' approaches criteria such as credibility, transferability, dependability and confirmability. This debate originated before the emergence of mixed methods as an almost dominant mode of social scientific inquiry that it has now become. The need for common standards across different methods was less pressing when quantitative and qualitative approaches were seen as a parallel or even competing rather than an integrated practice.

6.8 The distinction between quality criteria advocated by these different research and evaluation traditions are not as clear-cut as it might at first appear. For example, the criteria 'transferability' as used by Lincoln and Guba (1985) can be understood as very close to external validity given its focus on generalisation. Similarly 'confirmability' which focuses on the investigator's decisions (and potential biases) can be seen as one way of getting as close as possible to reliable or replicable findings. The Cabinet Office study (by Liz Spencer and colleagues) also suggests a degree of equivalence between these approaches as in the table below.

[22] Underpinning these arguments are different paradigms – naturalist and scientific or within a more fine-grained template: positivist, post-positivist, interpretivist, constructivist, realist etc.

Table 6.1: Lincoln and Guba's *naturalistic* criteria (source Spencer et al 2003)

Aspect	***Scientific term***	***Naturalistic term***
Truth value	Internal validity	Credibility
Applicability	External validity or generalisability	Transferability
Consistency	Reliability	Dependability
Neutrality	Objectivity	Confirmability

6.9 Seale adopts a similar version of equivalence. He argues for reflexivity as a way of 'showing the audience of research studies as much as possible of the procedures that have led to particular set of conclusions' (Seale 1999: 158). This can be seen as another way to bridge reliability and dependability. The notion of the 'responsiveness' of the researcher to circumstances they encounter in the field and being sufficiently engaged to be able to support analysis and justify interpretations (Morse et al 2002; Stake 2004) can also be seen as a way of reframing rather than abandoning reliability and validity.

6.10 Responsiveness could also be seen to encompass conducting an evaluation so as to encourage dialogue between commissioners and evaluators; and between beneficiaries or other stakeholders. For example Giel Ton (2012) puts forward a 'three step process for improving rigour in impact evaluation'[23]. This includes in Step 1 developing 'specific questions, in collaboration with our clients'; and in Step 3 'maximize learning' by engaging with stakeholders so that they are able to compare particular IE's findings with others.

6.11 Morse and her colleagues argue that responsiveness opens the way to auditing or verifying how and why research decisions are taken. This focuses attention on the research *process* and *procedure* rather than on *methods* alone. Verification strategies that aim to improve or steer the course of an investigation can become 'mechanisms used during the process of research to incrementally contribute to ensuring reliability and validity and thus the rigour of a study'. Transparency as a standard appears to capture much of this debate on verification and audit. This is an important feature of this way of thinking about quality in research and evaluation: like the ToR for this study it is as much concerned with the way the investigation is conducted as with making judgements about the quality of the outputs – by which time it is too late to correct poor decisions.

6.12 Ton (2012) provides an interesting example of bridging QA concepts whilst being eclectic in drawing on different designs and methods. Following Shadish, Cook and Campbell (2002) a research group based at Wageningen University took four types of validity (statistical conclusions validity; internal validity; construct validity; and external validity) and for each anticipated 'threats to validity'. As the table below shows, each of these threats was then addressed through the use of additional 'mixed' methods.

[23] Step 1 'Refines the evaluation questions based on the intervention logic'; Step 2 'Anticipates validity threats to the expected type of conclusions'; and Step 3 seeks to 'Maximize the scope for comparative research'.

Table 6.2: Summary of Additional Methods to Counteract Threats to Validity (from Ton 2012)

Type of Validity Threat	*Main Threat*	*Additional Mixed Methods*	*Result/Observation*
STATISTICAL CONCLUSION	Selection bias between treatment and comparison group.	Case-based statistics to maintain case integrity in group comparisons.	Instead of measuring and comparing averages of impact, we identify types of responses related to types of contexts and types of constellations of factors. These typologies are refined/validated in focus group discussions with key stakeholders.
INTERNAL	Attribution in complex systems.	Process tracing based on significant experiences in resolving agency dilemmas in collective action.	Evidence of ways that organisations change their organisational capabilities by collective marketing activities is gathered, with thick descriptions of key moments to do so. The evidence underpins claims that experience with value-added activities translates into learning and refined internal regulations and incentive structures.
CONSTRUCT	Measurement of organisational capabilities.	Repetition of measurement of the self-assessment procedure with differing panel composition in the same organisation.	The self-assessment procedure for qualifying the strength of farmers' organisations is cross-checked before assuming that it can be used as a monitoring device.
EXTERNAL	Diversity *in extremis.*	Structured case studies, with due attention to incentive structures (mechanisms) that limit opportunistic behaviour.	By focusing on behavioural incentives for internal control, instead of functional diversity in economic activities, common challenges of organisations are explored and solutions presented with a defined generalisation domain

6.13 One conclusion from reviewing the literature on assessing the quality of research and evaluation is that it *is* possible to use overarching standards such as reliability and validity across different designs and methods. Of course these standards will need to be operationalised in terms of criteria that are suited to the designs in question. Furthermore QA for IE will still need to focus through a particular IE lens. However this is more a question of supplementing overarching QA approaches and standards rather than abandoning them.

6.2 Existing QA Systems

6.14 In order to form a clearer view of QA processes we overviewed a range of existing QA systems developed for research and evaluation. There is no shortage of frameworks to assess the quality of research: the study by Spencer and colleagues (2003) reviewed 29 frameworks geared to qualitative research alone. Our much less comprehensive overview concentrated on QA approaches that were intended specifically for international development such as the OECD-DAC Quality Standards for Development Evaluation, DFID's own Quality Assurance Template and Europe Aid's quality control checklist. However we also considered approaches used outside of international development such as the guidance for using RCTs when evaluating complex interventions produced by the Medical Research Council (MRC 2008).

6.15 Some approaches to quality that we considered are explicitly geared to 'Impact Evaluations' – e.g. the NONIE Guidance, but other guidance and standards were not. For example, some guidance was intended to be used with any qualitative methods, as is the case with the UK Cabinet Office guide to 'Quality in Qualitative Evaluation' (Cabinet Office 2003) or the 'Methodology Checklist: Qualitative studies' produced by the National Institute of Clinical Excellence (NICE 2009). On the other hand guidance and protocols from the Campbell and Cochrane collaboration (e.g. Campbell Collaboration 2001) – admittedly concerned with systematic reviews - are more geared to quantitative methods.

6.16 An initial conclusion is that there are already many examples of QA guidelines in existence and it would be wise to build on these rather than to proliferate or duplicate what already exists where they are applicable.

6.2.1 Why QA matters?

6.17 These sources give an overview of purposes and intentions of QA and allow us to address the question: 'why does QA matter?'. The concern in this study's ToR, to ensure that methods can 'withstand criticism' is present in many of the sources of guidance reviewed. Thus the Cabinet Office rationale for its 'framework' was to ensure that in the context of evidence-based policy, government could be sure that 'research and evaluation is of the highest standard'. Other sources emphasise avoiding 'bias' or misrepresentation of evaluation results. This underpins much Campbell and Cochrane guidance, for example, the OECD-DAC Quality Standards and Europe Aid 'checklist' both aim to 'improve the quality of development evaluation processes and products' and reflect a wider concern to improve practice by influencing what evaluators do.

6.18 The table below derived from the report to the Cabinet Office (2003) by Liz Spencer and colleagues summarises the main issues in quality assurance for qualitative evaluation and how they can be addressed.

Table 6.3 Key Quality Issues and Concerns in Qualitative Evaluation

(Based on Spencer et al, 2003 pages 71-72)

Underlying concern	*Ways of ensuring quality (Examples)*
The defensibility of the approach	• A clear logic of enquiry; • Clarity of questions posed; • Rationale for questions; • Responsiveness to real life context; • Fitness for purpose.
The rigour of conduct	• Collection of in-depth data; • Careful recording of data; • Contextual documentation; • Systematic and thorough analysis; • Explication of conceptual and analytic process; • Auditable documentation.
The relationship of the researcher to the researched	• Ethical behaviour (e.g. gaining consent); • Involvement of participants in study; • Reflexive awareness of investigators' role; • Open and empathetic fieldwork skills; • Recognition of different subjective perspectives.
The credibility of claims	• Triangulation; • Validation by informants/respondents; • Peer review; • Consideration of alternative explanations & negative cases; • Balanced presentation of evidence; • Demonstrating links between data and conclusions.
The broader contribution of the study	• Relevance and utility to policy; • Ongoing involvement of potential users in planning and discussing recommendations; • Timeliness; • Clear reporting and active dissemination; • Linking findings to broader research and theory.

6.19 These issues also suggest different audiences for QA: the evaluation practitioner (including those who commission evaluations) facing technical challenges; and the policy-maker and public who want to be reassured that they are dealing with a 'quality' product. In the international development environment, where different 'partners' are working together and share evaluation results, there are also 'partnership' considerations. For example the OECD-DAC Quality Standards are also intended to 'increase development partners' use of each others' evaluation findings.

6.2.2 Approaches to QA in practice

6.20 Most of the sources reviewed are consistent with the broad approaches to QA that were outlined in the above discussion of the literature.

6.21 First it should be noted that with the exception of Europe Aid which is more oriented to 'Quality Control', most approaches follow the life-cycle of an evaluation. Thus DFID has separate QA Templates at 'Entry Level' and 'Exit Level' and the OECD-DAC Quality Standards are expected to be 'used during the different stages of the evaluation process': and is organised into sections on 'purpose, planning and design' and 'implementation and reporting'.

6.22 Second, there are certain key terms that recur in most standards and QA systems. These include terms such as 'validity', 'rigour', 'robustness', 'reliability', 'impartiality' and 'credibility'. However how these terms are used varies considerably. Within particular methodological 'camps' there is a tendency for specific meanings to be attributed to these concepts that are consistent with that schools assumptions. Participatory evaluators, 'Realists' and followers of quasi-experimental methods each have their own lexicon. Thus for evaluators who use standardised statistical methods 'reliability' is often taken to mean replicability of the evaluation provided the same methods are followed by others. However for qualitative evaluators 'reliability' can be more about being clear about how findings were reached: through what methods, with what assumptions and aided by what procedures. At a certain level of analysis an explicit 'audit trail' can also lead to what some call replicability but for others this means the repeated use of a specified method or technique.

6.23 This diversity in the way terms are used argues for pitching any QA approaches for IE that are to be applicable to different designs and methods at a sufficiently inclusive level to avoid discriminating in favour or against any particular design.

6.24 Third, the content of QA whether described as a 'guideline' or a 'standard' is mainly operationalised as a set of questions that evaluators, managers or policy makers should ask themselves ('How has this evaluation involved stakeholders?', or 'Has the design that is used been justified and explained?'). The underpinnings for these questions cover:

- <u>Norms</u> – i.e. assumed values or principles. These might include the need to involve stakeholders, working with partners and the adherence to ethical standards of behaviour.

- Conformance – i.e. did the evaluation address the evaluation questions and the aims specified in the ToR.
- Processes – i.e. how was data collected, is it clear how data was analysed and conclusions reached, is the choice of design made explicit.
- Techniques – i.e. were techniques or methods correctly used, for example, was the sample big enough, were statistical techniques appropriately applied, were interviews transcribed etc.

6.25 From the standpoint of IE it is the latter two aspects of 'processes' and 'techniques' that appear most useful to defend studies from criticism and minimise risks of bias. It is noteworthy that across most existing QA guidance the majority of attention has been paid to processes – how to do things rather than to technical criteria. This is probably because of the sheer diversity of possible guidance that would have to be included if technical criteria that encompassed a full range of possible methods – for example econometrics, experiments, ethnography and case studies - were to be included in generic guidance.

6.26 Fourth, existing QA systems are consistent with some of the main distinctions in the literature outlined earlier. For example criteria such as *validity* are sometimes used in technical ways: does a particular method adequately describe or measure what it intends to describe or measure?. This can be extended to distinctions between internal and external validity, addressing the problem of generalisation. It can also be used as a process term: 'Has this evaluation argued convincingly that the methods it uses are likely to adequately represent the 'object' of evaluation?'. Similarly for some practitioners 'rigour' means adherence to their preferred techniques and protocols whilst for other evaluators it means evidence of rigorous thinking, and the deployment of a defensible and logical argument (Bamberger and Rugh 2008). This latter understanding makes it easier to include a diversity of designs and methods and ensure a level playing field.

6.27 Fifth, in terms of the potential scope of QA (i.e. choice of appropriate methods; their proper application; and drawing legitimate conclusions), existing QA systems appear to be unbalanced. Most focus on the proper application of designs and methods once these have been chosen. They are not concerned with choosing between methods – there is more likely to be a presumption that a particular design is 'best'. Drawing conclusions is addressed more routinely but mainly in terms of whether conclusions can be supported by data and analysis. Consideration of value judgements is less common although the normative criteria dimension in many sets of 'standards' does provide a basis for guidance on value judgements.

6.3 Implications from the above discussion

6.28 A number of implications can be drawn from the above discussion for the kinds of standards and QA systems that would be appropriate to a broad range of IE designs and methods:

- Existing QA systems cover much of the territory relevant to IE. There may be some specific elements that are not but much existing guidance will be transferable across.

The design decisions are how far to rely entirely on existing standards and QA systems or whether to assemble a dedicated subset – this will partly depend on the needs of stakeholders in QA and partly on the adequacy with which existing QA systems deal with impact.

- The main purposes of any QA system should be to ensure that IE evaluations can be defended and justified; whilst at the same time encouraging good practice among evaluators.

The design decisions are: how QA should be inserted at all stages of the evaluation cycle; and how to ensure that there is both an 'encouraging good practice' element and an externally and independent element that can assure policy makers that designs are defensible.

- Good practice should meet both technical and processual criteria: focusing on designs and methods used; and *how* they are applied in practice.

The design decisions are: how to find a level of analysis that it is inclusive of different designs and methods drawn from different schools so that no acceptable approaches are unduly favoured or discriminated against; and how to accommodate the diversity of designs and methods in relation to technical standards of good-practice.

- QA for IE should pay attention to the choice of methods and reaching evaluative judgements as well as to how to apply methods.

The design decisions are: how to incorporate design choice; and evaluative judgements into considerations of quality for IE.

6.4 A framework for assessing quality of IE

6.29 It is not within the scope of this study to fully specify a QA system for IE across all possible designs. However the above discussion does suggest what a framework for QA that is oriented to IE in development settings and is qualitative as well as quantitative inquiry would look like.

6.30 The suggested framework is in three parts, addressing:

- How an IE is conducted;
- The technical quality of designs and methods; and
- The international development context – normative, ethical and institutional.

6.4.1 Standards about how an IE is conducted

6.31 This sets out 'process standards' that are concerned with how an evaluation is conducted. It includes with standards previously discussed such as Transparency, Reliability and Robustness. The table 6.4 follows the logic of the life-cycle of evaluation. It begins with the 'choice of designs and methods', then 'application of designs and methods' and finally 'drawing legitimate conclusions'. Many of the questions that are asked apply to all kinds of evaluation although some are specific to IE evaluations.

Table 6.4 How an IE is conducted: Standards and Criteria over the Evaluation Life-cycle

(Reliability, Robustness and Transparency)

Choice of designs & methods ***Reliability***	Are designs and associated methods put forward that are established, well documented and able to be defended? Do the chosen designs take into account Evaluation Questions and intervention attributes? Are they able to explain how an intervention contributes to intended effects for final beneficiaries? Do the EQs allow for success and failure (positive and negative effects) to be distinguished?
Proper application of designs and method ***Robustness***	Are the ways that designs and methods are applied clearly described and documented? Does the application of designs and methods and subsequent analysis follow any protocols or good practice guidelines? Is the evaluation team knowledgeable about the methods used?
Drawing legitimate conclusions ***Transparency***	Do conclusions clearly follow from the findings? Has the evaluation explained the effects of the programme? How are evaluative judgements justified? Have stakeholder judgements been taken into account when reaching conclusions? Are the limitation of the evaluation and its conclusions described?

6.4.2 Technical standards for IE designs and methods

6.32 The second part of a framework is concerned with 'Technical Standards' for methods and designs that mainly relate to what is usually labelled validity and rigour. These standards specifically concern IE. They ask how the designs and methods chosen address the three main characteristics of IE following the definitions and discussions contained in this report. We have argued that IE should:

- Address the contribution made by an intervention;
- Provide a clear causal link to effects;

- Offer an explanation of how a programme worked.

6.33 For each of these categories a set of criteria are suggested as generic questions that could be asked of any IE design.

Table 6.5 Technical standards and criteria to judge the quality of IE designs and methods

(Validity and Rigour)

Contribution	***Explanation***	***Effects***
Is the design able to identify multiple causal factors? Does the design take into account whether causal factors are independent or interdependent? Can the design analyse the effects of contingent, adjacent and cross-cutting interventions? Are issues of 'necessity', 'sufficiency' and probability discussed?	Does the evaluation make it clear how causal claims will be arrived at? Is the chosen design able to support explanatory analysis (e.g. answer how and why questions)? Is theory used to support explanation? (e.g. research-based theory, Theory of Change), if so how has theory been derived? Are alternative explanations considered and systematically eliminated?	Are long term effects identified? Are these effects related to intermediate effects and implementation trajectories? Is the question 'impact for whom' addressed in the design?
Please attach any protocols, guidelines or quality assurance systems used in connection with this design Please also provide previous reports or publications that illustrate how this design has been used previously for IE purpose.		

6.34 This could also be backed up by documentation regarding the design concerned: including examples of technical standards used with any particular design; and evidence that the design has been used previously in related ways (see final row of table). This would allow a more design specific set of QA guidelines to be developed over time. For example, evaluators could be required to fill in a pro forma along these lines when they submit proposals justifying their use a particular design. A follow-up version could also be used when reports are submitted. Together with documentation (protocols, guidelines, previous examples of use) lodged when IE proposal were submitted, this would allow for a 'registry' of designs to be assembled that was more design specific.

6.4.3 Normative criteria: standards related to the international development context

6.35 The third part of a QA framework recognises that we are dealing with IE in the specific normative context of international development. This context can be understood in terms of:

- Country-based criteria;
- Ethical criteria; and
- Institutional criteria.

6.36 Country-based criteria follow from the standards and norms agreed in the international community, many of which were discussed in Chapter 2 and are covered in OECD-DAC guidelines and the Paris Declaration. Ethical criteria partly follow from these broader norms although many of them would be similar to good-practice in most evaluations. Institutional criteria derive from the institutional arrangements that are common in development programmes including the types of policy instruments used and the aid modalities that are employed.

Table 6.6 Normative standards and criteria in the international development context

Country-based criteria	***Ethical criteria***	***Institutional criteria***
Have country-based stakeholders (government and civil society) been actively involved and consulted in formulating evaluation questions? Have country based administration and information systems been used as far as possible? Has the evaluation been set into the country context and other country based evaluation taken into account?	Have the evaluators made explicit their interests and values as they relate to this intervention? Have arrangements been put in place to monitor and remedy bias or lack of impartiality? Have confidentiality and risks to informants been taken into account? Have feedback and validation procedures that involve stakeholders been specified and used?	Are there any joint or partnership arrangements in place – joint evaluations, consortia involving local partners? In what ways has the evaluation contributed to evaluation capacity building in-country? What has the evaluation done to feed into policy making both among donors and in partner countries?

Conclusions

6.37 This section has outlined a framework for QA in IE that will need to be further developed and integrated with other existing QA systems that DFID and other development agencies already use. It should be regarded as indicative only. However it does integrate a range of issues and approaches described in the literature on QA for evaluation and research and builds on a range of existing guidelines developed in different policy domains.

6.38 It is understandable that broadening the range of IE methods and designs will highlight concerns that policy makers will have about quality and being able to justify and defend evaluation findings. This chapter has sought to develop a QA framework that will allow policy makers to have confidence in IE provided certain standards and criteria are in place. At the same time it is important to reinforce a concern for quality in the evaluation community. This is why the proposed framework integrates quality considerations at every stage of the conduct of an evaluation. This is intended to encourage evaluators to think through and make explicit their assumptions and decisions. In this way QA systems can support learning and become a vehicle for quality improvement as well as quality checking.

Chapter 7: Conclusions and Next Steps

Introduction

7.1 This chapter revisits the study's conclusions in relation to the ToR noting both what has been achieved as well as the study's limitations. Many of these limitations follow from the available time and resources given an ambitious agenda. Others are the inevitable consequence of a study undertaken by academics and consultants who can never be fully aware of all the constraints and requirements of policy and operations. It is for this reason that many of the next steps identified below would require close engagement by practitioners and policy makers to take IE to the next stage.

7.2 It is already clear however that capacity in various forms is a major problem that will need to be addressed. An important part of the recommended 'next steps' concern capacity – amongst the commissioners of IE; within the evaluation community; and for other stakeholders. At a country level in particular the implications of the study's conclusions for stakeholders and for country-based evaluators are considerable in a post Paris Declaration era.

7.3 A further limitation is inbuilt into what IE can reasonably promise and deliver. It is not easy to establish causal links between interventions and their effects when we move beyond simple programmes. IE using most methods is demanding in terms of skills and other resources and the results can never be certain or accurately predictive. This suggests the need to be extremely clear and selective about the circumstances in which IE is justifiable in terms of 'Value for Money'. These circumstances are likely to continue to be unusual and atypical.

7.4 The chapter is in three main parts. The first brings together in one place the study's conclusions prefigured in earlier chapters. The second indicates the limitations of the study in terms of what was not fully covered and the implications this has. The third recommends next steps that follow from the study and its conclusions.

7.1 Ten conclusions: What the study has found

7.5 The ToR defines the objective of the study as: 'To produce an assessment of a range of designs and methods appropriate to rigorous impact evaluation of a broad range of development programmes'. To this end the study was expected to review evaluation designs and methods; consider existing examples of IE; match methods to types of interventions; and draft quality standards.

7.6 The following 10 conclusions summarise what the study found:

1. ***IE can best be understood as a form of causal analysis that links an intervention with effects for beneficiaries.***

7.7 This study has put establishing a cause and making a causal claim at the heart of IE. There are various ways to draw casual inferences associated with different methodological schools and paradigms. There is only limited consensus across the social sciences although each design24 has its acknowledged strengths and weaknesses.

7.8 It is more useful to focus on causality rather than on any single design or causal logic. For example, counterfactuals are at the core of experimental designs and are more widely useful as a general way of thinking about policy options. However associating all of IE with counterfactuals in order to attribute results to programmes would risk defining IE in an overly method-specific way.

2. *Explaining cause and effect is also a necessary part of IE in many policy settings*

7.9 Policy makers who are concerned with policy learning, generalisation and scaling-up look for explanations as well as causal analysis. In most cases IE will also involve explanatory analysis – answering 'how' and 'why' alongside 'what' questions. Explanation beyond the single case raises questions of the external validity of findings.

7.10 Explanations are not needed in all evaluations. If the purpose of an evaluation is to establish whether a particular intervention has worked in a particular setting and the interest is in accountability rather than policy learning explanation may not be required.

3. *IE should as far as possible reflect the normative dimension of development aid*

7.11 A great deal of effort in development has gone into debating and negotiating norms that should govern the relationship between donors and the recipients of aid. This study has therefore built on the OECD-DAC definition of impact.

7.12 The aid relationship as defined in the Paris Declaration and other international agreements also has implications for IE. For example, country leadership and using countries' own systems affects access to information; IEs need to strengthen rather than undermine governance and administrative capacity; poverty reduction and supporting basic equalities focuses attention on 'impact for whom'; and the decentralisation of programme leads to renewed interest in joint evaluations and participatory approaches.

4. *Appropriate IE designs should match the evaluation question being asked and the attributes of programmes*

7.13 IE designs need to be appropriate to the programmes being evaluated and the evaluation questions being asked. We do not think that IE designs should be classified as experimental and quasi experimental on the one hand and all other 'alternatives' on the other.

7.14 Reviewing existing evaluations that were described as being concerned with 'impact' it was clear that evaluation questions were often not clearly articulated and were sometimes completely absent. There was also very little systematic analysis of how programme attributes should shape how evaluations are conducted. Sometimes poor quality evaluations were the result of inappropriate designs and methods being applied to programmes for which they were unsuited.

[24] A design is understood as clusters of methods linked together by common causal and explanatory logics.

7.15 A focus on *appropriate* designs is also consistent with an integrated framework for design choice – which can include a full range of methods including statistical and experimental designs in suitable circumstances.

5. There are a range of potential designs based on a number of causal inference logics that are suitable for use in IE

7.16 The study has identified a range of possible IE designs that could be useful to evaluate the impact of development programmes. Such designs include experimental, statistical, theory based, case-based and participatory approaches.

7.17 Many of the more innovative designs are at present poorly applied in development evaluations. This suggests both quality assurance and capacity deficits which will need to be remedied if new designs are to be taken forward.

7.18 Many of these designs require high level evaluation skills and are resource intensive. This reinforces the importance of selectivity and clear criteria for when IE is deployed.

6. The strength of causal inference is inversely correlated with the scope of the programmes being evaluated

7.19 Critics of extending the range of IE designs often do so on the basis that these designs may only be able to support 'plausible' causal claims rather than 'prove' – and even measure - how a cause led to an effect. This criticism is valid but stems from the capabilities of designs in the face of programmes with different attributes.

7.20 Designs and methods applied to narrowly specified interventions can support strong causal claims but as the scope and scale of an intervention increases the strength of causal claims is reduced. The reality is that many contemporary programmes are not narrowly specified: they are ambitious, broad in scope and made up of many sub-programmes. Policy makers may therefore have to accept a trade-off between strong causal inference and relevance.

7.21 There are some remedies to this dilemma: combining methods to strengthen causal inference (see conclusion on mixed-methods below); and applying 'tests' to evidence generated in broadly specified programmes (see conclusion on QA below). In some circumstances it may even be possible to break down large-scale programmes into smaller parts and evaluate these. But this may not help assess the impacts of the larger programme with which a particular sub-part is a part.

7. Most interventions are a 'contributory cause' and part of a causal package making it unusual for an intervention to cause a development effect on its own

7.22 Interventions rarely produce development results on their own. Their potency depends on how they combine with other factors. These may include other country-based policies and programmes, hence the importance of policy coherence; institutional and cultural pre-conditions; and exogenous factors like an economic cycle or world trade. This conclusion has implications for the kinds of evaluation questions that need to be asked in IE. More generally the study identifies the methodological implications of different evaluation questions.

7.23 Theory-based and case-based approaches are especially suited to unpicking 'causal packages' - how causal factors combine - and what might be the contribution of an intervention. However such approaches are not good at estimating the quantity or extent of a contribution.

8. Combining methods is a useful strategy both to increase confidence in IE findings and help compensate for the weaknesses of particular methods

7.24 Mixed methods are now commonplace. In the past the label was mainly applied to combining qualitative and quantitative methods but nowadays can also include combinations within quantitative and qualitative categories. Support for mixed methods has increased with the erosion of simple distinctions between quantitative and qualitative data; and the emergence of quantitative but not statistical methods to analyse qualitative data.

7.25 There are few ground-rules for how methods should be combined although distinctions can be made between 'parallel' and 'sequential' combinations of methods. The first is consistent with what is commonly described as 'triangulation' and the latter to applying different methods at different stages of an evaluation. For IE purposes the main concern needs to be with the implications of combining methods for strengthening causal inference. To that extent the criteria for combining methods should be derived from the causal logics at the level of designs rather than at the level of methods used within a design.

9. Programmes have attributes that are shaped by what they try to do and the contexts in which they are situated – and these matter for IE

7.26 Ambitious policy goals make many programmes complex. Strengthening governance, fragile states and democratic institutions; or helping to mitigate climate change, requires multi-dimensional interventions, many of which are in policy areas about which relatively little is known and which results will only become clear in the long-term. Their contexts may be unsafe as well as uncertain. Post Paris Declaration delivery and implementation will usually be through extended chains: via partner country governments or via arms-length agents such as funds and resource centres or jointly with other donors. The study identified particular attributes of programmes that have implications for IE designs.

7.27 Comprehensive and multi-dimensional programmes make deciding on units of analysis for an evaluation challenging. One common error is breaking down interconnected interventions into component parts so as to make them evaluable by particular methods and then generalising about the entire programme. Criteria for when to treat programmes as whole-systems and when to decompose them into subparts are needed. In this regard analysis of the extent of vertical and horizontal integration can be helpful

10. A QA (Quality Assurance) framework is needed to assure policy makers that IE findings are defensible and to encourage evaluators to improve IE design and implementation

7.28 One of the main barriers to extending the range of IE designs is policy makers' concerns that evaluation findings are reliable and defensible. The study concluded that it is possible to use common QA system across different designs and methods. This has become even more necessary given the now-common practice of combining methods and the way contemporary methodologies are breaking down traditional distinctions between quantitative and qualitative methods.

7.29 A three-part framework for QA is proposed that covers: the conduct of an evaluation over the life-cycle; technical criteria to judge IE designs and methods; and normative criteria that follow from the way the aid relationship is understood post Paris, Accra and Busan. This framework operationalises standards such as Reliability, Robustness, Transparency, Validity and Rigour in ways that can be applied to a broad range of designs. The framework is intended to be useful for those who commission evaluations whilst also encouraging evaluators to improve the quality of IEs.

7.2 Limitations of this study

7.30 There are a number of limitations of this study. Some are the result of time and resource constraints; some follow from decisions made during the course of the study in the face of alternative options; and some follow from the priorities included in the ToR. These limitations suggest future work in the IE area and also prefigure a number of specific recommendations that are put forward in the following section on 'Next Steps'.

7.31 Only a few cases of existing IE were analysed. This was because it rapidly became clear that the cases that had been collected from a cross section of development agencies offered very few examples of good practice on which future designs could build. This judgement was reinforced by the responses of development agencies which either noted that they did not undertake IE or identified IE entirely with experimental methods rather than a broader range of methods. It may well be that a more thorough trawl of development agencies and more extended follow-up would have yielded more positive examples. One way of sharing positive examples of IE in the future would be to accumulate a central registry of completed IE classified according to broad categories of designs

7.32 Because of the relative weakness of available case material more effort was put into reviewing literature. These literatures spanned many subjects including: evaluation practice, aid effectiveness, social science methodologies, philosophy of science and complexity theory. It was not possible within the time and resources available to conduct a thorough review of these diverse literatures. In part the team relied on prior knowledge whilst the emphasis for new sources was to look for points of contact across literatures rather than to attempt an in-depth review within so many different specialisms.

7.33 The study team was made up of consultants and academics rather than development practitioners even though some team members had been practitioners in the past. Although team members participated in two workshops there were limited opportunities to engage with DFID HQ staff and even fewer opportunities to engage with DFID country-based staff. Those who expected a user-friendly report translated into everyday language will therefore be disappointed. There is undoubtedly scope for practitioners within DFID to take this report and turn it into a number of more accessible outputs. However there are limits to how far this will be possible because of what are likely to be continuing challenges of undertaking IE to a high standard when working with complex programmes. Some of the underlying arguments around causal inference and the technical difficulties of applying state-of-the-art methods will continue to demand highly specialised knowledge and experience. This raises issues of capacity development addressed below under 'next steps'.

7.34 The study concluded that a broader range of IE designs and methods exist and if suitably developed and applied would be able to extend the evaluation of impacts to programmes and contexts where this is currently difficult. However this study can only be regarded as 'proof of concept'. In practice designs and methods identified in the report will need to be tailored, refined and field tested. This is especially so when as is the case with some of the IE approaches identified designs are immature, still being used mainly by their originators and a small group of enthusiasts and have not been routinized in any way. Field testing a small number of demonstration projects is considered in the section on 'next steps'.

7.35 The ToR made a clear distinction between experimental and non-experimental approaches to IE; and between qualitative and quantitative methods. In the course of the study it became clear that what were needed were *appropriate* rather than *alternative* designs. Indeed the team was strongly advised by participants in workshops and by peer reviewers to develop an integrated framework that could be applied across all IE designs and methods. This was attempted: for example the classification of designs and the QA framework were both intended to accommodate all designs and methods. However the team was not able to give the same attention to experimental and statistical approaches as it did to less well known approaches. It would have been useful to subject what many consider 'traditional' approaches to IE (RCTs, quasi experimental and statistical designs) to the same critical scrutiny as was applied to newer approaches.

7.36 It was noted in the ToR that relevant approaches to IE were unlikely to be confined to international development evaluation. Scans of evaluation journals confirmed this. IE methods using theory based evaluation approaches are particularly evident in public health; whilst realist evaluation approaches seem strongest in health and criminology. Other methods and approaches were not necessarily tied to evaluation. For example, QCA grew out of and remains strongest in comparative political science research. Emerging methods such as 'Agent based modelling' appear strongest in areas like climate change and other environmental fields; and systems dynamics and other forms of modelling are scattered across engineering, urban spatial analysis and corporate and product planning. This study cannot claim to have adequately covered all of these topics. The dispersed way in which innovative designs emerge suggests that some kind of awareness/scanning initiative could be useful in future.

7.3 Next steps

7.37 This section identifies a number of specific actions that follow on from this study and which if implemented would serve to ensure that a broader range of IE methods would be adopted. These actions are under two headings:

- Policy and practice development.
- Capacity development.

7.3.1 Policy and practice development

Disseminating material and conclusions

7.38 There are various ways of disseminating the results of this study – including preparing briefing notes, presentations to senior management, incorporation into staff training sessions and presentation at conferences. A systematic dissemination plan would support a wider understanding, appropriate take-up and further development of the ideas and methods identified in this report.

Preparing guidance notes for programme and field staff

7.39 Not all of this report can be translated into practical guidance or 'how to' material such as HTNs; but many parts can be. Such guidance can best be prepared by in-house evaluators who understand programming and evaluation commissioning. The aim should be to prime staff sufficiently so that they can understand when different designs are appropriate and how to commission relevant IEs with appropriate ToRs and appropriate competences in evaluation teams.

Adapting QA systems to focus on IE

7.40 The study proposed a framework for QA with criteria and standards that would focus specifically on IE. This should be considered alongside existing QA systems such as the entry level and exit level QA templates that DFID already uses and seen as a complement to existing QA systems rather than as a substitute. The proposed three part framework outlined in Chapter 6 aims to make IE designs defensible for policy makers whilst also encouraging transparency, responsiveness and good practice among evaluators. The proposed framework has been conceived as 'inclusive', able to accommodate, for example, both qualitative and theory-based designs as well as experimental and statistical designs. This avoids the danger of some existing IE standards which are better suited to a less inclusive range of designs and methods. Any more IE oriented standards would also need to be discussed and further developed with DFID's partners in OECD-DAC and indeed through civil society partnerships such as Programme Partnership Arrangements (PPA). However this should not prevent DFID applying and testing an extended QA framework that is complementary to what has already been shaped by discussions with partners.

Developing criteria for when to implement innovative IE designs

7.41 The cost and difficulty of many IE designs has been noted at several points in this report. Even at the stage of field testing new IE designs some criteria are needed to decide the merits of undertaking an IE and whether the investment is justified. For example, criteria could take into account budget size and assessed risk; likelihood of scaling up in the future; innovativeness of programmes; extent of pre-existing evidence and experience in the programme area concerned etc.. Criteria should also take into account the building blocks of this report, i.e. evaluation questions, available designs and programme attributes. This would, for example, minimise the risk of initiating IEs in circumstances where the evaluation question that policy makers wanted answers to were not IE questions; or where the attributes of the programme made it unlikely that available designs and methods could be applied.

Identifying pilot sites for field testing

7.42 Given DFID's level of decentralisation to country and regional offices potential sites for field testing should be identified in dialogue with decentralised offices. They will for example be most knowledgeable about the willingness and capacities of country based counterparts to cooperate in innovative approaches to IE.

Conducting and learning from pilots

7.43 A limited number of pilots will help clarify the strengths, weaknesses and usefulness of an extended range of IE methods and designs. Provided that pilot sites were selected with different evaluation questions and different programmes in mind, their implementation should be monitored to ensure that the experience is captured and lessons learned.

7.3.2 Capacity development

Capacity development within DFID

7.44 There is a need for internal link-persons within DFID to act as advocates and sources of information on IE. More general incorporation of content from this report into staff development, training and away days would also ensure wider awareness of IE designs in addition to those designs that are currently understood and supported in DFID's staff development activities.

Networking with evaluators, researchers and consultants

7.45 Many of those with relevant knowledge and skills are not currently engaged in IE evaluations in international development. Necessary skills are often in the research rather than the evaluation or consulting community. Various forms of networking could raise awareness and mobilise resources so that they could be called on when needed. Existing network include those linked to major evaluation societies such as European Evaluation Society; and others are associated with practitioners such as 'Better Evaluation', the 'Big Push Forward' and NONIE could also be useful. A consortium of bilateral and multi-lateral development agencies and foundations would also offer a structured platform for future IE related capacity development initiatives.

Supporting high quality resources and expertise

7.46 To be able to deliver on the potential of many of the IE designs identified in this study high quality expertise will be needed. People and networks might be identified through conferences and workshops and DFID may wish to set up an advisory panel of relevant experts. Ideally the kinds of resources and promotion for an extended range of IE designs and methods are needed to mirror the support resources available to experimental and statistical evaluations through 3ie and J-pal.

Networking with stakeholders and partners

7.47 Development agencies work closely together. New ideas about IE need to be discussed with other donors and development agencies in order to raise awareness and support innovation. Dissemination via existing networks such as the OECD-DAC, and Nordic Plus would be most cost effective for other donors. Partner country governments and NGOs also need to be kept informed as at present whilst there are useful resources available to support the use of experimental methods in IE at a country level, fewer resources are available to encourage a wider range of designs and methods.

Bibliography[25]

Abbott, A D (2001) *Chaos of Disciplines*. Chicago. Chicago University Press

Aminzade, R (1993) Class Analysis, Politics, and French Labor History in *Rethinking Labor History,* Berlanstein L (ed). Urbana and Chicago: University of Illinois Press, 90-113.

Bamberger M and Rugh J (2008) A framework for assessing the quality, conclusions validity and utility of evaluations. Paper presented at AEA conference, Baltimore 2007

Bamberger M (2010). 'Institutionalizing Impact Evaluation' in *From Policies to Results. Developing Capacities for Country Monitoring and Evaluation Systems.* M. Segone (ed) New York: UNICEF, 189-217.

Bamberger M, Rugh J and Mabry L (2006). *Real World Evaluation: Working under Budget, Time, Data, and Political Constraints.* Thousand Oaks CA. Sage.

Bamberger M, Rao V and Woolcock M (2010) *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development*, Policy Research Working Paper, 5245: The World Bank.

Bennett A (2010) Process Tracing and Causal Inference. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Brady H E and Collier D (eds), 207–19. Lanham, MD. Rowman and Littlefield.

Brady HE (2002). *Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory. Annual Meetings of the Political Methodology Group.* University of Washington, Seattle, Washington. http://www.polmeth.wustl.edu/media/Paper/brady02.pdf.

Brady HE and Collier D (eds) (2004) *Rethinking Social Inquiry: Diverse Tools, Shared Standards,* Lanham, MD, Rowman and Littlefield

Byrne D (2009) Case-Based Methods; Why We Need Them; What They Are; How To Do Them. In Byrne D and Ragin C (eds) *The SAGE handbook of Case-Based Methods*, London: Sage.

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

Cadot O, Fernandes A M, Gourdon J and Mattoo A (2011) *Impact Evaluation of Trade Interventions: Paving the Way*, Policy Research Working Paper, 5877. Washington: The World Bank.

Campbell DT and Stanley JC (1966) *Experimental and Quasi-Experimental Designs for Research* Chicago, Illinois: Rand McNally.

Caracelli VJ, Green JC and Graham WF (1989) Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis*. 11 (3).

Cartwright N (2007) *Hunting Causes and Using Them: Applications in Philosophy and Economics.* Cambridge University Press.

Cartwright N and Munro E (2010) The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice* 16: 260-266.

Cartwright N and Hardie J (2012) *Evidence-based Policy: Doing It Better. A Practical Guide to Predicting If a Policy Will Work for You.* Oxford University Press

Cilliers P (2001) Boundaries, Hierarchies and Networks in Complex Systems. *International Journal of Innovation Management* 5(2): 135-147.

Collier D (2011) Understanding Process Tracing. *Political Science and Politics* 44 No.4 (823-830)

Collier D, Brady HE and Seawright J (2010) Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Brady H E and Collier D (eds), 161-199. Lanham, MD. Rowman and Littlefield.

Cooper and Geyer (2007) *Riding the Diabetes Rollercoaster: A new approach for health professionals, patients and carers.* Oxford: Radcliffe Publishing Ltd.

Dart J and Davies R (2003) A dialogical, Story-based Evaluation Tool: the Most Significant Change Technique. *The American Journal of Evaluation*. 24 (2): 137-155

[25] Bibliography includes sources referred to in the paper *Models of Causality and Causal Inference* annexed to this report.

Davidson EJ (2009) *Causation inference: Nuts and Bolts*. A Mini Workshop for the ANZEA Wellington Branch. Wellington. http://davidsonconsulting.co.nz/index_files/pres/causation-anzea09.pdf.

Davies HTO, Nutley SM and Smith PC (eds) (2000). *What Works? Evidence-Based Policy*. Bristol: The Policy Press.

Deaton, A (2010) Instruments, Randomization and Learning about Development. *Journal of Economic Literature* 48 (June 2010): 424–455

Denis J-L, Hebert Y and Langley A, et al. (2002) Explaining diffusion patterns for complex health care innovations. *Health Care Management Review* 27: 60-73

Denzin NK (2010) Moments, Mixed Methods, and Paradigm Dialogs. In *Qualitative Inquiry* 16 (6) 419 –427

Dowe P and Noordhof P (2004) *Cause and Chance: Causation in an Indeterministic World*. Abingdon: Routledge.

Edgington D (2004) Counterfactuals and the benefit of hindsight. In *Cause and Chance: Causation in an Indeterministic World*. Abingdon: Routledge.

Ellerman D (2005) *Helping People Help Themselves: From the World Bank to an Alternative Philosophy of Development Assistance*. Ann Arbor: University of Michigan Press.

Elster J (1998) A plea for mechanisms. In Hedström P and Swedberg R (eds) *Social Mechanisms: an Analytical Approach to Social Science Theory*. Cambridge University Press.

Forrest J (2007) Evolutionary and Behavioral Characteristics of Systems. In *System Concepts in Evaluation: An Expert Anthology*. Williams B and Imam I (eds). Point Reyes CA, Edge Press/American Evaluation Society.

Forss K, Marra M and Schwartz R, (eds) (2011) *Evaluating the complex: Attribution, contribution and beyond*. New Brunswick and London: Transaction Publishers.

Frissell CA, William JL, Warren, CE and Hurley, MD (1986) Hierarchical Framework for Stream Habitat Classification: Viewing Streams in a Watershed Context. *Environmental Management*, Volume 10, Issue 2, pp.199-214

Funnell, S and P Rogers (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Fransisco: Jossey-Bass.

George AL and Bennett A (2005) *Case Studies and Theory Development in the Social Sciences*. Cambridge Mass. And London: MIT Press.

George AL and McKeown TJ (1985) "Case Studies and Theories of Organizational Decision Making." *Advances in Information Processing in Organizations* 2: 21-58

Gilbert N and Troitzsch KG (2005) *Simulation for the Social Scientist*. Maidenhead: Open University Press

Glennan S (1996) Mechanisms and the Nature of Causation, *Erkenntnis*, 44: 49-71.

Glouberman S and Zimmerman B (2002) *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?*, Discussion Paper No. 8: Commission on the Future of Health Care in Canada. http://www.plexusinstitute.org/resource/collection/6528ED29-9907-4BC7-8D00-8DC907679FED/ComplicatedAndComplexSystems-ZimmermanReport_Medicare_reform.pdf.

Guba E and Lincoln Y (1981). *Effective Evaluation*. San Francisco: Josey-Bass.

Guba E and Lincoln Y (1989. *Fourth Generation Evaluation*. Newbury Park CA: Sage

Hall N (2004) The intrinsic character of causation. In Dean Zimmerman (ed) Oxford Studies in Metaphysics, volume 1, pages 255–300. Oxford University Press.

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

Hempel CG (1965) Aspects of Scientific Explanation and Other Essays in *The Philosophy of Science*. New York: Free Press

Herrell JM and Straw RB (2002) *Conducting Multiple Site Evaluations in Real-World Settings*. San Francisco, CA: Wiley.

Howe K (2004). A Critique of Experimentalism. *Qualitative Inquiry*, 10 (4): 42-61.

Hume D (1739). *A Treatise on Human Nature*

Hummelbrunner R (2010) Beyond Logframe: Critique, Variations and Alternatives. In *Beyond Logframe: Using System concepts in Evaluation*. Fujita N (ed) Tokyo: FASID: 1-34.

International Initiative for Impact Evaluation (2008) Founding Document for Establishing the International Initiative for Impact Evaluation. http://www.3ieimpact.org/strategy/pdfs/3ieFoundingDocument30June2008.pdf

Jonas N, Jonas H, Steer L and Datta A (2009). *Improving Impact Evaluation production and use*. Occasional paper 300: Overseas Development Institute (ODI). http://www.odi.org.uk/resources/docs/4158.pdf

Katila M,. Williams PJ, Ishengoma R and Juma S (2003).*Three Decades of Swedish Support to the Tanzanian Forestry Sector*. Stockholm: Department for Natural Resources and Environment, SIDA.

King G Keohane RO and Verba S (2004). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

Kirsh SR, Lawrence RH and Aron DC (2008). Tailoring an intervention to the context and system redesign related to the intervention: A case study of implementing shared medical appointments for diabetes *Implementation Science* 2008, 3:34

Lakoff G and Johnson M.(1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books

Leeuw F and Vaessen J (2009). Impact Evaluations and Development: NONIE Guidance on Impact Evaluation.

Lewis D (1973). *Counterfactuals*. Harvard University Press

Lieberman E (2005). Nested Analysis as a Mixed-Method Strategy for Comparative Research. *American Political Science Review* 99(2): 435-452.

Lincoln Y and Guba E (1985). *Naturalistic Enquiry*. Beverly Hills: Sage

Machamber PL, Darden and Craver CF (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1): 1-25.

Mackie JL (1965). Causes and Conditions. *American Philosophical Quarterly*, 2(4): 245-64

Mackie, JL (1974). *The Cement of the Universe: A Study of Causation*. Oxford University Press.

Mahoney J (2008). Towards a Unified Theory of Causality. *Comparative Political Studies* 41(5): 412-436

Marini MM and Singer B (1988). Causality in the Social Sciences. *Sociological Methodology*, 18: 347-409

Marra M (2011). Micro, Meso and Macro Dimensions of Change: A New Agenda for the Evaluation of Structural Policies. In *Evaluating the Complex: Attribution, Contribution and Beyond*. K. Forss, Marra M and Schwartz R (eds). New Brunswick NJ: Transaction.

Menzies P (1989): Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science*, 56: 642–663

Miles M and Huberman M (1994). *Qualitative Data Analysis: An expanded Sourcebook*. Thousand Oaks CA: Sage.

Mill, JS (1843). *A System of Logic*

Miller JH and Page SE (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press

Moehler D (2010). Democracy, Governance and Randomized Development Assistance. *The ANNALS of the American Academy of Political and Social Science*, 628 (1): 30-46.

Morell J (2010). *Evaluation in the Face of Uncertainty: Anticipating Surprise and Responding to the Inevitable*. New York: The Guilford Press.

Morse J, Barrett M, Mayan M, Olson K and Spiers J (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods* 1(2), 1–19.

Mosley, Paul (1987). *Foreign Aid: Its defense and reform*. University Press of Kentucky.

MRC (2008) *Developing and Evaluating Complex Interventions: new guidance.*

National Audit Office (2003) *Getting the Evidence: Using Research in Policy Making*. NAO. London

National Instititute of Clinical Excellence (2009). *Methodology Checklist: Qualitative studies.*

NONIE Subgroup 2 (2008). NONIE Impact Guidance. http://www.worldbank.org/ieg/nonie/docs/NONIE_SG2.pdf

Nutley SM, Walter I, and Davies HTO (2007). *Using Evidence: How research can inform public services.* Bristol: The Policy Press.

OECD-DAC (2002). Glossary of Key Terms in Evaluation and Results Based Management Paris: OECD. http://www.oecd.org/dataoecd/29/21/2754804.pdf.

OECD-DAC (1996). *Shaping the 21st Century: The Contribution of Development Cooperation*. Paris: OECD. http://www.oecd.org/dataoecd/23/35/2508761.pdf

OECD-DAC (2010). *Quality Standards for Development Cooperation*. DAC Guidelines and Reference Series. http://www.oecd.org/dataoecd/55/0/44798177.pdf

Page S (2006). Path Dependence, *Quarterly Journal of Political Science*, 1: 87-115.

Pawson R (2006). *Evidence-Based Policy: A Realist Perspective*. London: Sage Publications.

Pawson R (2007). *Causality for Beginners. NCRM Research Methods Festival 2008*, Leeds University. http://eprints.ncrm.ac.uk/245/.

Pawson R and Tilley N (1997). *Realistic Evaluation*. Thousand Oaks and London: Sage.

Porter J (ND). *Complexity Science and International Development*. UK Collaborative on Development Sciences Report. http://www.ukcds.org.uk/_assets/file/publications/Complexity%20and%20International%20development%20-%20%20Final%20report%20(2).pdf

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

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

Ragin CC and Becker HS (eds) (1992). What is a Case? Exploring the foundations of social inquiry. Cambridge: Cambridge University Press.

Rihoux B and Ragin C (eds) (2008), Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques. Thousand Oaks and London: Sage.

Ravallion M (2009). Should the randomistas rule? *The Economists' Voice*. 6 (2): 1–5.

Rogers EM (2003) *Diffusion of Innovations*, Fifth Edition, New York: Free Press.

Rogers P (2008) Using programme theory to evaluate complicated and complex aspects of interventions." *Evaluation* 14(1): 29-48.

Rothman, K.J. and S. Greenland (2005). Causation and Causal Inference in Epidemiology. *American Journal of Public Health*, Supplement 1, 95(S1): 144-150.

Ryan GW (ND) What Are Standards of Rigor for Qualitative Research? Rand Corporation. http://www.wjh.harvard.edu/nsfqual/Ryan%20Paper.pdf

Seale C (1999) *The Quality of Qualitative Research*. London: Sage Publications.

Shadish WR, Cook TD and Campbell DT (2002). *Experimental and Quasi Experimental Designs for Generalized Causal Inference*. Boston, New York: Houghton Mifflin Company.

Scriven M (1976). Maximizing the power of causal investigations: The modus operandi method. *Evaluation studies review annual*. V. Glass, Ed. Beverly Hills, CA, Sage Publications. 1**:** 101-118.

Spencer L, Ritchie J, Lewis J and Dillon L (2003). Quality in Qualitative Evaluation: A framework for assessing research evidence. Government Chief Social Researcher's Office. London: Cabinet Office.

Stacey R (1992). *Managing the Unknowable*. San Francisco: Josey-Bass.

Stake RE (2004). *Standards-Based and Responsive Evaluation.* Thousand Oaks, CA: Sage Publications.

Suppes P (1970). *A Probabilistic Theory of Causality*. Amsterdam: North Holland Publishing Company.

Tarrow S (2009). Bridging the Quantitative-Qualitative Divide. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards,* Brady HE and Collier D (eds). Lanham, MD: Rowman and Littlefield.

Ton G (2012). The mixing of methods: A three step process for improving rigour in impact evaluations. *Evaluation* 18 (1): 5-25

Weber M (1906) *Selections in Translation* (ed) W.G. Runciman. Cambridge University Press.

Williams B and Imam I (eds) (2007). *Systems Concepts in Evaluation: An Expert Anthology*. Point Reyes, CA: Edge Press/American Evaluation Society.

White H (2009). *Theory-Based Impact Evaluation: Principles and Practice*, Working Paper 3: International Initiative for Impact Evaluation (3ie). http://www.3ieimpact.org/admin/pdfs_papers/48.pdf.

White H (2010). A Contribution to Current Debates in Impact Evaluation. *Evaluation* 16(2): 153-164.

Woolcock M (2009). *Towards a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy*. Working Paper 73, University of Manchester: Brooks World Poverty Institute.

World Bank (1998). Assessing Aid: What Works, What Doesn't, and Why.

Yin RK (2003). *Case Study Research: Design and Methods, 3rd Edition*. Thousand Oaks CA:Sage.

Models of Causality and Causal Inference

Barbara Befani[1]

A review prepared as part of the DFID study:

'Broadening the Range of Designs and Methods for Impact Evaluation'

[1] With inputs from Nicoletta Stame, John Mayne, Rick Davies, Kim Forss and Elliot Stern

Introduction

The notion of causality has given rise to disputes among philosophers which still continue today. At the same time, attributing causation is an everyday activity of the utmost importance for humans and other species, that most of us carry out successfully in everyday life, outside the corridors of academic departments. How do we do that? And what are the philosophers arguing about? This chapter will attempt to provide some answers, by reviewing some of the notions of causality in the philosophy of science and "embedding" them into everyday activity. It will also attempt to connect these with impact evaluation practices, without embracing one causation approach in particular, but stressing strengths and weaknesses of each and outlining how they relate to one another. It will be stressed how both everyday life, social science and in particular impact evaluation have something to learn from all these approaches, each illuminating on single, separate, specific aspects of the relationship between cause and effect.

The paper is divided in three parts: the first addresses notions of causality that focus on the simultaneous presence of a single cause and the effect; alternative causes are rejected depending on whether they are observed together with effect. The basic causal unit is the single cause, and alternatives are rejected in the form of single causes. This model includes multiple causality in the form of single independent contributions to the effect. In the second part, notions of causality are addressed that focus on the simultaneous presence of multiple causes that are linked to the effect as a "block" or whole: the block can be either necessary or sufficient (or neither) for the effect, and single causes within the block can be necessary for a block to be sufficient (INUS causes). The third group discusses models of causality where simultaneous presence is not enough: in order to be defined as such, causes need to be shown to actively manipulate / generate the effect, and focus on how the effect is produced, how the change comes about. The basic unit here – rather than a single cause or a package – is the causal chain: fine-grained information is required on the process leading from an initial condition to the final effect.

The second type of causality is something in-between the first and third: it is used when there is no fine-grained knowledge on how the effect is manipulated by the cause, yet the presence or absence of a number of conditions can be still spotted along the causal process, which is thus more detailed than the bare "beginning-end" linear representation characteristic of the successionist model.

1. Simultaneous presence of cause and effect: the successionist view

This paragraph covers some of the more traditional accounts of causality – based on both regularity / invariance of patterns and counterfactual thinking. Because the basic causal unit considered here is the single cause, most quantitative, variable-oriented methods are based on this model; including regression[2], experiments (RCTs) and quasi-experiments..

1.1 Regularity

Historically, the first modern account of causality revolved around the observation of regularities: if potential cause C and effect E are always found together, then either C causes E, or E causes C. The assumption is that a true cause does not work by accident, but operates constantly and regularly, producing the same effect over time and in different settings (hence its characterization as "lawlike").

[2] Regression and analyses of correlation are based on Mill's Method of Concomitant Variation, which can be regarded as an extension of the two methods that will explored in detail here: the Method of Agreement and the Method of Difference.

While labeling this a "successionist" view of causation, Ray Pawson (2007) provides us with a visual representation: the "omnipresent arrow" inevitably leading to the effect.

$$\mathbf{C} \qquad \longrightarrow \qquad \mathbf{E}$$

What matters in regularity is the simultaneous observation of two separate entities, while the description of the causal connection (the nature of the "arrow"), or the process leading from C to E remains unknown; what happens in-between cause and effect, what the cause does in order to produce the effect, is kept closed inside what much literature has called the "black box". In Hume's words: "we can never penetrate so far into the essence and construction of bodies as to perceive the principle, on which their mutual influence depends" (Hume, Treatise II, III, I).

In the successionist view, the cause is both necessary and sufficient for the effect: sufficient because all events where the cause is observed also present the effect; "we may define a cause to be an object, followed by another [the effect], and where all objects similar to the first are followed by objects similar to the second" (Hume, Enquiry, VII, II). But also necessary in that "if the [cause] had not been, the [effect] never had existed" (Hume, Enquiry, VII, II).

We may thus redesign the arrow to account for the biunique character of the relation cause-effect:

$$\mathbf{C} \qquad \longleftrightarrow \qquad \mathbf{E}$$

In mathematical terms, we could say there is an isomorphism from the set of causes to the set of effects.

1.1.1 How causation is claimed: by agreement

Within the regularity framework, causation is claimed through observation of regular co-presence of both cause and effect. In order to claim necessity, the cause must always be present whenever the effect is. And in order to infer sufficiency, the effect must always be present whenever the cause is.

In his "Method of Agreement" (A System of Logic, 1843), John Stuart Mill provides a procedure to establish necessary and sufficient causal links by systematically comparing events on a number of characteristics. If the cause is found to be co-present / connected with two different effects (in two different events), the cause is rejected on the grounds that it is not sufficient (sometimes it produces one and sometimes the other). Similarly, if two different causes are observed together with the same effect (in two different events), then both causes are rejected on the grounds that neither is necessary for the effect. However, if a single cause is always observed together with the same effect, and all other elements can be rejected as potential causes because they are observed in only some events but not all, then it can be inferred that the relation is causal. Below two events are compared and implications are drawn.

$C \;\; E^1 \mid C \;\; E^2$ => cause is rejected because it's connected to two different effects (not sufficient)

$C^1 \; E \mid C^2 \; E$ => cause is rejected because it's connected to effect only sometimes (not necessary)

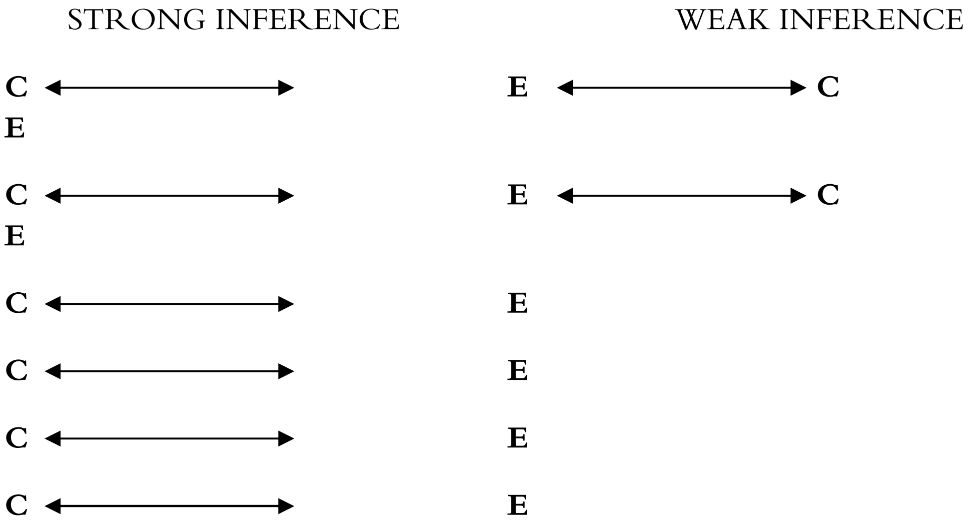
$C \;\; E \mid C \;\; E$ => causal link is established but =>

- ⇨ f g h C E | i j k C E all other elements must be present in only one of the events compared, otherwise they could be part of cause or effect

The main problem with the method of agreement lies in checking for the difference of "all possible elements": in practice this requires comparing a high number of events taking place in a wide range of settings. Statistical causal modeling can be suitable for this, being based on the assumption that, after regularity between cause and effect is found in a certain number of cases, the residual characteristics can be attributed to "chance" and thus must necessarily vary in a way that cannot be causally connected to the effect.

Because we cannot reach absolute certainty of inference, we have developed ways to evaluate the "quality" (likeliness) of the causal link: as in statistical modeling, within the more general regularity framework, the frequency of association between cause and effect strengthens the causal assumption; and theoretical relationships / associations / models that apply to a high number of cases are "more reliable" in their "explanatory power" than theories applying to a small number of cases or only one case.

In other words, **the strength of the causal association** increases with the number of cases where conjunction between cause and effect is observed; and finding cases where the cause is not simultaneously observed with the effect weakens the inference. In regularity causation, the closer we are to a "law" (see also Hempel's deductive-nomological model, Mill's Method of Concomitant Variations and Davidson's "dose-response" proportionality), the better.



In **impact evaluation**, the regularity approach is useful when the knowledge gap one wants to fill concerns the number of beneficiaries that present given characteristics, say, after an intervention. It does not provide answers on why the beneficiaries have these characteristics, nor on how they were developed following the intervention. Regularity does not allow the evaluator to trace the process leading from cause to effect, and the attribution of the impact, while shown to hold in many cases, lacks "depth" on how the causal association happens.

1.2 Counterfactuals

Although regularity and statistical significance are fundamental pillars of scientific enquiry, they only "vouch" for the causal claim and are never able to "clinch" it (Cartwright) with certainty: it's indeed impossible to claim to having considered all possible existing elements of reality in the comparison between events. In order to solve this problem, a method (and more generally, a way of reasoning) has been proposed in which the comparison is not done between highly different cases only sharing the cause and the effect but between identical cases differing only in cause and effect.
The roots of counterfactual thinking are to be found in Mill's Method of Difference (a.k.a. a double application of the Method of Agreement, see Ragin). Weber (1906) is being influenced by this model when he argues that unlike historical science, social science ought to be answering questions like "what would the course of history have been like if Bismarck had not decided to go to war in 1866": by comparing a real, factual situation (Bismarck decided to go to war) with a counterfactual one (Bismarck decided not to go to war) one should be able to imagine the difference between the consequences of Bismarck's real decision (war) with the counterfactual consequences of Bismarck's counterfactual decision (no war), in order to estimate the "impact" of Bismarck's decision on the course of history.

Counterfactual analyses share several properties with studies aiming to prove regularity:

- both focus on the simultaneous presence of a single cause with an effect, without enquiring into the nature of causal relation (the process, or the "arrow");
- both see the cause as both necessary and sufficient to produce the outcome

$$\begin{tikzcd} \mathbf{C} \arrow[r, leftrightarrow] & \mathbf{E} \end{tikzcd}$$

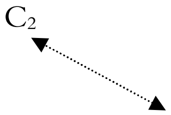
The property of sufficiency has been dropped by Lewis (1973), who argued that in order for a causal link to subsist, the following had to hold: a) when C occurs, then E occurs; b) when E occurs, then C occurs; and c) when C does not occur, then E does not occur. Lewis argued that the fourth proposition "when E does not occur, C does not occur" did not need to hold: in other words, the cause is not sufficient, because it can occur also when the effect does not occur; however it is still necessary, because whenever the effect occurs so also does the cause (but, as we will see below, this stratagem is still unable to properly address the multiplicity of causation).

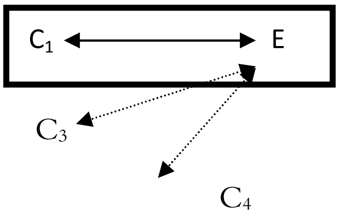
$$\begin{tikzcd} \mathbf{C} & \mathbf{E} \arrow[l] \end{tikzcd}$$

1.2.1 How causation is claimed: by difference

Apparently, counterfactuals have the advantage of needing only two events (as opposed to infinite) to infer causation; however, this advantage is only apparent because those two events need to be identical on an infinite number of elements except cause and effect.

The cause is isolated through the careful choice of the two events to be compared:





In Mill's Method of Difference, causality of C with regard to effect E is claimed in the following way:

f g h i j k | f g h i j k C E => C is the cause of E (or E the cause of C)

The above two events are compared and C is the only "new entry" in the second event: all the other elements f g h i j k l m are present in both.

When other factors are present in only one event along with C and E, we cannot infer causation; for example in:

f g h i j k | f g h i j k L C E => either L or C could be causes of E.

While f, g, h, i, j and k are rejected on the grounds that they are present in both events, L cannot yet be rejected. In order to reject L, too, we need to find the case f g h i j k L.

As with the regularity approaches, the **strength of inference** through counterfactuals increases as the number of alternative causes we are able to reject increases; the higher the number of elements that can be shown to be equal in the two events, the better.

f g h | f g h C E => **WEAK INFERENCE** (i, j, k, l, m and n haven't been rejected yet)

f g h i j k l m n | f g h i j k l m n C E => **STRONG INFERENCE** (many more causes have been eliminated)

A famous strategy used to find a specific event presenting a number of specific factors without C and E is the experiment in controlled settings (see paragraph 3.1). In **impact evaluation**, a number of strategies are used to design experiments and quasi-experiments: RCTs, statistical matching, regression discontinuity, difference-in-difference, etc. These are usually considered the strongest, most rigorous and most robust methods to attribute a given result to an intervention. However, even when the known threats to experiments are controlled for (see paragraph 3.1) and the causal association covers a high number of cases, a knowledge gap remains on the characteristics of the association between a given factor and a given effect. Like the regularity approach discussed above, counterfactuals do not provide answers as to what happened "between" the alleged cause and the effect; e.g. on what the "true" cause was at the micro, in-depth level.

1.3 Critiques to the successionist view

The main critiques that have been addressed to the successionist view concern direction of causation, the nature of the causal relation, and the interaction with other causes. Indeed, neither the regularity and counterfactual approaches – although enjoying different comparative advantages in rejecting alternative explanations – enlighten on these aspects specifically.

1.3.1 Direction of causation

With the partial exception of Lewis's stratagem and a number of sophisticated econometric techniques checking for the goodness of fit of various models, none of the approaches addressed above provides convincing ideas, methods or criteria to deal with the direction of causality. In most problems, C and E could be either causes or effects: there is a biunique relation (isomorphism) between the set of causes and the set of effects: only one effect for each cause, and only one cause for each effect. In simple (e.g. uni-variate) regression, the dependency relation is purely contingent: $y = ax + b$ could instantly become $x = y/a - b/a$. In multivariate modeling, the functional form can be changed so that y becomes one of the independent variable and one of the x-es becomes dependent[3].

Similarly, the mere comparison of two identical events differing only in one feature, although it can draw attention to the simultaneous presence of the two different states and lead observers to investigate whether one transformed (into) the other, does not provide any insight on the details nor the direction of this transformation.

In order to remedy, Hume (1739) has proposed the criterion of temporal precedence: when confronted with two events that are one the cause and the other the effect, the event preceding the other is to be identified with the cause, and the event preceded by the other with the effect:

$$\begin{tikzcd} \mathbf{C,\ t_0} \arrow[r] & \mathbf{E,\ t_1} \end{tikzcd}$$

However, temporal precedence has its limitations. First, it is not always easy to locate events temporally with a high degree of precision: Weber's argument that the Protestant Reform caused the development of capitalism requires that the Reform precedes Capitalism; however, it is not easy to say with precision when the Protestant Reform became fully institutionalized, nor when did Capitalism, and thus whether the former preceded the latter (Brady 2002). Secondly, even when events are located correctly and precisely along a time scale, the actual cause might take place in a different point of the time scale than when those events happened. For example: when women leave full-time schooling in order to get married, we don't know whether marriage is the cause or the effect of their decision to leave school, even though school leaving always precedes marriage. They could have decided to get married either before leaving school, or later, even though marriage comes always last on the time scale (Brady 2002, Marini and Singer 1988). In this case the actual cause is not "school leaving" but "decision to get married". In symbols:

$$\begin{tikzcd} \textbf{School leaving, } \mathbf{t_0} \arrow[r] & \textbf{Marriage, } \mathbf{t_1} \end{tikzcd}$$

$$\begin{tikzcd} \textbf{Decision to get married, } \mathbf{t_0} \arrow[r] & & \textbf{School leaving, } \mathbf{t_1} \\ & \textbf{Marriage, } \mathbf{t_2} & \end{tikzcd}$$

$$\begin{tikzcd} \textbf{School leaving, } \mathbf{t_0} \arrow[r] & & \textbf{Decision to get married, } \mathbf{t_1} \\ & \textbf{Marriage, } \mathbf{t_2} & \end{tikzcd}$$

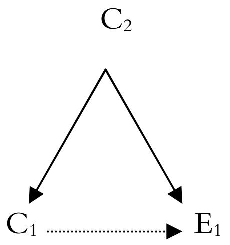
[3] A number of techniques can be used to assess the goodness of fit of the different models, providing information on direction; however this "directionality" is limited to quantitative relationships among variables and is based on the regularity with which a number of numerical contributions add up to a total effect.

In **impact evaluation**, sometimes interventions are part of a causal flow that can present different shapes, including recursive: as in when they are made possible / implemented thanks to the presence of skills or demands that the intervention itself is supposed to generate or improve. Just because the intervention happened before a result, it doesn't mean it has *caused* it: the "real" cause can be something that produced both the result *and* the intervention. Which brings us to the next paragraph.

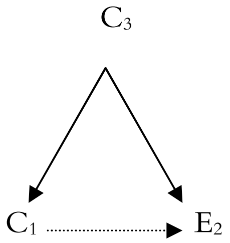
1.3.2 Correlation or causation?

Every social scientist knows that correlation is not the same as causation: and that when we say that C is correlated with E, there might be a third factor which actually causes both C and E. This problem – also known as spurious correlation – is approached in statistical modeling through the assumption of independence of variables and the isolation of the "pure", independent contribution of each single factor in the model.

This category of models solves the problem of correlation and opens that of independence, which we will deal with below. Here we will address the problem of correlation by remarking that, while counterfactuals apparently do not have this problem because all variables in the model are "held constant", in fact the action of the cause could be dependent on a particular context that both the treatment and the control cases are embedded in. Therefore, even when both terms of the comparison are perfectly equal, in fact their behavior might be linked to some contextual feature C_2 that influences the reaction of both; and affects C_1 on one hand and E_1 on the other.



What the counterfactual can thus illuminate on is the effect in a particular situation (E_1); which cannot be generalized to a context C_3 where the behavior / effect of the same cause might be different (E_2).



Cartwright notes that counterfactuals are able to tell us whether something works "here", in a particular context, at a specific time and place; but fall short when we ask whether the same works "somewhere else", at another time and place; let alone everywhere and all the time. In **impact evaluation**, even when experimental techniques like randomization ensure equivalence of the two terms of comparisons, the contextual characteristics in which both control and treatment groups are embedded in might influence their response to the intervention, thus impairing the external validity of the experiment (see paragraph 3.1).

In sum, the causal relation inferred through both regressions and counterfactuals can be spurious.

1.3.1 Can causes be independent / unconditional?

Even in situations where the above-discussed threats are controlled, the successionist view is unable to conceive of causes as behaving differently in different settings, because it postulates causal independence and the interaction of causes is usually not addressed. Causes are modeled as independent forces whose behavior is not influenced by context, experimental settings, or other causes. In other words, they are mostly **single** causes acting **independently / unconditionally**. Their power / influence is constant, does not depend on other factors nor varies according to them, and is defined in terms of coefficients of concomitant variation: that is, their causal power defines how much the value of the effect will change, on average, due to the presence of one more unit of the cause.

The coefficient is the closest the successionist approach has to a description of the causal process (the "arrow") and – being an average – is constant throughout the replications of the phenomenon: the relation between cause and effect (embodied by the coefficient) is thus not allowed to change according to the context. The context, if considered, is modeled in turn as an independent force with its own "adding-up" power to the "total". The process is conceived as being linear.

Multiple causation in the successionist view

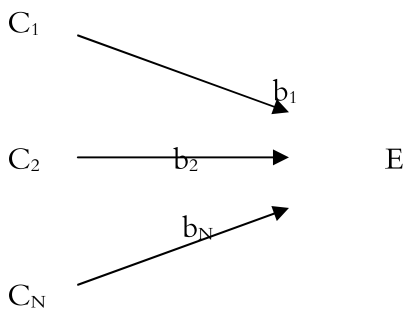
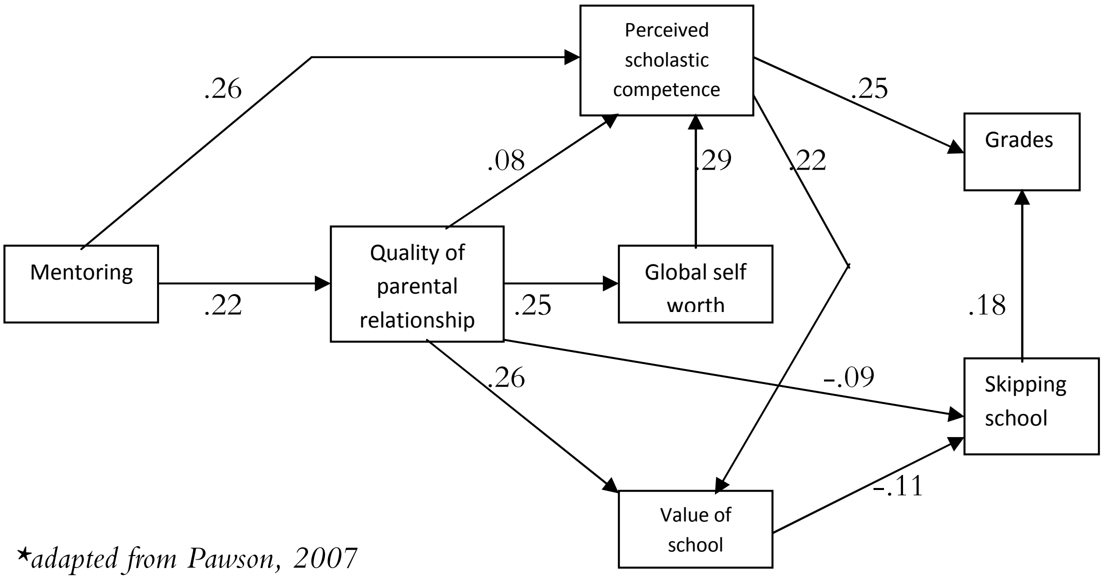


Figure 1: Path model of direct and indirect effects of mentoring*



**adapted from Pawson, 2007*

But context can in fact change the power of other causes in ways that it does not necessarily make sense to add up and make an average of: as the figures above show, C_1 ends up not being sufficient for neither E_1 nor E_2. Does this mean C_1 is a cause of neither E_1 nor E_2? The successionist view does not help in deciding whether we should still call C_1 a cause of E_1 or E_2 because it is suited to provide information on average quantities and is not tailored to perform an analysis of necessity or sufficiency of single causes or packages or multiple causes; as are the configurational methods expounded in the next paragraph.

2. Co-presence of multiple causes: necessity and sufficiency

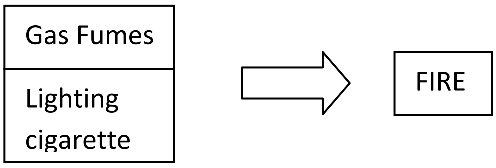
The above-mentioned Methods of Agreement and Difference can be considered logically equivalent but have different comparative advantages: when two events seem highly different, one will tend to spot their similarities; conversely, when they are highly similar, one tends to spot their differences. But this choice varies across individuals: when confronted with the same event, an individual might think of a highly similar event and spot the difference, while another might prefer to compare it with a highly-different one and spot the similarities. In a maternal health intervention in Tanzania, an expert in maternal health will tend to observe the differences between this and many other similar interventions she has knowledge of; while someone with, say, country-expertise in Tanzania will tend to spot a small number of similarities between this and the many different interventions from all sectors he has knowledge of.

But even when we compare events we want to explain with an identical event where the cause did not take place, say a similar area in Tanzania that did not receive assistance for maternal health, this comparison (or control) event might be *equal in different ways* to the treatment one. The similarities identified between the two events might differ across individuals and situations. In other words, these similarities are "chosen" according to the available evidence and knowledge; and different evidence / knowledge might be available to different individuals.

Randomistas and counterfactualists will immediately think that this is an argument to suggest controlling for a higher number of factors; but this is not the point. The point is not that the highest number possible of characteristics should be compared, but rather that in everyday life we are not interested in the average effect of causes when we attribute causality: we mostly just easily figure out what caused what, when, and depending on the situation we attribute different causes to similar effects.

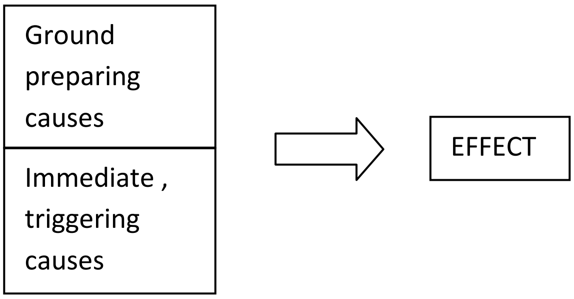
For example, consider the situation of a person lighting a cigarette at home in the presence of gas fumes accidentally leaked from the heating system and thus causing a fire. Because s/he has often lit a cigarette at home without causing a fire, s/he will tend to think of the control event "lighting a cigarette at home in the absence of gas fumes" and blame the presence of gas fumes for the accident. But if the same accident happens at a gas station, the same person will tend to think of all the other times she was at a gas station, with gas fumes around, and didn't light a cigarette: with this control event in mind, she will blame the act of lighting a cigarette for the accident.

Now, what is the real cause of the event "fire"? The immediate cause is lighting the cigarette, but it's obvious that the fire would not have happened without the gas fumes: the same person in fact blames both factors albeit separately and in different situations. The answer is that both are causes of fire, because fire would not have happened without either: so, although no cause taken singularly is sufficient, both are singularly necessary and jointly sufficient for fire to happen.



One could observe a difference between the causes, in that gas fumes are a sort of "background", ground-preparing cause, while the cigarette lighting is the immediate cause of fire. This distinction is even more evident when inferring causation for historical events. Suppose two historians need to evaluate the causal significance / impact of the assassination of the Archduke of Austria in relation to the onset of World War I. One is an expert in political instability, the other an expert in assassinations. The latter will immediately think of a number of assassinations that had negative consequences and claim that the assassination of the Archduke did indeed cause WWI. The former will think of comparable situations of political instability, all of which eventually led to a war, with or without an assassination, and claim that the assassination was not really a cause of WWI, but political instability was (Brady 2002).

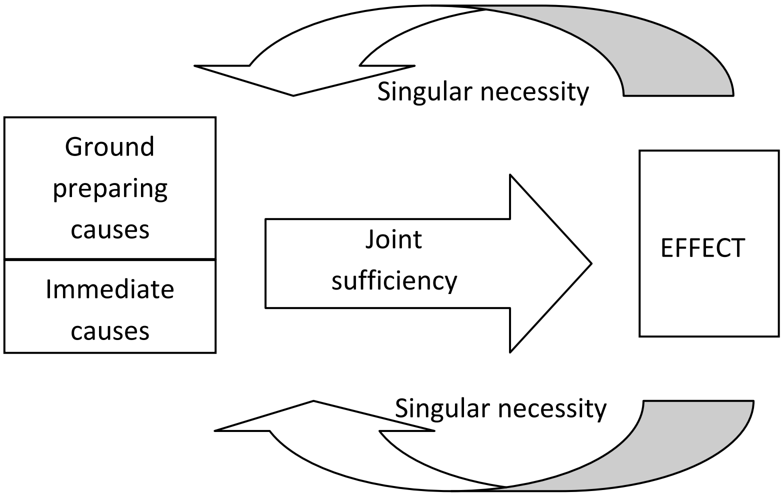
Both are right: they just focus on different types of insufficient causes. The assassination expert is focusing on the immediate cause of WWI, while the political instability expert is focusing on the "ground-preparing" causes. The assassination is akin to cigarette lighting; and political instability to gas fumes.



Historical processes, including development, can be described as the accumulation of a number of ground preparing causes, "lit" now and then by "sparks" that cause leaps and bounds, e.g.. in the development level of countries and regions. Ground preparing causes alone are not sufficient: they need a "trigger" without which they can't produce the effect; but the trigger alone is not sufficient either, and would not "perform" without the other causes.

Interventions can sometimes prepare the ground, or other times provide the trigger and achieve effects immediately. For example, immunization immediately reduces infection rates; and direct food assistance immediately reduces the number of starvation deaths. However, not all causes act so quickly: in order to acquire a good education a child must have books, good teachers, some incentives to invest time and energy in studying, and ensure that she attends school regularly. In some cases results might be poor because only 2 or 3 causes are present; but when the third or maybe fourth cause is present results might start improving – not necessarily in a linear way, but possibly in a "from no to yes" kind of way. From poor performance to excellent performance. **Impact evaluations** that attribute the effect entirely to the last cause on the grounds that it was the only one always or inevitably present in all cases risk overestimating its power, while at the same time underestimating the importance of the other, ground-preparing causes, without a combination of which the trigger would have not performed.

The ground preparing causes are thus not sufficient but are nonetheless necessary: gas fumes flying in the air do not catch fire without a spark; but when the fire happens, the presence of gas fumes (or of other combustible agents) is always detected. The immediate causes are also not sufficient but necessary: lighting a cigarette in a gas fume-free environment does not cause a fire (unless some other underlying condition is present), but when the fire happens, so is the cigarette lighting detected.



This notion of causality is relevant to **impact evaluation** because although intermediate outcomes might not be sufficient to produce a specific effect:

- They might be proven to be necessary
- The combination of intermediate effects with other supporting, ground preparing and / or sustainability factors can be shown to be *jointly sufficient* for impact.

Back to the schooling example: if evaluated with the regularity approach, the situation described by the table below would conclude that all causes on average have a similar effect on performance, ignoring the fact that – far from having an independent influence – none of them is able to have an impact without ALL the others being present.

When taken one by one with a counterfactual approach, each cause would appear miraculous, because everything else staying the same, each cause would be capable of increasing performance from low to high. But in a situation where an intervention would provide, say, only books, then only the two blue-colored cases would be compared: what would matter in this case is that "everything else stays the same", not what the other causes are. The results would be entirely attributed to books, ignoring the fundamental contribution of teachers, regular attendance and incentives.

Books	Good teachers	Incentives	Attendance	Performance
YES	YES	YES	NO	LOW
YES	YES	YES	YES	HIGH
YES	NO	YES	YES	LOW

YES	YES	YES	YES	HIGH
NO	YES	YES	YES	LOW
YES	YES	YES	YES	HIGH
YES	YES	NO	YES	LOW
YES	YES	YES	YES	HIGH

When looking at the bigger picture, one soon realizes that "everything else staying the same" is not enough informative in this case, and once the interaction of causes is disentangled and understood, it becomes clear that without many other "ground preparing causes" what the intervention provided (books) would not be able to reach the critical mass / diversity of resources needed to "climb the step" and finally impact performance. The above case is much better understandable with a non-linear, discrete / step-like, necessity-sufficiency analysis that with a successionist, linear / continuous approach to causality looking for attribution to a single cause or to estimate the independent effect of single causes.

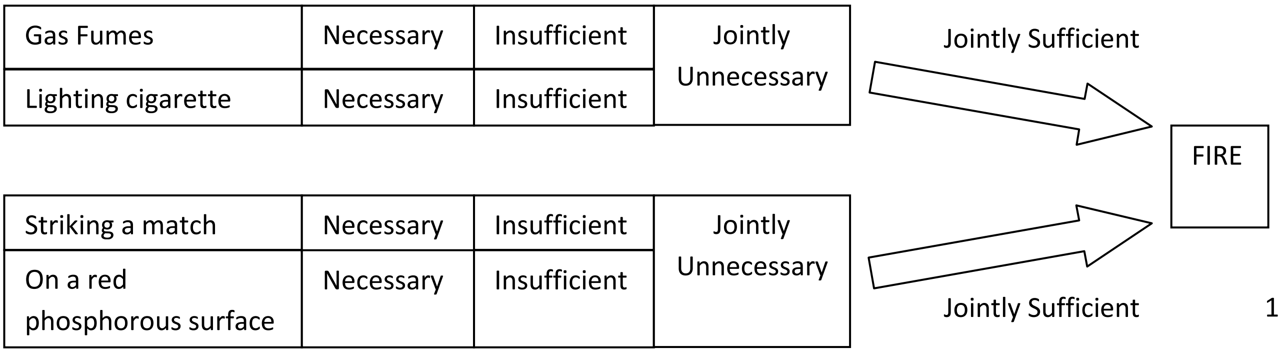
At the theoretical level, non-linear multiple causality is exemplified by a famous type of cause introduced by John Mackie (1974).

2.1 The INUS cause and the "causal field"

In an attempt to remedy the inadequacy of the successionist view in accounting for the effects arising from interaction of causal agents, John Mackie introduced in 1965 the notion of "causal field", exemplifying the idea that the link to be studied is not between the effect and a single cause, but between the effect and a causal package: a "block" of single causes that might not have an independent influence on the effect. In 1974 the same author theorizes a special type of cause called the INUS: an insufficient (I) but necessary (N) part of a causal package, which is in itself unnecessary (U) but sufficient (S).

Fire cannot only be caused by gas fumes and cigarette lighting, although they are jointly sufficient for it. It can also be caused by, for example, striking a match on a red phosphorous surface. Each of these four causes is an INUS: none of them is sufficient for fire; but each of them is necessary for a combination to be sufficient for fire. The match in itself does not light a fire, but neither so does the red surface: none of them alone are sufficient, and both of them need the other to produce fire. In other words they are jointly sufficient in that they are part of a sufficient combination. This combination, however, is not necessary for fire to happen: fire also happens when cigarette lighting devices being activated meet gas fumes.

THE INUS CAUSE



Gas Fumes	Necessary	Insufficient	Jointly Unnecessary
Lighting cigarette	Necessary	Insufficient	

Striking a match	Necessary	Insufficient	Jointly Unnecessary
On a red phosphorous surface	Necessary	Insufficient	

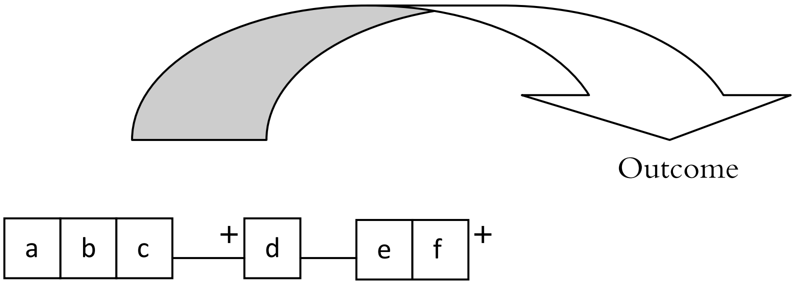
The INUS cause has been associated with contribution analysis (NONIE 2), and the idea that an intervention strategy might not be necessary to achieve a result which can be obtained in multiple ways; however, an intervention can be shown to be a necessary component of a sufficient strategy, and thus be shown to "cause" it, in combination with other factors. When considering the different paths that produce the result, it can also be argued that the one including the intervention – although not the only one / not necessary – is preferable (e.g. because it speeds up achievement or has lower costs in terms of social justice) or the only possible one in a specific context (e.g. with low economic performance, poor institutional capacity, etc.) (see Rogers in Forss, Marra and Schwartz eds., 2011)

2.2 Configurational causation

The analysis of necessity and sufficiency thus reveals that many causes come in "packages": pairs or vectors of causes, and are unable to produce the effect unless they are combined with other causes; and that the same effect could be produced by several different combinations of different single causes.

In what is called the configurational view of causation (Pawson 2007, Ragin 1987, Rihoux & Ragin 2008), a cause is thus identified with a configuration / constellation of conditions, or more precisely with a combination of single causes producing the same effect. Different combinations may lead to the same outcome; and similar combinations may lead to different outcomes, because individual conditions can affect the outcome in opposite ways, depending on what other factors they are combined with.

Figure 2: Configurational causation



2.2.1 Critiques of configurational causation: component causes of mechanisms

Configurational causation solves some problems connected with the successionist view but not all. Causality is no longer essentially single and independent, and is now seen as properly multiple and conjunctural; however the problem of establishing the direction of causation is not yet satisfactorily addressed. Moreover, the correlation problem is replaced by a different "association" problem: it is no longer a matter of observing the simultaneous presence of cause and effect, but the simultaneous presence of multiple single causes and the effect. The black box has become partially transparent and some characteristics of the causal process begin to emerge (we know more than just beginning and end, we do have some in-between elements), but the understanding is not yet fine-grained and combinations need to be interpreted. We have started to "peek in" but still haven't opened the black box. We can glimpse some characteristics of the shaft of the causal "arrow" but we likely need more details in order to establish direction and connect specific parts of the cause and specific parts of the effect in a finer way (see pairing and pre-emption below).

From the schooling example we know that in order to achieve good results in school, children must have books, attend regularly, have incentives to invest time and energy in studying, and good teachers. However, a list of causal conditions does not provide information on *how* those conditions are connected, what should come first or later, and whether there are any synergies between factors, how does one help the other, etc. A better description of the causal chain resembles an argument of the kind "books increase the chances of reading written text and thus improve reading ability" or "good teachers speed up the learning curve by adapting their method to the learning possibilities of children", "families provide incentives by to the child by collaborating with teachers", and so on. Mechanisms (see next paragraph) describe at a high level of detail how each cause or package of causes manipulates / changes the situation in order to produce the effect.

3. Manipulation and Generation

While everyone might know the recipe to a meal, just having some ingredients on the table or in the fridge does not make the meal. Someone must actually put together the ingredients in a certain way to obtain the final effect – the meal. While the configurational view of causation sheds some light on necessity and sufficiency (the recipe), this paragraph focuses on a notion of causality that informs on how the ingredients must put together; following what order and techniques. In other words, we will address the specific processes of mixing, mashing, flavouring, cooking, etc. that can make different meals out of the same ingredients depending on how someone mixes them together.

This should provide enough clarity on the problems of asymmetry and direction: a.k.a. it's not the meal that makes the ingredients, but some individual that uses the ingredients to make the meal. It does so in two ways: first by exploring the notion of manipulation and controlled experiment; and later by introducing the concept of mechanism as an actual description of the causal process taking place between cause and effect (the "arrow").

3.1 Human agency: claiming causation through intervention

In the first paragraph, causation was described as the simultaneous presence of cause and effect: the presence of two separate, detached entities at the same time in the same place. A number of authors, however, describe causality as "forced movement: […] the manipulation of objects by force; […] the use of bodily force to change something physically by direct contact in one's immediate environment" (Lakoff and Johnson, 1999). Rather than being detachedly / shyly "just" present together in the same event, causes "bring, throw, propel, lead, drag, pull, push, drive, tear, thrust, or fling the world into new circumstances" (Brady 2002). The emphasis here is on intervention and agency, and "the possibility that the failure to engage in manipulation will prevent the effect from happening" (Brady 2002).

Indeed, counterfactuals are sometimes offered as "explanatory laments" or as "sources of guidance for the future" (Brady 2002), such as "If he had not had that drink, he would not have had that terrible accident". In a situation where a person chooses a new, faster road to get home from work, and gets their car hit by a drunk driver going too fast, the person will tend to think "next time I'll take the old road on the way home", identifying the cause of the accident with her/his choice of itinerary; rather than think "there should be stricter laws on drunk driving", identifying the cause of the accident in the wider legal system. Research has shown that individuals tend to infer causal power to those events that can be manipulated: "when considering alternative possibilities, people typically consider nearby worlds in which individual agency figures prominently" (Brady 2002).

The idea of manipulation permeates the entire world of public policy interventions: by allocating resources and planning and implementing a number of activities in a certain context we hope to "change" something. And **impact evaluation** aims at understanding whether we have succeeded and provoked, *caused* some effect. The two main ways to evaluate our causal power are a) organizing experiments and b) put our actions under the microscope in order to *see* causality at work – e.g. what part / detail / component / cell / molecule of our action changed what part / detail / component / cell / molecule of the affected reality to produce the effect (biologists do both at the same time). We address the former way in this paragraph and the latter way in 3.2.

Manipulation is an important addition to counterfactual thinking in that it can solve the asymmetry problem: in laboratory and controlled-settings experiments, the scientist can manipulate the cause, activating and deactivating it at will, and collect empirical evidence on the consequences. In randomized experiments, the researcher administers the treatment to only one of the two identical groups, and collects evidence on the treatment consequences. In these cases there is no doubt on the direction of the causal arrow: however, manipulation does not protect causal inferences from other risks.

Experiments are indeed subject to three main forms of criticism: lack of external validity, limited applicability (e.g. threats to internal validity) and pre-emption.

3.1.1 Critique #1: Lack of External Validity (Accidentality)

Causation inferred through experiments can be considered "accidental" because it might be independent of any law ("possible worlds are not considered", Brady 2002) and does not provide evidence on the behavior of the cause outside of experimental settings (Cartwright 2012, Rothman 2005, Howe 2004). Its value as "guidance for the future" is doubtful when, even though one successfully avoids taking the new, faster road on the way home, perhaps more drunk drivers are around because that road happens to be full of bars and pubs, and accidents are still likely to happen. Similarly, although human agency can successfully intervene in determining that no cigarette be lit, (see example above), gas fumes in the house are still dangerous because other sparks can easily fly (gas cookers in the kitchen, heating stoves, etc.).

Generalization based on experiments stand on the (sometimes risky) assumption that the relevant contextual conditions in which the experiment takes place will remain unaltered throughout the reality targeted by generalization (e.g. basic biologic properties of the human body that do not change through individuals and interact with a given drug always in the same way). As Cartwright puts it, "experiments tell us that an intervention works here" but "we need something else" to extrapolate that finding and deduct that the same intervention will work "somewhere else".

3.1.2 Critique #2: Threats to Internal Validity

Although RCTs are good at removing selection bias through randomization, they are still exposed to post-selection differential change: after selection, change might take place in a differentiated way in the two groups. In particular (Scriven 2008):

- Groups might be differentially influenced by their awareness of being part of the treatment or the control group (a.k.a. the Hawthorne effect);
- Subjects in the treatment group might leave the experiment for reasons related to the treatment, which are not the same reasons that subjects in the control group would leave the group for (a.k.a. differential attrition / mortality);
- The two groups might interact and influence each other in an asymmetrical way: for example subjects in the control group might start imitating subjects in the treatment group (a.k.a. diffusion effect).

3.1.3 Critique #3: Pre-emption

While successfully solving the direction problem, manipulation (or full-on experimentation) does not account for those causes that could have acted, but where prevented from acting by the cause that actually operated. The following example is often presented in the philosophical literature (Brady 2002): a man walks across a desert with two enemies chasing him. The first enemy puts a hole in his water bottle. The other enemy, not knowing about the first, puts poison in his water. Manipulations have certainly occurred, and the man dies on the trip. Now, the first enemy thinks he caused the man's death by putting the hole in the bottle. Similarly, the second enemy thinks he himself caused the death, by putting poison in the water. In reality, the water dripping out of the can might have prevented the cause "poison" from acting: and the man might have died of thirst rather than from poisoning. So the counterfactual "if the second enemy had not put poison in the water, the man would have survived" doesn't hold, because the man would have died anyway of thirst.

Let us assume the man died of thirst. The closest possible world to the one that has actually happened is one where poison would have killed the man anyway: so a well-constructed experiment on the impact of a hole in the bottle on the man's health would conclude that the hole did not kill the man, because *ceteris paribus* the "control" man also died. But this is incorrect because the "treatment" man did in fact die from the scarcity of water caused by the leak, so the hole did have an impact, and is certainly able to kill another man in the same situation.

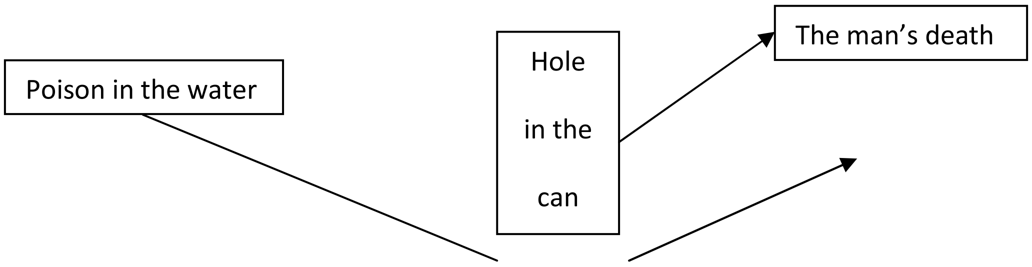
In policy interventions, targeted groups sometimes have different strategies / opportunities and before acting they wait to see whether they have been selected for a certain treatment. If they are not selected, they will act in a way that compensates for non selection, and perhaps might achieve similar results to the selected. In this case an **impact evaluation** that compares treatment and control groups might conclude that the intervention had no net effect or was not "necessary" for the result; but it cannot conclude that it did not have some causal power in producing the effect, because the treatment group achieved the result through the intervention, benefiting from the resources provided by it and not by choosing and alternative strategy like the non-selected / control groups. It's not possible to say that the intervention did not cause / produce the outcome, but it's possible to say that while exerting its causal power, it prevented another potentially effective cause from operating.

Similarly, when an intervention is implemented in contexts with different previous and current interventions interacting with it, it might seem to work differently, e.g. only in some contexts, because its operation might have been "displaced" (prevented) by other interventions that acted first.

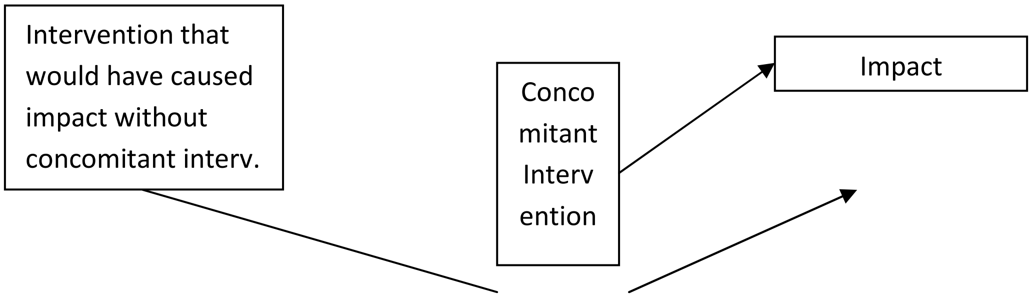
Similar concepts are the "crowding out effect" in economic and social science theory (see for example Elster 1998), and path dependence (Page 2006).

Metaphorically, when pre-emption operates the causal arrow is "broken" and displaced by another causal arrow:

Pre-emption (crowding-out, path dependence): the man in the desert



Pre-emption (crowding-out, path dependence): impact evaluation


Intervention that would have caused impact without concomitant interv.
Conco mitant Interv ention
Impact

In the above case it might appear easy / attractive to dismiss the intervention as useless or irrelevant. When administrations are interested in the pure effect (e.g. whether the man dies or not, or whether an intervention worked in a specific time and place or not, without knowing why), experiments might rank the intervention low or zero on impact score; however, an intervention might not work in a specific setting where other interventions were taking place, but work well in other, less-crowded settings. This does not mean that the intervention doesn't have the causal power to produce the effect; it just means that it needs specific conditions to do it. Discovering how the intervention produces the effect might explain why it works better in some cases than others.

3.2 Generative causation: the description of the causal mechanism

Pre-emption is part of a more general problem that all approaches to causal thinking expounded so far are unable to address: the pairing problem. In an example above, although school leaving is consistently found to precede marriage, it cannot be considered a cause of marriage: it's difficult to argue that people get married because they suddenly find themselves out of school. At the same time, no one gets married without having previously decided to marry: so the decision to get married can be considered the "real cause"; and a closer examination reveals that the decision can be taken either before or after leaving school.

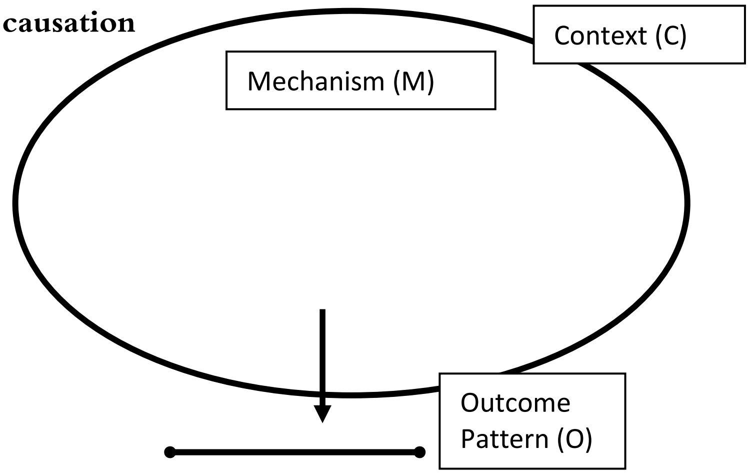
In many cases, even if constant conjunction between cause and effect is proved; even if the cause is articulated into a configuration; even if direction is established and even if the ineluctability of the effect is demonstrated, we still don't have enough information to properly "pair-up" cause and effect. Even if we know the man in the desert would have died anyway, we still can't attribute that particular death to a particular cause (leak or poison). We might know that an intervention clearly works, but we don't know what specific parts of it actually do the trick and what parts are irrelevant or less important. In other words, we know that a generic walk in the desert caused the death but we still don't know what specific part of that walk (or that cause) exactly caused the effect: and in order to obtain this kind of knowledge we need an approach to causality that allows for a detailed description of the causal relation / chain / process / arrow that generates the effect: e.g. that explains how the effect actually comes / is brought about.

An autopsy, e.g. a literal digging into the "inner mechanism" that brought the man to his death, helps reject the poisoning explanation and declare that the man died of thirst. A close examination of the water bottle might also rule that the water would have run out of the tank before the poison would properly diffuse. A closer examination of an intervention that is proven to have an effect on a variable might shed light on what components, what activities were actually essential and why and what were not.

This approach is important depending on the use we want to make of **impact evaluation**. If we need it to justify public spending, or for accountability purposes, then a simple link (or an unopened black box) might suffice. But if we need it to improve response, or to extend it to other areas, then we need to know the details of the actual working gears: how can we improve the mechanism or adapt it to / make it work in other areas? What gears are actually important / relevant? What supporting factors are needed in order for the gears to work? Will these supporting factors be present in other areas?

Pawson provides an account of the generative approach as the one enquiring about the "how": how the causal association comes about. "[Within the generative causality framework] the causal explanation provides an account of why the regularity turns out as it does. The causal explanation, in other words, is not a matter of one element (X), or a combination of elements (X1.X2) asserting influence on another (Y), rather it is the association as a whole that is explained. Accordingly, Figure 3 removes the ubiquitous causal arrow and replaces it with the dumbbell shape representing the tie or link between a set of variables. It becomes the causal arrow. And what it explains is the uniformity under investigation. It explains how the outcome pattern is generated. It explains the overall shape of the outcome pattern and thus the causal arrow penetrates, as it where, to its core." (Pawson 2007)

Figure 3: Generative causation

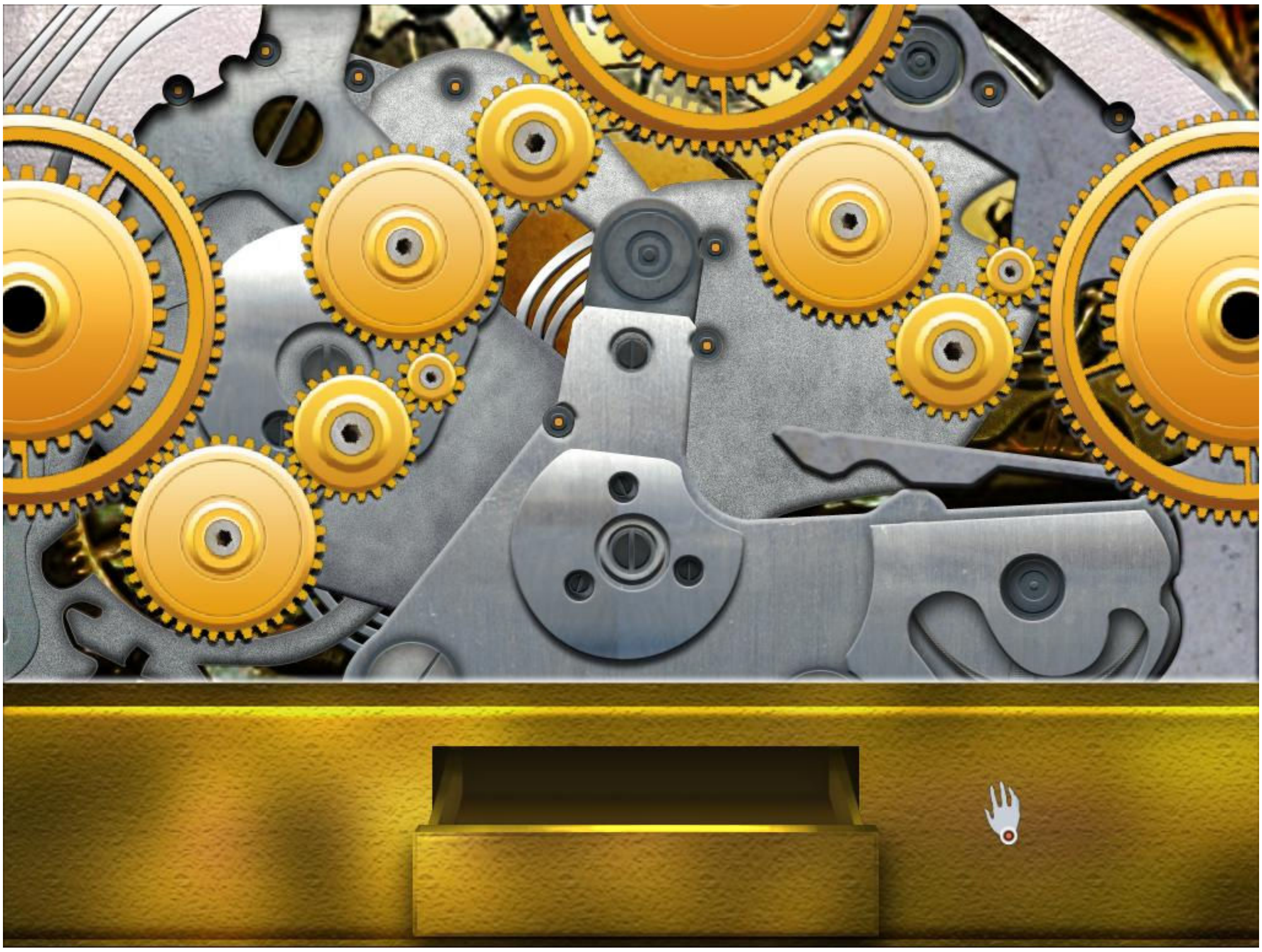


Other theorists claim that mechanisms are "entities and activities organized such that they are productive of regular changes" (Machamber, Darden and Craver 2000) and "causal chains" that provide a more fine-grained and more satisfactory explanation than black box regularity (Elster 1998).

3.2.1 How causation is claimed: digging deep

Glennan (1996) stresses the regularity property of mechanisms, arguing that "two events are causally connected when and only when there is a mechanism connecting them" and "the necessity that distinguishes connections from accidental conjunctions is to be understood as deriving from an underlying mechanism". At the same time, other authors write that mechanisms "misbehave" (Pawson 2007): we don't always know which one will be activated when; even though when they are, we can recognize it after the fact (Elster 1998.

These apparently contrasting accounts can be reconciled by the fact that mechanisms are "meso" entities that come in different "sizes" and belong to different levels and layers: micro-mechanisms are small parts of higher-level, "bigger" mechanisms (macro-mechanisms or systems). The law-like regularity mentioned by Glennan thus refers to higher-level mechanisms, which are often described in terms of and represented by a causal chain (or intersections of causal chains): or an assemblage of lower-level mechanisms, that have roles in different chains, and might play different roles in each, like gears of different shapes and sizes in a clock or in a manufacturing machine. This is why micro-mechanisms can contribute to different outcomes, depending on the chain they are operating in and the place they occupy in the chain. But at the same time, the whole chain – or an even bigger group of intersections of various causal chains – produces the result in a more "law-like" than accidental way (e.g. long pathways and complex systems).

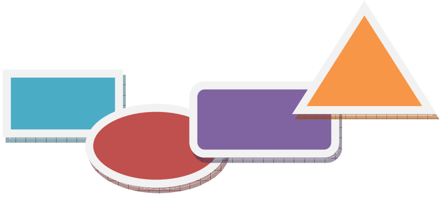


Cartwright claims that, like experiments, causal chains “clinch” conclusions: e.g. they are proper laws. Although lawlike statements are rarer in the social sciences than in the natural sciences, a fairly articulate and comprehensive (social science) causal chain is able to reject an enormous amount of alternative explanations for an effect, many more than a successionist inference. In fact, while successionism eliminates single causes one by one, a fine-grained explanation containing the same number of single causes rejects a higher number of alternative explanations, because chains with single causes variously combined are different entities even when they include the same causes. Order of conditions and relations between conditions also matter. Single conditions are not linked directly and independently to the effect: rather the whole assemblage is. Therefore, in the case of generativism, by “rival explanations” we do not mean rival single causes or rival combinations, but rather “rival chains” (Campbell in Yin 2003).

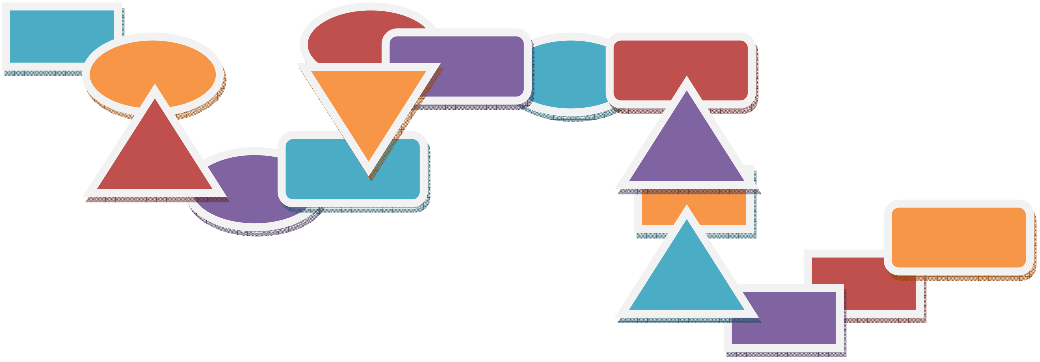
3.2.2 Quality of inference

How fine-grained does the explanation need to be? When is it fine-grained enough? The fit-to-purpose criterion holds here. Even in hard sciences, there is no model or theory depicting absolute truth – theories and paradigms hold until new, better theories and paradigms that unlock a wider, better range of applications are discovered. Usually, the more fine-grained the explanation is, the more in-depth the understanding of the micro-mechanisms that are present in different parts of the chain/system, including their relations; and the greater the possibilities of generalizing the findings to situations presenting similar conditions.

WEAK INFERENCE



STRONG INFERENCE



Determining the causal contribution of several elements helps bring into focus the contribution of the intervention – by subtraction. The more we know about how many factors influence the outcome, the less we have left to learn about "all other possible elements / alternative explanations". In a reality with a finite number of elements / explanations, considering an ever increasing number of elements eventually leads to rejecting all possible alternative explanations, and thus leads to certainty of inference according to the method of agreement.

3.2.3 Mechanisms have parts: component causes and complete mechanisms

Because they must provide an account of how causality works, mechanisms are often complex and / or complicated objects with several parts. Agent-based modeling illustrates the complexity of macro mechanisms emerging from a high number of micro-mechanisms being activated at the agents' level (Gilbert and Troitzsch 2005, Miller and Page 2007). Rothman and Greenland (2005) avoid the micro-macro distinction and call micro-mechanisms "single component causes". Each component cause is a necessary part of a complete causal mechanism that is sufficient for the effect. The same effect can be achieved also by other causal mechanisms, which may or may not have some component causes in common.

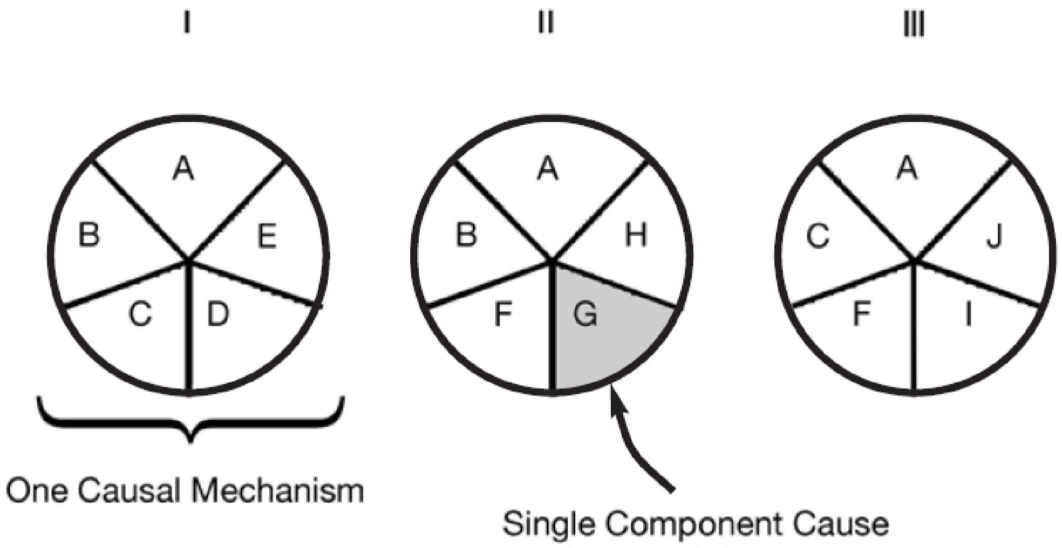


FIGURE 1—Three sufficient causes of disease.

The above – with its considerations of necessity and sufficiency ("completeness") – is reminiscent of configurational causation: the difference is that in this case the linkages among the conditions that constitute the combinations are expressly described and understood, and configurations are not allowed to be mere sets of simultaneously present causes.

Concluding remarks

This review has attempted to highlight the strengths and weaknesses of different models of causality and causal inference. In theory, the most refined way to understand causality is the causal mechanism. The problem with the causal mechanism is that it might be difficult to untangle in all its intricacies, particularly in a complex systems framework. While we try to identify / describe all the steps in a causal chain, or describe the functioning of a micro-macro interaction, agent-based mechanism producing emergent macro outcomes, regularities and counterfactuals (e.g. Mill's methods) help ascertaining relevant correlations / associations that, although rough, might signal mechanistic causal relations.

Configurational methods refine the association between cause and effect by "peeking in" in the black box and spotting the presence / absence of a number of elements / conditions, preparing the ground for mechanism identification and theory development. This is an iterative process, with configurations in turn displaying some characteristics of mechanisms and serving as instruments for generalizing / testing theory across cases with limited diversity. Successionism is then eventually able to generalize / test the theory across a high number of similar cases, when theory is enough advanced to allow for the reduction of cases to sets of independent variables.

In conclusion, lessons can be learned from all models of causation and causal inference, and might be summarized in 3 criteria that a causal claim / link needs to meet. Whenever possible, a causal claim / link needs to:

1. Include observation of simultaneous presence of cause and effect. Simultaneous presence might mean:
    a. constant conjunction through several cases / events, while other elements / potential causes vary;
    b. double observation of conjunction cause-effect and conjunction no cause-no effect in two identical situations differing only in cause and effect.
2. Include an analysis of necessity and sufficiency of the cause with regard to the effect:
    a. under what conditions / in what combinations with other causes the effect is produced / unlocked / triggered
    b. When the cause is absent, what other causes / combinations of causes produce the effect
3. Include a detailed, theoretically-informed and theoretically-consistent description of the causal relation / force / chain / "arrow", or how the cause produces the effect.

The following table summarizes several of the arguments presented in this paper.

	Mere Co-Presence			**Active Causality**	
	Of independent causes and effect		**Of causal packages**	**Accidental**	**Regular**
	Regularity	**Counterfactuals**	**Configurations**	**Manipulation**	**Mechanisms**
Major problem solved	Lawlike generalization	Single-cause validity	necessity / sufficiency of packages and single causes within packages	Direction	Pre-emption and pairing
Inference Design	Agreement	Difference	Systematic Comparison	Experiments	In-depth examination (microscope)
Inference Method	Regression, statistical modeling, observational studies	Natural EXPs, Quasi-EXPs w/ control, observational studies	Qualitative Comparative Analysis (QCA)	RCTs, Laboratory experiments in controlled settings	Discovery, construction and refinement of substantive theory
The causal process or "arrow"	Simultaneous presence of "ends"	Simultaneous presence of "ends"	Conditions included between ends	What is the head / nock (direction)	Description of shaft (including head and nock)

DEPARTMENT FOR INTERNATIONAL DEVELOPMENT

DFID, the Department for International Development: leading the UK government's fight against world poverty.

Since its creation, DFID has helped more than 250 million people lift themselves from poverty and helped 40 million more children to go to primary school. But there is still much to do.

1.4 billion people still live on less than $1.25 a day. Problems faced by poor countries affect all of us. Britain's fastest growing export markets are in poor countries. Weak government and social exclusion can cause conflict, threatening peace and security around the world. All countries of the world face dangerous climate change together.

DFID works with national and international partners to eliminate global poverty and its causes, as part of the UN 'Millennium Development Goals'. DFID also responds to overseas emergencies.

DFID works from two UK headquarters in London and East Kilbride, and through its network of offices throughout the world.

From 2013 the UK will dedicate 0.7 per cent of our national income to development assistance.

Find us at:
DFID,
1 Palace Street
London SW1E 5HE

And at:
DFID
Abercrombie House
Eaglesham Road
East Kilbride
Glasgow G75 8EA

Tel: +44 (0) 20 7023 0000
Fax: +44 (0) 20 7023 0016
Website: www.dfid.gov.uk
E-mail: enquiry@dfid.gov.uk
Public Enquiry Point: 0845 300 4100
If calling from abroad: +44 1355 84 3132