bond
FOR INTERNATIONAL DEVELOPMENT

Impact Evaluation

A Guide for Commissioners and Managers

Prepared by Elliot Stern for the Big Lottery Fund, Bond, Comic Relief and the Department for International Development, May 2015

Acknowledgements

I would like to acknowledge the team that prepared the initial report for the UK Department for International Development (DFID) on which this guide builds. This team included Nicoletta Stame, Kim Forss, Rick Davies, Barbara Befani and John Mayne as well as the author of this guide. Two members of this original team (Barbara Befani and John Mayne) have also acted as 'sounding boards' during the preparation of this guide.

The impetus for this guide came from a 'cross-funders group' interested in helping decision-makers within civil society organisations and those that fund them to better understand how to commission, manage and use impact evaluations. Members of this cross-funders group drawn from Bond, Comic Relief, the Big Lottery Fund and DFID have offered helpful advice and support throughout drafting.

As part of the drafting process a workshop was convened by Bond that brought together those involved in over 20 CSOs together with funders to explore the IE context as they understood it and to suggest issues that the guide should cover. Those who participated in this workshop also contributed examples of IE related material that has further informed this guide.

Elliot Stern
May 2015

Big Lottery Fund
Big Lottery Fund (BIG) is one of the largest grant-making organisations in the UK and is responsible for distributing 40 per cent of all funds raised for good causes by the National Lottery. BIG distributes funds to both UK and International charities, voluntary and community sector organisations.

Bond
Bond is the UK membership body for organisations working in international development. We work to influence governments and policy-makers, develop the skills of people in the sector, build organisational capacity and effectiveness, and provide opportunities to exchange information, knowledge and expertise.

Comic Relief
Comic Relief is a major grant-making charity based in the UK which gives grants to both UK and International charities, with the aim of bringing an end to global poverty.

Department for International Development
The Department for International Development (DFID) is the ministerial department leading the UK's work to end extreme poverty.

Contents

1. Introduction and scope 2

2. What is impact evaluation? 4

Defining impact and impact evaluation 4

Linking cause and effect 5

Explanation and the role of 'theory' 7

Who defines impact? 7

Impact evaluation and other evaluation approaches 8

Main messages 9

3. Frameworks for designing impact evaluation 10

Designs that support causal claims 10

The design triangle 11

Evaluation questions 11

Evaluation designs 13

Programme attributes 14

Main messages 15

4. What different designs and methods can do 16

Causal inference: linking cause and effect 16

Main types of impact evaluation design 20

The contemporary importance of the 'contributory' cause 21

Revisiting the 'design triangle' 21

Main messages 23

5. Using this guide 24

Drawing up terms of reference and assessing proposals for impact evaluations 25

Assessing proposals 25

Quality of reports and findings 27

Strengths of conclusions and recommendations 28

Using findings from impact evaluations 29

Main messages 29

Annex 30

1. Introduction and scope

All those who are involved in practical development work whether nationally or internationally face demands for 'impact evaluation'. Funders, stakeholders and the public at large want to know that funds are used to good effect: that they achieve results and improve the lives of people and their communities.

Impact evaluation (IE) seeks to demonstrate that intended results follow from programme activities whether directly or indirectly. Whilst evaluation of development programmes is nothing new, the focus on impact has been given greater urgency by resource constraints and political demands for more accountability and transparency. These demands come not only from funders but also from those affected by development programmes – often the most poor and marginalised – who want to know that greater resources, rights and services will genuinely follow from their engagement with development actors.

Against this background, various approaches to IE are advocated – many accompanied by claims by experts that theirs is the best or only way. One of the problems

faced by those who need to decide how to approach demands for IE, is that it is often presented as a technical or methodological question only accessible to experts or researchers. To some extent this is true but the main arguments, logics and choice-points are more accessible. This is because the choice of IE designs should be based not on advocacy for particular methods but on practical considerations that face those who commission, manage and fund development programmes. These policy-makers and managers need to decide what they hope to get out of an evaluation, how this relates to the kinds of programmes or initiatives they are involved with, and what are the realistic capabilities of designs and methods on offer. This is the starting point of this guide, the purpose of which is to support managers and commissioners of impact evaluations to better manage the entire process from drawing up terms of reference, selecting contractors, steering evaluations and utilising evaluation results. The guide also argues that relying only on traditional approaches to IE does not fit well with the kind of customised, complex, locally engaged and often sensitive programmes that non-governmental organisations (NGOs) and civil society organisations (CSOs) undertake. A broader range of designs and methods are needed.

This 'design guide', as the title suggests, starts from the assumption that:

- Evaluation design is a vital stage in the overall impact evaluation process. If neglected, it will have negative consequences down the line in terms of the relevance, validity and usability of evaluation outputs.
- It is important for those who commission, manage and use Impact Evaluations to have access to frameworks and guidance. These allow them to ask the right questions of the specialist evaluators who will in the end do the IE work that is needed.

The audience for this guide are those who:

- Draw up IE terms of reference
- Have to assess IE proposals that cross their desks
- Manage and steer ongoing IEs
- Wish to assess the strength of conclusions and recommendations reached by those conducting IEs
- Need to develop new programmes and policies that are 'evidence-based', ie, learn lessons from completed IEs

In depth evaluation and methodological expertise is not assumed in this guide – rather readers are expected to have familiarity with evaluation issues and challenges; and with the demands of socio-economic development programmes. The guide signposts more specialist sources and references, but is mainly interested in equipping practical managers in the development sector with enough knowledge to allow them to have meaningful conversations with technical experts.

This guide builds on a major report funded by the Department for International Development that was published in 2012: *Broadening the Range of Designs and Methods for Impact Evaluations*. That report, which including annexes exceeded 120 pages, was intentionally more technical and more geared to evaluation specialists rather than managers and practitioners. The 2012 report provides an additional point of reference for those wishing to further deepen their understanding of IE [1]. Some readers of this guide will undoubtedly wish to cross-refer to sections of the earlier report to pursue some issues in greater depth and this is signposted in the text.

1 See: http://r4d.dfid.gov.uk/Output/189575/

2. What is impact evaluation?

This chapter aims to help readers identify what is distinctive about impact evaluation. It sets IE into the wider setting of 'evidence-based policy'; introduces some of the important methods-related debates that surround IE including the position of experimental methods and the role of theory in support of explanation. The chapter concludes by arguing that Impact Evaluation is not completely separate from other kinds of evaluation. IE is only one part of a bigger picture, and in development settings in particular, has to draw on various evaluation traditions in order to do its job well.

Defining impact and impact evaluation

There are two main ways in which 'impact' and its evaluation has been defined. The first focuses on content and the second on methods. The best known example of a content definition of 'impact' in the international development field can be found in the OECD/DAC lexicon: "…positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended."

This definition:

- Stresses the search for **any** effect, not only those that are intended
- recognises that effects may be positive and negative
- recognises 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 effects
- focuses on the longer-term effects of development interventions

Methodological definitions tend to be focussed, more narrowly. The World Bank poverty/net website defines Impact Evaluation in terms of attribution: "…assessing changes in the well-being of individuals, households, communities or firms that can be attributed to a particular project, programme or policy."

Howard White of 3ie, an institution specialising in IE, defines it explicitly within an experimental and counterfactual logic: "…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)

Comparing the content and methods ways of defining IE illustrates why IE thinking has moved away from sole dependence on experiments. Experimental methods are concerned with **intended** rather than unintended effects; assume **direct** links between interventions and outcomes; address **primary** rather than secondary effects; and usually look to evidence in the **short-term** rather than the long-term. This latter is especially important as in many development settings effects are not known when programme funding ends, only becoming clear over a much more extended timescale. Most counterfactual methods on the other hand focus on the short-term, which is likely to capture only a sub-set of programme results.

However, criticism can equally be made of any other method or family of methods – all do some things better than others. (See chapter 4 for a fuller discussion of the strengths and weaknesses of different designs and methods.) The key message is that we need to start with what we want to know about programmes rather than a particular tool-kit. What we want to know is what 'caused' the effects of development programmes through the best methods available.

Linking cause and effect

Discussions in the evaluation community about methods, counterfactuals and 'quality' have helped refocus evaluators' attention on causal analysis. Simply put, answering the question, 'Did this programme make a difference or would changes have occurred anyhow?' matters. It has been argued that some evaluators and commissioners of evaluation have paid insufficient attention to what are variously called 'impacts', 'results', and 'effects' even though this is a question that various stakeholders quite reasonably want answers to.

IE grew out of what became known as the 'evidence-based policy movement' (EBPM). This movement emphasises that policy should be evidence-based and able to demonstrate and where possible measure 'results', 'value for money' and 'effectiveness'. IE became an important means to provide the evidence that policy makers required to show that their policies 'worked'.

EBPM itself was built on foundations in 'evidence-based medicine' with a long history of pharmaceutical trials using experimental methods, mainly randomised control trials, to demonstrate effective treatments. These trials set out to identify causal patterns, to 'attribute' particular health outcomes to particular therapeutic interventions. In the early days of EBPM, studies and evaluations conformed to a similar methodological template and there is still a tendency for some to identify IE with experimental methods.

However, as EBPM has matured, it has been increasingly apparent that no one methodological tool-kit can appropriately evaluate all kinds of policies and programmes [2]. For example, these may:

- Be inherently difficult to measure – cultural changes around equality and human rights, greater empowerment and participation in governance or strengthening civil society are all socially constructed and have qualitative as much as quantitative outcomes.
- Have causal pathways – what evaluators call 'theories of change' – that lead from programme to outcome that are often complex, little understood and hard to unravel, making them unsuited to analysis through the experimental manipulation of single causal factors.
- Be relatively small scale and not provide the numbers of cases needed for statistical analysis. This is made even more difficult when development programmes are quite sensibly 'tailored' to take account of their very different contexts, depriving evaluators of a standard intervention to compare, control for or measure.

Of equal importance is that policy makers have shifted from a largely 'accountability' purpose of evaluation to one that also prioritises learning. They have therefore become interested in understanding **why** and how programmes succeed or fail, as well as **whether** they succeed or fail, in order to improve current programmes and replicate them with confidence in the future. Explanatory questions of the **why** and **how** variety, have been important drivers of the diversification of methods used in IE.

Despite this diversification IE has retained a cause and effect focus throughout its evolution: IE tries to link policy causes with policy results. It may no longer do this by looking for single causal factors to which effects can be 'attributed,' but the enduring and distinctive characteristic of IE, is that it tries to find out whether a policy or programme as a cause can be linked to identifiable and intended effects.

However, it also needs to be remembered that not all evaluations place the same emphasis on cause/effect relations. Evaluations that are purely accountability-driven whilst intended to demonstrate and measure results do not have to be centred on the links between cause and effects. An indication or association between effects and programmes as probable causes will often be sufficient. There are many good examples of these kinds of 'indicative' IEs which although they do not demand such stringent designs as are advocated in this guide, are nonetheless of great value when what we want to know is whether the balance of evidence suggests that a programme is having an effect [3].

2 The implications of programme characteristics for IE methods is discussed below in chapter 3.

3 UNICEF, (2011) Inter-Agency Guide to the Evaluation of Psychosocial Programming in Emergencies. New York: United Nations Children's Fund. http://www.unicef.org/protection/files/Inter-AgencyGuidePSS.pdf
The One Love Campaign in South Africa: What has been achieved so far? http://www.comminit.com/hiv-aids/content/onelove-campaign-south-africa-what-has-been-achieved-so-far

Explanation and the role of 'theory'

We have noted that nowadays IE is concerned both to demonstrate and measure effects and as often as not also to explain – and to answer 'how' and 'why' questions. This raises an important distinction in evaluation and in IE in particular: that between causality and explanation. You might draw a conclusion (or causal inference) from an evaluation that funding for education programmes for girls led to or 'caused' higher family income in a particular community. However, when it becomes evident that similar educational programmes do not always lead to the same result in all places, people start to ask 'why'?

Although explanation is not always a priority in IE, it often is. This is why 'theory' has become part of the evaluators' dictionary which it was not when evaluators were only expected to judge the success and failure of policies. In IE, as in scientific research, explanation ultimately relies on good theories. Opening up the 'black box' that connects 'causes' and 'effects' requires different kinds of analysis, which is what 'theories of change' and 'programme theory' (discussed further in Chapters 3 and 4) are intended to support. Developments in IE have also made evaluators aware that they need to draw on broader community, social and economic theories in order to interpret complex and often confusing or even contradictory data.

Who defines impact?

Various words in evaluation have similar meaning to impact. Most commonly evaluators talk of results, outcomes and effects fairly interchangeably with impacts. As the above discussion suggests, impacts can be direct or indirect, short or long term, primary and secondary, positive or negative. All of this underlines that defining impact is an important first step in most IEs and that putting together such a definition can be quite difficult.

One difficulty is that different evaluation actors and stakeholders may view impact quite differently. An impact may be:

- The effect as intended by policy makers and programme planners or as experienced by intended beneficiaries and others
- An immediate experience or a more enduring change in circumstances or capacities
- At the level of individuals or communities or institutions

The language of 'impacts' implies a passive 'voice' for beneficiaries or so-called target populations: outsiders administer treatments to those who face problems which are themselves often defined by outsiders. Whilst this language may have been reasonable in the settings where IE first appeared in international development (eg large scale international programmes addressing immunisation needs) it is less appropriate in the kinds of community and local settings in which community-based organisations (CBOs), CSOs and NGOs operate. In these settings programmes are often jointly planned or at the very least there are strong participatory inputs at the planning and implementation stage. Furthermore there is an expectation of continuity and sustainability, which itself assumes that the results of programmes, if they are truly to have impact, have to be owned by those they are intended to benefit.

The bias throughout this guide is therefore to assume that stakeholders in general and those directly affected by programmes more particularly should have a strong voice when defining what constitutes impact. It is argued that those affected by programmes should have a privileged voice in formulating and defining impacts; and that stakeholders continue to have a central role in feeding into and validating how data is interpreted, conclusions are reached and recommendations are framed.

Impact evaluation and other evaluation approaches

There has been a tendency for those who are interested in IE to present their work as quite separate from other evaluation approaches. This is dangerous because most evaluations including IE face very similar problems, such as:

- **Being clear about what is being evaluated.** Measures and indicators have to be true representations of the 'object' of evaluation. What is often called 'construct validity' relies on understanding the world-view and experience of programme participants and stakeholders. It necessarily draws on participatory evaluation approaches, sometimes seen as the antithesis of many currently used approaches to IE.
- **Ensuring that programmes are implemented with impact in mind.** This is partly about assessing whether programmes are getting through to those for whom they are intended. In development settings this can be critical with marginalised or hard-to-reach groups. This highlights the importance of process evaluations alongside impact evaluations if we are to distinguish between 'programme' and 'implementation' failure.
- **Addressing the normative and ethical problems that development policies always raise.** These range from ensuring that policies do no harm through to ensuring that those who benefit are those who are most in need. Many development programmes are value based – supporting the very poor, promoting women's rights, supporting inclusive governance. Any evaluation including IE which ultimately helps stakeholders make judgements about 'value', always has to consider the underlying values that inform judgements about success and what counts as 'good' development.
- **Distinguishing the 'programme theory' of policy makers and the 'theories of change' of how the programme works in practice.** Such theories provide a set of hypotheses against which the reality of programmes can be tested: this was supposed to happen: did it?

Theory based approaches also focus attention on different contexts, an essential requirement for generalisability or 'external validity'. This guide advocates the use of 'theory-based' approaches as one useful approach to IE, but this is a more general point. Even if one was following a counterfactual approach to IE, theory is essential for generalisation beyond a particular programme evaluation.

- **Knowing whether a programme builds or risks undermining capacity.** This is a common source of unintended, negative programme consequences. It is not unusual for major development programmes to ignore pre-existing resources, networks and capacities – which are essential for sustainability and indeed for accessing good data and monitoring progress on the ground. More dangerously, ignoring existing capacities makes it possible that they could be damaged when new programmes are introduced.
- **Like all evaluations IE also has to deal with ethical and quality issues.** These can variously concern relationships with informants and fieldsites; providing feedback; clarifying the rights and 'ownership' of evaluation outputs; ensuring confidentiality and avoiding endangering participants; and maintaining the independence of the evaluation, such that it is not captured by any one interest group.

Taking a 'broad' approach to IE will be discussed below in terms of combining designs and methods. Recognising that those engaged in IE have to address the same problems as most other evaluators implies a different kind of 'broadening'. To be a good impact evaluator it is not enough to understand the technicalities of causal inference alone. Specific IE skills should be seen as supplementing rather than substituting for the broader and more routine understandings that evaluators always depend on. Similarly 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. IE can be expensive and is not always needed. Deciding when a fully-fledged IE approach is justified is an important consideration for evaluation commissioners.

Main messages

- IE is part of the wider 'evidence-based' policy movement that emphasises value-for-money and 'results'. It therefore fulfils an accountability purpose for funders and policy makers by making programme workings more transparent. But IE can also contribute to learning by helping us understand how to do things better and more reliably in future.
- IE is distinctive because of the emphasis it places on demonstrating that it is programme actions and interventions that **cause** effects. However, this is not easy to do given the nature of development programmes. No one methodological approach is best or even sufficient on its own, which is why we need to draw on a broad range of approaches and methods for IE.
- For policy makers and programme managers who want to improve programmes, scale-up or replicate, attributing effects to causes will not be enough; they will also need to explain the effects. This is why theory is important in IE because without theory you cannot explain. Explanatory approaches such as theories of change also highlight the importance of context and make it possible to address questions of generalisability beyond a particular programme evaluation.
- IE is not separate from the rest of evaluation. It relies on many of the same skills and approaches that are central in most evaluations. A broadly based approach to IE needs to be built on a number of different evaluation traditions: it is not just about causal or even explanatory analysis even though this is central. IE also needs to draw on participatory, process-oriented, qualitative, ethical and other research traditions and bodies of knowledge.

3. Frameworks for designing impact evaluation

This chapter considers the design choices that those who commission and manage IE have to make. It concentrates mainly on the kinds of design choices that support causal claims which are at the core of what is distinctive about IE. A 'design triangle' is introduced that highlights the interdependence of evaluation questions, programme attributes and the capabilities of different methodologies. The chapter mainly discusses methodological choice. There is a fuller discussion of IE, the capabilities of methods and designs in the next chapter.

Designs that support causal claims

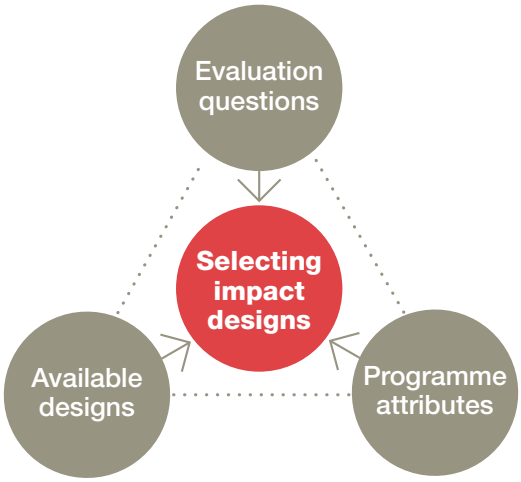
Designing evaluations requires making clear choices about many things including, for example: the purpose of an evaluation; the resources needed; required skills; ethical guidelines; data collection and analysis procedures; and how to encourage evaluation use. Getting these choices right at the beginning is essential to ensure any evaluation will be of good quality. This chapter concentrates on design choices that are specific to IE: that will ensure that it is possible to say something about cause and effect, and that will be credible and defensible when the evaluation makes these causal claims.

3. Frameworks for designing impact evaluation

This chapter is about methodological design, but not simply about methods and techniques. The aspects of 'design' discussed here refer to the underlying logic that links together sets of methods and techniques. Statistical evaluations and case studies for example may both use questionnaires, observational data and administrative records but the underlying logic that allows them to say something about causality is quite different.

The design triangle

Working out the best design for making causal claims in any single IE is a crucial planning decision. Although there is no mechanical way to make these decisions there are some logical steps to go through that help inform decision-making. The following 'design triangle' suggests what these steps are.



This diagram suggests that three factors have to be taken into account when deciding on a suitable IE design: the kinds of evaluation questions you want answers to; the 'attributes' of the programmes you want to evaluate; and the capacities of available designs. The layout of this triangle emphasises that many of these decisions are interconnected. So the kinds of evaluation questions that can be asked partly determines the selection of designs but also has to take account of programme attributes in understanding the kinds of questions that can be answered. For example, is the programme being implemented in many settings, allowing for comparative case analysis; or are large numbers of people involved so that statistical analysis is possible?

Although questions, designs and programme attributes are interconnected they are considered in turn below, but with interconnections noted along the way.

Evaluation questions

Different commissioners of evaluations will ask different types of 'impact' questions, or even more likely a different mix of such questions. Some may want precise answers to precise questions; others will want to understand whether a programme has had any kind of effect at all; and others will be most interested in the explanations for what happens. The table below lists four typical questions that IEs ask.

Table 1: Four typical questions in impact evaluation

1	To what extent can a specific 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?

When thinking about the designs that may be able to answer these kinds of questions we immediately have to consider what kind of programme is being evaluated.

To what extent can a specific impact be attributed to the intervention? This first question suggests the classic counterfactual/experimental approach. But in order to go down that path the preconditions for viable experiments have to be in place.

For example we need to be sure that the programme has a primary cause and a primary effect [4], because that is what experiments work with. Similarly we need to be able to create a control group or comparator, because experiments and other counterfactually based designs (eg quasi experiments) require some kind of comparison or 'control'. As is suggested in the next chapter there are designs and methods that can help identify specific 'impacts' when experiments are not possible. These may be weaker in their ability to precisely measure effects, but may be better able to demonstrate that some kind of causal connection is occurring. Choosing these methods follows from the nature of a programme. For example, a programme or intervention that does not have a primary cause and a primary effect cannot be compared with a virtually identical programme in a similar setting, and will not be suited to experimental methods. In many complex programme settings it is fruitless to demand accurate measurement under all circumstances.

Did the intervention make a difference?
Increasingly nowadays this second question is what policy makers are most interested in. This is because particular programmes are often just one part of the picture. NGOs and CSOs work together; national governments have their own programmes and the efforts of local communities and businesses will have as much influence on results as the programmes of development agencies or NGOs. Identifying your **contribution** and recognising the **contribution** of others is more realistic than searching for evidence of sole **attribution**.

As is further discussed in the next chapter, this is also consistent with methodological developments in the social sciences that focus on multi-causality, 'causal packages' and 'contributory causes'. These developments rest on the understanding that 'causes' may be necessary but not sufficient of themselves to lead to a change. It may even be that there is more than one way to achieve a similar objective, in which case there may also be more than one possible 'necessary' causal factor. And yet none will be sufficient without other 'supports'; and what may sometimes be necessary may indeed be unnecessary in other circumstances [5].

The third and fourth questions in the above table fall into the 'how' and 'why' or explanatory category. As noted in chapter 2, if the purpose of an evaluation is purely accountability to show that results have been achieved then causal analysis will be enough. But when the aim is to learn so as to improve success or to replicate programmes elsewhere then explanations are needed. It is in these circumstances that theory becomes important. But again the attributes of a programme, including what is known about its implementation, can lead to different ways in which theory is used. For example, in areas where there has been much previous experience and research there is likely to already be a body of theory – hunches and hypotheses about what works, when and how – then an evaluation can be set up to 'test' this programme against this pre-existing theory. In areas where less is known and there is little theory then an evaluation will have to develop its own theory. This could be by reconstructing the 'theories' of the programme designers/policy makers or possibly by developing new ones based on careful observation and analysis of what happens during and after programme implementation.

4 For example, an improvement in nutritional content of diet affects childhood illness, rather than improvements in family income, public health services and diet together lead to improvements in school-attendance, which itself is partly affected by diet.

5 See Section 4.3 in Stern et al (2012) for a fuller discussion of necessity and sufficiency.

As the word 'theory' is used quite loosely in evaluation it is worth holding on to the following distinctions: **Pre-existing theory** is derived from research and prior experience; **explicit programme theory** is based on the starting assumptions of programme planners (although hopefully also rooted in some pre-existing knowledge); and **grounded theory** only begins to emerge once a programme is being implemented or is underway. In all the senses of the word, theory can be used both to guide action (eg programme design and implementation decisions) and provide hypotheses or propositions that can be further refined or tested during the course of an IE.

Evaluation designs

When thinking about evaluations we often consider the merits of combining methods. Thus, mixed methods will combine quantitative and qualitative or more than one quantitative or more than one qualitative method. This will strengthen confidence in conclusions when they are based on several different sources of information gathered in different ways, therefore avoiding the risk of what researchers sometimes call 'instrument effect'[6]. What the logic of IE designs underlines, is that in IE in particular it may be mixed **designs** rather than mixed **methods** that are most useful. Often what are required are several well-chosen designs, each of which will use a variety of methods, and be tailored to answer the various IE questions posed by evaluation commissioners and other stakeholders.

Few evaluations ask a single question; they usually want to both assess impacts **and** explain what works where and when. Or they both want to judge the contribution of a programme **and** identify lessons that might make further replication of a programme elsewhere likely to succeed. This is one reason why few evaluations stick to a single design, preferring instead to combine designs. For example, they may combine an experiment to assess and hopefully quantify impacts attributable to a programme; a participatory design to ensure validity, relevance and targeting; and comparative studies of 'cases' to better understand the implications of different contexts.

In complex programmes it can be useful to identify different levels or scales of activity such as: national, regional, county and municipal; or society, local communities, households and individuals. In these circumstances different designs can be 'nested' with some designs addressing the more inclusive units of analysis and questions; and others addressing more limited units of analysis and questions within the overall scope of the evaluation. For example, a statistical survey of administrative data may be used to describe national trends; a quasi-experimental design might compare results in different municipalities; and case studies could be used to examine causal pathways and mechanisms.

6 This occurs when all or part of the results of analysis could be explained by limitations or biases inherent in particular methods being used.

Programme attributes

The shape, form, location, purpose, inter-relationship and life-cycle of programmes vary enormously. It is unsurprising then that these 'attributes' also affect IE design. Of course there are some programmes that are easy to understand: inoculating babies or providing mosquito nets may face implementation problems but there is no doubt what the intervention is and the expected results are obvious and relatively easy to assess. This is partly because there is a substantial body of medical research that leaves little room for doubt. However, many programmes are more complex:

- They overlap with other interventions with similar aims
- They are made up of multiple and diverse 'interventions' and projects
- They are customised to a local context and therefore non standard
- Often they work 'indirectly' through several 'agents' each having their own goals
- Likely impacts are long term
- They are in areas of limited understanding/ experience
- They work in areas of risk or uncertainty
- Intended impacts are difficult to measure, possibly intangible

These kinds of attributes reinforce the relevance of IE designs that can deal with multiple causality and diverse contexts. However, these and similar programme attributes may require:

- **Decisions about what is the unit of analysis.** In a multi-intervention programme is it each separate intervention, or all together and how to take account of interactions between interventions? If there are various programmes with similar aims can they be evaluated separately or must they be looked at as a set?
- **Developing theories of change[7].** This is especially difficult in areas where little is known; and in extended implementation chains as is common when delivering programmes through 'agents'. The challenge is how to analyse linked but separate theories of change.
- **Taking account of unpredictability and 'emergence'.** When programmes have long-term impact trajectories and even more so when they operate in areas where little is known, evaluation plans have to be flexible, possibly staged and able to refocus when necessary.

Programme attributes not only have implications for designs and methods, they also have implications for evaluation questions. If programmes have results that are difficult to measure it may not be sensible to ask precise attribution questions of the net-effect variety. On the other hand if a programme is a one-off and unlikely ever to be replicated, as can be the case in certain humanitarian emergency or fragile state programmes, there may be less urgency as to whether the programme will work elsewhere.

7 Theories of change is a process for developing a common view among stakeholders of how change is expected to happen in a project or programme, and to articulating assumptions.

Main messages

The main messages of this chapter which considered evaluation design were:

- Evaluation design is always important but IEs raise their own special design challenge: how to link cause and effect and how to support causal claims. Systematically reviewing evaluation questions and programme attributes alongside methodological capabilities is one way to design better impact evaluations.
- Simple IE questions like 'Did it work?' are becoming more difficult to ask when programmes overlap with other programmes; and are influenced by other development actors and their activities or policies. A more useful question in these circumstances is: 'Did the programme make a difference?' The growing interest in contribution of programmes alongside attribution stems from today's more complex development landscape.
- Explanatory questions are appropriate when one purpose of an IE is improvement or replication. Explanation requires theory and IEs have to be aware of the very different starting points in terms of available theory across development programmes. When theory exists an IE can test a programme against this pre-existing theory. When it does not, then an IE that wants to explain will have to develop its own theory – eg theories of change – based on what happens on the ground.
- Although some programmes consist of interventions that can be understood as simple causes with straightforward 'impacts', most cannot. Development programmes in particular are often made up of multiple interventions, face considerable uncertainty, may have to change direction as new problems and processes 'emerge' over extended time spans; and deal with outcomes that are in part at least difficult to measure. These kinds of attributes have implications for the evaluation questions that can be asked as well as the kinds of designs and methods that are suitable.

4. What different designs and methods can do

This chapter takes us further into the practicalities of IE design. It explores some of the basic ways causality is understood and the main families of designs and methods that are available to those conducting IE. Of course all designs have their strengths and weaknesses, which is why combining designs and methods is so important in many real-world IEs. These strengths and weaknesses are also discussed here as well as different rationales and strategies for combining designs and methods. The chapter builds on the earlier discussion of Evaluation Questions and programme attributes. It includes examples of different designs, how they can be combined and frequently encountered challenges.

Causal inference: linking cause and effect

If the essence of IE is the ability to describe, measure and understand how programmes as intended causes lead to consequences, then we need to know how the link between cause and effect can be made. Establishing this 'link' is often described as making a 'causal claim' or establishing the basis for causal inference. There is more than one way of going about this in evaluation as in scientific research; just as there are also many different ways of classifying these designs. One common distinction is between those causal claims that depend on controlling the intervention and those over which we have no control, for whatever reason, and must therefore rely on observation. We can decide to deliver a literacy programme or a water distribution system, but we cannot control many other things that matter in development.

This could be because we literally cannot control some things (as with the weather); or because we do not know enough (as with how to 'control' post conflict reconstruction); or because it would be unethical (if it involved experimenting on people). Where control is possible, experimental designs that depend on the manipulation of causal factors rather than observation, come into their own. However, even here the evaluation question being asked and the attributes of the programme concerned may override such a preferred design choice. For example, this would be the case if the evaluation is asking 'how' and 'why' questions.

What follows is a classification of the foundations for causal inference into four main approaches that builds on a substantial review of the literature [8].

- **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.
- **Multiple causation** that depends on combinations of causes that lead to an effect – the inference basis for 'configurational' approaches to IE including qualitative comparative analysis (QCA) and contribution analysis.
- **Generative causation** that depends on identifying the 'mechanisms' that explain effects – the inference basis for 'theory based' and 'realist' approaches to IE.

In addition to these primary types of causal inference, participation is often central to development programme design and implementation.

Participatory approaches can also be seen through a causal lens even though the main justification of participation is often value-based rather than relying on causal logic. First, as has already been argued, the voice of programme participants, stakeholders and intended beneficiaries are essential to identify the impacts of a programme (often described as **construct validity** in methodological terms). Second, there are well-established theories suggesting that programmes are likely to be more successful when those involved have ownership and commitment to programme goals[9]. Third, there are well-rehearsed arguments in the philosophy of science[10] that 'the intentions of actors (actions based on reasons) constitute one source of causality' even though it is 'only a part of an explanation, because human actions also interact with structures not in the control of human agents' (Stern 2008[11]). Because of the importance of participation in development programmes, this form of causation, labelled 'actor agency' is also included in the summary Table 2 below.

8 This classification simplifies an immensely difficult area but is intended to help practitioners and managers – the audience for this Guide – rather than to comprehensively explore methodological debates.

9 See David Ellerman (2006) Helping People to Help Themselves. University of Michigan
See: Donald Davidson: Actors Reasons and Causes. Journal of Philosophy, 60 1963; Reasons, Causes, and Action Explanation Mark Risjord Philosophy of the Social Sciences, Vol. 35, No. 3

10 See: Donald Davidson: Actors Reasons and Causes. Journal of Philosophy, 60 1963; Reasons, Causes, and Action Explanation Mark Risjord Philosophy of the Social Sciences, Vol. 35, No. 3

11 Thematic Study on the Paris Declaration, Aid Effectiveness and Development Effectiveness http://www.oecd.org/development/evaluation/dcdndep/41807824.pdf

Table 2: Design approaches, variants and causal inference

Design approaches	Specific variants	Basis for causal inference
Experimental	RCTs Quasi experiments, Natural experiments	**Counterfactuals:** the difference between two otherwise identical cases – the manipulated and the controlled; the co-presence of cause and effects.
Statistical	Statistical modelling Longitudinal studies Econometrics	**Regularity:** 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	**Generative causation:** Identification and confirmation of causal processes or 'chains'. Supporting factors and mechanisms at work in context.
Case-based	**Interpretative:** Naturalistic, grounded theory, ethnography Structured: Configurations, QCA, within-case-analysis, simulations and network analysis	**Multiple causation:** 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.	**Actor agency:** 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.)

Each of these main causal approaches has requirements, that is, conditions under which they do and do not apply; potential strengths; and potential weaknesses. For example:

- 'Regularity' requires high numbers of diverse cases. Without this it is not possible to capture sufficient diversity (or difference).
- Counterfactuals are good at answering the question: 'Has this particular intervention made a difference here?' But they are weak on answering generalisation (external validity) questions: 'Will it work elsewhere?'
- Multiple causalities are good at dealing with moderate levels of complexity and interdependence but not at unpicking highly complex and highly interdependent combinations of causes.
- Generative causation is strong on explanation but weak on estimating quantities or extent of impact.
- Experiments and regularity/statistical association approaches work best when causal factors are independent of each other, but not if various causal factors interact with each other.
- Neither experiments nor statistical models are good at dealing with contextualisation – taking account of cultural, institutional, historical and economic settings.

It is unusual for evaluators or even researchers to make explicit the basis on which they make causal claims. This is because most evaluators and researchers come from particular methodological traditions and take for granted what they know best. This makes it especially important for those who commission evaluations to have their own ways of assessing what they need, and to look for the kinds of skills that meet their requirements[12].

12 It is also important for those who commission evaluations to ask the evaluators to explain their causal claims satisfactorily. See discussion of ToR content in chapter 5.

Main types of impact evaluation design

As already noted, different approaches to causal inference are associated with different designs even though there is not always a one-for-one association. The main designs useful for IE are:

- **Statistical:** where large numbers of cases – populations, small businesses and so on – and characteristics of these cases (variables) are analysed.
- **Experimental:** where different but similar situations are compared to situations when an intervention is or is not present.
- **Theory based:** where what happens is compared with pre-existing theories or causal pathways identified during an evaluation.
- **'Case-based':** where different cases (or case-studies) are analysed and sets of case characteristics (configurations) are compared in relation to outcomes.
- **Participatory:** where the judgements and experience of stakeholders and beneficiaries are best able to identify the most relevant theories of change and meaningful outcomes from among several possibilities.
- **Synthesis-based:** where the results of a number of evaluations are combined in order to reach a judgement based on cumulative findings.

These 'big' categories of design can take a number of specific forms. For example, experimental designs can include 'quasi-experiments' where the level of control over the programme setting is less than required by a fully randomised trial (RCT) and a control group is used rather than randomisation. Theory-based evaluations encompass Realist evaluation, Contribution Analysis and Process Tracing. These variants of the main design types and the basis for causal inference on which they depend are summarised in Table 3 below.

Although the main designs identified above will be familiar to most readers of this guide, a number of innovative or emergent methods will not be. Examples of these include:

- Theory based evaluation
- Realist Evaluation
- Qualitative Comparative Analysis
- Contribution Analysis
- Process tracing

A brief introduction to these five examples is included in an Annex to this Guide together with some further reading for those who want to deepen their knowledge further.

The contemporary importance of the 'contributory' cause

One of the striking developments in the social sciences in recent decades has been the growing interest in complexity and multiple-causality. This thinking is now becoming more prominent in evaluation practice and IE. Programmes are increasingly viewed as 'contributory' causes – one factor among many, part of a 'causal package'. Programme success depends on what else is going on or has gone on around them. This can be contrasted with 'attribution' based logics, a feature of counterfactual/experimental approaches. Theory-based and case-based designs such as 'Realist evaluation', Contribution Analysis, QCA, Network Analysis and Process Tracing all help evaluators to better address multiple causality. Some kinds of modelling, such as agent-based-modelling also contribute to this expanding 'toolkit'.

The idea of the 'contributory cause' (see Annex) is particularly relevant for socio-economic and international development. Contemporary understandings of development emphasise the importance of mobilising not only the resources of external development agencies such as foreign governments, regional banks and NGOs but also national governments, civil society, CBOs, municipalities and local communities. Aid is also seen as only one source of development funding among many. In these circumstances it becomes increasingly difficult for development actors to say 'we did this on our own'. As suggested earlier, a far more common evaluation question nowadays is: 'Did we make a difference?' And the required answer is that a programme can be shown to be a necessary contributory cause in a particular programme setting.

Revisiting the 'design triangle'

Chapter 3 introduced the 'Design Triangle' that highlighted the connections between evaluation questions, IE designs and the attributes of programmes being evaluated. This chapter has reinforced the linkage between evaluation questions and designs including their underlying causal logics. The chapter has therefore signposted some important IE design choices. For example:

- If an evaluation wants to attribute a net impact to an intervention then experiments are indeed your best bet. This requires of course that you have enough control to 'manipulate' the intervention (separate out the 'treatment' from the 'control' group); you are clear that there is a primary cause and a primary effect that you are interested in; and there are enough cases to support statistical analysis.
- If on the other hand an evaluation wants to know whether a programme has contributed to desired change or any other kind of change – has 'made a difference' – some kind of theory-based or case-based design is necessary. This requires a degree of prior (theoretical) understanding of how a programme works and is connected to other 'contributing' causal factors[13]. It will also be strengthened if there are a number of cases that can be compared with each other.
- If the evaluation is interested in explanations – answering 'how' and 'why' questions – theory is again needed, whether pre-existing or purposefully developed. The kinds of theories and associated designs required will be those that can unpick contextual factors that might have causal potency, and identify other things going on that could also influence outcomes and impacts. These might for example include participatory as well as theory-based designs such as Realist evaluation[14].

The final bullet point highlights that IE designs are rarely pure types: hence the importance of hybrid designs (and combining methods) as described in chapter 3.

13 Although an IE may also develop its own theory in the course of an evaluation based on an emergent theory of change or on an elaboration of the initial assumptions articulated in the starting 'programme' theory for the programme.

14 See Annex.

4. What different designs and methods can do

Table 3 summarises some of the main methodological design implications of different evaluation questions. (This is further elaborated in chapter 4 of Stern et al (2012) and in the Appendix to that report by Barbara Befani.)

Table 3: Summarising the design implications of different impact evaluation questions

Key evaluation questions	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, eg contribution analysis Case-based designs, eg 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 and 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 and Contribution Analysis 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 eg typologies of context Clusters of causal packages Innovation diffusion mechanisms	Participatory approaches and some Experimental and Theory-based approaches Natural experiments Realist evaluation Synthesis studies

Main messages

The main messages in this chapter are that different designs offer different possibilities for linking cause and effect. Causal inference is crucial for IE and the chapter emphasises:

- There is more than one way of linking programmes as causes with impacts in IE, as there is in scientific research more generally. These different grounds to make a 'causal claim' underpin different IE designs. Choosing between different designs (or combinations of designs) depends partly on the extent of control over programme implementation that is feasible and desirable – alongside programme attributes.
- Design choices also have to be considered in terms of some basic pre-conditions; and the kinds of evaluation questions being asked and the attributes of programmes. Much of what is contained in this chapter takes us back to the 'design triangle' introduced in chapter 3, which emphasised the need to keep in balance questions, designs and the attributes of programmes.
- The main categories of IE design – statistical, experimental, case-based, participatory and synthesis-based – come in a number of variants and sub-types. These are variously able to answer different evaluation questions and respond to different programme attributes. This chapter again underlines the importance of combining designs and methods – pure types are rarely sufficient.
- Many programmes nowadays are 'complex', containing multiple interventions and variously implemented in different contexts. This is the result both of the ambition of many development programmes and the evolving nature of the aid and development architecture with its emphasis on combining international, national, civil society and community interventions.
- These complex programmes are what drive demands for designs that are able to analyse 'contributory causes'. Counterfactual and some statistical designs are best suited to programmes where there is one primary cause and effect of interest. Other designs, especially theory and case-based and certain kinds of modelling, are better able to accommodate multiple causes and multiple outcomes and impacts.

An overarching message is that there is no one single best design. Those who commission IEs need to be aware that evaluators are most comfortable with those evaluation approaches with which they are familiar. A key part of the IE design process is to choose designs and methods that best fit the questions being asked and the specific possibilities and constraints of the programme under consideration, and only then to choose evaluators with understanding of these designs and methods.

5. Using this guide

As noted in the opening chapter, this guide is intended for those who commission, manage and use impact evaluations. In this concluding chapter two main scenarios are considered. The first is at the beginning of an evaluation when the terms of reference (ToR) for an IE is drawn up and proposals have to be assessed. The second is at the end of an evaluation when the quality of reports must be judged, and conclusions, recommendations and lessons have to be extracted.

Most of the material presented in this chapter has been introduced previously. But this chapter highlights how thinking about IE design can be applied in practical situations by commissioners and managers.

Drawing up terms of reference and assessing proposals for impact evaluations

The IE design process usually begins with a ToR. These could be drawn up by CSO managers accountable to funders for money spent; by funders themselves; or by Headquarters or decentralised offices of development agencies wishing to learn lessons from innovative practice. An IE could be conducted by external evaluation specialists, or possibly by an internal unit within a commissioning body. Whatever the circumstance, a ToR sets out the expectations of commissioners and key issues that evaluators need to address in their proposals.

The ToR for an IE share many requirements with other evaluations. For example, those commissioning any evaluation will need to decide:

- whether an IE is justified
- the size of the budget
- timescales and deliverables
- team composition and structure
- quality assurance arrangements required
- ethical issues such as risks for those affected by the evaluation

These decisions will also need to be made by those drawing up ToRs for IEs – and will have implications for the strength and quality of subsequent evaluation design. However, the focus here is on the main issues distinctive to IEs. These include:

- How to identify impacts
- Taking account of previous knowledge
- The overall purpose of an evaluation – which determines evaluation questions
- Programme attributes, including architecture, scale and complexity
- Whether the context or setting supports a contribution or attribution approach
- Whether measurement of impacts is wanted or possible given available and potential data

Table 4 below elaborates these issues, the underlying rationale and implications for ToRs.

Assessing proposals

Proposals should be assessed in terms of the main issues identified in the ToR. Overall, commissioners need to be confident that a proposal for an IE will be able to link the results of a programme to the activities and interventions that the programme made possible, whether on its own or jointly with other causal and contextual factors. Proposals should indicate the means through which an evaluation will link programme causes and effects.

To summarise the above and the content of the ToR table, an assessment checklist at the proposal stage should include:

- Have impacts been identified and understood?
- Are stakeholders going to be involved in validating these impacts?
- Has existing knowledge about this kind of programme, including ToCs, been taken into account?
- Are programme purposes understood and evaluation questions clearly stated?
- Has the proposal shown how IE design is able to link cause and effect and answer evaluation questions?
- Is the proposed design consistent with programme attributes and the simplicity or complexity of the programme?
- Is the timing of the IE consistent with the likely trajectory of intended change?
- If the programme is complex are the proposed methods able to disentangle more than one cause?
- Are proposals putting forward measurement of impacts consistent with the kind of programme data available and collectable; and the designs and methods to be used?
- Have protocols and methodological guidance used in connection with the proposed design, where these exist, been cited and used?
- Have examples of reports or publications that illustrate how this design has been used previously for impact evaluation been provided (this may include examples of work by the proposal team or others)?

Table 4: Drawing up terms of reference for impact evaluations

Design issues	Specific questions	Rationale	Implications
Identifying impacts	How should programme impacts and effects be identified?	Conceptualising and identifying impacts is difficult, and sometimes data is unavailable. When to assess impacts and which impacts affect whom, are also design issues. Stakeholders' participation helps identify valid impacts.	Proposers should indicate how they understand and will identify impacts – including impacts for different groups. Commissioners should indicate data availability problems.
Building on what is known	Is there already substantial knowledge about how these kinds of programmes work, perhaps a credible theory of change?	If much is already known there might both be risks of duplication and waste; and advantages building on existing knowledge.	Proposers should demonstrate familiarity with current state of evaluation/research knowledge and indicate how this will shape their use of theories of change.
The overall purpose of the evaluation	What kind of use for whom is envisaged – demonstrating past effectiveness; scaling-up and replication; improvement; learning for future policy and practice?	Purposes of IE may differ. It is important to identify main purposes as this determines evaluation questions and choice of methods able to answer these questions.	Proposers should be expected to discuss how overall purpose connects with evaluation questions – and show an awareness of design and method implications.
Programme attributes, scale and complexity	Is the programme made up of a single intervention or several? What is the programme 'architecture'?	Programme attributes constrain the choice of IE designs and methods. Multi-level or decentralised programmes offer opportunities for nested designs.	Proposers should be asked to demonstrate understandings of programme attributes and the implications for designs and combinations of designs.
Context and contribution	How important is context and how far are different causal and contextual factors likely to influence impacts?	Programmes that are open to multiple influences – complex, embedded rather than simple and self-contained – will need to focus on the contribution of programme interventions rather than attribution.	Proposers should be asked to discuss the programme context including the importance of multiple causal factors; and how this relates to a contribution or attribution focus.
Measurement and extent	Does the IE set out to measure how much of an impact a programme has had – and is this feasible?	Sometimes it is possible to assess contribution but not extent (how much?). Whether the programme has impacts for large numbers of households, or few will also determine the possibility of statistical designs and methods.	If appropriate, proposers should be asked to discuss their approach to measurement and extent.

Quality of reports and findings

Whatever the ToR and proposal, the strength of an IE will only become clear at subsequent report stages. Inception reports usually offer a first opportunity to assess more detailed design specifications. For example, the amount of time allocated to different parts of the evaluation; and the match between skills of team members and specific analytic activities should be clearer by the time inception reports are submitted. These can be asked for at that time.

When substantive reports that include findings are produced, it is worth revisiting the ToR framework and proposal assessment checklists in the first instance. However, a checklist at this stage can be more focussed. For example, commissioners and managers should ask:

- Does the report make it clear how causal claims have been arrived at?
- How have different types of theory been used – testing programme assumptions or building on wider research? Has new theory been developed?
- Is the report clear about when and where impacts can be observed?
- Does the report convincingly identify contextual and causal factors and take them into account?
- Is the chosen design able to support explanatory analysis (answering how and why questions) if this was required?
- Is there a consistent link between evaluation questions asked, overall design, data collection and analytic methods used?
- Have alternative explanations that do not depend on programme effects been considered and systematically eliminated or accounted for?
- Have beneficiaries and other stakeholders been involved in scoping the evaluation and validating and interpreting results?
- Are the ways methods were applied and data collected clearly described and well documented?

A positive answer to these questions makes an IE reliable and defensible. They incorporate some of the key elements that a researcher would call validity, robustness, rigour and transparency. A final judgement on these qualities cannot only be up to commissioners; they also require third party peer reviews. However, the above checklist suggests a common language that both commissioners and external reviewers can use.

It is also worth considering the inclusion of these kinds of assessment questions in a ToR package. This would ensure that evaluators were aware from the beginning how their work was going to be assessed and provide them with a template for continuous self-monitoring. There are a number of different tools or standards available for checking the quality of evaluations and evidence (see for example Nutley, Powell and Davies 2013.[15]) Bond's Evidence Principles and checklist is another such tool which was designed for use in international development, and provides a means of assessing evidence quality irrespective of the specific evaluation design and method used [16].

15 Sandra Nutley, Alison Powell and Huw Davies. What constitutes good evidence? The Alliance for Useful Evidence 2013. http://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf

16 http://www.bond.org.uk/effectiveness/principles

Strengths of conclusions and recommendations

The strength of conclusions depends on various factors, for example:

- The soundness of the IE design – a central concern in all the frameworks and checklists described in this chapter.
- The way that design and associated methods were implemented, the reason why transparency and 'auditability' of methods and data are important.
- Consistency between the conclusions drawn and the evidence base and designs on which these conclusions are based. For example, a theory-based design cannot quantify impact on its own. Neither can a counterfactual-based design predict what might happen in a different setting on its own.
- The scope of evidence: What kinds of judgements does the evidence support? For example, the IE of a specific programme cannot be used to judge an entire class of similar programmes across different settings.
- The judgement of the evaluators – conclusions rely on judgement as there is rarely an automatic link between evidence and conclusions. Hence, the importance of evaluators making their criteria and often their values explicit.
- Evaluators acknowledging the limitations of all designs and methods; and the innate difficulties of going beyond probability and plausibility.

Different commissioners have different expectations of how far evaluators should go in making recommendations. Policy commissioners often take the view that evaluators are not sufficiently knowledgeable about policy contexts to make sensible recommendations. However, in the voluntary sector, and especially when evaluations are commissioned and partly specified by potential users of findings, it can be argued that evaluators who mainly work with NGOs and CSOs do have enough background knowledge. A sensible middle-ground is to expect evaluators to put forward recommendations, based on 'sense-making' discussions and workshops with key users of findings. Subsequently, commissioners may need to situate these recommendations into a wider body of organisational or policy knowledge.

Consideration of the validity of recommendations should take account of:

- The connection between recommendations and conclusions
- The strength of evidence that fed into conclusions
- The criteria and values used to justify conclusions
- The input of stakeholders into a validation process
- The extent to which conclusions are supported by a more extensive evidence base, such as previous evaluations, syntheses and research

Using findings from impact evaluations

In terms of the subsequent use of IE outputs, this will be partly anticipated by the initial purpose of the IE (see Table 4 above). However, experience suggests that direct or what is often called 'instrumental' use of an evaluation is unusual – even if often hoped for. This may of course happen when an IE was commissioned to feed directly into a specific decision. For example, if a programme is looking for further funding or a looking to scale-up a pilot programme. Whether an evaluation has direct action implications or not, particular IEs will add cumulatively to what is known about types of programmes in types of contexts. Evaluation use will be strengthened by an accumulation of convincing findings. Hence the importance of evaluation syntheses, that systematically collate a broad body of evidence around common development priorities.

It is also often the case that the relevance and utility of findings from a single IE will only become obvious at some point in the future when new and similar circumstances occur. This underlines the importance of disseminating evaluation findings widely within policy and practice communities; and investing in knowledge management systems that can make accessible what is known for future relevance and use.

Main messages

The main messages of this chapter are that quality, validity and defensibility of an IE have to be followed through during every stage in an IE cycle. This starts with ToRs through to the assessment of proposals, inception reports, and to final reports and evaluation use. In particular:

- Designs and associated methods must have the ability to link cause and effect and answer particular evaluation questions.
- Scope, definition of impacts, conclusions and recommendations need to be validated with stakeholders.
- Data should be collected and analysed in transparent and auditable ways.
- Conclusions and recommendations should be consistent with IE designs and the evidence these designs produce.
- Caution is needed about basing evaluation use on any single IE.
- Use and the reliability of IEs will be strengthened if integrated with evidence from research and other IEs. This requires investment in knowledge management and synthesis reviews that collate what is known to encourage timely use in both the longer and shorter run.

ANNEX: CONTEMPORARY METHODS

Various contemporary methods, some of which are only just beginning to be used by evaluators, have been referred to in this Guide. (See for example Table 2 in chapter 4 in particular). These include:

- Theory-based evaluation
- Realist evaluation
- Qualitative comparative analysis
- Contribution analysis
- Process tracing

Each of these is briefly introduced in this Annex together with additional source material for those readers who want to explore methods in greater depth.

Theory-based evaluation

"There are some core features of the TBE approach that appear consistent across the main accounts of the approach:

- Opening up the black box to answer not simply the question of what works, but also why and how it worked. This is key to producing policy relevant evaluation.
- Understanding the transformational relations between treatment and outcomes, as well as contextual factors.
- Defining theory as the causal model or theory of change that underlies a programme.
- Having two key parts: conceptual (developing the causal model and using this model to guide the evaluation); and empirical (testing the causal model to investigate how programme cause intended or observed outcomes).
- Being issues led, and therefore, methods neutral."

Carter, R. (2012), Governance and Social Development Resource Centre, University of Birmingham http://www.gsdrc.org/docs/open/HDQ872.pdf

See also:

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

Mayne J. and Stern E. 2013. Impact evaluation of natural resource management research programs: a broader view. ACIAR Impact Assessment Series Report No. 84. Australian Centre for International Agricultural Research: Canberra. http://aci Stame, N. (2004). Theory-based evaluation and varieties of complexity. Evaluation, 10(1), 58-76.ar.gov.au/files/ias84.pdf

Realist evaluation

"Realist approaches assume that nothing works everywhere or for everyone, and that context really does make a difference to programme outcomes. Consequently, policy-makers and practitioners need to understand how and why programmes work and don't work in different contexts, so that they are better equipped to make decisions about which programmes or policies to use and how to adapt them to local contexts. Consequently, realist evaluation does not ask 'what works?', 'does this work?' or (retrospectively) 'did this work this time?' A realist research question contains some or all of the elements of 'how and why does this work and/or not work, for whom, to what extent, in what respects, in what circumstances and over what duration?"

Gill Westhorp (2014) Realist Evaluation:
An Introduction. Methods Lab Overseas Development Institute London

http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9138.pdf

See also:

Ray Pawson, 2002. The Promise of Realist Synthesis, in Evaluation the International journal of theory, research and practice

Downloadable at https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp4.pdf

Dieleman, Wong and Marchal https://www.abdn.ac.uk/femhealth/documents/Realist_methods_workshop.pdf

Qualitative comparative analysis

Qualitative comparative analysis (QCA) is an approach to systematic cross-case comparison. It establishes what factors, common across cases, can explain similar outcomes; or what factors could explain different outcomes. Unlike most methods intended to draw generalised lessons across cases, QCA does not look at variables in isolation. It focuses on combinations or configurations of factors within single cases; and allows generalisation only to the extent that these holistic combinations are preserved.

Although QCA establishes an association between a 'dependent' condition (the outcome) and a number of 'independent' conditions, the aim of QCA is not measuring correlation, or understanding how much a given variable "adds" to the outcome for each addition unit; but rather establishing a) what are the necessary conditions for an outcome and b) what are the sufficient combinations of conditions for the same outcome. Causal necessity means that an outcome is required: it can never be observed without the presence of certain conditions. Sufficiency means that the combination is good enough to produce the outcome and does not need any other requirement.

QCA is appropriate to identify the preconditions and make sense of the diversity in results across small numbers of cases when there are several but not many causal factors. It is not appropriate when the explanation is only one case.

See also:

Charles Ragin: What is Qualitative Comparative Analysis http://eprints.ncrm.ac.uk/250/1/What_is_QCA.pdf

Compass Website: http://www.compasss.org/wpseries/allWPdate.htm

Tim Blackman, J Wistow, D Byrne (2013) Using Qualitative Comparative Analysis to understand complex policy problems Evaluation, International journal of theory, research and practice http://oro.open.ac.uk/37540/2/5C07E325.pdf

Contributory causes and contribution analysis

"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 lead to 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."

"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."

Broadening the Range of Designs and methods for Impact evaluation Stern et al 2012 (pp40 & 41). http://r4d.dfid.gov.uk/Output/189575/

See also:

John Mayne, Contribution Analysis: Coming of Age? In Evaluation 18.3 July 2012. Special Issue: Contribution Analysis

SOCIAL SCIENCE METHODS SERIES, Guide 6: Contribution Analysis. Scottish Government http://www.scotland.gov.uk/resource/doc/175356/0116687.pdf

Process tracing

"This approach was first developed in 1979 and was fleshed out comprehensively in George and Bennett's Case Studies and Theory Development in the Social Sciences (2005). Process tracing centers on dissecting causation through causal mechanisms between the observed variables, primarily in case studies. In essence, the focus of process tracing is on establishing the causal mechanism, by examining the fit of a theory to the intervening causal steps. Theorists using process tracing ask 'how does "X" produce a series of conditions that come together in some way (or do not) to produce "Y"?' By emphasizing that the causal process leads to certain outcomes, process tracing lends itself to validating theoretical predictions and hypotheses.

Despite often focusing on only a single case, process tracing is a useful tool for testing theories. Researchers must examine a number of histories, archival documents, interview transcripts, and other similar sources pertaining to their specific case in order to determine whether a proposed theoretical hypothesis is evident in the sequence of a case (George and Bennett, 6). Looking at these sources in terms of the sequence and structure of events can serve as evidence that a given stimulus caused a certain response in a case. Process tracing aims to ascertain the causal process linking an independent variable(s) to the outcome of a dependent variable, particularly in small-n studies. This method is particularly useful for looking at deviant cases and determining the specific factors that lead them to diverge from expected trends. While process tracing may not be able to exclude all but one theory in a given case, it can narrow the range of possible explanations and can disprove claims that a single variable is necessary or sufficient to produce an outcome."

Users Guide to Political Science: Government Department, Wesleyan University

http://govthesis.site.wesleyan.edu/research/methods-and-analysis/analyzing-qualitative-data/process-tracing/

See also:

Barbara Befani and John Mayne (2014) Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation http://onlinelibrary.wiley.com/doi/10.1111/1759-5436.12110/abstract

David Collier (2011), Understanding Process Tracing in Political Science and Politics, 44.

http://polisci.berkeley.edu/sites/default/files/people/u3827/Understanding%20Process%20Tracing.pdf

Process Tracing – Draft Protocol Oxfam

http://policy-practice.oxfam.org.uk/~/media/Files/policy_and_practice/methods_approaches/effectiveness/Process-tracing-draft-protocol-110113.ashx

bond
FOR INTERNATIONAL DEVELOPMENT

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

+44 (0)20 7520 0248
bond.org.uk

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