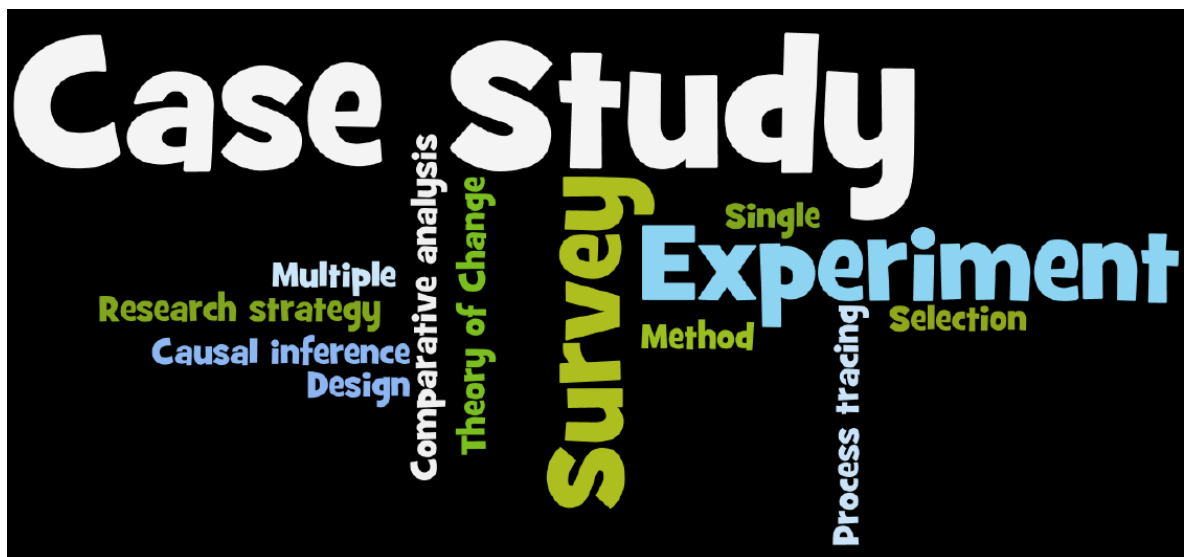




UNIVERSITY
of
GREENWICH | Natural
Resources
Institute

The Strategic Use of Case Studies

in the Monitoring and Evaluation Systems of Sustainability Standards



Valerie Nelson and Adrienne Martin, Natural Resources Institute, University of Greenwich
7/1/2013

Contents

1. Introduction	3
2. Methodology.....	4
3. Case study research – an overview of current theory in the literature.....	5
3.1 Case study theory evolution	5
3.2 Strengths and weaknesses	9
3.3 Evaluation questions and different research strategies	11
3.4 Types of case study research	13
3.5 Case study design.....	14
3.6 Causation, attribution and contribution	20
3.7 Analytical strategies and techniques	21
3.8 Process tracing	24
3.9 Complexity	26
3.10 Multi-method collaborative research.....	29
3.11 Rigour, utility and quality.....	29
4. Practical Guidance	32
4.1 When do I use case study research?.....	32
4.2 How do I design case study research?	33
4.3 How do I select case studies?	38
4.4 How many cases are enough?	39
4.5 How do I collect case study evidence?	39
4.6 How do I collect data and which methods should I use?.....	41
4.7 How do I analyse the case study evidence?.....	42
4.8 How do I report my case study?	43
5. Conclusions	44
References	48
Appendices.....	50
Appendix 1:	50
Appendix 2: Analytic techniques.....	52
Appendix 4: Utilization focused framework steps.....	54

Acknowledgements

Many thanks are due to Kristin Komives and Marta Maireles-Gonzalez for their insights and in-depth feedback on the draft reports. We would also like to thank Rik Kutsch, Biko Nagara, Kate Kilpatrick and Vidya Rangan for their support for the study and the insights they have provided for this study.

1. Introduction

This short study has been commissioned by ISEAL to support standard system members to improve their monitoring and evaluation by increasing their understanding of the potential role of case studies. The defined goal of the study is as follows:

“to help sustainability standard systems make strategic choices about which and how many ‘case studies’ to engage in as part of their M&E system. Sustainability standards would like to know how best to choose case studies so as to maximize what they can learn and credibly demonstrate from the collection of case study results”.

The study has been commissioned by ISEAL as part of the ‘Demonstrating and Improving Impacts’ (DIPI) project¹, funded by the Ford Foundation. ISEAL, through the DIPI project, seeks to support the sustainability standard systems which are their members, in their efforts to develop and/or improve their monitoring and evaluation systems, in line with the ISEAL Impacts Code.

The study TOR (see appendix 1) explains that: the ISEAL Impacts Code requires the standard system members to identify their sustainability objectives, define their theory of change, select indicators to track and assess progress, data collection and reporting, but also *additional* outcome and evaluation activities (either conducted by the standard systems or by external researchers), which are shared with the organization and wider public. The outcome and impact evaluations are intended to be in-depth studies which answer specific research questions, and to show what has changed, why, and for impact evaluation, the attribution to the work of the standard. These will be focused on a subset of the system by necessity, according to ISEAL. While focus on a particular case allows for more in-depth data collection and analysis, the extent to which findings can be generalized is unclear and the findings are more easily ‘dismissed or downplayed’ because they represent a single case. This consultancy was therefore designed by ISEAL to help sustainability standard systems to make strategic choices about which and how many ‘case studies’ to undertake as part of their overarching M&E system.

The key study questions are as follows:

Box 1: Research questions

1. How many cases are “enough”? What is (or what determines) the added value of an additional case?
2. Does the answer to this question differ if the cases apply different versus parallel methodologies?
3. Are more cases / is better (geographic) coverage of cases always better? What are the advantages/disadvantages of repeating studies of one or a small set of cases or locations, versus attempting to cover more cases or locations in the collection of studies?
4. What are the advantages or disadvantages of choosing “easy” versus “hard” cases (cases where one would suspect that certification would easily have the desired impact versus cases where having any impact would be very difficult)?
5. Should random sampling ever play a role in the choice of cases? Or does strategic choice of cases

¹¹ See ISEAL website: <http://www.isealalliance.org/our-work/collaborative-learning/demonstrating-impacts>

always make more sense?

6. What should standard systems think about when choosing cases to achieve maximum learning value for their own organisations? Or to provide the most convincing evidence of the results and impacts they produce? Is there any trade-off between learning and demonstrating?

7. How should standard systems talk about the results of a study of a single location or case? What can and can't they claim? To what extent is the answer determined by the methodology used in the case study as opposed to the inherent limitations of looking at results or impact in one particular case? (How) does that change with multiple, carefully chosen cases versus a single case?

8. Are there any general guidelines for the design of case study research that would help increase the learning and demonstration value of the case study or the collection of case studies? (independent of the exact methodology used for the case study)

This report is structured as follows: Section 2 details the study methodology; Section 3 reviews the literature and reflects on the opportunities and challenges for standard systems in using case study research; Section 4 provides guidance and specific recommendations for standard systems on how to use case study research within their M&E systems. The appendices include the terms of reference, and more detailed analysis of particular aspects of case study research.

2. Methodology

The main envisaged activities were: i) reviewing the methodological literature about the use and selection of case studies and drawing on own experience; ii) holding a workshop session to be held at ISEAL impacts workshop at the University of Greenwich; iii) interviews with at least 4 sustainability standard systems to; iv) careful study of the use of (and potential uses of) case studies by two selected standard systems & recommendations; v) production of training webinar and materials.

As well as a review of the literature, a session was held within a larger ISEAL workshop (November, 2013) giving the consultants an opportunity to explain the study and to gather feedback from the standard system representatives present. Four case study organisations were selected in discussion with ISEAL – GoodWeave; Union of Ethical Biotrade; Marine Stewardship Council & Fairtrade, based on the following criteria: i) standard systems of different size were chosen; ii) different types of standards were chosen in terms of their focus (social and/or environmental) and level of operation (enterprise, community to value chain to market); iii) their interest in participating in the study and time availability. The four cases sought to unpack the vision, theories of change and (planned) M&E systems for each standard system, and to tease out their ideas for using case studies (in the future and how they have been used in the past), and their current structure and reach of their standard and universe of certified entities. Of the four standard systems, two (GoodWeave and Union of Ethical Biotrade), were chosen for more in-depth collaboration and learning via Skype and email exchanges. A representative from the Marine Stewardship Council was also interviewed. Through an on-going and previous collaboration, the consultants were able to draw on some insights on some first-hand case study research experience with Fairtrade International.

This report is intended as a guidance document to explain applicable theory on case study research to standard systems and to provide them with specific recommendations, including illustrations of the theory and recommendations with examples. Interactions with two standard systems have been undertaken to enable the consultants to learn from their experiences and to provide guidance to the particular standards on the use of the case studies are a second output. A recorded webinar is planned as a recording for use in training.

3. Case study research – an overview of current theory in the literature

3.1 Case study theory evolution

This section explains how case study theory is evolving, to orient standard systems in the theory, which they can then apply in their Monitoring and Evaluation (M&E) to maximum effect (i.e. strategically). Tailored practical guidance is provided in section 4.

Case study research in the social sciences has a very long history. However, it fell out of fashion with the spread and development of statistical approaches and formal modelling in the 1960s and 70s, and with **sustained criticism of case study approaches** (George and Bennett, 2005). A key criticism of earlier case studies, is that they do not allow for systematic comparison and ‘orderly cumulation’ (of knowledge and evidence), because they were essentially descriptive and lacking in a theoretical basis (Macridis and Brown, 1955 cited in George and Bennett, 2005). Only in the last thirty years have scholars formalized case study methods and linked them more closely to the underlying arguments in the philosophy of science (George and Bennett, 2005).

Box 1 Definitions

‘The case study approach – the detailed examination of an aspect of a historical episode to develop or test historical explanations that may be generalizable to other events...A case can be defined as ‘an instance of a class of events’, with a ‘class of events’ being a phenomenon of scientific interest, such as revolutions, types of governmental regimes, kinds of economic systems, or personality types that the investigator chooses to study with the aim of developing theory (or ‘generic knowledge’) regarding the causes of similarities or differences among instances (cases) of that class of events’ (George and Bennett, 2005, (p17-18).

‘A case study is an empirical enquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident. The case study enquiry ... copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as a result relies on multiple sources of evidence, with the need to triangulate data, and as a result benefits from the prior development of theoretical propositions to guide data collection and analysis” (Yin, 2009, p18)

Essentially case study research is a **research strategy**, just like experiments, a history or a simulation (Yin, 1994). None of these are linked to a particular type of evidence or method of data collection. **Statistical approaches, modelling, and case study research all share a similar ‘epistemological logic’ and all include empirical research, but each has different methodological logics:** *‘Epistemologically, all three approaches derive observable implications from these theories, they test these implications against empirical observations or measurements, and they use the results of these tests to make inferences on how best to modify the theories tested. Methodologically, these three methods use very different kinds of reasoning regarding fundamental issues such as case selection, operationalization of variables, and the use of inductive and deductive logic’* (George and Bennett, 2005, p6). This implies that they can be **complementary in fact rather and the task is matching the most appropriate method with the task in hand and alternative methods should be drawn upon to compensate where there are limitations** (George and Bennett, op cit). Case study methods are somewhat intuitive, but have only been more formalized in recent years (George and Bennett, 2005).

There remains a great deal of **confusion in case study research**, with:

- conflation of case study research with specific types of evidence (e.g. qualitative)

- conflation of case study research with specific research methods and methodologies (e.g. ethnographic)
- a lack of understanding of whether and how case study research fits with or differs from other research strategies
- a lack of understanding of the relative strengths and weaknesses of case study (and other forms of) research strategy.

In fact, a case study does not necessarily imply a particular type of evidence and can employ either qualitative or quantitative evidence (Yin, 1981). Rather than thinking about specific methods or tools, **case study research is a kind of research strategy – the other kinds of research strategy available and potentially complementary are experiments, histories, simulations etc.** Stern et al, (2012) identify five different research strategies specifically in relation to impact evaluation, namely: experimental (e.g. RCTs), statistical, theory based (e.g. realist or mechanism-based designs), case based (e.g. ethnography – but not only) and participatory. The actual methods and techniques can overlap within the different strategies (e.g. use of interview data, focus group discussions, statistical methods), but what varies is the *fundamental logic* being employed, although the classification is rarely tidy in practice. Most investment has been in the first two, but a broader range of approaches are needed and available, especially for more complex and difficult to evaluate programmes (Stern et al, *ibid*). Statistical methods ‘excel at estimating the generalized causal weight or causal effects of variables’ and formal models employ ‘rigorous deductive logic’ to develop ‘both intuitive and counter-intuitive hypotheses about the dynamics of causal mechanisms’. Case study methods are also somewhat intuitive, but have only been more formalized in recent years (George and Bennett , 2005).

At a broader level, Stern et al (*ibid*) suggest you can distinguish between ‘true experimental designs’ based on control or ‘manipulation’ and other ‘observational’ studies. Generally, the former are thought to be the approaches which allow researchers to draw causal inferences. **Randomized control trials, for example, can indeed produce strong causal claims (where conditions can be manipulated), but it is possible for many different types of design (qualitative and quantitative), where correctly designed and implemented, to support causal and explanatory analysis** (Stern et al, 2012). Each approach actually has a range of strengths and weaknesses, including their fundamental approaches to causation.

The decision about which research strategy to choose depends upon three things (Yin, 2014):

- research question
- extent of control the researcher has over actual behavioural events
- the degree of focus on contemporary as opposed to historical events.

Case study research are generally thought of as being useful to answer ‘how and why?’ questions, and where in-depth research is needed using a holistic lens. It does not require control over actual behavioural events. Surveys are useful to answer research questions on the ‘who, what, where, how many, how much?’ and does not require control over behavioural events. Finally, experiments are useful for answering ‘how and why?’ questions, but do require control over behavioural events (e.g. which portion of a population will receive medication and which will not). Both surveys and experiments focus on contemporary events. Quasi-experimental methods are used when an experimenter cannot manipulate behaviour but in which the logic of experimental design still may be applied. However, when applying such methods to organisations, rather than just individuals, selection bias has to be dealt with large numbers of observations and advanced statistical techniques such as Propensity Score Matching.

In evaluation situations, randomized field trials have been put forward by some scholars (e.g. Boruch and Foley, 2000) to be used in virtually all evaluations, even complex community initiatives. Following the design of laboratory experiments in field trials designs allows for strong causal inferences to be drawn according to some proponents, but others caution against this. Quite often development interventions seek to achieve community level outcomes (not just at the individual level), but most field experiments would not be able to support the participation of a sufficiently large number of communities to overcome the severity of the subsequent statistical constraints' (Yin, 2014, p13). There is also a series of practical concerns, the best known being spill-over effects (when the 'no-treatment' sites adopt components of an intervention of interest before the end of the field experiment) and changes required at a system level within the experimental community creating variability (Yin, 2014, p14).

Robert Yin is one of the foremost writers about case study research and provides clear guidance to researchers seeking to improve their understanding. He suggests that case study research is defined by its scope and its main features. See box 2 below.

Box 2 Defining case study research: scope and features

A two-fold definition of a case study encompasses its particular scope and key features

In terms of its **scope**:

- 'A case study is an empirical inquiry that:
 - o Investigates a contemporary phenomenon (the 'case') in depth and within its real-world context, especially when
 - o The boundaries between phenomenon and context may not be clearly evident'

Experiments set out to separate a phenomenon from its context, and it is represented by only a few variables. Survey designers can try to tackle a phenomenon and its context, but the number of items in a questionnaire always has to be limited to fall safely within the allotted degrees of freedom (usually constrained by the number of respondents who are to be surveyed). In contrast case studies can tackle 'real-world cases' and contextual conditions are very much part of the analysis. They also tackle contemporary events, whereas histories are looking at the past. Case studies have the following **features**:

- A case study inquiry:
 - o Copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result
 - o Relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result
 - o Benefits from the prior development of theoretical propositions to guide data collection and analysis.

Source: Yin (2014, p16-17)

There are different applications of case study research in evaluation, where commissioning agencies may set the scope of the study (Yin, 2014).

- To explain presumed causal links in real world interventions that are too complex for survey or experimental methods;
- To describe an intervention and the real-world context in which it occurred;

- To illustrate certain topics within an evaluation (also descriptive)
- To enlighten those situations in which the intervention being evaluated has no clear, single set of outcomes.

It is possible to conduct single or multiple case study research. (Some authors suggest that comparative case methods are a subset of the latter).

But one of the key differences in the recent literature is whether case study research is about single cases, and studies (single or multiple) which use a standardized set of research questions, with a theoretical basis and seeking to cumulate information and make causal inferences.

Case study research is defined by some as the **internal examination of single cases** (George and Bennett, 2005; Stern et al, 2012). The case study method is a kind of research that ‘concentrates on one thing, looking at it in detail, not seeking to generalize from it. When you do a case study, you are interested in that thing in itself, as a whole’ (Thomas, 2011, p3). Any phenomenon can be studied (e.g. a person, group, institution, event, period in time, relationships amongst students etc), but it is about one unique thing and so generalization is not possible and nor is sampling as used in other research strategies, except in the sense of finding the subject of the study and justifying that choice (Thomas, 2011). A choice or selection of cases is made, but not a sample, and there is no expectation that it represents a wider population. If the inquiry were repeated by different people at a different time, then similar findings would result. **The ‘quality’ of a case study depends less on ideas of sample, validity and reliability and more on the conception, construction and conduct of the study** (i.e. the initial idea, the ways that you choose your case, the thoroughness with which you describe its context, the care you devote to selecting appropriate methods of analysis and the nature of the arguments you deploy in drawing your conclusions’ (Thomas, 2011, p?). The strength of the case study is being able to ‘drill down further’ and to create a ‘three dimensional picture’ (Thomas, 2011). Multiple studies can be conducted and comparisons drawn, but this still does not represent any kind of representative sample and the aim is still to delve deeper and to take a holistic approach. One of the key strengths of case studies is the closer connection to reality that is enabled by eschewing methodological formulae and instead focusing on a critical, creative approach to problem solving (Thomas, *ibid*).

‘clearly we cannot use just one person’s experience or a single set of uncorroborated observations as the basis for [generalisation]. That is why such trouble has been taken in many kinds of research to establish the sample as being representative of the wider population. If we want to generalise we need to make the basis of our generalisation clear: what gives us grounds to make the claim that we can generalize? What we are talking about with a case study, though, is a different kind of inquiry from those where generalisation is expected to follow. Here we are talking about understanding why and how something might have happened or why it might be the case. The assumption in a case study is that, with a great deal of intricate study, looking at our subject from many and varied angles, we can get closer to the ‘why’ and the processes of inquiry and analysis differ from those in many kinds of social research (for we are not working from samples that enable us to generalise), assumptions about the use of evidence do not change. With a case study there is still the assumption that we must collect good evidence and lots of it’ (Thomas, 2011).

The Table 1 below shows the differences according to Thomas (2011) between case study research and experiments and surveys.

Table 1: Case study compared to other forms of enquiry

	Case study	Experiment	Survey
--	------------	------------	--------

Investigates...	One case or a small number of cases	A relatively large number of cases	A relatively large number of cases
Data collected and analysed about...	A large number of features of each case	A small number of features of each case	A small number of features of each case
Study of...	Naturally occurring cases where the aim is not to control variables	Cases where the aim is to control the important variables	Naturally occurring cases selected to maximise the sample's representativeness of a wider population
Quantification of data	Is not a priority	Is a priority	Is a priority
Added by G. Thomas...			
Using...	Many methods and sources of data	One method	One method
Aiming to..	Look at relationships and processes	Look at causation	Look for generalisation

Source: Thomas, 2011, p10..adapted from..case study compared to other forms of enquiry (Hammersley and Gomm, 2000)

In another school of thought, the **'strongest means of drawing inferences from case studies is the use of a combination of within-case analysis and cross-case comparisons within a single study or research programme** (although single-case studies can also play a role in theory development)' (George and Bennett, 2005). By employing certain designs, there can be systematic causal analysis of cases (often viewed as complex systems) (Stern et al, 2012). Instead of placing the focus on causal analysis of variables taken out of their specific context, some scholars now focus on variables in the context of the case (Stern et al, 2012). Currently, in evaluation there is limited use of this need opportunity for causal analysis². In this approach, the aim is not to **try to 'universalize' the findings, but there is interest in generalizing beyond the single case'** (Byrne, 2009 cited by Stern et al, 2012, p. 28) This is done by making generalizations 'under certain conditions' and 'identifying clusters or subsets of cases about which it is possible to make similar causal inferences'.

Case study research is only one sub-set of qualitative research – the former involves theory development prior to the conduct of any data collection, whereas other methods such as ethnography and grounded theory set out without prior theoretical development (Yin, 2014). Similarly, George and Bennett (2005, p.18-19) distinguish between case study methods and qualitative methods, with the latter being about 'cumulative and progressive generalizations about social life and seeks to develop and apply clear standards for judging whether some generalizations fit the social world better than others' (a relatively positivist approach). But they also say that comparative methods are different from case study research (with the latter involving both 'within-case analysis and comparison across a small number of cases').

3.2 Strengths and weaknesses

The strengths and weaknesses of case study research depend upon your approach to case study research itself. There is a wide range of discussion on this in the literature (refs, e.g. Colorado table below), but it is most useful to concentrate on recent literature, because there has been significant innovation in this field.

Table 2 Patterns Strengths & Weaknesses of Case Studies (summarized from Colorado State University Guidelines)

Strengths

² Proponents include George and Bennett (2005), Ragin (2008) and Byrne and Ragin 2009).

Flexibility: Comparatively flexible method of scientific research

Better able to deal with creativity, innovation and context compared to statistical analysis which is better for homogeneity and routine
 Less prescriptive as exploratory rather than predictive. More freedom to address issues as they arise, rather than needing to identify all possible outcomes prior to the experiment

Emphasis on context

Gathering as much data about a single or small number of subjects, information on particular contexts = 'Deep data' or 'thick description'. Gives data a more human face and concrete aspect, and can complement more abstract, statistical data.

Weaknesses**Inherent subjectivity**

Criticized as too subjective in implementation presentation and evaluation of case study research. Some even say it is pseudo-scientific. Researchers seen as having deviated from their academic disciplines and having insufficient precision (i.e. quantification), objective and rigor (Yin, 1989). Relies on personal interpretation of data and inferences, results not generalizable, difficult to test for validity, rarely offer a problem-solving prescription. Reliance on one or two subjects for cognitive extrapolations risks inferring too much from what might be circumstance.

High investment

Too detailed for many large-scale research projects which look at a subject pool of many thousands. Level of resources required for in-depth analysis can be high compared to large scale surveys collecting less in-depth data.

Ethical considerations

In many educational case studies these are financed by people who have, either directly or indirectly, power over those being studied and those conducting the investigation. This conflict of interest can hinder the credibility of the study. The personal integrity, sensitivity and possible prejudices/biases of researchers need to be considered. Personal biases can creep into how the research is conducted, alternative methods use, and the preparation of surveys and questionnaires. Criticism that researchers change direction during a study and leave unknown gaps and biases, and should report preliminary findings to reduce possible bias.

Reliability, validity and generalizability

Many attacks made based on a lack of reliability, validity and generalizability. Responses should include (Merriam 1985): i) prolonging data gathering processes on site to improve accuracy of findings based on concrete information; ii) triangulation e.g. using a variety of data sources, case cluster method; iii) conduct member checks (i.e. talk to your subjects, for active corroboration); iv) collect referential materials (e.g. collect additional documental support); v) Engage in peer consultation (prior to composing the final draft of the report, researchers should consult with colleagues in order to establish validity through pooled judgement. It may not be possible to tackle generalizability issues, but case study research can be judged credible and confirmable, if not valid and reliable (Merriam, 1985).

Becker et al, (1994-2012)

In the 1960s and 1970s the distinction was drawn by most scholars between the study of a small number versus a large number of instances of a phenomenon, with case studies characterized as being 'small-n' studies in contrast to 'large-N' statistical studies. In other words the critical difference being employed is the number of cases – which means that the latter are always considered the ideal and inherently better – yet this **prism of statistical methods obscures the advantages of case studies and other methods in answering particular types of questions** (George and Bennett, 2005). Increasingly, there is a school of thought that promotes case study research on its own terms and suggesting that it can be very useful particularly as complement to other research strategies, including in relation to systematic causal analysis³.

George and Bennett (2005) write extensively on how **case study research is valuable for testing hypotheses and theory development** (precisely where statistical methods and formal models are

³ Earlier definitions presented the case as a 'phenomenon for which we report and interpret only a single measure on any pertinent variable' (David Eckstein, p85 cited by George and Bennett, 2005), but this has been widely rejected and encourages a misapplication by scholars of the 'degrees of freedom' problem of statistical methods, with the conclusion that case studies provide no basis for evaluating competing explanations of a case.

weak), because of their potential for addressing causal complexity, achieving high conceptual validity, their strong procedures for fostering new hypotheses and their value as a useful means to closely examine the hypothesized role of causal mechanisms in the context of individual cases. They discuss or respond to identified weaknesses of case study research: case selection bias; identifying scope conditions and 'necessity'; lack of representativeness; single case research designs; and potential lack of independence of cases.

3.3 Evaluation questions and different research strategies

In evaluation different questions may be asked and each implies different assumptions, requirements and methodologies. Stern et al, (2012) suggest some of the different research strategies which could be employed to answer the different evaluation questions including the place of case study research.

Table 3 Different Causal Patterns

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

3.4 Types of case study research

There is an extensive literature with many studies seeking to provide different categorizations of types of case studies. One of the best known typologies distinguishes between descriptive, exploratory and explanatory case studies. See Table 4 below.

Table 4 Three types of case study

	Case study		
	Descriptive	Exploratory	Explanatory
Number of cases	One	One	More than one
Aim	The primary interest is the case itself	Give insights into an issue or refine a theory	Test a theory or a framework
Scope	In-depth understanding of a particular case	Understand a particular case in order to help understand a wider issue.	Compare cases to identify causes and explain outcomes

Source: Adapted from Stake (1994) and Yin (2009).

A **descriptive case study** presents a complete description of a single phenomenon within its context. The primary research objective is to understand more about the case itself, because it is exceptional, or atypical in a way that makes it interesting. For this reason, it has also known as an **intrinsic** case study (Stake, 1994) since the main interest is on the features of this particular case that make it interesting or unique.

Example: the Colombian coffee-growers organisation CENICAFE is a descriptive case study of a remarkable farmers' organisation. This case study describes CENICAFE as a "a textbook example" of farmer participation and agricultural sustainability".⁴ (Orr, 2009).

An **exploratory case study** is where the research objective is to give insights into an issue or to refine a theory. Here the case itself is secondary. The primary research objective is to use the case to understand more about a particular problem or issue. For this reason, it is also known as an **instrumental** case study (Stake, 1994) since it is interesting not just for its own sake but for the light it sheds on a wider issue.

Example: Helen Todd made case studies of 20 women who had been members of the Grameen Bank for 10 years in order to test "some of the theoretical positions that have been taken on the Grameen Bank". Specifically, the case studies explored the view that credit alone was not enough to alter women's position in the absence of other inputs.⁵ (Orr, 2009).

An **explanatory or collective case study** is one where the research objective is to identify causes and explain outcomes. Typically this involves using several cases and drawing comparisons between them. (Another name for this approach is the **cross-case comparison**). Here the main interest is in

⁴ J. W. Bentley and P. S. Baker (2000). The Colombian Coffee Holders Federation: Organised, Successful Smallholder Farmers for 70 Years. AGREN Network Series No. 100. January. http://www.odi.org.uk/networks/agren/papers/agrenpaper_100.pdf

⁵ Helen Todd (1996). *Women at the Center: Grameen Bank Borrowers After One Decade*. (Dhaka: University Press Limited).

using the case studies to test a theory or generalisation and reaching general conclusions about the subject of study.

*Example: an explanatory case study by ODI compared 16 NGOs in Asia and Africa to reach general answers to three questions: How good are NGOs at promoting technological innovation and addressing constraints to change in peasant agriculture? How effective are NGOs at strengthening grassroots and local organizations? How do donor pressures influence NGOs and their links to the state?*⁶ (Orr, 2009).

However, there is not a clear consensus, with different authors proffering different categorizations of case studies. Thomas (2011) provides a table summarizing some of these.

Table 5 Kinds of case studies, as defined by different authors

Author	Merriam (1988)	Stake (1995)	Bassey (1999)	de Vaus (2001)	Mitchell (2006) drawing on Eckstein, 1975)	Yin (2009)
Categories of case study	Descriptive	Intrinsic	Seeking a theory	Descriptive/explanatory	Illustrative	Critical
	Interpretative	Instrumental	Testing a theory	Testing or building a theory	Social analytical	Extreme or unique
	Evaluative	Collective	Storytelling	Single or multiple causes	Extended (over time)	Longitudinal
			Drawing a picture	Holistic, embedded	Configurative, idiographic	Representative Revelatory
			Evaluative	Parallel or sequential	Disciplined, configurative	
				Retrospective or prospective	Plausibility probes	

As explained in the previous section, some authors focus on case study research as a means of systematic causal analysis. George and Bennett (2005) set out the following kinds of case study research:

- **Atheoretical/configurative idiographic case studies** (descriptions, but which may contribute to later studies that build theory)
- **Disciplined, configurative** – use of established theories to explain a case (e.g. for pedagogical purposes)
- **Theory testing** (testing validity and scope conditions of single or competing theories)
- **Heuristic** – identify new variables, hypotheses, causal mechanisms or paths. Deviant and outlier cases useful here
- **Plausibility probes** (preliminary study on relatively untested theories)
- **Building block studies** of particular types or subtypes of a phenomenon. Identify common patterns or serve a particular kind of heuristic purpose.

Beach and Pedersen (2005) distinguish between theory testing, theory building and explaining outcomes in process tracing research.

3.5 Case study design

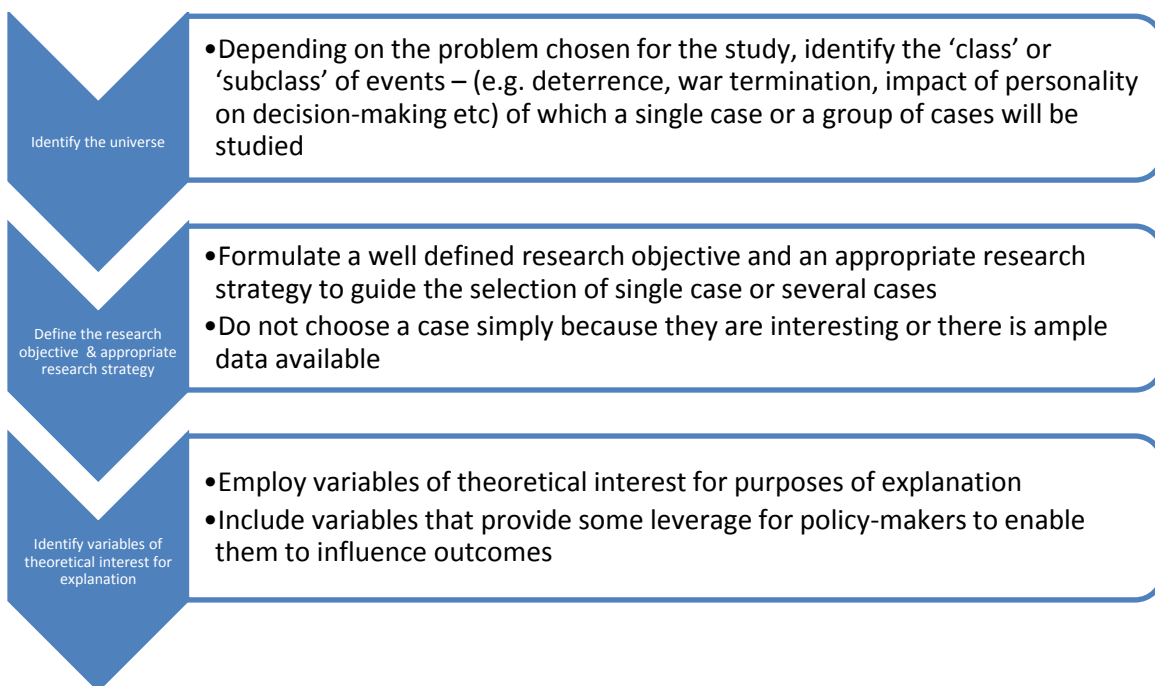
As explained in the previous sections there are different designs depending on the purpose of the study and the fundamental logic chosen.

⁶ J. Farrington and A. Bebbington with K. Wellard and D. J. Lewis (1993). *Reluctant Partners? Non-Governmental Organisations, the State, and Sustainable Agricultural Development*. (London: Routledge).

To allow for structured, focused comparison (whether an individual case or with the expectation that later cases will be added) George and Bennett, 2005 suggest a method of structured, focused comparison. This approach is likely to be useful for standard systems seeking to use case studies in their M&E systems to understand impact better.

The research is **'structured'** when researchers carefully write general, standardized questions reflecting the research objective and theoretical focus of the inquiry and these questions are applied to each case study to guide and standardize data collection, and thus making systematic comparison and cumulation of the findings of each of the cases possible. The **focus** in the study stems from having a specific research objective in mind and a theoretical focus appropriate for that objective – which shape data requirements. A single study cannot address all the interesting aspects of a historical event, for example, - a single event can be relevant for research on a variety of theoretical topics (George and Bennett, 2005). The overall aim of this approach is to be able to **produce generic knowledge on important policy problems and to draw the explanations of each case of a particular phenomenon into a broader, more complex theory**. The steps which are critical to achieve structured, focused comparison are set out in figure 1. below. Whatever the research strategy employed, George and Bennett (2005) advise that comparative case studies should be **centrally planned and conducted (ideally) or with strong central coordination of collaborative studies**.

Figure 1 Initial design process for case study research



Source: Adapted from George and Bennett, 2005

In other approaches scholars seek to conduct a single case study for description or as part of the naturalistic or interpretative tradition, but for those that conduct studies for causal analysis it is thus important to identify the research puzzle or problem, develop research questions and the overall theory underpinning the analysis. Studies oriented towards causal analysis can include single or 'within case' analysis (to establish configurations), but also cross-case comparison (Stern et al, 2012). Whether conducting a single case study or multiple, standardized questions should be used to allow

for future studies to be comparable and to cumulate knowledge - even if they are conducted by different researchers at a different time. Multiple cases require significant time and resources.

In conducting multiple case study design with an aim of systematic causal analysis, common questions that arise are: 'How should the cases be selected?' and 'How many are needed?' There are some fairly complex questions to this – for example, see George and Bennett, (2005) and Beach and Pedersen (2013).

Yin (2014) distinguishes between ***analytic generalization*** and ***statistical generalization***. **The former is used in case study research to generalize beyond the level of the specific case, but not to universalize (as in statistical generalization) to the whole of a population.** In a single study Yin suggests that theory is employed and in multiple or cross-case analysis there is the use of replication logic to guide selection

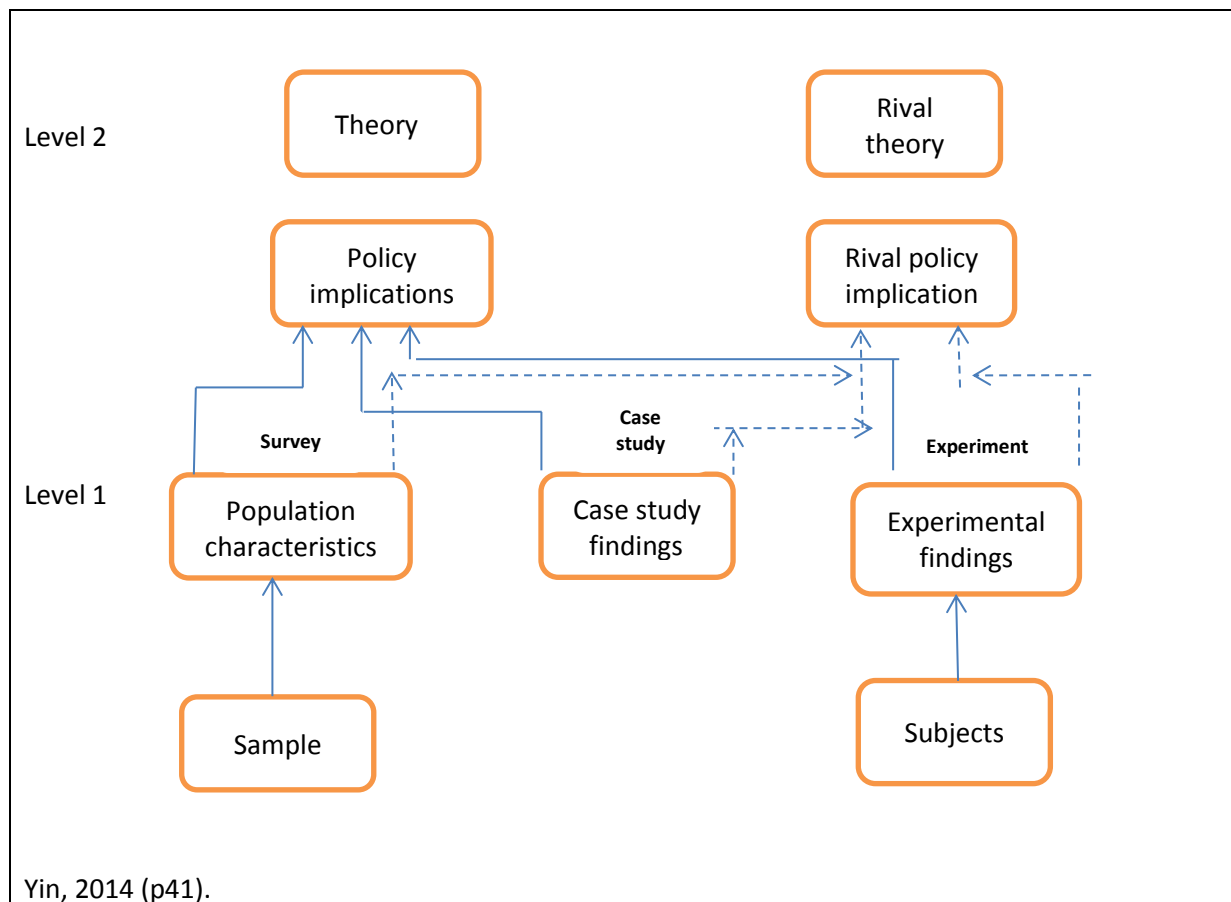
Box 3 Statistical and analytic generalization

Statistical generalization: An inference is made about a population (or universe) on the basis of empirical data collected from a sample of that universe... This method of generalizing is commonly followed when doing surveys...A fatal flaw in doing case studies is to consider statistical generalization to be the way of generalizing the findings from your case study. This is because your case or cases are not 'sampling' units and also will be too small in number to serve as an adequately sized sample to represent any larger population.

Analytic generalization is based on either: a) corroborating, modifying, rejecting or otherwise advancing theoretical concepts that you referenced in designing your case study; , or b) new concepts that arose upon the completion of your study.

Source: Yin, 2014, p41

Figure 2 Making Inferences: Two Levels



Replication as an approach to generalization (i.e. external validity) essentially the same approach that is used in multiple experiments – i.e. a **‘replication’ design rather than a sampling design is used**. In multiple experiments when a scientist uncovers a significant finding from a single experiment, they conduct second, third or fourth studies. Some of these replications may attempt to duplicate the exact conditions of the original experiment, but others might alter one or two experimental conditions considered unimportant to the original finding, to see whether the finding could still be duplicated. Only with such replications would the findings of the original experiment be considered robust. The same logic holds for multiple case study research (Yin, 2014, p57).

Cases are carefully selected so that they provide literal replication (predicts similar results) or theoretical replication (predicts contrasting results but for anticipatable reasons) (Yin, 2014, p57). If it is possible only to conduct two or three studies, this is more likely to be for literal replication, whereas if six to ten are conducted, then there is a possibility of including different patterns. If all the cases turn out as predicted, these 6 to 10 cases, in the aggregate, would have provided compelling support for the initial set of propositions. If the cases are in some way contradictory, the initial propositions must be revised and retested with another set of cases’ (Yin, 2014). The important thing according to George and Bennett (2005) is that this process ‘cumulates’ – i.e. it builds up evidence. **The selection of cases in this replication procedure should reflect a theoretical interest**. This is very different to a sampling logic which is not appropriate for case study research, because: a) case studies are not the best method of assessing the prevalence of a phenomenon; b) a case study would have to cover both the phenomenon and its context, with a large number of

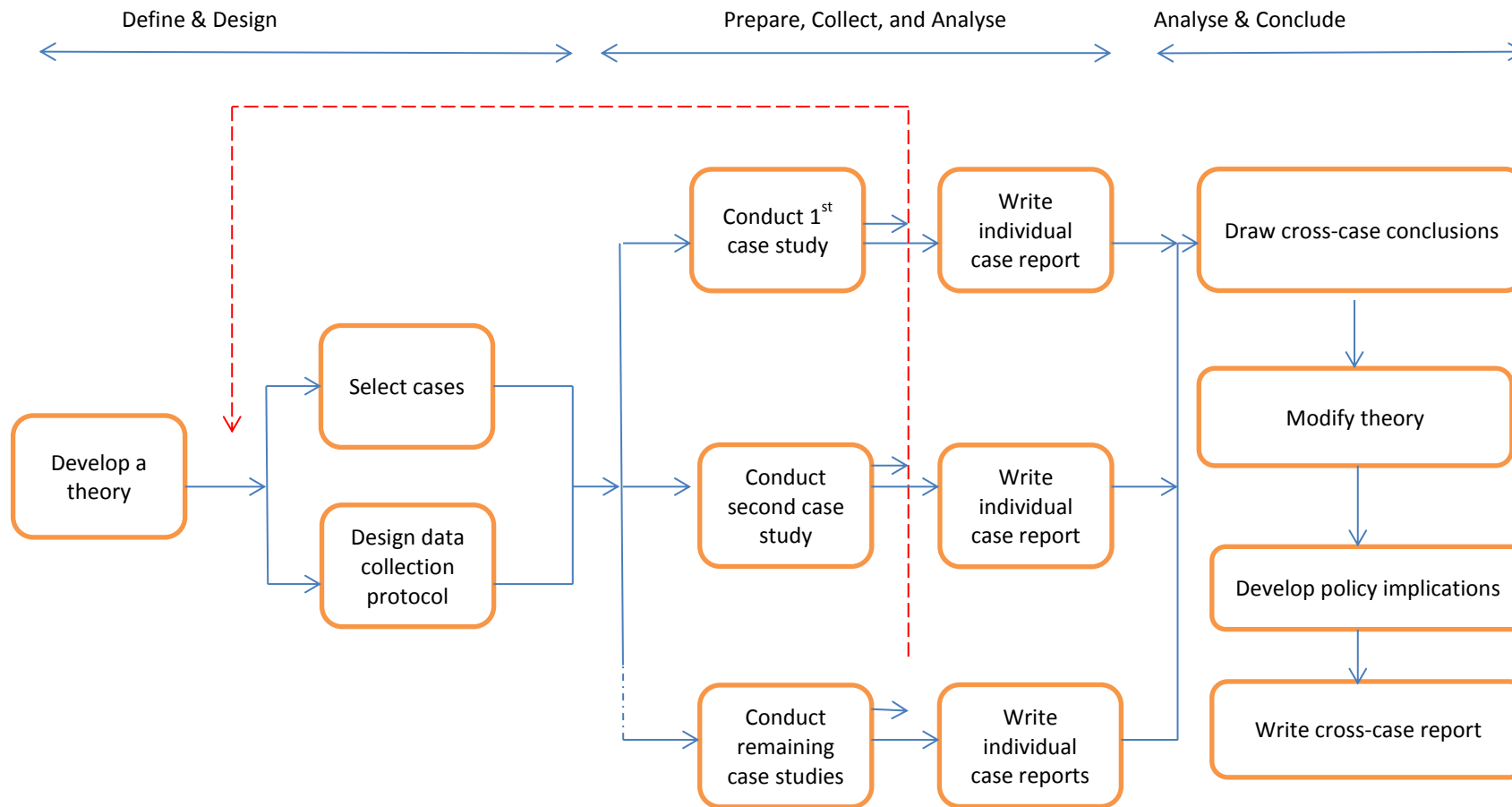
potentially relevant variables, requiring a large sample of cases which would only allow for superficial investigation; and c) some topics cannot be empirically investigated using a sampling logic. See Figure 3 below. The actual number of cases is therefore discretionary, and the stronger the possible rival explanations the more cases that could be included (Yin, 2014, p61).

To achieve a high quality case study design means responding to four tests (Yin, 2014, p47):

- i) **construct validity** (identifying correct operational measures for the concepts being studied);
- ii) **internal validity** (seeking to establish a causal relationship, whereby certain conditions are believe to lead to other conditions, as distinguished from spurious relationships);
- iii) **external validity** (defining the domain to which a study's findings can be generalized);
- iv) **reliability** (demonstrating that the operations of a study – such as the data collection procedures – can be repeated with the same results).

To achieve construct validity involves identifying specific concepts and operational measures. Further, collecting multiple sources of evidence, establishing a chain of evidence and engaging with others to review the draft case study report are three tactics suggested by Yin (2014) to ensure construct validity. Internal validity issues are relevant in explanatory studies where the aim is to establish causal relationships. Four analytical tactics can be used to try and tackle internal validity threats – pattern matching, explanation building, addressing rival explanations and using logic models. For standard systems looking to use case studies to strengthen their understanding of how and why certification leads to impact, this might include: Pattern matching – e.g. if the predicted values for the predicted outcomes of a standard system have been found and alternative values for alternative outcomes have not been found, this provides the basis for strong causal inferences to be made.

Figure 3 Multiple Case Study Procedure



Source: Cosmos Corporation, Yin, 2014, p60

3.6 Causation, attribution and contribution

A research design is shaped by the theory of causation being employed. There are a number of different theories of causation, with regularity frameworks and counterfactual logics being the most well-known.

In statistical approaches, ‘regularity frameworks’ are used, and this requires large numbers of observations, but they are not good at dealing with contextualisation or explaining how and why questions. ‘Counterfactual frameworks’ are employed in experiments, and are strong on internal validity (i.e. whether an intervention has made a difference or not), but they are not good at dealing with contextualisation or at answering questions relating to generalization. ‘Generative causation’ is the approach used in ‘realist evaluation’ (e.g. Pawson and Tilley, 2004), in which the mechanisms that explain effects are identified. These are strong on explanation but weak on estimating the extent of impact. Finally, ‘multiple causation’ frameworks are used to identify multiple causes that lead to an effect (known as ‘configurational’ approaches) (Stern et al, 2012). See table 6 below.

Table 6: Different theories of causation and their strengths and weaknesses

Different approaches	Theory of causation	Requirements, Strengths & Weaknesses - Examples
Regularity frameworks	the frequency of association between cause and effect – the inference basis for statistical approaches to IE	Requires high numbers of diverse cases to capture sufficient diversity (or difference). Work best when causal factors are independent. Not good at dealing with contextualisation.
Counterfactual frameworks	the difference between two otherwise identical cases – the inference basis for experimental and quasi-experimental approaches to IE	Experiments are good at answering the question – has this particular intervention made a difference here. But weak on generalisation. Work best when causal factors are independent Not good at dealing with contextualisation.
Multiple causation	combinations of causes that lead to an effect – the inference basis for ‘configurational’ approaches to IE	Good at dealing with limited complexity and interdependence, but not at unpicking highly complex combinations
Generative causation	identifying the ‘mechanisms’ that explain effects – the inference basis for ‘theory based’ and ‘realist’ approaches to IE	Strong on explanation but weak on estimating quantities or extent of impact

Source: Stern et al, 2012

Table 7 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.

Source: Stern et al, 2012

Contribution Analysis is an approach which allows for an assessment of a project or programme to a particular outcome. It helps to unpack what difference an intervention has made or is making and what is the relative role of other interventions and external factors. This is simpler to implement where a theory of change has already been articulated, and it provides a platform for confirming (providing evidence and a line of reasoning from which a plausible conclusion can be drawn) or revising the assumptions embedded within it – although it does not necessarily help to identify previously unidentified or rival explanation frameworks. There are six key steps (Mayne, 2008): i) Set out the attribution problem to be addressed; ii) Develop a theory of change and risks to it; iii) Gather the existing evidence on the theory of change; iv) Assemble and assess the contribution story and challenges to it; v) Seek out additional evidence; vi) Revise and strengthen the contribution story. This approach is useful because of the complexity involved in assessing impact in relation to standard systems, with their multiple causes and outcomes, and because there is increasing incidence of multiple certification.

3.7 Analytical strategies and techniques

Part of a rigorous case study design, lies in the approach to analysis of evidence. Yin (2014) usefully outlines four general strategies: i) relying on theoretical propositions; ii) working the data from the ‘ground up’; iii) developing a case description; iv) examining plausible rival explanations.

Table 8 Analytical strategies

Analytical strategy	
Relying on theoretical propositions	Use the theoretical propositions that led to your case study. The original objectives and design of the case study were based on theoretical propositions, which reflected a set of research questions, literature review and new hypotheses or propositions.
Working the data from the ‘ground up’	An inductive strategy, with key concepts emerging from a close examination of the data, not from prior theoretical propositions. (See guidance from proponents of ‘grounded theory’, Corbin and

	Strauss, 2007; Glaser and Strauss [1967] where codes are assigned to the data, each code represents a concept or abstraction of potential interest.
Developing a case description	Another useful strategy (also one that can be employed if the first two are not feasible/working) is to organise the case study against a descriptive framework. Ideas may emerge from an initial literature review, to influence the design of data collection instruments.
Examining plausible rival explanations	Define and test plausible rival explanations – can be done in combination with the other three analytic strategies. Initial theoretical propositions (strategy 1), working from the ground up (strategy 2), may produce rival inductive frameworks; and case descriptions (strategy 3) may involve alternative descriptions of the case. There are various types of rival explanation.

Adapted from Yin (2014)

Where studies have been designed based on clear theoretical propositions – such as a theory of change, as ISEAL members have been developing – it is possible to synthesize all the evidence from different sources to compare against the original propositions.

Another increasingly used strategy is the examination of rival plausible explanations. A frequent hypothesis in evaluations are that the ‘observed outcomes are the result of a planned intervention. The simple or direct rival explanation would be that the observed outcomes were in fact the result of some other influence besides the planned intervention and that the investment of resources into the intervention may not actually have been needed’ (Yin, 2014, p140). Thus, awareness of the direct rival is needed early on to ensure data collection includes the collection of evidence on it – and this should be vigorous rather than so limited as to actually being more of a reason to reject them (Patton, 2002, cited by Yin, 2014). There are two types of rival explanations – those that are essentially issues with the implementation of the study – or ‘craft’ rivals, and substantive or ‘real-world’ rivals, and the latter should be carefully identified *before* data collection. **The more rival explanations that the analysis tackles and rejects the greater the confidence there is in the findings** (See table 9 below).

Table 9

Type of Rival	Description of Examples
Craft Rivals:	
1. The Null Hypothesis	The observation is the result of chance circumstances only
2. Threats to Validity	e.g. history, maturation, instability, testing, instrumentation, regression, selection, experimental mortality, and selection-maturation interaction
3. Investigator bias	e.g. ‘experimenter effect’, reactivity in field research
Real-world Rivals	
4. Direct Rival	An intervention (‘suspect 2’) other than the target (Practice or Policy) intervention (‘suspect 1’) accounts for the results (<i>‘the butler did it’</i>).
5. Commingled Rival	Other interventions and the target intervention both (Practice or Policy) contributed to the results (<i>‘it wasn’t only me’</i>).
6. Implementation Rival	The implementation process, not the substantive intervention, accounts for the results (<i>‘did we do it right?’</i>
7. Rival Theory	A theory different from the original theory explains the results better (<i>‘it’s elementary, my dear Watson’</i>).
8. Super Rival	A force larger than but including the intervention accounts for the results (<i>‘it’s bigger than both of us’</i>).
9. Societal Rival	Social trends, not any particular force or intervention, account for the results (<i>‘the times they are a-changin’</i>).

Source: Yin, 2000b cited by Yin, 2014

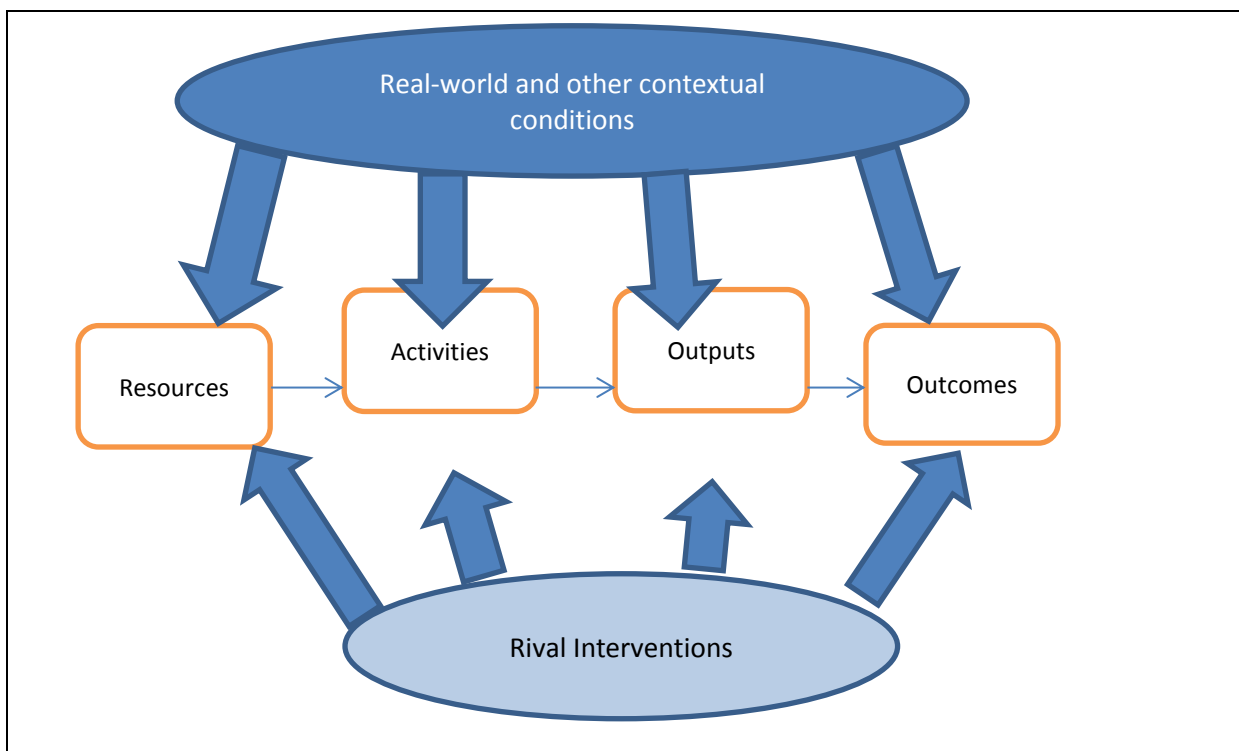
There are also different techniques which can be used to support analysis of the data collected. Five techniques can be employed as part of any of these analytic strategies. These techniques are important for tackling internal and external validity challenges. See box 4 below. The analytic techniques which can be used include: i) pattern matching; ii) explanation building; iii) time-series analysis; iv) logic models; v) cross-case synthesis (Yin, 2014, p142-170). For more details see appendix 2.

Box 4 Internal and External Validity

Internal validity refers to the strength of a cause-effect link made by a case study, in part determined by showing the absence of spurious relationships and the rejection of rival hypotheses. **External validity** refers to the extent to which the findings from a case study can be analytically generalized to other situations that were not part of the original study (Yin, 2014, p239).

Contextual factors may be external to the case selected for study, but it may have a strong influence on the intervention's outcomes. Neilson and Pritchard (2009) in their study of South Indian tea and coffee industries, also highlight that the process of implementation of sustainability standards is not one of imposition from above, but should more be seen as a '*co-production*' of outcomes.

Figure 4 Attending to contextual conditions and rivals



Source: Yin (2014, p163).

3.8 Process tracing

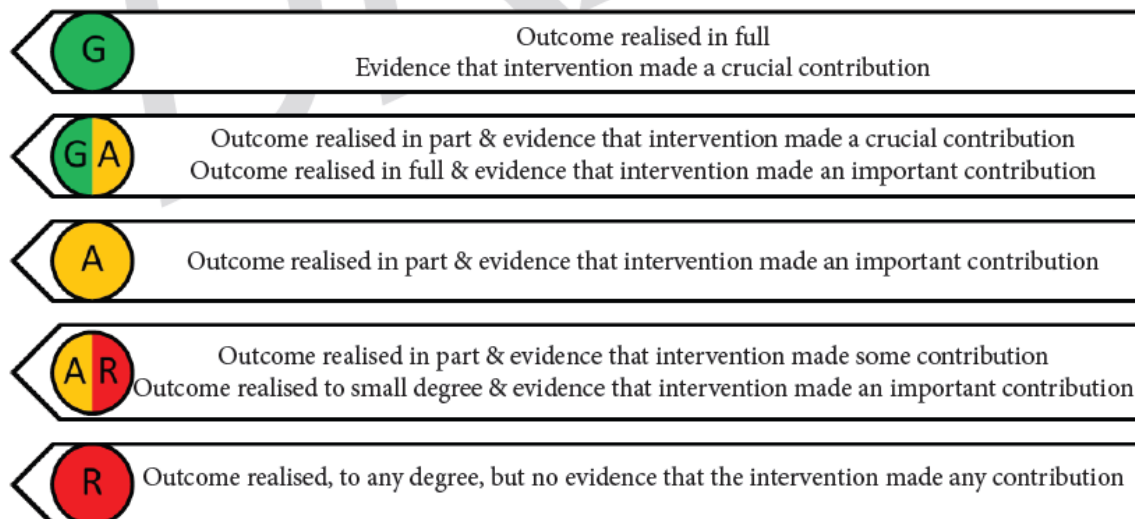
Process tracing methods 'are tools to study causal mechanisms in a single case research design' (Beach and Pederson, 2013, p2).

A straightforward outline has been provided by the International NGO Oxfam, which has been fairly active in developing an approach to assessing programme impact and effectiveness. They have developed a process tracing 'draft protocol', which provides some easy guidance on conducting process tracing. They identify some key steps – although significant iteration is needed between steps:

1. Undertake a process of (re)constructing the intervention's theory of change, in order to clearly define the intervention being evaluated – what it is trying to change (outcomes), how it is working to effect these changes (strategies/streams of activities) and what assumptions it is making about how it will contribute to these changes (key assumptions).
2. Work with relevant stakeholders to identify up to three intermediate and/or final outcomes considered by stakeholders to be the most significant for the evaluation to focus on (central to the interventions theory of change, and useful for learning/forward planning).
 - Systematically assess and document what was done under the intervention to achieve the selected targeted outcomes.
 - Identify and evidence the extent to which the selected outcomes have actually materialized, as well as any relevant unintended outcomes.
 - Undertake 'process induction' to identify salient plausible causal explanations for the evidenced outcomes.
 - Gather required data and use 'process verification' to assess the extent to which each of the explanations identified in step 5 are supported or not supported by the available evidence.
 - Write a narrative analytical report to document the above research processes and findings
 - Summarize aspects of the above narrative analysis by allocating project/campaign 'contribution scores' for each of the targeted and/or associated outcomes.

Source: Oxfam..ref

They also provide a scoring system which indicates how far an Oxfam programme has been successful or not in achieving its outcomes, combined with the level of evidence available.



There are also other more sophisticated texts providing guidance on process tracing – see George and Bennett (2005) and Beach and Pedersen (2013). The latter provide guidance on process tracing in social science as a method for studying causal mechanisms linking causes with outcomes. They are part of this new wave of social science scholars that are working with theories of causal mechanisms and linking causes with outcomes. Beach and Pedersen (2013) argue that process tracing is actually not just one method, but there are three different variants: a) theory-testing; b) theory-building; and c) explaining-outcome. The differences are explained in table 10 below. These are fairly sophisticated approaches. If the aim is to make strong ‘within-case’ inferences Beach and Pedersen (2005) suggest using process tracing methods, if the aim is cross-case inferences, they suggest using other methods including Qualitative Comparative Analysis or combinations of process tracing with comparative methods.

Table 10 Three variants of process tracing

	Theory-testing	Theory-Building	Explaining-Outcome
Purpose of analysis – research situation	<i>Situation one</i> – Correlation has been found between X and Y, but is there evidence that there exists a causal mechanism linking X and Y?	<i>Situation two</i> – Build a plausible causal mechanism linking X:Y based on evidence in case	<i>Situation three</i> – Explain particularly puzzling historical outcome by building minimally sufficient explanation in case study
Ambitions of study	Theory-centric	Theory-centric	Case-centric
Understanding of causal mechanisms	Systematic (generalizable within context)	Systematic (generalizable within context)	Systematic, non-systematic (case specific) mechanisms and case-specific conglomerates
What are we actually tracing?	Single, generalizable mechanism	Single, generalizable mechanism	Case-specific composite mechanism that explains the case
Types of inferences made	1. Parts of causal mechanism present/absent 2. Causal mechanism is present/absent in case	Observable manifestations reflect underlying mechanism	Minimal sufficiency of explanation

Source: Beach and Pedersen (2005, p21).

3.9 Complexity

Case study research is important for answering how and the why questions. Increasingly, Stern et al (2012) suggest that cases are seen as ‘complex systems’ in their own right. We find that it is useful to analyse the sources of complexity in a particular programme or project (i.e. in its attributes) to inform the design of research or evaluations. This is because some study topics may be more complex than others. For example, it has often been noted that some types of intervention, such as the distribution of malaria nets, are somewhat less complex to evaluate than those involving multiple types of intervention. For an analysis of sources of complexity undertaken in the preparatory stages of a Fairtrade coffee study see table 11 below.

Stern et al, (2012) argue that evaluation in development, increasingly involve complex situations and phenomena. It is important to recognize that the attributes of a particular programme (in combination with the particular evaluation questions in hand) have a bearing on the impact design. Thus situations of complexity may require different causality frameworks and thus choice of methods and tools.

To achieve an appropriate impact design for a Fairtrade coffee study we have explored sources of complexity in order to inform the impact design process. We found a wide range of sources of complexity which challenge the use of counterfactual logics and quasi-experimental techniques. Essentially, the FLO trajectory of change for POs and individuals may be stepped, it has central components, but also has variable and customized elements, has multiple pathways to impact, interdependencies (e.g. other certifications, other NGO partners to POs etc) and there are emergent properties (the FLO system evolves and changes over time, as does the context in which the standards are implemented). All of this adds to complexity and reduces the ability of an evaluation to create a robust counterfactual. We therefore have suggested the use of the theory of change not only to identify what has happened along the logical chain, but to explicitly explore with stakeholders rival explanations. We will conduct research with comparison groups to fully explore rival explanations and to avoid the common ‘positive bias’ of short-term qualitative studies that do not reach comparison groups. However, there may still be an element of survey work with counterfactual groups to provide answers to the ‘what’ and ‘how much’ type questions – i.e. not only mixing methods but also mixing **impact designs**.

FLO attributes are identified in the table below, along with implications for available designs for complexity in these dimensions (the latter according to Stern et al, 2012).

Table 11 Defining the attributes of FLO certified Fairtrade

Attribute	Explanation of attribute	FT characteristics	Design implications
Bounded or Embedded (less complex is bounded; embedded is more complex)	If you bound a system for analysis too narrowly you can miss the sources of influence and feedback that can lead to instability.	Fairtrade operates across a wide range of value chains and countries.	When programmes have relatively open boundaries thus need more whole system analysis, not partial sub-systems. High levels of uncertainty – need for more real time monitoring & impact tracking. Programmatic & contextual uncertainty & risk varies.
Linear causality & trajectories (Less complex is predictable and linear; more complex is unpredictable & non-linear)	Linear systems where outputs are proportionate to inputs and effects are understandable in terms of a preceding cause, but in social systems there are iterative and disproportionate properties. What does the impact trajectory look like? Is it steady, or could there be decrease before increase or a step-function? Need to make these explicit	FT producer support works with groups to achieve certification. There are continuing improvement features of FT as well.	Assess the 'trajectory' of intended change Ensure timing of evaluation matches this trajectory Monitoring systems and real time evaluation for rapid feedback
Level of standardisation & diversification (greater complexity where more diversified)			Assess relevance of local programme variant Input from programme stakeholders Data collection more complex should be participatory
Centrally specified vs locally customized		Standard systems have centralized standard content and auditing system, although there is variation within the standards in FT for different products and in audit implementation. AND there is local customization of inputs in i) the actual producer support provided from HQ or the liaison officer; ii) additional inputs and influence of FT plus actors such as CafeDirect; iii) growing markets and networking elements are variable and evolve (emergent properties)	Process track and compare different ToCs Meta-evaluations and comparisons of different programmes – QCA /metrical designs
Mechanisms are universal/different causal mechanisms are employed in different contexts		The main FT causal mechanisms remain the same but there is variance (networking, producer support etc). The standards vary sometimes by product, but at a broad level the essential mechanisms remain the same	Realist synthesis Develop typologies of contexts Study mechanisms in context – realist/mechanism-based designs Configurational designs, QCA (qualitative comparative

		– however, at different points in time some mechanisms may not be functioning (e.g. when the world market price exceeds the Fairtrade Minimum Price for an extended period)	analysis)
No of causal strands : single or few (less complex) versus multiple strands (more complex)	More causal strands (e.g. in comprehensive programmes of targeted programmes) there are more pathways to analyse & possible synergies and interactions (interdependencies)	FT has multiple causal strands	Process trace different causal strands & how they interact Identify alternative causal paths to same goal Identify 'blockages', virtuous and vicious cycles.
Homogeneous (less complex) vs nested (more complex) systems			Systems mapping to identify zones of greater or lesser complexity Differentiate designs accordingly
Pre-identified (less complex) versus emergent (more complex) effects	Programmes transform or evolve over time as the context changes. Development processes are inherently uncertain and change over time - (systems notion of emergence) Also some goals can be achieved more quickly than others so this requires an evaluative decision about best timing for an evaluation	FT standards change in a structured fashion with consultations and reviews at set periods in time. Sometimes the reviews lead to bigger changes than others (E.g. recent review of hired labour strategy and consultation on hired labour). Sometimes there are more far reaching changes – e.g. New Standards Framework 2011. Less standardized aspects such as producer support and growing markets are also more inherently emergent In terms of timescales – different impacts might be achieved on different timescales (e.g. income impacts might be felt quicker than overturning embedded inequitable gender relations)	Staged designs to adapt to evolving TOC System analysis & monitoring Real time evaluation to track how goals are redefined.

3.10 Multi-method collaborative research

Multi-method collaborative research is increasingly seen (e.g. Stern et al, 2012) as desirable. The complementarity of case studies, statistical methods and formal models is becoming more evident (George and Bennett, 2005). The different methods can be used in a single study or sequentially (George and Bennett, 2005):

- Statistical analysis can help to identify outliers or deviant cases, and case studies can then investigate why these cases are deviant, perhaps leading to the identification of omitted variables;
- Case studies can also explore the possible causal mechanisms behind the correlations or patterns observed in statistical studies, providing a check on whether correlations are spurious or potentially causal and adding details on how hypothesized causal mechanisms operate.
- Alternatively, when case studies lead to the specification of new variables or the refinement of concepts, statistical studies can explore whether these new variables and concepts are relevant to a larger population of cases;
- Formal models can be tested in case studies to see if their hypothesized causal mechanisms were in fact in operation and the variables and concepts developed through case studies can be formalized in models.

However, the levels of sophistication involved in all three, means that collaboration between researchers is probably needed to sustain the top level cutting edge theoretical and empirical knowledge – thus individuals may be specialists in one area but need to be conversant in alternative approaches, be aware of their respective strengths and limitations, and be able to make informed reading of their substantive results (George and Bennett, 2005). **Too often research strategies are inappropriately criticized and even dismissed, because in fact their sophisticated underpinnings and value are not appreciated.**

There is further debate to be had, however, about where and when such methods can be implemented in a participatory process or not, and to what extent (ranging along a continuum).

3.11 Rigour, utility and quality

The question of rigour in social science and evaluation is important, but also utility (Patton, 2008; 2010) and quality are important characteristics. The idea that case study research is not rigorous is ill-founded and misses the particular strength of case study research. Of course any research can be done badly, including experimental and quasi-experimental approaches, but it is important to judge what each approach can offer on its own merits. In case study research there are now sophisticated theory and techniques to ensure rigour.

Bamberger et al (2010) argue that **'rigor' is not determined solely by the use of a particular method, but rather the appropriateness of the 'fit' between the nature of the problem being assessed and the particular methods (singular or in combination) deployed in response to it, given prevailing time, political, financial, ethical and logistical constraints.** Both Bamberger et al (2010) and Stern et al (2012) point to the increased blurring between qualitative and quantitative approaches in impact studies. Participatory generation of numbers or quantification, for example, has been part of participatory research for many decades, but has not been seriously scaled up in most evaluations, although this might be beginning to change with new publications and some donors indicating greater interest. Similarly, software has been used to code qualitative data to

make it more manageable for analysis in ‘within case’ examples (Stern et al, 2012), but some recent innovations are attempting ‘across case comparisons’ using computer aided tools, such as Sensemaker⁷ and some attempts at participatory statistics that aggregate data gathered from individuals using PRA tools in group settings (Holland, 2013). Bamberger (ibid) provides a useful recasting of the distinction between ‘**data collected from structured, closed-ended questions and non-structured, open-ended modes of enquiry**’. How far different approaches are participatory is another –and important - question of course, being fundamental to international development in the 21st century.

Although some scholars argue that credible evidence can only be generated by experimental and quasi-experimental designs, the Stern et al (2012) paper is important, because it has been commissioned and published by DFID working paper and makes a clear case for the alternatives to such designs which may be different, but can have **equivalent robustness and credibility**.

There are many different aspects of rigour. Zelik, Patterson and Woods (2010) propose a ‘**rigor attribute model**’, which has eight different attributes (see box 5 below). Their approach stems from a recognition that a rigid adherence to prescribed standards can lead to a failure of ‘intelligence’. Risks of shallow analysis can be reduced where analysts initiate strategies opportunistically in the analysis process – although these may reduce efficiency, they increase accuracy. Their model sets out various attributes of a rigorous research design and provides relevant indicators for each of low, medium and high rigour. For example, for hypothesis exploration low rigour exists where there is no consideration of alternative hypotheses to high rigour where there is an analysis of alternative explanations via the direct evaluation of specific hypotheses, incorporation of external perspectives, and revising hypotheses as new data are collected (Zelik et al, 2010, p5). In relation to information synthesis, high rigour is achieved when the researchers are ‘reflexive’, i.e. ‘attentive to the ways in which their cognitive processes may have hindered effective synthesis’ (Zelik et al, 2010, p 9).

Box 5 The Rigor Attribute Model

- Hypothesis exploration (the construction and evaluation of potential explanations for collected data);
- Information search (the focused collection of data bearing upon the analysis problem)
- Information validation (the critical evaluation of data with respect to the degree of agreement among sources)
- Stance analysis (the evaluation of collected data to identify the relative positions of sources with respect to the broader contextual setting)
- Sensitivity analysis (the evaluation of the strength of an analytical assessment given possible variations in source reliability and uncertainty)
- Information synthesis (the extent to which an analyst goes beyond simple collecting data in ‘putting things together’ in a cohesive assessment)
- Specialist collaboration (the extent to which substantive expertise is integrated into an analysis)
- Explanation critique (the critical evaluation of the analytical reasoning process as a whole, rather than the specific details).

Zelik, Patterson and Woods (2010)

⁷ Sensemaking by the company Cognitive Edge has garnered attention, through the use of computer software to identify patterns in collected micro-stories which are then coded by participants using a framework of questions identified by the researcher.

However, as well as rigour, **utility is also an important attribute of evaluation activities**. Patton (2008) sets out an approach called 'utilization-focused evaluation' or UFE, which proposes that usefulness of findings and of the process itself to intended users (e.g. informing decisions and improving process) should be the guiding factor in design and implementation. Primary intended users should be identified and engaged from the start of the process to ensure their uses can be identified and this should frame all other decisions about the process. Evaluator's should not make decisions independently of the intended users, but should facilitate decision-making amongst them, because this builds the groundwork for use – i.e. it builds ownership and increases the likelihood that the findings will be used (Patton, 2008). In a checklist of key steps, Patton (2013) outlines key steps (see appendix 3) and provides practical guidance (see also Patton, 2012).

Irene Guijt (blog post 'battlefields of rigour', March 29 2012⁸) argues that instead of focusing only on rigour, evaluators should take into account other dimensions of *quality* in evaluation. Amongst some groups of evaluation professionals, 'accuracy' wins out over 'utility', and in other cases it is the reverse. Accountability is often seen as being incompatible with learning, because it is usually seen as involving pre-defined deliverables, but argues that this should not be the case – particularly in situations of complexity which demand individuals and organisations to have 'respond-ability' (Guijt, undated). Stern et al (2012, p46) also suggest that accountability and learning should 'not be seen as alternative evaluation purposes'. They identify a **contradiction between the attention paid to causal questions by donors aimed at learning how and why development interventions make a difference (e.g. accountability for development assistance demands and activities) and donors that seek to 'ensure effectiveness and Value-for-Money – and to learn from specific decentralized experiences to inform their broader policies'**. The answer to this contradiction is participatory designs that engage the recipients of aid in an evaluation as part of mutual accountability and to support learning (Stern et al, *ibid*).

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. 'Joint evaluations' between donors and partner countries as a form of participatory design is proposed and participatory approaches that involve stakeholders in programme implementation and learning through evaluation. *Hybrid* designs are possible 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....**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'** (Stern et al, 2012, p46-47).

To answer the question about whether initiatives can be replicated elsewhere thus requires four things: i) A joint commitment to learning and mutual accountability; ii) Identification of different implementation strategies, that could be expressed as 'causal packages', identified in dialogue with partners; iii) Generic understanding of contexts e.g. typologies of context so that extension/scaling up can be systematically planned; iv) Building in innovation and diffusion processes so that successful 'configurations' in different contexts could be extended (Stern et al, 2012, p48).

⁸ <http://bigpushforward.net/archives/1432>

However, the increased demand for certain types of evidence, which interprets evidence as controlled comparisons and experimental or mixed methods undoubtedly creates distance between researchers and intended 'beneficiaries'. It is important for organisations that claim to promote farmer participation and empowerment to fit their practice with this stated goal and their ethos.

4. Practical Guidance

This section draws heavily on the guidance of Robert Yin (2014), 'Case Study Research: Design and Methods', as well as the other scholars highlighted in the theory section above. It also draws on our own experience of conducting case study research for standard systems and in other settings, and from the examples of the three standard systems that participated in this research. Given that many standard systems will be using case studies for impact assessment, it is also useful to gather up-to-date guidance on evaluation (See for example, 'Better Evaluation' Rainbow Framework for basic guidance materials, 2013). Appendix 4 provides a visualization of the case study process proposed by George and Bennett (2005).

4.1 When do I use case study research?

There is no single right answer to this – it depends on your purpose and research questions (and in reality, to some extent, your worldview). To choose a research strategy, first of all requires classification of the type of research question you have. If your questions focus on the 'how' and 'why', then case study research is likely to be relevant. Further, if there is a need for more 'in-depth' analysis, then again case study research is likely to be appropriate.

- Case study research is useful to answer 'how' and 'why' questions, particularly, when there is no control of 'behavioural events' (e.g. when a group becomes certified).
- A survey might be more appropriate for answering questions about 'who, what, where, how many, how much' (e.g. how many producers have benefited from increased incomes, how much have their incomes risen). It does not require control of behavioural events.
- An experiment is useful for answering questions about 'how and why', but you have to be able to control events. This is unlikely to be relevant to standard systems, where private sector actors usually decide themselves whether to seek certification and may do so at different points in time. Quasi-experimental research is used when the experimenter cannot manipulate behaviour, but the logic of experimental design still may be applied.

Some 'what' questions are exploratory where as others relate more to measurement (e.g. how much, how many?). For example, i) 'What can be learned from the study of Rainforest Alliance certified farmers?' is an example of an exploratory question and different research strategies are possible (E.g. experimental, survey or case study); ii) to understand incidence and frequency is more likely to require a survey research strategy. But where the aim is to understand 'how' and 'why' Rainforest Alliance certification has had an impact on certified farmers, then this involves understanding the operational links and mechanisms by which Rainforest Alliance makes a difference and case study research is useful. To understand the scale and reach of impact, therefore, you might use a survey method, but to understand how and why the impact has been achieved (or not), then case study research is very useful.

To define research questions it is helpful to conduct a literature review in order to develop sharper and more insightful questions (Yin, 2014).

Remember all research strategies have different strengths and weaknesses – there is not a hierarchy in which one is better than another. Standard systems may wish to use case study research alone or in combination with other research strategies (such as surveys). The purpose may be as part of the on-going monitoring or evaluation data collection and analysis, or it might be as thematic, exploratory studies aimed at providing answers to a particular thorny question (e.g. child labour drivers and patterns). Here we are focusing mainly on the use of case studies in monitoring and evaluation. Although data may be collected on similar indicators in baseline, monitoring and evaluation activities (e.g. on performance and impact indicators) monitoring allows for more probing of difficult questions or filling of research gaps, and baseline and evaluation exercises may have more extensive analysis of impact.

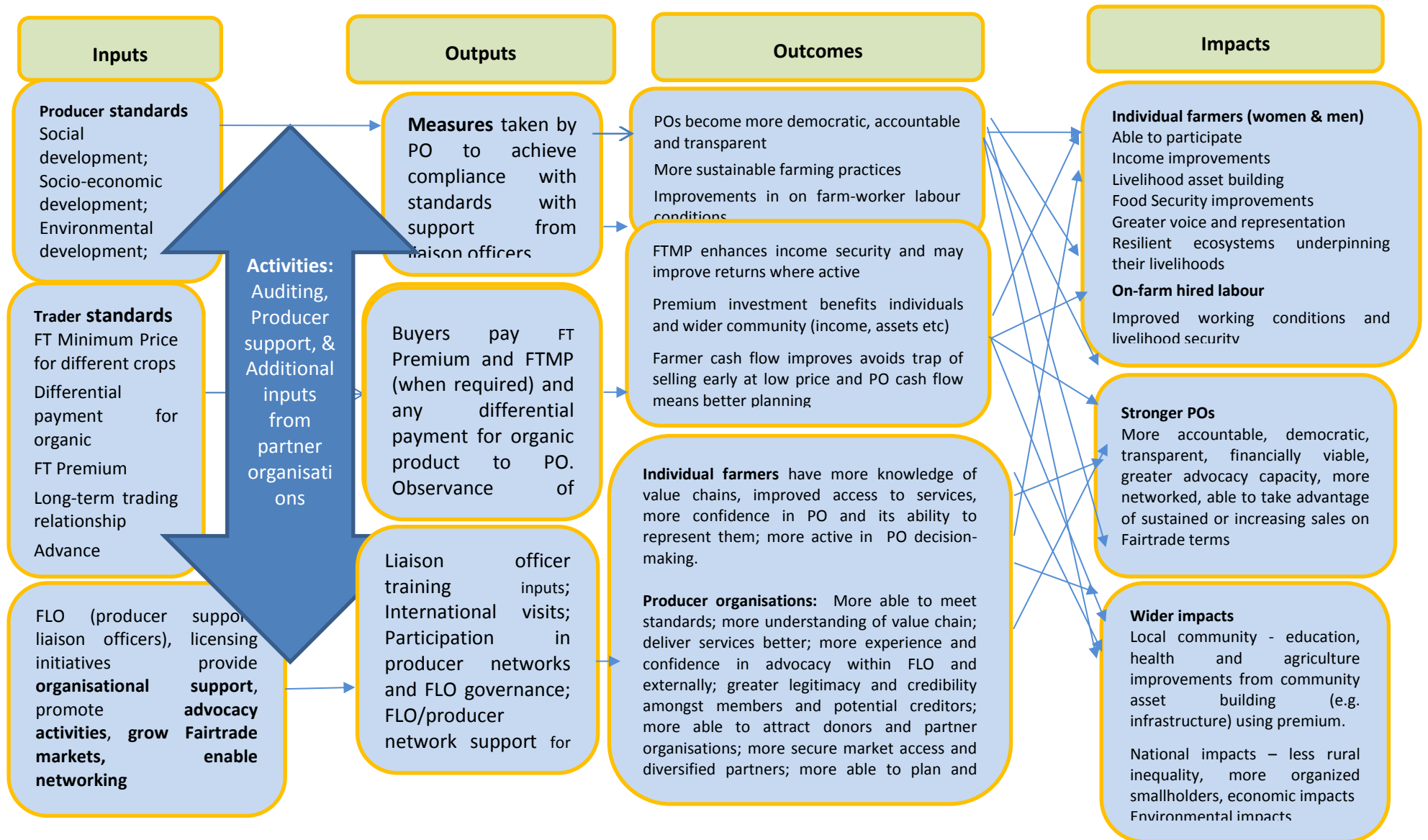
Case study research is not only for exploratory pieces, but can be used to test or build propositions or theories – i.e. they can be explanatory or causal case studies. In selecting cases it can be useful to ‘define your universe’, i.e. to create a typology of the value chains, or producers, or producer groups with which the standard system works. However, to produce such a typology might require further study. The selection of cases should reflect the theory or theoretical propositions that have been formulated (e.g. that Fairtrade certification has a positive impact on producer wellbeing; Fairtrade certification has different types and extent of impact on different kinds of producer organisations). To define the different producer organisations it might be useful to do some exploratory case study research. But to then test a theory would require explanatory case study research.

4.2 How do I design case study research?

A research design ensures that the questions studied, the data deemed relevant, the data collected and the analysis of the results, are all coherent. (In an evaluation context, an appropriate design must be found that fits with answering the evaluation questions and the programme’s attributes).

Firstly, **formulate and classify the research questions** (‘who’, ‘what’, ‘where’, how etc) as this gives preliminary guidance on what kind of research strategy to use.

A second step is to **formulate theory and theoretical propositions**. Unlike ethnography and grounded theory where the study is not based on hypotheses, in case study research there is some theory development prior to implementation. This theory shapes the rest of the study (e.g. choice of indicators, research tools and methods, type of analysis). The only exception in case study research is for exploratory studies, which may be more open and less structured. Setting out the theory and propositions then starts to provide guidance on what evidence should be gathered. A theoretical proposition is the (hypothetical story) about why acts, events, structure and thoughts occur – in the case of impact studies by standard systems the theory of change encapsulates the theoretical propositions about what (hypothetical) activities and inputs lead to what outcomes and impacts.



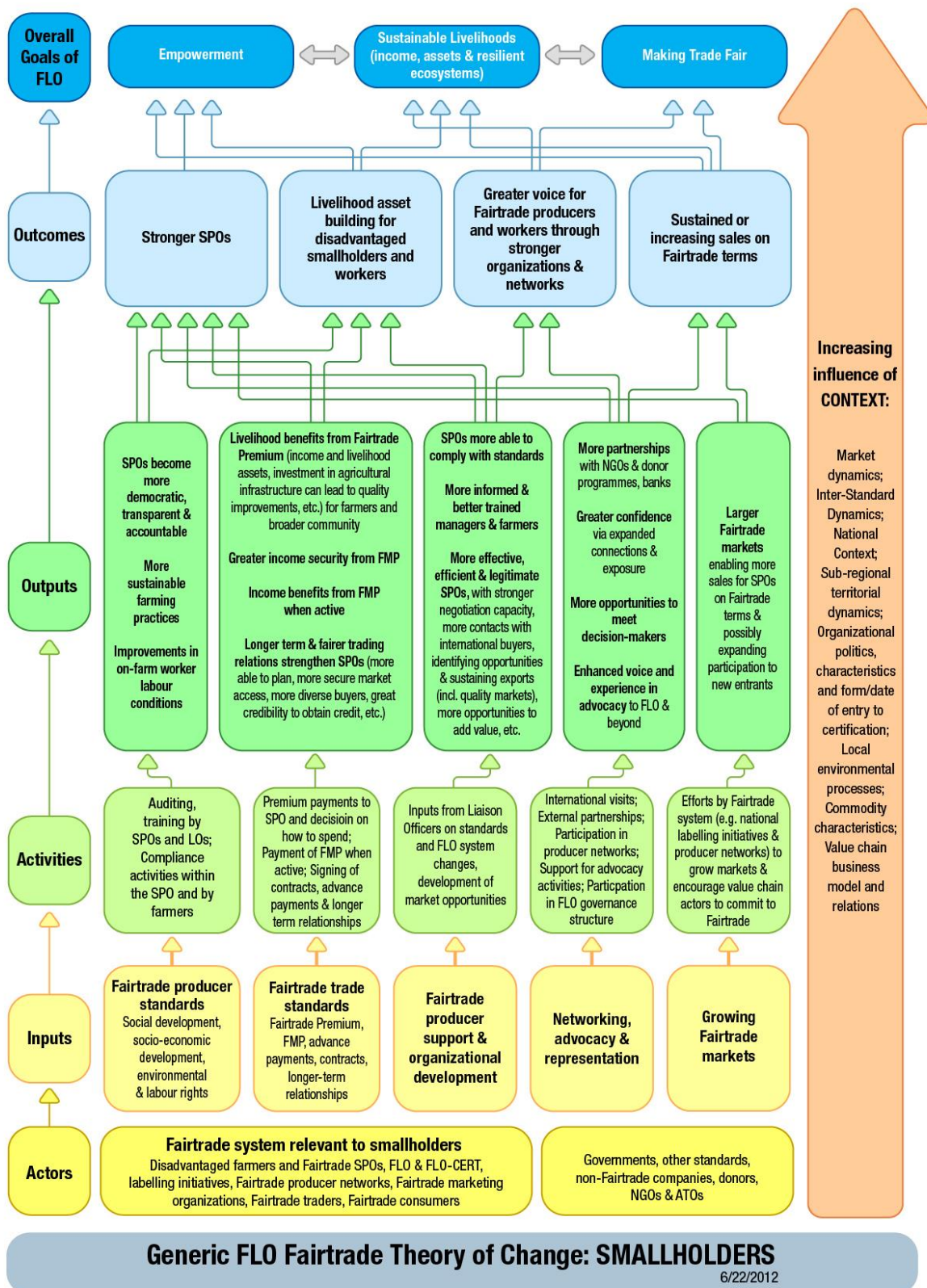
Thirdly, **the case to be studied needs to be defined and bounded**. The case to be studied may be individuals (e.g. child labourers, workers, farmers), but may also be organisations (such as producer organisations), or sectors, specific events, or sets of relationships (e.g. value chains). The 'unit of analysis' is chosen depending on the research question: is the question about the impact of certification on individual producers, for example, or it is about change at organisation level or across a whole industry or sector?

Where a study focuses on individuals, the aim would not be to collect all information about the individuals in the study, but to gather information relevant to the questions that have been outlined and the specific propositions or theory. For example, individuals that participate in certification experience or benefit (outcomes and impacts) from activities, inputs and outputs. Because defining the unit of analysis can be tricky, and can lead to confusion and ambiguity (Yin, 2014) suggests discussing the selection with a colleague.

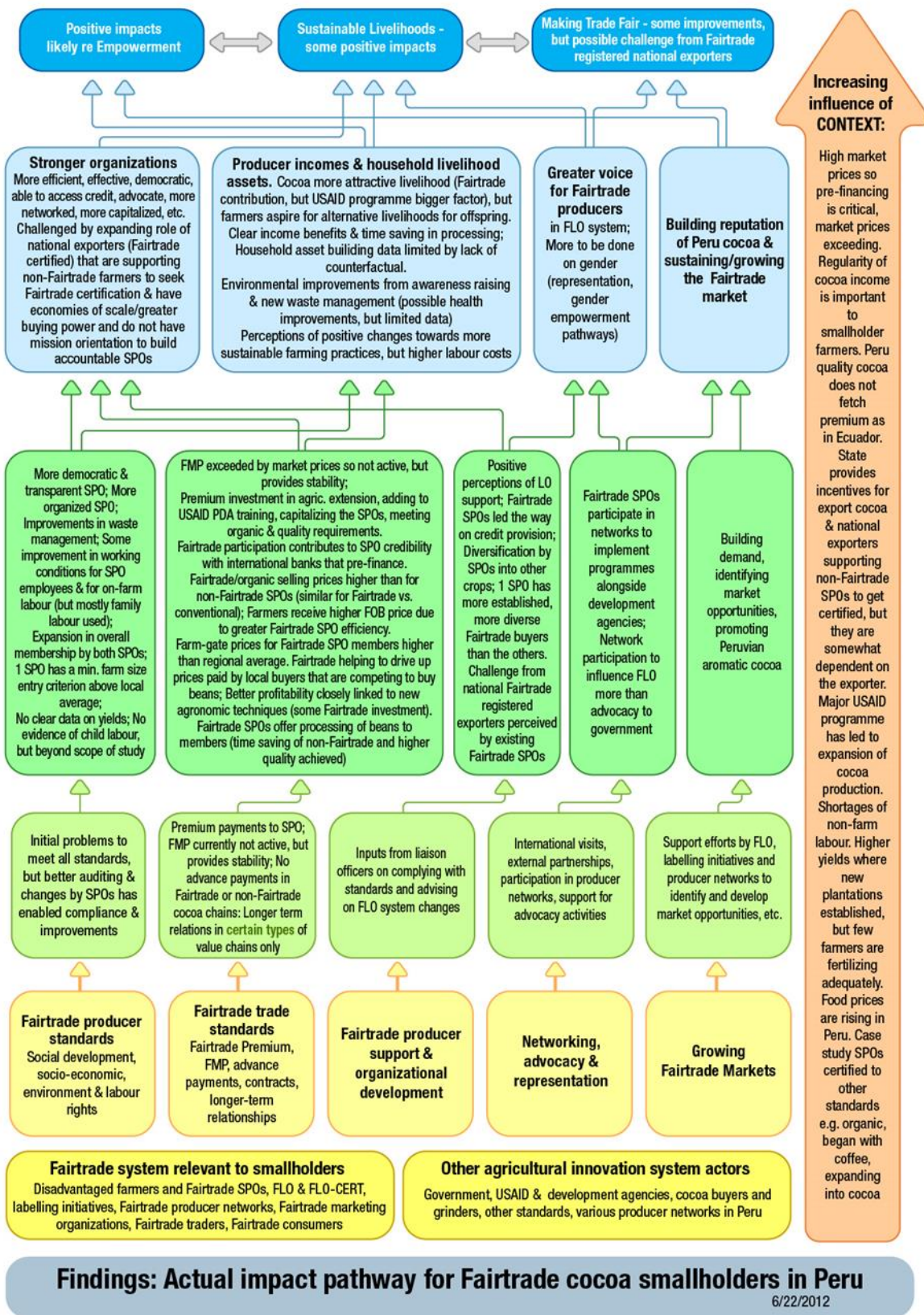
Providing clear boundaries for the case is important in the research design phase. For example, what are the time boundaries – defining the estimated beginning and ending of the case. For impact case study research it is perhaps useful to consider when individuals or groups became certified. It is also important to consider if there was engagement with a farmer or producer organisation in the months or years leading up to certification. Quite often standard systems engage with such groups to provide support to achieve compliance – this means changes will be occurring before official certification is achieved. In fact in our experience, many changes occur then and this can be the critical phase – even where continuous improvement is expected or some benefits accumulate over time. If the unit of analysis is a small group, define who is in the group and who is without. In some situations, there can be fluidity. For example, in cocoa farming in Ghana, members of a certified cooperative may leave for one season and rejoin the next, depending upon other opportunities for sale. But often for certified groups there is a membership register or process and this provides the main delineation of the unit of analysis. Clearly defining and communicating decisions made in research design (e.g. definitions of the case and unit of analysis) can help in future case study research – by the standard organisation or others – and the accumulation of knowledge.

Fourthly, **be clear on the choices on offer for analytical strategy so that the research design supports the chosen and appropriate approach**. Yin (2014) calls this 'linking data to propositions and establishing the criteria for interpreting findings'. There are various approaches (pattern matching, explanation building, time-series analysis, logic models and cross-case synthesis – but the choice should be determined by the research questions and propositions. Please see section xx for more information.

If causation frameworks and analytical strategies are chosen in impact design which supports process tracing then this can be considered one of the methods for use. For standard systems, process tracing will be useful in order to establish the casual links and to test whether they have actually occurred in reality.



Source: Laroche, Jimenez and Nelson, 2013



Source: Laroche, Jimenez and Nelson, 2013

By engaging with stakeholders this can increase the trustworthiness of the analysis – i.e. is there consensus or disagreements about whether x has caused y. This process helps to guide the development of indicators, tools and methods and informs the analytic strategy – i.e. by comparing the findings of the case against the hypothetical theory of change.

However, process tracing on its own does not necessarily support a focused analysis of rival explanations. It is important to systematically work out what the **rival explanations** might be – and to consider if this is appropriate as part of the impact design and then to formulate data collection needs. For example, if certified producer's incomes have increased is this the result of the inputs from Fairtrade or Rainforest Alliance certification, or is the result of another development programme, of changes in world market prices etc. Gathering data on rival explanations could be done either through the researchers conducting a series of key informant interviews, but explicitly exploring different explanations, or in a workshop setting, which allows stakeholders to discuss and debate the different options. The ability to do this depends on the resources available, as it will be important to carefully select who can attend – if chain-wide stakeholders are invited then this could be expensive. Conducting such a workshop would also allow for contribution analysis with stakeholders.

The theory of change can include **different levels** – e.g. it could be about an individual person, firm or organisational level, or programme level (Yin, 2014). For standard systems there are often different levels included within the single theory of change, and so data should be collected for each.

Ensure that **contextual conditions** are also considered in the analysis in a systematic way. Similarly to exploring rival explanations, external factors should be identified either through key informant interviews or in a workshop setting.

Mixed method designs involve the use of different methods to answer the same research questions and enable the collection of complementary data, potentially providing a rich analysis. Quite often mixed methods designs involve both qualitative and quantitative data collection. However, mixed method design is more complex (requiring skills in different methods and analysis) and generally requires more resources. It is very important to consider from the start what information the different methods will gather so that gaps in the data will not emerge later. If resources are available it can be useful to sequence methods – e.g. conducting some preliminary qualitative interviews to inform questionnaire design, conducting questionnaire surveys and then further qualitative interviews and methods to explore the questionnaire finding and see what the reasons might be for emerging patterns and trends.

It is very important to remember that utility is important not just rigour and accountability. For some different audiences/stakeholders this can mean statistically valid data generated through controlled comparisons, because they want to convince donors or certain sections of academia. For others, such as producer organisations utility and learning would more likely to be delivered by their engagement in design and analysis and a joint participatory learning process throughout. Case study research at the top level is not about sampling, surveys and experiments, but there is the possibility to incorporate this at a lower level within an embedded design (essentially mixing designs not just methods).

4.3 How do I select case studies?

Single case studies can provide important insights, but where resources allow, multiple case designs are more desirable than single designs – because they provide the possibility of direct replication. A research design that encompasses different cases – i.e. cross-case design – can support stronger

findings, just as multiple experiments would. Each study can be conducted separately and then the findings aggregated in a separate exercise, or they can be coordinated and conducted by the same team.

Tabulating the data from the individual cases according to one or more uniform categories can allow a probing of whether the different cases share similar profiles (instances or replications of the same type) or should be considered contrasting cases. If there are easily identifiable, embedded units at a lower level these can be sampled (e.g. a sample of primary level farmer societies that are linked up at higher secondary levels in a cooperative structure) or selected using cluster technique. The overall design is still considered an **embedded case study design**, rather than a holistic design in which only the global nature of an organisation is considered (see Yin, 2012 for more details).

Typology development can support a rigorous design – i.e. by enabling the researchers to choose replications (literal or theoretical).

4.4 How many cases are enough?

In terms of deciding how many cases to select, the logic used is not the same as sampling and so instead the question becomes about the number of replications you need or want in the study. The answer is discretionary and not formulaic – in other words it is up to the standard system. In many ways the more the better – but resources will ultimately be a constraining factor. Similarly, the ability of the researchers to manage multiple studies is relevant. Several scholars strongly advise that the research is centrally coordinated. One author even suggests that the same person conducts all the studies. This may not be feasible, but a robust case study protocol is needed to ensure that each team works in a coordinated fashion and is clear on what is required. If one or more of the cases are treated differently this undermines the ability to draw strong comparisons and creates a headache for those trying to bring the findings together. This process can be tricky, because it does not so easily support a collegiate approach between researchers – so it is doubly important to involve all the team in the initial design phase to build shared ownership of and clarity about the chosen design. However, again resources and planning are needed to make this happen.

4.5 How do I collect case study evidence?

The main preparatory stages are as follows.

Desired skills and values – case study research requires a high level of skills than survey data collection because the procedures are not routinized.

Yin (2014) states that the key skills as a researcher are to be able to: ask good questions, be a good listener, stay adaptive, have a firm grasp of the issues being studied, avoid biases, know how to conduct research ethically (e.g. refer to recognized quality standards). Specifically, standard systems should:

- carefully construct terms of reference (with stakeholder comment) to ensure that the research team have sufficiently good skills and experience.
- ensure that the budget matches the ambition of the study and vice versa, and allows for researchers of the proposed calibre to be involved.
- consider opportunities and strategies for evaluation capacity building in relation to the standard
- tender for researchers in a transparent process to ensure perceived independence.

- identify clear quality standards and guidance on research ethics – agree with researchers early on how issues such as child labour might be addressed, obtaining informed consent etc

Training

Ensure training is conducted so all team members understand the basic concepts, terminology and methodological issues (Yin, 2014). Specifically, standard systems should ensure that the researchers conducting the study:

- invite all the research team to discuss research question, design and contribute to the research study protocol and methodology.

Develop a protocol

This increases the reliability of the study, especially where multiple cases are being conducted. It guides the researchers in data collection. 4 main sections: overview (questions, propositions, objectives, etc); data collection procedures; data collection questions; guide for the case study report (Yin, 2014). Specifically, standard systems or researchers could:

- develop and obtain comments from advisors and stakeholders.
- include their hypothetical theory of change, research questions and propositions in terms of reference;
- outline data collection questions, sources and a draft report outline.

Screen candidate cases for the study

Where there are a small number or possible candidates, Yin (2014) suggests defining a set of operational criteria that will serve to qualify or exclude cases and the selection of cases that best fit your (literal or theoretical replication) (Yin, *ibid*). Where there are a large number of eligible candidates – a two-stage screening procedure is needed and selection of quantitative data, devising of relevant criteria for stratifying or reducing the number of candidates to 12 or fewer if possible and then selection using operational criteria.

For standard systems it is likely that there will be a large number of eligible cases for impact studies.

- Firstly, gather quantitative data (e.g. on when groups were certified, geographical location) as a first screen and exclude some (e.g. that have only recently been certified, or choose a spread or a specific region).
- Then select a number of countries for coverage to reduce numbers and devise a final set of criteria to guide selection of farmers or producer organisations.
- It is possible to randomly sample at lower levels, but selection using robust and transparent criteria linked to theory and propositions is also acceptable for replication.

Conduct a pilot case study

It is advisable to conduct one pilot case first to test the protocol and to make revisions to further case studies. For standard systems, the ability to do this depends on resources and planning, as well as make-up and locations of the study team(s), but Yin (2014) advises that conducting one pilot study first is important to enable testing, learning and improvement in follow-on studies.

4.6 How do I collect data and which methods should I use?

There is not space to detail the many methods which can be used to collect data. There are also many different sources of evidence (documentation, archival records, interviews, direct observations, participant-observation, physical artifacts).

Remember, some methods are more participatory than others in the sense that they might involve farmers, workers or traders in discussions that they can follow (e.g. using visual methods and avoiding some literacy issues). **Using a more participatory method is not the same as adopting a participatory process** – which is more about whether the entire process is driven by participants (rather than external people) and is of value to them, supporting learning and action by them in areas that fit with their interests.

Some key principles to guide the process of data collection are outlined by Yin (2014). Below we summarize each of these and provide specific guidance for standard systems and associated researchers.

Use multiple sources of evidence and triangulate the findings in the analysis.

Triangulation strengthens a case study (data triangulation, investigator triangulation, theory triangulation, methodological triangulation). It is not enough to gather multiple sources of evidence, the important thing is to develop **converging lines of inquiry** that can strengthen findings (as in navigation) (Yin, 2014). Real corroboration occurs when your findings are supported by more than a single source of evidence, but this does require sufficient skills as it can be complex to manage (Yin, *ibid*).

Standards should:

- Ensure that multiple sources of data are gathered.
- It is useful to develop a table that shows the data sources expected to generate findings on different research questions and how the data might fit together (or converge).
- Use the theory of change in an impact study to identify not only relevant indicators along the chain, but to indicate likely sources of data for each indicator (e.g. types of evidence – standard system databases, audit reports, interviews, workshops and focus group discussions etc).

Create a case study database.

Keep data or evidence base separate from researcher reporting, so that critical readers can inspect raw data (Yin, 2014).

- This is good practice, but it raises issues of confidentiality for standard systems. Some data cannot be shared except in an aggregate form, because there are commercial sensitivities and because anonymity may have been promised.
- Ensure these issues are clearly worked out in the early stages and then followed in practice throughout the study.
- Careful management of data, including copies of tabulated qualitative notes and statistical databases and analyses is important.
- It is advisable to train the research team clearly in how field notes should be written up – too often rich data is lost. While there should be space to probe unusual answers, where there are disagreements etc it is also the case that checklists will be devised for focus group discussions and key informant interviews and the team need clarity on how to write this up.

- There is new software that can be employed to analyse large quantities of qualitative material. Mobile phone and smartphone software is also available to speed up data inputting and analysis in surveys. Database development and plans for statistical analysis are also needed.

Maintain a chain of evidence

Yin (2014) suggests maintaining a chain of evidence between questions, protocol, specific sources of evidence, databases, case study report). Reports should illustrate how the chain of evidence has been implemented (e.g. indicate evidence sources of key findings in the text, make available raw data, have a detailed explanation of the methods used, include a copy of the protocol and questions).

- Standard systems should encourage this approach in their case study research by indicating this in terms of reference and referring to their theory of change.

Exercise care when using data from electronic sources

Some generic guidance is provided by Yin (2014) pertaining specifically to the use of electronic sources of data, including practical tips about cross-checking and how to conduct a websearch.

Finally, data collection requires a choice of tools and methods that fit the research questions and impact design.

Standard systems should seek to build up a body of evidence about their impact, by keeping track of different studies and building capacity to be able to judge its quality. By reviewing the existing body of work in assessing the impact of standard systems (as this has now increased in coverage and scope), there is the opportunity to build up the evidence base even where research is done by completely different researchers and research teams. There is still scope for innovation, but reference to previous impact designs (including causal frameworks, analytical strategies etc) will help in this process. It is also important that research teams situate their work in the body of literature (academic and more practitioner oriented) to support this accumulation of evidence and understanding.

The suite of different methods available is very wide ranging. Some methods of particular current relevance are: process tracing (and ideas on replication, theory testing etc), elimination of rival explanations, contribution analysis, outcome mapping. These are all relatively new methods which should support rigour and utility in standard system monitoring and impact activities involving case study research.

4.7 How do I analyse the case study evidence?

Focus on the questions of the case study protocol and establish what evidence has been collected to answer each one, rather than starting with the data (Yin, 2014). This is extremely useful advice, as otherwise the risk is that researchers can become overwhelmed. A clear analytic strategy is needed and should have been established in the case study design. Four different strategies have been suggested by Yin (2014), including relying on theoretical propositions, working your data from the 'ground up', developing a case description, and examining rival plausible explanations. For standard systems using case study research for evaluation purposes, it is likely that the researcher will use a combination of these – comparing theoretical propositions to the actual findings, exploring how far

plausible rival explanations can be excluded or are part of the analysis (and maybe quantifying stakeholder perceptions of their relative contribution to impact).

Time-series analyses may have been chosen as part of the impact design. This provides an analytic strategy which enables comparison of change over time. They also provide an opportunity to explore whether there are certain sequences of events that might be pre-requisites and to cover different conditions at different points in time.

The theory of change will be useful in this process, because the data will be collected together to analyse whether specific cause-effect-cause-effect patterns have indeed been found in reality. Again there is scope for stakeholder participation in this process, because this only increases opportunities for learning, but also the rigour of the analysis by drawing on a diversity of stakeholder perspectives.

Consider the use of software and technology to support analysis: computer-assisted-qualitative data analysis software (e.g. NVivo, Atlas.ti) can be used for both text and video based data, but requires skills in coding. The textual data can be from open-ended interviews or large volumes of written materials. This software can assist with categorizing data and identifying patterns, but unlike statistical analyses you cannot use the software's outputs themselves as if they were the end of your analysis. Instead you will need to study the outputs to determine whether any meaningful patterns emerge...developing a rich and full explanation or even a good description of your case', but to answer the 'how and why' questions, requires much more post-computer thinking and analysis (Yin, 2014, p134). Sensemaker is also a type of software that can be used in a methodological process to support the gathering of rapid micro-narratives, which respondents then categorize themselves and give meaning to. Some of the questions asked of respondents about their micro-narratives can include questions about causation. However, this technique has to be complemented with other methods to answer causal impact type questions, but it can be used in an iterative fashion with participants discussing emerging patterns in an exploratory process (ref Guijt and Pyburn?).

4.8 How do I report my case study?

Several reporting principles are identified by Yin (2014) and are useful to guide planning for case study reporting. Below we outline the implications and guidance for standard systems.

Audiences

- Identify the range of audiences for the findings (e.g. within the standard system, donors, media, academics and development practitioners, different types of private sector actors, study participants).
- Orient the reporting to the needs of different audiences – perhaps even consider different types of outputs (e.g. a summary document, main report, powerpoint). Consider the use of infographics and of tabulation of findings or visualisations of hypothetical and actual empirical findings – but remember what is required of researchers must fit with resources.

Varieties of case study compositions

- For multiple case study reports it can be more straightforward to produce one main synthesis document, with single cases not being reported (this can save time for researchers especially if the audience is unlikely to be interested in the detailed findings on each case. The research teams for each case study can contribute their findings but not necessarily write up a full individual report). But where each case is covered, then either include as separate chapters or produce as separate reports, but follow a common format.

- There are different structures available for reporting (e.g. chronological, comparative, suspense, linear-analytic and unsequenced etc). Most likely to be of use for standard systems is to set out the evidence for each part of the theory of change – explaining if the evidence does or not does not match what was expected, whether rival explanations are involved and what their contribution is, contextual factors shaping outcomes, with a conclusion on the overall evidence of impact.

Procedures for composing a case study report

- Consider issues of anonymity, review processes – set these out clearly from the start.
- Consider sharing the report internally with peers, including the initial outline and then the draft report.
- Allow sufficient and realistic time for gathering comments from stakeholders in an external review and comment process.
- Ensure there are resources budgeted for publication and review processes and for participatory processes of validation and feedback in appropriate forms to farmers, workers and community members, producer organisations etc.
- Ensure there is clear agreement with researchers about publication – it is good practice to be transparent. The timetable for publication should also be part of the initial terms of reference – i.e. the commissioning agent and researchers should commit to a specific publication date and media before embarking on the study.

Speculations on the characteristics of an exemplary case study (extending beyond the report itself and covering the design and content of the case)

According to Yin (2014) it is worth considering what ‘ideal’ case study research might look like to encourage high quality research. For example, he suggests that the study should tackle significant issues and be of public interest, the report must be completed. It should consider alternative perspectives and demonstrate sufficient evidence, and be engagingly composed and disseminated.

- Standard systems/researchers should include a section should be included in the methodology chapter that explains any limitations of the design and implementation.

5. Conclusions

Although case study research has been around almost as long as social science research, this is a dynamic field of enquiry, and in recent years, sophistication has grown regarding the use of case study research to explore causal mechanisms. Increasingly specialist knowledge is needed to conduct case study ‘state-of-the art’ research, just as it is required for statistical and modelling techniques. There are also but there are also issues of ethos and development philosophy – for example, how far the approaches chosen are consistent with a participatory approach.

The users and audiences for case study research may vary: Are farmers, workers, fisherfolk, for example, seen as primary intended users – this may imply are more participatory approach, or at minimum greater engagement in the evaluation process which should be oriented towards learning. Standard systems are important users of the findings as it is they that will or should make initial changes to their theory of change or practices in order to improve impact. However, there are also important external audiences. Private companies are likely to be interested in evaluation findings,

because sourcing certified products is part of their overall reputation management as well as sustainable sourcing strategy. The media often report on alleged abuses, failures, and (less frequently) on success stories – their work would and should be informed by robust and independent evaluation studies, but these have been lacking to date. Donors are interested in whether standards and labelling are an effective strategy for tackling poverty and enhancing ecosystem services, as well as in terms of accountability for resources invested. In recent years they have invested funds in impact evaluation of standard systems (e.g. in the Netherlands and the UK).

The academic and evaluation community is an important audience for case study research conducted or commissioned by standard systems for the purposes of M&E. In the evaluation community there are different strands of professionals with differing epistemologies and approaches. Within academia there is a growing interest in understanding and researching standard systems. Blackman and Rivera (2010) reviewed a number of impact studies and deemed only a small handful to be credible based on whether they include *credibly constructed counterfactuals*. However, this criticism fails to recognize the value and potential of case study research.

There can be differences between research in general and evaluation in particular. In research it is usually the case that researchers set their own research questions, whereas in the latter policy makers and commissioners set the questions (Stern et al, 2012). The purpose of an evaluation can include both learning and demonstration of impact (Stern et al, *ibid*), but while many (and increasing) studies try to fit both purposes, there are also occasions when the client of an evaluation will indicate a clear preference for one or the other. Questions also arise as to who the study is for. In evaluation there are particular issues of utility, involvement of stakeholders, acknowledgement of country ownership, ethical dilemmas of working with very poor and marginalised people (Stern et al, 2012). The (increased) demand for evidence, by donors and governments in particular, has (and does) in many quarters mean a certain type of impact evaluation which it is assumed involves experimental and quasi-experimental techniques. However, there is also increased interest in mixed methods (e.g. Bamberger, 2010?) and a broader set of impact designs which are still rigorous (Stern et al, 2012).

A key issue for standard systems, which are operating in the full glare of public scrutiny, is being able to learn from failures. This is a key part of any learning process, but there are clearly reputational risks. The temptation is therefore to select more positive stories for inclusion in studies, but for rigorous evidence it is essential that a robust set of selection criteria for choosing cases is established, but also that this process is *seen* to be independent for evaluation purposes, via the development and use of clear protocols for commissioning research and clarity on the respective roles for researchers and client, as well as on publication dates and form, anonymization issues etc. If researchers are given the chance to select cases, then this supports perceived independence of findings, but the selection should be based on a clear set of agreed criteria. Otherwise there is a risk that even if cases are selected for diverse reasons, if they are selected by the standard body that the choice can then be questioned – and this, even if completely unwarranted - can affect the credibility of the entire study.

Clear agreements should also be made right at the outset for how producer organisations and groups will be engaged in the study and how findings will be shared and dissemination (and how this work will be funded). Where there is a need for translation of reports and time allowed for stakeholders to comment on draft reports, it is necessary to leave sufficient and realistic time and resources. Developing a learning and dissemination plan for each study is advisable from the start as this may lead to improved stakeholder engagement, mapping of audiences, planning of participation and feedback etc.

It is the case that however balanced a study may be, that the findings will be mis-reported, which could damage reputation or even sales. It is therefore a good idea to plan ahead for the publication process. Standard M&E systems will have greater credibility if they do publish all findings – but it is also reasonable that they have a chance to prepare responses that can be shared at the same time as the report is published.

It is also important to match resources to budget. The temptation for commissioning agencies to outline a wide ranging terms of reference, but if resources are limited in comparison this will affect the overall quality of the study and the ability of the researchers to deliver. It is generally that case that more ambitious studies are less likely to be successfully delivered. Many scholars writing on case study design suggest that it is good practice to clearly delineate the boundaries of the study and to be focussed.

It is likely that there are a variety of stakeholders with expectations of a study – and this requires careful management. If a broad-ranging study is outlined by a client, but limited resources given to researchers, and stakeholder expectations not kept in check, then it is highly likely that the study will not be satisfactory for some audiences.

Stakeholders should be involved in commenting on the terms of reference prior to their publication and the process of refining and designing the research will benefit from stakeholder engagement. At the same time, if this is an in-depth process it might be advisable to envisage a two-step process in which the first stage is specifically devoted to design and consultation of stakeholders, followed by an implementation phase. Contracting should fit this process.

Because of the increased demands for rigour (and utility) of evaluation studies, this requires greater skills of researchers to achieve the necessary level of sophistication of design and ability to implement the study, as well as to accompany field researchers, and experience in guiding writing up and analysis. Too often the analytical and write-up phase is under-budgeted. Where complex statistical techniques are being employed this requires significant time and resources – although increasingly, there are opportunities to use mobile phone software to speed up data collection and basic analysis, at least for simpler questionnaires. It is important to consider including resources in budgets for translation of reports and for dissemination activities.

It is quite possible that M&E staff commissioning researchers in standard bodies may have limited experience of managing research and consultants. It is important that capacity is built up amongst relevant staff in order to ensure that a successful study is undertaken. Understanding, for example, when the commissioning agency should play a role and when it should step back and allow researchers to conduct the study is important – micro-management can be demoralising and time consuming for researchers, but having and monitoring a clear set of milestones and timetable is also important to ensure that researchers are on track.

Tendering for researchers to conduct a study is a good way of achieving a transparent system of selection and for commissioning agencies to view a range of proposals in order to choose the one that best meets their requirements. It can be time consuming for commissioning agencies and for researchers, but it is a fairer process generally. Where a very complex design is likely to be required, it is possible that the tendering process is two-stage, or that a more detailed design is only required at a later stage. Adding a page limit to tender proposals and curriculum vitae is probably advisable. It is important that commissioning agencies provide a clear set of criteria for judging proposals and that the terms of reference are not overly ambitious given resources available. It is also advisable that an indicative budget or budget ceiling is provided to guide researchers as to the likely scale of the study envisaged. Researchers can still compete on value-for-money, as well as other criteria, but the proposals are more likely to fit with the vision of the client.

As well as recognizing the increased sophistication required in M&E including case study research, but also in experiments, surveys etc and frequently in being able to mix designs and methods, (and matching the budget to this), it is also important to consider a longer-term plan for building evaluation capacity. It is possible to identify research partners in different regions of the world, for example, that could be involved on a partnership basis to provide evaluation services. Tendering for each research study may mean that a diverse range of potential researchers can be identified and this could support innovation, but it is also the case that significant capacity building is needed and that learning occurs during each study which can be built upon.

Finally, an interesting question arises as to the judgement in an evaluation about what success should look like. An organisation or standard system should be judged against what it sets out to achieve, but it is not always clear the scale of intended impact. Although case study and other research strategies provide evidence about impact, there is still a need to debate and engage with stakeholders **about how much impact is expected to constitute 'success'**. For example, if poverty impact or producer empowerment is a goal, how much should the standard system be responsible for achieving? For example an evaluation might show that there has been a positive impact on poverty for producers in certified organisations, but some scholars/professionals then question whether this is sufficient. Are farmers expected to have escaped poverty, to have moved several rungs up a poverty ladder (and who defines what this ladder is)? Further, the approach of a standard system may be judged relatively effective, but there could still be other ways of tackling a perceived problem which are more effective. This is partly a question of divergences in development philosophy and while a consensus cannot be achieved, it is worthwhile *reflecting* on what the alternatives might be as part of a solid learning strategy.

References

- Bamberger, M., V. Rao, M. Woolcock (2010) 'Using mixed methods in monitoring and evaluation: Experiences from International Development'. World Bank, Policy Research Working Paper 5245.
- Becker, B., P. Dawson, K. Devine, C. Hannum, S. Hill, J. Leydens, D. Matuskevich, C. Traver, and M. Palmquist (1994 - 2012), Case Studies. Writing@CSU. Colorado State University. Available at: <http://writing.colostate.edu/guides/guide.cfm?guideid=60>
- Better Evaluation (2013) 'Planning an Evaluation: Using the Rainbow Framework', May 2013.
- Collier, D. (2011) 'Understanding process tracing'. PS Political Science and Politics 44, No.4, :823-30.
- David, M. (Ed.). (2006) 'Case study research'. Four Volumes. London: Sage.
- George AL and Bennett A (2005) *Case Studies and Theory Development in the Social Sciences*. Cambridge Mass. And London: MIT Press.
- Guijt, I. (undated) 'Accountability and Learning'. SNV.
http://www.snvworld.org/sites/www.snvworld.org/files/publications/21_accountability_and_learning_-_exploding_the_myth_of_incompatibility_between_accountability_and_learning_-_irene_guijt_.pdf
- Holland, J (2013) 'Who Counts: the power of participatory statistics' Practical Action.
- Laroche, K, R. Jimenez, V. Nelson (2013) 'Assessing the impact of Fairtrade in Peru'. NRI report, commissioned by Fairtrade International and Max Havelaar, Belgium.
- Mayne, J. (2008) Contribution Analysis: An approach to exploring cause and effect, ILAC methodological brief, available at http://www.cgiar-ilac.org/files/ILAC_Brief16_Contribution_Analysis_0.pdf
- Orr, A., (2009) 'The Case Study Method in M&E: A Guide for Research Into Use'.
- Oxfam (undated) 'Process Tracing: Draft Protocol'. <http://policy-practice.oxfam.org.uk/blog/2013/02/~media/C396B507E01C47AB880D7EEF9ECCD171.ashx>
- Patton, M.Q. (2008) 'Utilization Focused Evaluation'. Sage Publications.
- Patton, M., Q. (2013) 'Utilization-Focused Evaluation (U-F E) checklist'
http://www.wmich.edu/evalctr/wp-content/uploads/2013/03/UFE_checklist_2013.pdf
- Patton, M (2012) 'Essentials of Utilization-Focused Evaluation'. Sage Publications: Los Angeles.
- Pawson, R. and Tilley, N. (1997) *Realistic Evaluation* Sage Publications.
- Thomas, G. (2011) 'How to do your case study: A guide for students and researchers'. Sage Publications: London.
- Stern, E., N. Stame, J. Mayne, K. Forss, R. Davies, B. Befani (2010). 'Broadening the range of designs and methods for impact evaluations'. DFID working paper. Report of a study commissioned by the Department for International Development.
- Yin RK (2003). *Case Study Research: Design and Methods, 3rd Edition*. Thousand Oaks CA:Sage.

Yin, R.K. (1994) 'The Case Study Crisis: Some Answers', *Administrative Science Quarterly*, Vol. 26, No. 1. (Mar., 1981), pp. 58-65. URL

Zelik, D. J., E. Patterson, and D. D. Woods (date) 'Measuring attributes of rigor in information analysis'.
http://csel.eng.ohio-state.edu/zelik/research/Rigor_files/ZelikPattersonWoods_MeasuringRigor_1.pdf
<http://links.jstor.org/sici?sici=0001-8392%28198103%2926%3A1%3C58%3ATCSCSA%3E2.0.CO%3B2-H>.
Administrative Science Quarterly is currently published by Johnson Graduate School of Management, Cornell University.

Yin (2014) 'Case Study Research: Design and Methods', Sage Publishing. Fifth edition.

Appendices

Appendix 1:

A guidance document explaining applicable theory regarding case study research, making specific recommendations for standard systems, and illustrating the theory and recommendations with examples for the sustainability standards studied for this project.

A set of specific recommendations for the two standard systems studied in depth as part of this assignment.

A 1-hour recorded webinar explaining this material, to be used for training purposes.

Among the questions the consultant should address in the work are:

☐ How many cases are “enough”? What is (or what determines) the added value of an additional case?

☐ Does the answer to this question differ if the cases apply different versus parallel methodologies?

☐ Are more cases / is better (geographic) coverage of cases always better? What are the advantages/disadvantages of repeating studies of one or a small set of cases or locations, versus attempting to cover more cases or locations in the collection of studies?

What are the advantages or disadvantages of choosing “easy” versus “hard” cases (cases where one would suspect that certification would easily have the desired impact versus cases where having any impact would be very difficult)?

☐ Should random sampling ever play a role in the choice of cases? Or does strategic choice of cases always make more sense?

☐ What should standard systems think about when choosing cases to achieve maximum learning value for their own organisations? Or to provide the most convincing evidence of the results and impacts they produce? Is there any trade off between learning and demonstrating?

☐ How should standard systems talk about the results of a study of a single location or case? What can and can't they claim? To what extent is the answer determined by the methodology used in the case study as opposed to the inherent limitations of looking at results or impact in one particular case? (How) does that change with multiple, carefully chosen cases versus a single case?

☐ Are there any general guidelines for the design of case study research that would help increase the learning and demonstration value of the case study or the collection of case studies? (independent of the exact methodology used for the case study)

Proposed Timetable & Allocation of Days

Initial interviews with ISEAL and 4+ sustainability standard systems selected by ISEAL or participation in launch workshop to deepen understanding of their standards, vision, current and anticipated future use of case study research in the M&E system	Early Nov start*	2 days
--	------------------	--------

Refinement of the study questions (e.g. general challenges and questions to be considered) and review of existing theory and guidance that could be applied to the challenges/questions	End November	2 days
Detailed examination of two different cases, in collaboration with representatives of the sustainability standards	End December	2.5 days per case (5 in total)
Development of draft findings and recommendations for each sustainability standard	Mid January 2013	3 days
Preparation of a guidance document outlining the general challenge/question; explaining the applicable guidelines and/or theory; and proposing general guidance for sustainability standards (with illustrations from specific cases)	End January 2013	4 days
Preparation / delivery of training webinar	End Jan 2013	2 days
TOTAL		18 days

**Workshop/interviews depend on availability of standard systems to some extent.*

Appendix 2: Analytic techniques

Analytic technique	Details
Pattern matching	<p>Comparing an empirically based pattern (i.e. based on the findings of the case study) with a predicted one made before data collection (or several alternative predictions). This is also called congruence method (George and Bennett, 2005). If the empirical and predicted patterns are similar this can strengthen internal validity. For explanatory studies patterns may be related to i) dependent and ii) independent variables (for descriptive studies pattern matching can be useful if a pattern of important descriptive conditions was identified prior to data collection).</p> <p><i>Non-equivalent, dependent variables:</i> If for each outcome identified, the initially predicted values have been found and at the same time alternative patterns of predicted values have not been found, then strong causal inferences can be made. If the results fail to show the entire predicted pattern, then the initial proposition is questioned. The strength of the findings can be increased with <i>theoretical replication</i> – i.e. the selection of two (or more) cases in a multiple-case study because the cases are predicted to have contrasting findings, but for anticipatable reasons. (<i>Literal replication</i> involves the identification and actual study of two or more cases).</p> <p><i>Rival independent variables as a pattern</i> The use of rival explanations is good not only as a general analytic strategy, but is also an example of pattern matching for independent variables. In such a situation several cases may be known to have had a certain type of outcome and the case study has focused on how and why this outcome occurred in each case. Rival theoretical explanations are identified and articulated in operational terms. The independent variables may involve or be measured by several or many different types of characteristics or events, each assessed with different measures and instruments. The case study focuses on the overall pattern of results and the degree to which the empirically based pattern matches the predicted one.</p> <p>More precise measures help to strengthen the case study. Where there is very limited precision, avoid postulating very subtle patterns – deal with gross mismatches or matches, because the interpretation is less likely to be challenged.</p>
Explanation building	<p>This is a kind of pattern matching, but the procedure requires more detailed explanation. Also known as process tracing and it is mainly relevant to explanatory case study research. Often explanation building occurs in narrative terms – explanations should reflect some theoretically significant propositions to offset the lack of precision. A series of iterations are involved: i) making an initial theoretical statement or an initial explanatory proposition; ii) comparing the findings of <i>an initial case</i> against such a statement or proposition; iii) revising the statement or proposition; iv) comparing other details of the case against a revision; v) comparing the revision to the findings of a <i>second, third or more cases</i>; vi) repeating this process as needed. The final explanation may not have been fully stipulated at the start of the study and is thus different from the pattern matching described above. To avoid threats such as loss of focus or creeping selective bias, there should be frequent checking back to the original purpose of the study, engaging external colleagues as ‘critical friends’, and continually examining alternative possible explanations, using a case study protocol, establishing a case study database, and following a chain of evidence.</p>
Time-series analysis	<p>Similar to the data collection in experiments and quasi-experiments. Simple time series may involve only a single dependent or independent variable, although with a large number of time data points. The ability to trace changes over time is a major strength of case studies (not being limited to cross-sectional or static assessments of a particular situation) although the available data points may only be a truncated segment of a broader (and opposing) trend). The essential logic of a time-series design is the match between the observed (empirical) trend and either of the following: a) a theoretically significant trend specified before the onset of the investigation, or b) some rival trend, also specified earlier. Complex time-series involve cases where the trends are far more complex, e.g. with mixed patterns over time. A non-linear model is needed. The strength of the case study is in developing a rich explanation for the complex time series. Greater complexities also</p>

arise when a multiple set of variables – not just a single one – are relevant to a case study and when each variable may be predicted to have a different pattern over time.

Chronological sequences can be compiled to trace events over time. More than a description, such an analysis can allow comparison of the actual chronology with that predicted by an explanatory theory (producing findings on necessary sequences, contingencies (some events must be followed by others), relevance of intervals of time before an event can occur, certain time periods are marked by classes of events that differ substantially from those in other time periods.

If the actual events of a case study (carefully documented and analysed) have followed one predicted sequence of events and not those of another compelling, rival sequence, then the single case study can again become an initial basis for causal inferences. Comparison to other cases can strengthen inferences.

Logic models

Increasingly useful in case study evaluations and studying theories of change. The logic model stipulates and operationalizes a complex chain of occurrences or events over an extended period of time. The events are staged in repeated cause-effect-cause-effect patterns, whereby a dependent variable (event) at an earlier stage becomes the independent variable (causal event) for the next stage. When developed collaboratively, the logic model can help a group more clearly define vision and goals and establish how the sequence of programmatic actions will (in theory) accomplish the goals. The use of logic models consists of matching empirically observed events to theoretically predicted events. Joseph Wholey (1979) was at the forefront of developing logic models as an analytic technique, tracing events when a public program intervention was intended to produce a certain outcome or sequence of outcomes. The intervention could initially produce activities with their own *immediate* outcomes, which could in turn produce *intermediate* outcomes, and these in turn would produce final or *ultimate* outcomes. This approach could be combined with the strategy of using plausible rival explanations – i.e. considering rival chains of events, as well as the potential importance of spurious external events. Depending on the number of cases the links can be analysed qualitative or quantitatively. There are different kinds of logic models depending on the unit of analysis – individual-level; firm or organisational level logic model, program-level. It is important to highlight the *transitions*, not just the activities in logic models - collect data on the transitions, not just the events. Further, attend to contextual conditions as an integral part of the model. Although external to the case, such conditions and rivals might in fact be found to strongly influence the intervention's outcomes, possibly outweighing the effects of the resources and activities supported by the intervention.

Cross-case synthesis

For analysis of multiple cases only and can be utilized whether the individual case studies were performed independently or as a pre-designed part of the same study. Word tables can be created displaying data from the individual cases according to one or more uniform categories. Such an array allows probing of whether different cases appear to share similar profiles and deserve to be considered instances (replications) of the same 'type' of general case. Alternatively, the profiles may be sufficiently different that the cases deserve to be considered as contrasting cases. A predicted similarity or contrast may have been part of the original design of the case study and the findings based on the observed profiles provide confirm or disconfirm the original expectations and connect well to prior research etc. The examination across cases using word tables relies heavily on argumentative interpretation, not numeric tallies.

Case studies can be conducted within a case study – an embedded set of case studies may be conducted, with the findings of the overall study drawing on separate data from the broader or larger unit of analysis as well as the cross-case data from the multiple studies.

Adapted from Yin, 2014

Appendix 4: Utilization focused framework steps

Box xx: The Utilization-Focused Framework

1. Assess and build program and organizational readiness for utilization-focused evaluation
2. Assess and enhance evaluator readiness and competence to undertake a utilization-focused evaluation
3. Identify, organize, and engage primary intended users: the personal factor
4. Situation analysis conducted jointly with primary intended users
5. Identify and prioritize primary intended uses by determining priority purposes
6. Consider and build in process uses if and as appropriate
7. Focus priority evaluation questions
8. Check that fundamental areas for evaluation inquiry are being adequately addressed: implementation, outcomes, and attribution questions
9. Determine what intervention model or theory of change is being evaluated
10. Negotiate appropriate methods to generate credible findings that support intended use by intended users
11. Make sure intended users understand potential methods controversies and their implications
12. Simulate use of findings: evaluation's equivalent of a dress rehearsal
13. Gather data with on-going attention to use
14. Organize and present the data for interpretation and use by primary intended users: analysis, interpretation, judgment, and recommendations
15. Prepare an evaluation report to facilitate use and disseminate significant findings to expand influence
16. Follow up with primary intended users to facilitate and enhance use
17. Meta-evaluation of use: be accountable, learn, and improve

Source: Patton (2008)

Appendix 4: Visualization of the case study research process (George and Bennett, 2005)

Phase 1 – Research design
(objectives, design and structure of research are formulated)

Identify an important research **problem or ‘puzzle’** (Specify the problem (including a clear reasoned statement of what the research will contribute, reflecting gaps in current state of knowledge, acknowledging contradictory theories and noting inadequacies in evidence for existing theories)

Identify research objectives (6 different kinds of theory building research objectives). Avoid being over-ambitious: clearly focus the study (e.g. on a well-defined subclass having defined a typology or choosing a specific policy instrument, goal or contexts) and contribute to overall theory development and evidence. If comparing and contrasting 2 or more interventions also need clearly defined problem (e.g. could be explaining the causal role of a particular independent variable across cases)

Developing a research strategy: specification of variables.

Decide what will be the dependent (or outcome) variable to be explained or predicted
Decide what independent (and intervening) variable comprise the theoretical framework of the study
Decide which of these variables will be held constant (serve as parameters) and which will vary across cases included in the comparison.
(These may be changed during the study in case study research)

Case selection

A structured process, flowing from well-defined objectives – (p82 not clear) and selection occurs from a typology (either comparing instances of the same case or cases from different subclasses).

Describing the variance in variables

(trade-offs between parsimony and richness here)

Formulation of data requirements flowing from theoretical framework and research objectives, and general (not overly specific) questions that allow results across cases to be compared, cumulated and systematically analysed (although more specific questions can be added appropriate to individual cases), and the intended contribution to theory

Phase 2 – Case study implementation following the design

Phase 3 – analysis of the findings of the case studies and assessment of their contribution to achieving the research objective of the study

Figure 5 Visualization of the case study research process (George and Bennett, 2005)

