STRATEGIC USE OF CASE STUDIES BY SUSTAINAINABILITY STANDARD SYSTEMS

NRI PRACTICAL GUIDANCE NOTE

Valerie Nelson and Adrienne Martin
Natural Resources Institute, University of Greenwich
September 2013

1. INTRODUCTION

This document provides guidance on the design, selection and conduct of case studies in the context of monitoring and evaluation of standards systems. It explains the theory regarding case study research and the extent to which findings be generalised. Specific recommendations are made for standard systems, illustrated with examples from sustainability standards.

2. WHAT CAN BE CONSIDERED A CASE IN THE CONTEXT OF A STANDARD SYSTEM?

In case study research, the case or unit of analysis can be thought of as an instance of a class of events. A case study usually involves empirical data gathering, in a real world context and is a phenomenon of scientific interest. A case study can focus on a type of structure or system (e.g. a government or economic system), an individual person, on a personality type (in psychology), an organisation, a community or an event (e.g. a revolution). However, for standard systems using case study research as part of their overall monitoring and evaluation system, it is most likely to be one of the following:

- An individual participant – e.g. a farmer, a worker, a child freed from child labour
- A producer organisation (or a union of producer organisations)
- A supply chain
- A community
- An industry
- An environmental service

How you define your case depends on the nature and scope of how you understand change to come about (your *theory of change*) and which case you choose to focus on and the resources required, depend upon the most urgent monitoring and evaluation questions you would like to answer. For example, not all standards' theories of change will necessarily focus on a community level change, or will explicitly seek to achieve market transformation across an industry. Where these are stated aims, then it will be important to consider them for case study research. However, given that many of the standards are seeking to support change for individual producers or for environmental services, these are likely to be a primary focus for case study research.

Box 1: Fairtrade theory of change and impact studies

FLO Fairtrade seeks to support producer empowerment, sustainable livelihoods and to make trade fair, but does so via producer organisation development and changing traders' practices. Further, Fairtrade does not only seek to benefit individual producers, but to support development within the overall community and potentially locally and nationally. Thus, it is important for FLO not only to gather data on impacts for individual producers, but also for the communities of which they are a part. Fairtrade works now with both producers and workers – so there is variation at the primary participant level. There is also variation between the types of supply chains of which they are part. In the Fairtrade theory of change, achieving change for individuals requires a number of things – producer organisation development, growing markets, changing trader practices. Thus, case studies

which seek to establish impact need to consider whether and how effectively these things are being achieved and if these are leading to individual producer impacts – in different countries and supply chain types. Thus impact data is needed (at some stage) across all of these areas and a whole range of case studies become desirable. Choosing which to focus on with limited resources is a challenge. In such a large system as FLO Fairtrade - 70 countries, 1.3 million farmers and workers, 17 products – it is not possible to build an instant evidence base covering all dimensions. Instead evidence must be accumulated: i.e. studies should be replicated, with standardized questions, to support a building up of a robust picture of evidence over time. FLO is commissioning different studies for selected commodities and locations depending on the priority M&E questions they identify and to inform learning and support changes to increase impact. Now that FLO has developed a theory of change, future impact studies can increasingly add to the evidence base, while ensuring that lessons are taken up in each case to improve impact.

3. WHY WOULD YOU UNDERTAKE CASE STUDY ANALYSIS?

Case study research is useful to answer 'how and why' questions in evaluation. Has a standard system created impact and if so, how? Why has impact been achieved or why was it not achieved. Were other causal factors of as much, or greater importance?

Case study research is useful for complex, real-world situations where there may be more variables of interest than data points – and as a result it uses a holistic lens. Multiple sources of evidence are needed and the data must be triangulated or cross-checked to check it is accurate. Because of this, case study research benefits from the prior development of theoretical propositions to provide a guiding framework for the collection of data and analysis. This is particularly the case in evaluation studies which are trying to understand whether a project, programme or in this case standard system has been effective or not.

Case study research is essentially a research strategy which is relevant for complex situations where an in-depth understanding is required. Other possible research strategies include statistical surveys, experiments and modelling. These strategies can be complementary and the choice of strategy will depend on the task in hand. Surveys are useful to answer research questions such as 'who, what, where, how many, how much'. Surveys are used to gather data that can reveal patterns and trends: for example a standard system could conduct surveys of individual members of producer organisations (POs) to measure levels of satisfaction with PO performance. Surveys are useful for gathering numbers (e.g. the number of children reached by Goodweave and supported in education programmes, or the numbers of communities participating in equitable benefit sharing). However, surveys on their own are less useful for unpicking how and why changes are happening in any particular context.

Box 2: Case study, survey and experimental research

- Case study research is useful to answer 'how' and 'why' questions, particularly, when there is no control of 'behavioural events' (e.g. a group's decision to become 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 ', 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 – i.e. based on comparing situations 'with' and 'without' an intervention, still may be applied. The 'without' situation is known as the 'counterfactual'.

Experiments require control over behavioural events and therefore raise significant ethical concerns in terms of human rights in international development situations. Experimental impact evaluation such as randomized control trials emerged from clinical trials in medicine where one (treatment) group is given a new drug and another participates but are not given the drug, forming the control group. The study is conducted over a period of time and comparisons are drawn between the two groups at the end of the trial as to whether the drug has had an impact. This kind of 'with' and 'without' comparison is becoming more frequent in international development evaluation (associated usually with statistical surveys, although they can involve mixed methods and/or participatory quantification where resources allow).

It is also the case that in standard system situations there can be selection bias, as certification is applied to organisations rather than individuals. Some groups have already formed prior to adopting a certification and so no baseline can be developed; nor can some groups be prevented from seeking certification. Therefore quasi-experimental approaches have been tried. These use the logic of experimental design, but large numbers of observations are needed and specialised statistical techniques such as Propensity Score Matching, making them costly and fairly complex. As there is no control of the behaviour of the organisations in question it also possible that longitudinal studies can be undermined when selected groups change their strategy during the course of the study – adopting additional standards or dropping a certification and making the intended comparisons either much more difficult or even impossible. Case studies do not require control over behavioural events, and thus are well suited to private sector sustainability standard evaluation studies.

4. HOW DO CASE STUDIES FIT WITHIN THE OVERALL M&E APPROACH BEING SUGGESTED BY ISEAL

ISEAL suggest that there will be three levels in a standard systems' M&E system, namely: Level 1: System-wide monitoring; Level 2: Sampled monitoring; Level 3: Case study research.

Different sets of data will be collected at each level, because each level has its own purpose:

- For Level 1 data will be collected from across the system and the aim will be to provide a comprehensive picture of the standard system.
- For level 2 samples will be selected of producers, organisations, supply chains or communities to provide a representative picture. This kind of data will show major patterns of change, but it is less likely to be able to explain why those changes are happening – especially if counterfactual evidence is not gathered.

- For Level 3 cases will be selected to answer specific questions, for example where more detailed understanding of impact is required or a particular issue needs to be explored in more depth.

Case study research should draw upon the data from Levels 1 and 2, because this broader data can complement the rich data collected within case studies and help to inform the analysis. Without this data, case study research is also more challenging.

If the objective is to learn about impact and to demonstrate impact, then it is important to understand how to do good case study research. Descriptive case study research can be useful for exploring new situations, but it will not answer evaluation questions: to answer evaluation questions as part of case study research requires a conceptual and analytical framework based on a theory. The idea is to test the theory to see if (or under what conditions) the theory holds true.

5. HOW DO I DO CASE STUDY RESEARCH?

5.1 FORMULATE THE RESEARCH QUESTIONS

What do you need to know? If the questions are about 'how and why' impact has happened, then case studies are needed. If the questions relate more to 'what' impact, 'how much', then surveys may be needed. It is important to define specific research questions. It may not be feasible to evaluate the whole of a programme and instead it might be better to prioritize a specific mechanism (e.g. education of rescued child labourers, support to producer organisation strengthening).

i) **Develop a theory.** Standard systems are already developing or have completed formulating their theories of change. These are the theories which now need to be tested through impact studies. This is known as '*theory based evaluation'*.

The theory of change and the specific research/evaluation questions guide the whole study methodology. In theory based evaluation you articulate the theory of change (alternatively called the logic or results chain or impact pathway) first of all. This sets out how expected inputs to lead to expected outputs; how outputs are expected to lead to outcomes; how outcomes are expected to lead to impacts.

Box 3: Differences between theories of change and logical frameworks

Theories of change go beyond traditional logical frameworks, because they focus more explicitly on mapping the causal linkages along the chain. Like logical frameworks, they identify underlying assumptions and risks, but also – in more sophisticated ones – link these to specific transitions in the chain. They are less linear and can show multiple connections and feedback loops. Further, they can show the role of contextual factors – with influence increasing as you move from inputs to impacts. While criticisms are made of both when they are used in an inflexible manner by funders, theory of change thinking can support the development of a clearer consensus amongst stakeholders and those trying to implement a project or programme. They can be developed in a participatory way and this is becoming increasingly common in international development. As the intervention evolves they can be revisited as a monitoring tool to support learning and prompt action and the diagrams

can be adapted to show feedback.

By having a clear articulation of what is supposed to occur in an intervention, e.g. a standard system, it is possible to test the theory and to go back and change it to make it better. This approach can be used to support learning and to enable greater understanding of the different factors at play in any real world context. In a theory of change, the focus is always on the transitions (causal linkages between inputs, outputs, outcomes and impacts). Data collection strategies and methods then flow from the theory of change. Data is carefully gathered along the chain (called 'process tracing') to build up a case as to how and why impact has been achieved or not. Indicators are thus needed at each transition in the chain.

Case study research is a way to conduct this kind of theory based evaluation. It is particularly useful when you do not have the resources to analyse *every possible case* in sufficient depth – something which is the case in many standard systems which cover multiple countries, diverse commodities and supply chains, or where there are not enough cases to support statistical analysis. It also allows you to understand *why* impact has or has not occurred – which a survey cannot tell you.

There are techniques being developed to explore whether other contextual factors played a role in creating change and how much compared to the intervention – in this case the standard system. There is also the option of explicitly exploring alternative plausible factors causing change – known as the 'elimination of rival explanations' and scoring the contributions of different causal factors known as 'contribution analysis'.

Box 4: Participation in case-oriented comparative research

Facilitating a participatory process is not straightforward when adopting a fairly rigorous approach to impact assessment, involving standardized questions and a system wide theory of change. There is utility for standard systems, because such studies can support learning (and thus potential improvement of impact) and because they can be persuasive to wider audiences 'demonstrating impact'– which is something which is necessary for sustainability of the system. This relates to 'upward accountability'.

There are steps that standard systems can take and researchers to increase participation. By investing in a participatory process of developing the theory of change, as Fairtrade International has done, this potentially supports a more empowering approach to impact assessment from the start. This involves engaging *all* key stakeholders across the standard system. Global indicators are being established by standard systems and ISEAL, but there is also scope to support goal setting and indicators by producer organisations themselves and these can be included in monitoring and evaluation activities.

In conducting an evaluation it is also important to take steps to try and ensure that case studies are of use to producers and workers, themselves. This is not easy – there are tensions between learning and accountability objectives – and it does require resources which are often limited. Having committed to a standardized approach to enable cross-comparative analysis in case study evaluation, it is not possible to conduct a study that is driven *only* by producers in terms of design and methodology. Appreciative enquiry is easier in a non-comparative situation. But, it is possible to the orient the study towards the more participatory end of the spectrum by taking certain steps:

- Ensure that producer organisation managers are fully aware of the purpose of the study. They should be properly informed about the study, the methods to be employed and there should be frank discussion of how findings will be used and feedback processes'
- Space should be given within the research process to managers and individual members' views and perspectives. Methods such as participatory video, in particular, could be used to support communications and dialogue on impact issues within the producer organisation, including over time (supporting recall);
- Provide adequate time and resources for feedback of study findings to individual producer organisations. At minimum the research team provide feedback to the PO having conducted the fieldwork. Ideally, a learning alliance could be established in the region (e.g. including other certified producer organisation leaders, and other relevant stakeholders in local government, trade associations, traders, NGOs, trade unions, etc) with on-going support and facilitation to engage in impact issues at a landscape/territorial level and beyond vertical value chain actors;
- Those commissioning the study should commit to taking up recommendations and acting on lessons. Ideally, plans should be in place to develop a response to the study including clear action points.

In selecting methods for enquiry, it is also possible to prioritize participatory research techniques, which are more likely to be of value to farmers and workers – although they can take just as much time and require *more* skilled researchers than in questionnaire surveys. Participatory quantification is feasible – for example, in assessing poverty impact it is possible to undertake a participatory process in which local people define categories of poverty/wealth, and these indices are then used in wealth ranking of households in a particular community. However, this is a fairly complex process and it is resource intensive – it is not a cheaper alternative to large-scale questionnaires. Where a standard system commits to longitudinal case studies, it may be possible to use methods based on existing poverty data such as the Grameen Foundation 'Progress out of Poverty Index'. This draws on extensive existing national surveys on living standards, taking just 10 key indicators, which can then be used in a rapid survey. The findings are then linked to the national living standards surveys to provide a more robust picture of relative poverty. It is not possible, however, to use PPI in a snapshot impact study. Participatory gross margin analysis is a method which is currently being trialled in a Fairtrade coffee impact study (led by NRI). This is an approach which enables farmers themselves to better understand such calculations to inform their decision-making. The same study is also trialling force field analysis to understand organisational change processes comparing the past with now to reveal, alternative causal factors and their relative contribution to impact.

5.2 DEFINE THE CASE(S) TO BE STUDIED.

The case to be studied may be individuals (e.g. child labourers, workers, farmers), but may also be at an organisational level (such as producer organisations) if these are a critical part of the theory of change. Similarly, the case may be a sector, a specific event, or a set of relationships (e.g. value chains). The 'unit of analysis' is chosen depending on the research question: for example, is the question about the impact of certification on individual producers, 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 not. 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 re-join 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.

5.3   PLAN YOUR DATA COLLECTION & SELECT YOUR METHODS

The kind of data required will depend on the priority research questions and on the theory of change. Thus, it is likely that data needs to be collected on inputs, outputs, outcomes and impacts.

The choice of methods should follow the previous steps. Too often the choice of methods is conflated with study design – but in fact study design is a bigger thing and refers to the underlying logic of causality in evaluation. Once case study research has been chosen as the overall research strategy, it is possible – reflecting on the particular research questions – to select appropriate methods. The range of possible methods available is extensive.

Box 5: An array of methods

> There is a wide range of methods available to case study researchers. For example, to name just a few:
> - large-scale questionnaire, statistical surveys
> - participatory quantification
> - scorecard surveys
> - focus group discussions
> - force field analysis
> - strengths, weaknesses, opportunities, threats
> - participatory gross margin analysis
> - individual case studies
> - management interviews
> - semi-structured key informant interviews
> - stakeholder workshops

- elimination of rival explanations
- contribution analysis
- wealth-ranking
- theory of change mapping
- gender action learning
- value chain analysis
- collection of micro-narratives associated with software to support analysis of textual data (coded by participants themselves)

Of course many of these methods can be used in combination.

It is increasingly common to use mixed methods in evaluation, in order to answer a range of types of evaluation questions (what, where, how, when, why etc). Therefore, within a case study, it is possible to employ a range of methods (surveys, focus group discussions, key informant interviews, individual case histories etc.) to support triangulation and because some methods are better are answering certain questions than others. It is also the case that some methods tend to facilitate a more participatory approach than others – but much depends on the overall process and who is driving it, rather than the specific technique. It is worth remembering that mixed methods tend to require more resources than single method designs, but are much more likely to produce rich datasets.

Box 6: Mixed design and mixed methods in a Fairtrade coffee study

In setting up a Fairtrade coffee study the objective is: "*to provide a clear understanding and articulation of the impact of Fairtrade certification for coffee producers and their organizations, drawing on a range of cases that capture some of the important diversity within Fairtrade coffee producing contexts".* The key research questions are: What organizational benefits are there from Fairtrade certification for smallholder coffee cooperatives? How has the Fairtrade approach and strategies led to the benefits identified? Which approaches and strategies are working well, in which contexts, and why? What is the impact of organic and Fairtrade certification on producer organisations? What are the development and economic goals of the Fairtrade certified organizations? What are the impacts at individual household level? To what extent is Fairtrade certification enabling households to meet their own developmental goals / aspirations? What are the broader community impacts, if any? The study should generate recommendations for the Fairtrade system.

Because of the range of types of evaluation questions, a mixed design is needed (both theory based evaluation and counterfactual comparisons) and mixed methods (a questionnaire survey of individual farmers, focus group discussions, participatory gross margin analysis, stakeholder workshop, value chain interviews, management workshops etc).

Source: Nelson and Martin (coffee study research methodology), NRI report (Forthcoming)

Sequencing of different methods can also be valuable if there is sufficient time and resources. For example, conducting some preliminary qualitative interviews to inform questionnaire design, conducting questionnaire surveys and then further qualitative interviews and methods to explore the questionnaire findings and to see what the reasons might be for emerging patterns and trends.

5.4 PLAN YOUR ANALYTICAL STRATEGIES

It is important to understand the choices on offer in terms of analytical strategy – because this may influence research design. Analytical strategies are defined as 'linking data to propositions and establishing the criteria for interpreting findings' (Yin, 2014). Ideally, all of the core team members should fully understand the analytical strategies being used, because this enables all members to support the final analysis – including comparative analysis between cases. There are various options (pattern matching, explanation building, time-series analysis, logic models or theory of change, and cross-case synthesis).

For evaluations by standard systems, case studies should use theory of change analysis: Having carefully gathered data on each transition in the chain and possibly comparing with a counterfactual group, it is then important to establish 'within the case' did the hypothetical theory of change occur in practice, or did other outcomes and impacts emerge. A findings diagram can be produced to match the hypothetical theory of change, although these can be complex for standard systems, as they have multiple strategies and inputs.

The second stage is then to read across the individual cases and compare the similarities and differences in terms of whether expected impact chains occurred, where they did not, what were the unexpected outcomes and impacts. This will examine the primary causal factors – and whether the standard system played the most critical or only role, in the light of other factors in the mix (e.g. other development programmes and projects, a change in government policy, the spread of other sustainability standards, changes in market trends such as prices).

By engaging with stakeholders this can increase the trustworthiness of the analysis – i.e. Is there consensus or disagreements about whether factor x has caused outcome y? However, process tracing on its own does not necessarily support a focused analysis of rival explanations (these could be contextual factors or other interventions). 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. Gathering data on rival explanations could be done either through the researchers conducting a series of key informant interviews and 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.

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. Longer-term studies – few of which have been conducted to date – are important to capture cycles in markets and development trends which may shape the ability of a standard system to have an impact.

Consider the use of software and technology to support analysis: computer-assisted-qualitative data analysis software 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 facilitate the categorization of data and identifying patterns, , but 'developing a rich and full explanation or even a good description of your case', and answering the 'how and why' questions, requires much more post-computer thinking and analysis (Yin, 2014, p134). Other kinds 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.

5.5 FIELDWORK, DATA ANALYSIS AND FEEDBACK PROCESSES

The next step is to conduct the fieldwork. To support comparative analysis between multiple cases there should be a clear standardized set of questions and guidance on data collection, as well as strong central guidance and adequate team training. By developing a draft research guide or protocol this can ensure that different teams have all the right information. It should include clear guidance on how the teams should write up qualitative and participatory exercises and indicate responsibilities for data management.

Feedback should be considered at the study beginning – and adequate resources allocated. It is important that independent research teams and commissioning organisations have clear agreements about what will be published, when, and producer organisations and researchers should have agreements on how the data will be used, and what, if anything, will remain confidential or be anonymized.

Standard organisations should consider how they will take up the findings and plan activities that can support uptake – e.g. planning meetings, development of public responses etc.

6. HOW MANY CASES DO I NEED?

A single case is usually considered to be descriptive. In other words it describes what is happening (e.g. in relation to an individual producer, or an organisation, or a supply chain), but it is not possible to tell whether this will also be the case anywhere else, because it could be a unique, particular or context specific situation. A single case does not allow you to draw any comparisons. If, for example, you conducted a study of certification by one producer organisation in Peru to Rainforest Alliance certification compared with another that was not certified, the findings would allow the researchers to state the findings in that one example. There might be some insights at a wider level, particularly if questions about contextual factors and certification with managers and stakeholders have ranged beyond the individual case, but generally speaking you cannot generalize at all from a single case. A single case study may have other value, but it is not particularly useful for understanding impact, particularly given the breadth of conditions within which standard systems operate, their multi-country reach, the diversity of commodities involved.

In order to be able to generalize beyond an individual case, instead of conducting a single case study, comparative case study research can be undertaken. While it is not possible to generalize to the whole of a standard system, comparative case studies allow researchers to test the theory of change under certain sets of conditions.

By replicating the case in question, the theory is tested more than once – which adds robustness to the evidence. Therefore, theoretically speaking the more cases that are conducted the better, but

there could also be diminishing returns after a series of studies are conducted and resources are often a constraining factorBut each case, *if it uses standardized questions and the theory of change of the standard system*, can add to the evidence base.

If the expected outcomes and impacts do not occur, the idea is then that revisions can be made to the theory of change – supporting improvement.

It is very important to understand that in this approach, which is sometimes called 'case oriented research' we are not using the logic of sampling – where a portion of the whole is selected in a random procedure to represent the entire organisation, project or system. In this logic, we are focused on selecting cases from a typology which we test against the overarching theory. The question thus becomes the number of replications you need or, most likely, the number you can afford. The number of cases is therefore a matter of discretion, rather than the following of a formulaic. In many ways the more the better, but it is possible to start with some cases and build on this by adding in more later – or other researchers might be able to add to the evidence base by following the same standardized propositions in the theory of change. The ability of researchers to manage multiple studies is also 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.

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 children, farmers, supply chains 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.

7. HOW DO I SELECT CASES?

First of all, describe your universe (the standard system). By developing a typology – i.e. identifying the major lines of difference within a standard system – this provides a clear framework from which to select cases.   If major *types* or *categories* are identified, then cases can be selected from each for study (ideally more than one to allow for comparisons to be drawn). It is possible to identify sets of conditions under which impact can be achieved and this then strengthens the inferences which can be drawn.   Developing a typology involves identifying the main lines of difference. What are the main areas of variance within the standard system – is it commodities, geographical locations, types

of value chains? If all three are highly relevant then it may be less ambitious to try and curtail the boundaries of the study by focusing on one commodity but covering different geographical locations and types of value chains. If you then test the theory of change in six or eight countries involving the same commodity and types of value chains then you are more likely to draw insights about impact in coffee. If it is thought by those designing the research that in fact it is the types of value chains which have most influence on impact, then maybe the selection should focus on the types of value chains, selecting cases according to the different types, but should focus on a smaller number of countries, to reduce the complexity at country level. In this case you could draw more insights (than a single study) on how the type of value chain influences impact, although you could not completely generalize that in *every* situation where there is type of value chain x, that y happens. But the more studies the better and the broader data captured in the M&E system of the standard can also support generalization beyond the individual case.

It is possible to select a set of cases across the typology to establish if the theory holds true in *different* locations or situations. However, more cases are still advisable, as single cases chosen as an example of a type or category from the typology could be highly unusual or particular and the variation can make it difficult to draw conclusions.   A second option is to choose several cases that are similar to each other – to see if the theory of change holds true for this type. This approach provides more tests of the theory for a particular set of conditions, but this relies on having prior information for selection.

It is also possible to include repeats of cases where the findings are <u>not</u> expected – this is another form of counterfactual.

Box 7: Selecting cases

In a Fairtrade coffee study, we have begun by mapping the Fairtrade coffee system and have identified the main lines of difference occurring within Fairtrade coffee.   There is a wide spread of Fairtrade-certified producer organizations across the world. In order to select cases based on a clear typology of the universe of Fairtrade coffee, we first established the countries in which Fairtrade coffee is produced, in order to select the four countries of the study, which the budget would allow us to cover, and to allow for inclusion of all the three main geographic regions. The number of certified organizations varies greatly within each region – 268 in Latin America, 32 in Africa and 29 in Asia – Oceania). Although the overall distribution of Fairtrade coffee producing organizations is heavily weighted towards Latin America, this research recognizes that African and Asian coffee co-operatives have not been sufficiently covered in previous research in Fairtrade coffee, and aims to help redress that balance.

A matrix has been developed with the following criteria:
-    the number of certified Arabica ONLY producer organizations which have been certified prior to 2010 onwards) drawing on excel files shared by Fairtrade International;
-    Fairtrade certified production (drawing on excel files shared by Fairtrade International;)
-    Percentage of the Fairtrade market (drawing on excel files shared by Fairtrade International)
-    Fairtrade coffee exports for the top 10 countries (FLO Monitoring report, 2012).
-    % sold on Fairtrade terms (FLO Monitoring Report, 2012).

Selecting POs producing Arabica coffee (as guided by the client), POs producing both Arabica and Robusta, were excluded from the selection. As this is an impact study we identified a time period -

2010 onwards – as our best estimate of the amount of time in which it is possible to see impact, because this means that there is a sufficient period within which the PO and members can have experienced benefits and costs of participation in Fairtrade certification. From the analysis we identified the top countries for selection (covering at least one country from South America, North and Central America, Africa and Asia) and these were Indonesia, Tanzania, Peru and Mexico. Where relevant security concerns would be taken into account. Having chosen the countries we have then selected the producer organisations. This approach represents a new step for FLO commissioned impact studies. It builds an external perception of independence, but it does rely on having or being able to access information to support selection from the field. The team has sought to gather contextual information *before* selecting cases for the certified groups and the comparison non-certified groups. However, without funding for dedicated visits to gather this information, it is necessary to rely on FLO Producer Support Services staff and research team contacts and prior knowledge.   Obtaining information on non-certified organisations can be the most challenging. Rarely is information held centrally by one organisation and so organisations have to be contacted individually which is resource intensive. Questions of location within country also need to be taken into account for logistical and budgetary reasons, but also to be able to match certified and non-certified POs and farmers in terms of their agro-ecological contexts.

Constructing a counterfactual is becoming harder where certification is already widespread in an industry - as the remaining organisations which do not yet hold certification may have specific characteristics which confound the comparison (e.g. they are very new, or do not export directly themselves as the certified ones do). In Tanzania only one organisation was suggested in the focal zone selected that does not hold other certifications – but it used to be Fairtrade certified, still follows Fairtrade principles, and the Japanese buyer pays the premium but does not use the label. Thus, it is not a proper counterfactual We have sought to match the organisations in the study (e.g. in terms of size, export modalities and location).   For the counterfactual, non-certified comparison, the first choice in our study is: a) a non-certified producer organisation in the same zone which has similar features as the certified PO (e.g. size, export model). If this cannot be found, then the comparison group should be b) a sample of non-organized farmers – i.e. they sell individually to other buyers. In each case we need to decide what provides the best comparison – and document this for the final report.

8. HOW DO I DO COMPARATIVE CASE STUDIES?

The theory of change remains the guiding framework when conducting multiple case studies. The evaluation questions should flow from the theory of change. If assessing impact, the evaluation questions will revolve around the inputs, outputs, outcomes and impacts – and the defined indicators of these.

It is important to first of all conduct the 'within case' analysis. Does the theory of change hold true? Have inputs led to outputs, outcomes and impact or did they not? Were other factors more important? Having analysed the specific case, then comparative analysis can be conducted. Did the inputs lead to outputs, outcomes and impacts in each case? By reading across each analysis it is possible to make comparisons – did impact occur in each case? Did the same extent of impact occur? Were the same factors causing impact or different ones?

Box 8: Union of Ethical Biotrade

The Union of Ethical Biotrade (UEBT) standard was established in 2007. UEBT is a non-profit association, involving an association of companies and non-private sector organisations. It is committed to ethical sourcing of biodiversity and promotes 'Sourcing with Respect' of ingredients that come from biodiversity. Members commit to gradually ensuring that their sourcing practices promote the conservation of biodiversity, respect traditional knowledge and assure the equitable sharing of benefits all along the supply chain. The UEBT mission is the advancement of sustainable business growth, local development and biodiversity conservation. By 2020, UEBT aims to achieve a situation where ethically sourced biodiversity-based products and services make a significant contribution to local sustainable development, through conservation and sustainable use of biodiversity, as well as the fair and equitable sharing of benefits; UEBT aims to bring together actors committed to Ethical BioTrade, and promotes, facilitates and recognises ethical sourcing of biodiversity in line with the objectives of the Convention on Biological Diversity.

Measures seek to conserve biodiversity and reduce negative impacts in sourcing areas. Local development is supported through equitable sharing of benefits generated through the use of biodiversity – i.e. fair trade practices and sharing of benefits from biodiversity and traditional knowledge innovations. Practices supporting the rights of workers, local and indigenous communities are included in the standard and respect of rights linked to natural resources. Finally, practices are included which encourage the identification of risks in sourcing and takes into account the economic viability of companies.

UEBT has articulated its theory of change. Interestingly, UEBT explicitly states that it operates as a standard system, but also as an association and as an agent of change. UEBT has begun to extend its overall monitoring and evaluation framework, including developing key indicators for data collection. UEBT has decided to collect performance and some impact data annually from its members. Data reports from members will be verified in the UEBT audits, but self-assessment represents a key component of the system. This data should give a general idea of the scope of the impact of the Ethical BioTrade commitment of UEBT members and indication of direction of this impact. It plans to issue a public report on the data submitted by members in 2014 (after the members submit their reports in the 1st quarter of 2014).

In addition, UEBT plans to establish field sites for in-depth impact studies, to provide UEBT with a better understanding of the impact of its standard and to be better able to interpret the data submitted by all its members. Some external expertise will be engaged for this work, with support from ISEAL. The UEBT theory of change identifies medium term and longer-term outcomes (the terminology used in their TOC). A medium term outcome is the payment of equitable prices for natural ingredients. A longer term outcome is that benefit sharing contributes to local development. The in-depth case studies are intended to be conducted over a number of years and to provide insights on the latter, i.e. longer-term outcomes.

The sites are intended to provide feedback on the effectiveness of the measures in the UEBT standard for that particular site and these findings will inform future standard revision processes. Thus, there is a clear feedback loop – the theory is tested and changes will be made to maximize impact in each case, but also the findings will feedback into overall standard system theory of change revision processes.

There is therefore a clear vision in UEBT of the M&E framework with linkages and complementarity identified between the system-wide Level 1 data gathering and the Level 3 in-depth case studies. Level 1 involves mainly self-assessment, verified by audit reports, and this data will focus on medium-term outcomes and will enable 'how much' and 'where' type evaluation questions to be

answered, because the data will be wide-ranging across contexts and value chains. The in-depth Level 3 data, will be collected with support from external researchers and therefore providing an additional verification to demonstrate impact, focusing on the 'why' and 'how' questions and ensuring the exercise supports standard system learning.

In practical terms UEBT has begun to characterize its universe, developing a typology which includes the different supply chains which UEBT is working with. There is a wide variety of supply chains already covered by the UEBT system, and there are associated collection and cultivation practices. For example, Allanblackia or Genipa (which is collected from the wild), Phys and orange which are produced by smallholders, and orange and sugar cane produced by commercial farmers. Other criteria for selection include:
- importance in terms of volumes of sales (small, medium, large);
- ecosystem (e.g. high Andes; Humid forest Andean/Pacific; Amazon; Cerrado; dry forest etc.)
- level of development of commercial activity/supply chain maturity (e.g. initial, medium, mature). Some are in the initial stages, others are rapidly growing and some are mature.
- engagement with other sustainability standards. In some cases UEBT has been applied first, in other instances groups have had other social and environmental support projects and are now beginning with UEBT, and in other cases a group has worked with other standards and already fulfilled many of the UEBT requirements (e.g. Organic).

Another line of variation is being considered. In some situations the UEBT standard is being implemented in supply chains in an accelerated way. In other cases the intervention from UEBT has been mainly at the management system level in the company in question. Tracing the impacts of changes in management systems at a more central level across a range of supply chains is more complex as the theory of change is more extended and could be less direct, than in the former situation. Some studies are being conducted within partner companies by students to try and gather some information on how changes may be occurring within the company and the follow-on effects, but this work is at an early stage.

Some practical criteria are also relevant and will shape the final choice of sites:
- Availability of external funding, at least for baseline;
- Availability of partner organisation to help with the research;
- Interest of UEBT member to contribute to impact assessment by collecting information and sharing with the research organisation.

Next steps include the categorisation of the UEBT supply chains, selection of sites and local organisation and development of field testing methodology, additional indicators data collection plans and analysis. This will be followed by the baseline and further monitoring and impact assessments.

It is possible, once cases are selected, to identify a variety of producer organisations – i.e. to construct a multi-level case study. If only single producer organisations are covered, then there is not much scope for generalizing beyond that single situation.

Interestingly, it is not clear from current plans, whether the case studies are relying purely on generative causation (i.e. establishing 'within case' what generated change by tracing along the impact chain) or will be including counterfactual logics – i.e. matching the case studies groups to similar groups in the same site who are part of supply chains where there has been no UEBT involvement.

Various secondary impacts are identified in the theory of change, including: i) market

transformation; ii) access and benefit sharing awareness; iii) non-member companies apply Ethical Biotrade practices (spillover effects); and iv) consumer awareness about biodiversity sourcing increases. While consumer awareness can be explored through market research, market transformation requires a different type of research strategy, although theory based evaluation is very appropriate – establishing the reasons for impact requires tracing through the inputs, outputs, outcomes and impacts and facilitation of stakeholder debate on the reasons for changes in market behaviour, culture and practices. Similarly, evaluating increased awareness of access and benefit sharing requires careful tracing of changes in behaviour, knowledge and practices and stakeholder analysis of the reasons for changes to build a credible evidence argument. In terms of spillover effects at non-member companies, this again requires interviewing managers at non-member companies and key informants who are independent to understand if they are making changes and if so, what these changes are driven by (e.g. are they copying the changes made by UEBT certified companies, or have they made the changes for another reason). Triangulation of data (e.g. drawing from multiple sources and types of data) is important in all case studies to improve the trustworthiness of the findings and to reduce some forms of bias.

9. HOW DO I ANALYSE THE DATA?

The data is used to interrogate the theory of change – what has happened, when? Did outputs lead to outcomes? Did outcomes lead to impacts? Were there other unintended or unexpected changes? What caused the changes? If the findings indicate that the standard systems' inputs did lead to the outputs, outcomes and impacts as expected, this provides a 'within case' analysis. From this you can say that in this case, there was x impact, caused by y and z factors.

It is very important to design the study to gather the appropriate data. This means clearly setting out the research questions, and working out what data is needed to answer these questions. The data should, when put together, cover all of the research questions and the data should trace the impact chain. For example, it is very important to focus on the transitions – i.e. how do inputs lead to outputs. This is building the evidence – and the report should enable the reader to see how the evidence has been collected and analysed.

Box 9: Complexity in evaluation of market based mechanisms for sustainable development

"No-one ever got fired for demanding more rigour, but there are many reasons why it is not practical to have an evaluation of every programme using only statistical methods. One is cost; the expertise and sample sizes required are considerable. Another is that it is very difficult to measure the impact of programmes that aim to have market-wide impacts and spill-overs, in rapidly-changing circumstances with self-selecting partners, using randomised controlled trials and other statistical techniques.

The alternative is a robust approach based on results chains, as many M and E specialists are increasingly agreeing. As John Maynard Keynes said, "it is better to be roughly right than precisely wrong". In addition, programmes are finding that the robust approach of the Standard is a highly effective management tool, guiding the work to be more effective during implementation, in the light

of lessons being learned, which, ultimately, is more important than precise numbers".

Source: The DCED Standard for Results Measurement "Frequently Asked Questions"
 – or 'Why is it the way it is?', July, 2010, p3)

10. FINDING THE RIGHT TEAM AND BUILDING RESEARCH CAPACITY

To conduct case study research in evaluation it is important to have an experienced research team, because the procedures for data collection are not routinized. Competent researchers can ask the right questions, listen well, adapt to changing circumstances, avoid biases, be familiar with ethical practice and have a firm grasp of the issues being studied (Yin, 2014). Specifically, standard systems should carefully construct terms of reference (with stakeholder comment) to ensure that: i) the research team have sufficiently good skills and experience; ii) timelines are appropriate and clearly agreed between parties, with specific milestones and deliverables; iii) the budget matches the ambition of the study and *vice versa*; iv) opportunities and strategies for evaluation capacity building in relation to the standard are considered; v) tendering for researchers is conducted in a transparent process to ensure perceived independence – including scoring of bids against terms of reference requirements; vi) 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. Such terms of reference should set out the mechanisms by which the sustainability standard will learn from the findings (e.g. feedback processes, learning alliances, meetings and workshops, official responses) and there should be clarity about confidentiality, anonymization approach, and what will be published where and when.

BIBLIOGRAPHY

Collier, D. (2011) 'Understanding process tracing'. PS Political Science and Politics 44, No.4, :823-30.

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

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