# Evaluating the impact of sustainability standards

## Lessons learnt on research design and methods from three impact evaluations

MARCH 2017

iseal alliance

# CONTENTS

# ABBREVIATIONS

| | |
|---|---|
| BCI | Better Cotton Initiative |
| COSA | Committee on Sustainability Assessment |
| DIPI | Demonstrating and Improving Poverty Impacts |
| M&E | Monitoring and Evaluation |
| MPI | Multidimensional Poverty Index |
| NRI | Natural Resources Institute |
| PPI | Progress out of Poverty Index |
| PSM | Propensity Score Matching |
| RA | Rainforest Alliance |
| RCT | Randomised Controlled Trial |
| SAN | Sustainable Agriculture Network |
| ToC | Theory of Change |

# 1. Setting the scene: impact evaluation in the standards' world

Impact evaluation is broadly understood as a study process that systematically and empirically investigates the impacts of an intervention (Rogers, Patricia J., 2012)[1]. Embedded within this definition are core concepts that every evaluation grapples with – causation, comparison, counterfactual thinking, systematic enquiry and the nature and scope of empirical investigation. With the increased uptake of sustainability standards, the impact evaluation of these interventions is also on the rise. However, sustainability standards differ from other development initiatives in important ways – they are complex, market-based approaches that use a package of activities and interventions to produce a range of desired sustainability outcomes.

### What makes the impact evaluation of sustainability standards different and difficult?

To claim that impact evaluation of sustainability standards is different might sound trite, for surely every impact evaluation, irrespective of sector, region or scope is unique. Further still, to say that impact evaluation of sustainability standards is difficult might come across as a case of sour grapes or an admission of failure. However, with a recent spurt in impact evaluations on certification, especially in the agriculture sector, common methodological observations made by researchers about how sustainability standards are unique are beginning to emerge:

- *Market-based instruments*: Sustainability standards are complex market-based approaches that usually apply to production or trading practices of a specific product or sector. Their interventions are not geographically bounded and extend to entire product supply chains (such as coffee or cocoa) or whole sectors (such as forestry or fishery). They are adopted by market actors at different ends of the supply chains, from smallholder farmer to food retailer, but in very different ways. As instruments 'in the market', their uptake is often up to the market entity and beyond the control of the standard system, which makes shaping the intervention to embed evaluation practically very difficult. This makes conducting randomised control trials particularly difficult.

- *Diversity in implementation contexts*: Most sustainability standards, though not all, have global applicability and reach. This creates high diversity in the way they are adopted and implemented across different contexts, which ultimately results in high variance in what change they can or do bring out in various contexts. This creates two problems for impact evaluation – that of generalisability and that of conflicting results across different contexts.

- *Increase in uptake makes finding counterfactuals difficult*: Standard systems focus on getting the market to start using standards as the first step of implementation, without which impact evaluation would be impossible. However, when such uptake occurs at the level of an entire region or jurisdiction or landscape, no counterfactual to the intervention to use in an impact evaluation may actually exist. Researchers call this 'the missing counterfactual'. It is most noticeable in tea, coffee and cocoa sectors where standards adoption is so high that more than 90% of farm production or households in an area use these standards.

- *Multiple certification*: A related issue is that more than one sustainability standard may be operating in the same sector in the same region, resulting in entities (mostly farmers groups or producer organisations) being 'multiple certified' i.e. certified to more than one standard. With the same group of farmers receiving very similar 'interventions' from different standards, coupled with lack of clarity on upstream impacts through the supply chain, disentangling the specific impact of individual standards is getting more difficult.

- *Certified entities and sample size issues*: Given standards have little control over which entities adopt them, making sure the intervention group is of a size amenable to sample size selection criteria is difficult. One can encounter standards implemented with groups of 25 or 45,000 farmers which makes sample size determination difficult.

The specific design of standards (and the standards systems that implement them) requires customised approaches to impact evaluation with innovative designs and a creative combination of methods and tools to produce relevant and robust results.

[1] This paper adopts the ISEAL Impacts Code definition of 'impacts' as positive and negative long-term effects resulting from the implementation of a standards system either directly or indirectly, intended or unintended. For more on the varied definitions and interpretations of 'impact', read the ODI's 'What is Impact?' briefing paper (Hearn and Buffardi, 2016).

Impact evaluation of sustainability standards is still a nascent field of enquiry and much can be learned from the design and implementation of ongoing evaluations to inform future work. Researchers active in this field have already made important contributions to our methodological learning, notably in Blackman and Riviera (2010); Crosse, Newsom and Kennedy (2011 for the Resolve Report); Nelson and Martin (2014); Ton, Vellema and Ge (2014); and Milder et al. (2016).

This methodological paper from ISEAL aims to make a contribution to this body of work by sharing insights and lessons learned from three ongoing impact evaluations that completed their baseline in 2016 and are due for end line evaluation in 2019. Our observations are targeted primarily at standards systems, and specifically their monitoring and evaluation (M&E) teams, but it is hoped that researchers and others will find it a useful read. Ultimately, our hope is that this is not read as a paper purely about methods and approaches to evaluation but helps understand the systems that we attempt to evaluate at a more fundamental level. ISEAL welcomes all feedback to this paper and a keen encouragement to keep the methods debate alive.

The specific design and functioning of standards as development interventions requires **customised approaches** to impact evaluation with **innovative designs** and a creative combination of methods to produce relevant and robust results.

# 2. The three DIPI impact evaluations and the focus of this paper

The three impact evaluations in question are a part of the ongoing Demonstrating and Improving Poverty Impacts (DIPI) project that is working to understand the contribution that certification makes towards sustainable livelihoods and pro-poor development practices.

**Demonstrating and Improving Poverty Impacts (DIPI) Project**

Through support from the Ford Foundation, ISEAL and its members have been working to understand the contribution that certification makes towards sustainable rural livelihoods and pro-poor development practices. The first phase of the project from 2012-14 focused on agreement on a set of common indicators to track poverty impacts, a common research agenda, and the development of strong monitoring and evaluation systems. In the second phase from 2015-2016, to help contribute to high quality research on the impacts of certification, ISEAL commissioned three evaluations that aimed to answer the important question of whether certification improves the livelihoods of smallholders and contributes to poverty alleviation. The evaluations were commissioned with the dual objectives of generating usable data and findings about the contribution of standard systems to poverty reduction and of testing methodologies and promoting consistency and coordination in the approaches that ISEAL members use in assessing the poverty and livelihood outcomes and impacts of their systems. For more information on the project and outputs visit: http://www.isealalliance.org/online-community/resources/iseal-dipi-project-three-commissioned-impact-evaluations-factsheet-on-baseline-stud

The evaluations study the emerging long-term impacts of different sustainability standards in the agriculture sector from the pre-certification stage to three years after certification. Evaluations were conducted in three varied product-geography contexts, focusing on a different standard or set of standards but with a common research focus on supporting the livelihood and income of poor smallholder farmers. The three case studies are: **smallholder coffee production** in **Western Kenya** with **Fairtrade and UTZ** certified farmers' organisations (**research partner: Committee on Sustainability Assessment COSA**); **smallholder cotton production** in **southern India** through the **Better Cotton Initiative (BCI)** standard (**research partner Natural Resources Institute, University of Greenwich**) and **smallholder coffee production in Indonesia** certified to both the **4C (now Global Coffee Platform)** and **Rainforest Alliance/SAN (RA/SAN)** standards **(research partner: University of Sydney**).

Western Kenya

Smallholder coffee farmers

- UTZ and Fairtrade
- Research partner: Committee on Sustainability Assessment (COSA), with IITA

Photo © UTZ

Andhra Pradesh, India

Cotton

- Better Cotton Initiative
- Research partner: Natural Resources Institute, University of Greenwich

Photo © Better Cotton Initiative

South Sumatra, Indonesia

Smallholder coffee farmers

- 4C and Rainforest Alliance
- Research partner: University of Sydney with J-PAL (Poverty Action Lab) and collaboration with SurveyMETER

Photo © David Bonilla, Rainforest Alliance

*Source: ISEAL Alliance, 2015*

More specifically from a methodological standpoint, our objectives were to:

- Capture changes over time on selected households and certified entities from a pre-certification change
- Compare differences between certified and non-certified entities as the intervention progresses
- Differentiate between the effects of training and practice adoption versus the effects of certification itself
- Understand contextual factors that affect the changes that certification can bring about
- Test the causal chains of the sustainability standards that are the focus of the evaluations
- Use the ISEAL common core indicators developed in phase one of the DIPI project

This paper shares insights on what we have learnt so far from these evaluations after completion of their baseline stage[2]. To capture cross-cutting methodological learnings, it is useful to first understand the broad similarities and differences between the three evaluations as presented in the table below. It is important to note that although the main research focus of all three evaluations is the same - to understand the effects of certification on smallholder livelihoods and certified entities and the contextual factors that influence them - there are differences in the specific research questions in each case as influenced by the standards under study, geography and product contexts[3].

As the table below indicates, each of the three DIPI evaluations makes a specific contribution to our understanding of methods and approaches in this field of study. However, our attempt in this paper is not to delve into every specific tool, approach or indicator in these evaluations as these are detailed in the three individual research design papers that are available to read. Our attempt in this paper is to address fundamental conceptual questions that arise in the course of designing and undertaking impact evaluations of sustainability standards and to share cross-cutting learning from how the DIPI evaluations addressed them. These include questions such as: What is the role of theory-based evaluation in this field? How do we understand and study 'treatment' in the context of standards? What are selection effects and why should they matter? What constitutes credible counterfactuals? How and when can randomisation be achieved? What is meant by mixed methods? **We focus on these questions as we consider them to be fundamental concepts that every impact evaluation in this field will encounter, irrespective of standard, sector or geography. A focus on these questions will also help strengthen the approach and robustness of impact evaluations undertaken by standards.**

Consequently, this paper should not be treated as an exhaustive summary of all the methodological insights that the DIPI baselines have to offer. It also does not assess the extent to which the design and methods used in the evaluations answers the main research questions as that will only be fully possible after the completion of the end line stage in 2019. There is a lot of rich learning on specific topics such as sampling approaches, use of ISEAL indicators and specific tools such as surveys that this paper will not go into but that we hope can be discussed at a future point. The scope of this paper is limited to broad learning on evaluation design, unpacking fundamental concepts and tackling common challenges in conducting impact evaluations with practical recommendations for standard systems.

[2] The full baseline reports and accompanying research design documents can be accessed here: http://www.isealalliance.org/online-community/resources/iseal-dipi-project-three-commissioned-impact-evaluations-baseline-full-reports-and-

[3] See Annex 1 for a description of the specific research questions of each DIPI impact evaluation

**Similarities and differences between the research design of the three DIPI evaluations**

| Topic | Similarities | Differences | | |
|---|---|---|---|---|
| | | Coffee, Kenya | Cotton, India | Coffee, Indonesia |
| Research Design | Experimental or quasi-experimental evaluation design that captures the differences between certified and non-certified entities and households | Quasi-experimental design using difference-in-difference approach with a contribution analysis framework (with 2 treatment groups and 4 control groups) | Quasi-experimental RCT with a cluster-based approach along embedded with and theory-based evaluation approach. | Quasi-experimental cluster RCT (study one treatment - change from 4C to RA/SAN) and propensity score matching (the other treatment – 4C vs non 4C). |
| Use of theory | Theory-based evaluation aimed at testing specific causal pathways of the intervention in each case | Constructed causal chain for the intervention with evaluation focus on select indicators of interest | Complete theory-based evaluation approach with detailed theory of change (ToC) constructed for the intervention and analysis along ToC pathways | Simple ToC constructed for the intervention but analysis limited to select indicators |
| Research Methods | Mixed methods that include a combination of quantitative techniques and analysis supported by qualitative methods | Participatory rural appraisal, producer organisation survey, household survey, farmer focus groups and key informant interviews | Household survey, additional household panel survey, farmer focus groups, key informant interviews | Household survey, village case studies, survey of farmer perceptions about standards, stakeholder interviews |
| Household survey sample size | A large household survey as the main source of data collection for the evaluation | 696 (120 households from each of the 6 select producer organisations) with buffer for 10% attrition at end lines | 729 households (320 each for treatment and control households) with buffer for 35% attrition at end line | 1588 (979 households -RCT treatment + control; 609 households - 4C + control) |
| Poverty Analysis | Use of Progress out of Poverty Index (PPI) as an index for understanding poverty status of households | Income-based poverty measure; UN's Multidimensional Poverty Index (MPI); 696 (120 households from each of the 6 select producer organisations) | Income-based poverty measure benchmarked against poverty lines; Asset-based poverty measure based on an asset ownership index, Progress out of Poverty Index (PPI) | Simple income-based poverty measure; Progress out of Poverty Index (PPI) |

This paper discusses **fundamental conceptual questions** on designing and undertaking impact evaluations of sustainability **standards** and how to address them – based on 3 evaluations commissioned by ISEAL that completed their baseline in 2016.

# 3. Theory-based evaluation approaches: how useful are they for evaluating standards?

Although not a new approach, theory-based evaluation designs and approaches are gaining currency in the field of impact evaluation, especially in the context of development initiatives. The argument is that theory based approaches enhance the policy relevance and learning from evaluation by focusing not just on what changed but why and how it changed, in specific contexts, thereby providing a framework to interpret evaluation data and findings. For sustainability standards, especially those that are ISEAL members, theory-based evaluation approaches can be an attractive option for the reasons above, but also because many systems already have an articulated theory of change (or at least causal pathways) in place that explains their intervention logic (as part of their compliance with the ISEAL Impacts Code).

But what is the explanatory power of theory based approaches? How far can they support evaluation efforts and what are the challenges in adopting such an evaluation design? We share a few insights from the three DIPI evaluations, all of which used the theory-based evaluation approach, albeit in different ways.

## 3.1 What is theory-based evaluation?

A theory-based evaluation design is one in which the analysis is conducted along the length of the causal chain (or pathway) from inputs and interventions to impacts and where evaluation examines the links and assumptions in those chains (White, 2006). The first principle of using a theory-based evaluation approach is to map out the causal chain of the 'intervention' or 'treatment' that is under evaluation (White, 2009). In the case of the DIPI impact evaluations, all three research teams constructed causal chains and frameworks for the three evaluation study contexts (see Annex for these). These serve as useful individual examples of what a constructed theory or causal pathway for the purposes of impact evaluation could look like.

In the case of the DIPI India cotton study, the Natural Resources Institute (NRI) research team constructed a detailed theory of change for BCI's project in Adoni Mandal - the site of the intervention and evaluation. This was necessary in the absence of an existing theory of change for the BCI system and was built in close consultation with the global and local BCI teams and the implementation partner executing the project in Adoni. It lists the intended outputs and outcomes that mostly focus on production and farming changes based on the BCI Production Standard, whilst impacts focus on the intended livelihood related changes at the household level. The acceptance of this programme theory of change was an important step in the baseline stage of the evaluation and served as a useful mechanism to get project partners on the same page about the intervention and about the evaluation.

In the case of the DIPI Indonesia coffee study, the University of Sydney research team constructed a simple theory of change, listing only the most significant outputs and outcomes they were focusing on in the evaluation. The criteria are mainly environmental, owing to the focus of the RA/SAN standard that is the subject of the evaluation in Indonesia.

In the case of the DIPI Kenya study, the Committee on Sustainability Assessment (COSA) team developed a 'contribution analysis' framework to support attribution of results from the quantitative analysis. The framework follows an actor-approach, identifying the activities to be undertaken by various actors implementing the intervention. It emphasises changes at the producer organisation level and household level, reflecting the nature of certification by Fairtrade and UTZ and also highlighting market-related dynamics as that is significant in the Kenyan coffee context under study.

Despite the differences in these frameworks, it's worth highlighting that all three recognise three key components of standards in the ways they define 'inputs' or 'interventions': a) training and direct field support to farmers and farmers' organisations b) market linkages through certification, access to market and chain of custody benefits and c) assurance and independent auditing as a standalone intervention associated with standards. It's worth keeping this in mind as we explore the complex intervention landscape of standards in the next section.

The full merits and challenges of using these theory-based evaluation approaches can be comprehended only after the three DIPI evaluations are completed. However, there are learnings put forth by the three research teams even from the baseline stage that we explore here.

| The use of theory-based evaluation approaches in the field of sustainability standards | |
|---|---|
| Strengths & Merits | Challenges & Limitations |
| Helps establish and test a generic theory of change, or specific causal pathways associated with the standard system in an empirical way | Testing a standard system's full theory of change is often outside the scope of a single outcome or impact evaluation necessitating selection of focus pathways to test |
| Supports attribution analysis to disentangle the role of the standard system's intervention vis-à-vis other interventions | Can be limiting in ability to identifying unintended or negative consequences from a standard system's intervention as these are, by definition, not captured in a theory of change |
| Can highlight which particular assumptions hold true or not in given product-geography contexts, thereby improving the practical and policy relevance of evaluation results | The way a standard intervenes in a particular project or geographic context may often differ greatly from its generic theory of change and this can make generalisation difficult from a given impact evaluation |
| When used at baseline stage, can help identify weak links in the causal pathway/s that can be monitored through the implementation phase | Generic theories of change are often not time-sensitive i.e. they do not capture the timing of when change happens along causal pathway, even though this can often be a deciding factor in the results of an impact evaluation |
| | A theory of change is often static (at least in the short-term) and the nature of the intervention might change as the evaluation progresses. Capturing the dynamic nature of programme interventions is important and should be built into the study design |

## 3.2 When to use a generic theory of change and a project theory of change?

An increasing trend in theory-based evaluation approaches is of research teams themselves 'constructing' their interpretations of the intervention's ToC – usually at the project or local level. It is not essential for every impact evaluation using a theory-based approach to construct a specific, localised project theory of change, this is usually done when a generic theory of change does not exist. However, we are seeing increasing cases of evaluations constructing a specific project theory of change even in cases where a generic one exists. Does this simplify or complicate evaluation and does it help or hinder adoption of learning by practitioners?

In the DIPI evaluations, all three frameworks developed are, to some extent, constructed localised theories of the intended change pathways in each case study content. In the case of the two coffee studies, these frameworks drew from the standards system's generic ToCs but in the case of the BCI cotton study, this was constructed from start in the absence of an existing BCI ToC.

In this context, it is very useful to note the difference between the general theory of change of a standard system as a whole (such as it would be for RA/SAN, UTZ or Fairtrade systems as a whole) and the specific project theories of change that the researchers constructed for the purposes of the DIPI evaluations. The former usually depicts the holistic vision of what change the system hopes to make (at a global, generic level) and therefore what they wish to evaluate while the latter is closer to what happens on the ground with that standard, so sets a more accurate framework for evaluating that standard in that context.

The decision on whether to use a system's generic theory of change or construct a specific one for the evaluation context is an important one. The choice often comes down to the extent of context-specificity or generalisability that the evaluation wants to achieve and also how unique the particular intervention or study context is. If the nature of the intervention is highly unique in the study context, it would be more relevant and accurate to develop a localised ToC to use as the framework for a theory-based evaluation as using a generic one could be misleading. On the other hand, if the evaluation is testing generic causal pathways of a system in multiple sites in the same evaluation, the generic ToC might be the more appropriate framework to use. Also, contrasting their generic theory of change with project theories of change (where constructed) could generate useful learning for standard systems on how they operate across different contexts. Below we share brief thoughts on the pros and cons of using project theories of change for evaluation purposes. As such, our recommendation is that more thought should be given upfront by researchers and standards to this question of what kind of a ToC is adopted to guide theory-based evaluations.

**The pros and cons of using a project theory of change as the basis for evaluation**

| Pros | Cons |
|---|---|
| Provides a sharper and more accurate framework for the specific evaluation context as is closer to the ground reality and site of implementation | Can restrict the scope of the evaluation to very specific interventions of the standard that have played out in that project context, limiting generalisability of findings and claims |
| Useful in evaluations that are very specific to one region or one production context only | Not suitable for evaluations that compare certified entities situated in very different contexts or completely different sectors |
| If constructed for specific contexts, would generate more contextually true and relevant results for the system | Given project implementation is highly dependent on project partners at the field level, it could end up being a theory of how they work and create change rather than the standard system itself. |
| As the theoretical framework is rooted in the project context, such an evaluation might be better able to provide feedback on implementation quality and effectiveness and learning for improvement | Standards implement a range of interventions not all of which might be implemented in very study context to the same extent. It is likely that project ToCs focus on evaluating the particular interventions that standards implement in that study context. Such a focus could reduce insight on the comparative effectiveness of multiple interventions in the same context, one of the strongest advantages of theory-based evaluation. A possible solution is to avoid 'picking' interventions but just use the project ToC to contextualise them better. |

Despite these caveats, we do believe that theory-based approaches provide distinct opportunities to ISEAL member standards to test the effectiveness and impact of their interventions either generically or in specific contexts. For this, they first need to be embedded more deeply within our evaluation approach and designs than is currently the norm.

**Theory-based approaches** provide distinct opportunities to test the effectiveness and impact of sustainability standards in general and for specific contexts. This requires **embedding theory within our evaluation approach** and designs from the start.

# 4. Focusing on the factual: understanding ‘treatment’ and ‘intervention’ in the context of sustainability standards

Methodological debates on impact evaluation in development contexts tend to focus on the difficulties of identifying appropriate counterfactuals. This is true of emerging debates in the standards arena as well, but as the three DIPI baselines indicate, adequate and appropriate analysis of the ‘factual’ or ‘treatment’ can prove to be equally challenging. In summarising the experiences of the independent evaluation team at the World Bank, Howard White notes the ‘importance of the factual’:

> “While a well-constructed counterfactual is central to establishing impact, the importance of the factual – what actually happened – should not be overlooked. Constructing a picture of how the intervention has played out on the ground, which nearly always requires data from the treatment group alone and not from a comparison group, is essential to a good impact evaluation, shedding light on the findings."

White, 2006, p10

In this section we unpack why this is the case, what concepts need consideration and how we can achieve a better understanding of ‘treatment’ in the standards’ world. We try and navigate the challenges of answering two questions – impact of what and impact on what in the contexts of standards.

## 4.1 What is ‘treatment’ in the context of sustainability standards?

Borrowed from the field of medical experimentation, the term ‘treatment’ is used synonymously with the word ‘intervention’ in impact evaluation to mean “the specific project, program, design, innovation, or policy to be evaluated” (Gertler et.al, 2016, p329). Sustainability standards are complex mechanisms that can make understanding the nature of ‘treatment’ or the ‘factual’ difficult in studies that evaluate them.

The first point to note is that in the standards’ world, there is often no single treatment or intervention - what exists is a ‘package of interventions’ that is implemented very differently in different contexts. The ‘package’ model of intervention infuses complexity into the evaluation that then needs to untangle which part of the package made what difference, how much of a difference and to whom. Further, treatment is multi-layered and often goes beyond what the individual standard-setting body itself implements. The table below captures the nature of these ‘packages’ in each of the three cases as described by the DIPI research teams in their own words. Despite the differences in their approach to understanding and describing the interventions of the specific standard/s in each case, the descriptions give us a flavour of the complexity and multidimensional nature of ‘treatment’ in the case of sustainability standards impact evaluation.

The challenges in understanding the package of interventions associated with standards are, in some contexts, amplified by another factor – multiple certification. Multiple certification refers to the situation in which the entity (such as a farm, a farmers’ group, a plantation or a factory) is certified to meet more than one sustainability standard. Multiple certification is a reality in many agricultural sectors such as tea, coffee, cocoa and bananas, given the existence of multiple standards operating in these sectors. In the DIPI evaluations, two of the three evaluations are being undertaken in such contexts – Kenyan coffee organisations certified to Fairtrade and UTZ standards and Indonesian coffee farmers certified to both the RA/SAN and Global Coffee Platform (previously 4C) Standard. The prevalence of multiple certification adds complexity to evaluations as it adds to the ‘interventions landscape’ that the evaluator is trying to unravel and also makes attributing impact to the individual standards even more difficult.

| Sustainability standards as a ‘package of interventions’ – the case of the three DIPI evaluations | |
|---|---|
| BCI, Cotton, India | Reduction in pesticide use, proper and safe use (registered, labelled, non-use of banned pesticides), inter-crop, border crop (both know-how and do-how), soil test-based nutrient application, timely application, composting, deep ploughing, crop rotation, repeated inter-cultivation, green manure, mulching (sun hemp and diancha), residue management, plant population, gap filling with other legumes, drought management, flood management, water conservation, water-use efficiency measures; monitor land use /conversion as per national convention, flora and fauna, community biodiversity committee (composite of learning groups), clean harvest and storage solutions, hazardous work, alleviating discrimination, lead farmer development, farmer field schools and other extension approaches, collective marketing, financial linkages, certification process |
| Fairtrade & UTZ, Coffee, Kenya | Assisting small scale farmers in Western Kenya in improving yields, quality, and access to markets through training, certification, sustainable farming and better links to market; assisting participating farmers in adopting sustainable coffee production in order to protect the ecosystem and increase market access; assist farmers in attaining UTZ Certified and Fairtrade certification in order to ascertain traceability; improve efficiency and increase market access |
| RA/SAN and 4C, coffee, Indonesia | Training materials are developed by independent standard setting organisations, by partner NGOs, or by private sector companies, to convey to farmers the practices necessary to meet the standards. This usually involves an intensive initial training phase with weekly or monthly training for a few months, and then ongoing supplemental training over a 3 or 4-year cycle.<br>The auditing process. Third-party, independent auditors are hired to travel to the field and verify the extent to which farmers are actually complying with the practices dictated by the standards. This is usually done by randomly selecting a small subset of farmers in a farmer group for detailed auditing on an annual basis.<br>The 'certification' process. Farmer groups that comply with the standards will hold certification (or verification, depending on the program).<br>Marketing processes. These above processes are commonly associated with a new marketing channel and often the establishment of a new local-level buying station, whereby verified or certified farmers obtain certain market privileges. This new marketing channel involves price premiums at the farmer-level for both: i) certified / verified coffee; and ii) for higher quality coffee. |

## 4.2 Who does the ‘treatment’ in the context of sustainability standards?

The second important point to highlight is that not only do standards’ interventions come in packages but that they are implemented by many actors: the standard-setting body, its local staff, assurance providers, contracted implementation partners, the government or its agencies and in some cases by certified entities themselves. Understanding ‘treatment’ then involves understanding what the package is but also how the package is implemented (or not) in the evaluation context. This means that understanding the factual involves understanding what standard systems do, but more importantly, what they don’t do and what their implementation partners and others do as part of the ‘treatment’. Establishing these roles, responsibilities and local dynamics of project intervention is an understated but crucial part of an evaluation’s baseline stage. There are many instances in the DIPI evaluations on the practical difficulties of understanding what should be regarded as treatment in the case of standards and who plays what role in the treatment cycle.

In the case of Fairtrade/UTZ coffee certification in Kenya, a fundamental question the research team was trying to work out was what would count as ‘treatment’ in the context of group certification of farmers’ organisations that was being implemented by a local partner organisation (the local coffee marketing agency). Is the treatment the standard itself and certification process? Is the treatment all that the implementation partner does to obtain and maintain certification? Does it go beyond this to include all that takes place within certified organisations that could be linked in some way to their certified status? Where does treatment start and where does it end? There was an added difficulty in this context - as the farmers’ organisations were not yet certified (the point of the evaluation being to capture change from a pre-certification stage), they had limited interaction with the standard systems and so information had to come from the coffee marketing agency and their field staff.

In the case of the BCI cotton project in India, early conversations with the implementation partner revealed that there was a clear difference between what a typical BCI intervention would be and the specific nature of what was being done, and not done, by the implementation partner in this project site. This raised questions of how the local partner understood treatment and consequently how the study could evaluate activities that are included in a generic BCI intervention but were not being undertaken specifically by the implementation partner in this project and region.

In Indonesia, the intervention is being implemented by the biggest local coffee trading company and so details of the treatment had to be understood through conversations with them rather than the standards themselves. A multiple certification context meant that the research team had to understand the sequence, timeline and trajectory of both certifications, one of which was already adopted across the entire field site.

Some general learnings surfaced in all three evaluations - there are differences of opinion between standards' staff and their implementation partners on what activities are actually undertaken at field level as part of 'standards implementation' or 'certification'. There is often no knowledge within standard systems of what happens in the field in the pre-certification stage (as entities are still to be certified and formally enter their system), even though this is critical to an evaluation trying to map impact from the pre-certification stage. There is a further lack of clarity on the specific details, timing and sequencing of the activities undertaken after certification which, if not understood at baseline stage, stand to affect the evaluation results, especially with a theory-based evaluation approach. Few records are maintained of the specific activities that standards implement with certified entities in a dynamic implementation context with lots of agencies providing myriad forms of support and training. A lack of understanding of the factual makes the choice of an appropriate counterfactual even more difficult and can result in an irrelevant counterfactual being chosen for the study (explained in a later section).

The points made here highlight the fact that both those undergoing evaluations (standards) and those undertaking them (researchers) need to spend time in developing a clear picture of the intervention landscape in any given study context. In the table below we attempt to create one such hypothetical matrix drawing examples from the DIPI case studies. Developing a matrix of the intervention landscape for a standard system in the study site can be a very useful exercise at the baseline stage by helping to clearly define the 'treatment'. It is also key to understanding what is attributable to the standard itself and to what extent. This, when combined with a system's Theory of Change, provides a solid framework for analysis that combines general theory with local implementation reality. In multiple certification cases, more collaboration and knowledge-sharing between systems would be needed to develop a thorough understanding of which system does what in the same context or with the same target group.

This illustration highlights two important points on how we define treatment in the case of sustainability standards. The first is that treatment is not restricted to what is done by the standard-setting body or standards system alone. It often encompasses a much wider set of activities undertaken to achieve or maintain certification or achieve the system's intended change in that context. The second is that treatment doesn't happen in one go and is often iterative, one activity building on the next, thus necessitating a time dimension to understanding the treatment and potentially the impact that results from it. The fundamental lesson here is that standards systems need to pay close attention to defining the treatment in their evaluations and that researchers need to pay close attention to contextually understanding treatment in the early stages of research. On a more pragmatic note, establishing what standards and their partners do as part of implementing the intervention is difficult, takes time and requires varied tools capable of capturing the full range of activities and actors involved in standards' implementation. Researchers and standards would do well to budget more time and resources to understanding the treatment at the baseline stage.

**Illustration: an intervention matrix to understand a potential 'treatment' in the context of sustainability standards**

| Implemented by →<br>Part of 'treatment package' ↓ | Standard-setting body | Local partner (NGO / company / government agency) | Certified entity themselves | Auditing partner or assurance provider |
|---|---|---|---|---|
| Precertification preparation of farmers' group for certification (such as establishing market link) | - | - Prepares the farmers' group for certification | - Undertakes work involved to become certified such as establishing marketing or basic ICT or bank account etc. | - Precertification audit |
| Good agricultural practice training (incl. soil management, fertilizer and pesticide application) | - Write and set the standards<br>- Field staff translate the standard and train the partner | - Trains the lead farmers on the standards' requirements<br>- Conduct field-based learning experiments | - Lead farmers implement and train other farmers<br>- All farmers understand and adopt (or not) the practices | - Audits if the practice in the field is as per the requirements of the standard on the basis of a sample of farmers |
| Formation of learnings groups and community development | - | - Form learning groups of farmers prior to certification | - Join the learning groups on a voluntary basis | - |
| Alleviating gender and child discrimination | - Stated in the standard's Theory of Change but not in the Standard itself | - | - | - |
| Access to market and collective marketing | - Generalised support through entry into system and certification OR<br>- Specialised support such as building a supply chain for the entity from scratch | - Establishing links to government marketing agencies or local exporters or builds a local buying station | - Marketing outreach undertaken the certified entity themselves such as establishing a website, going on tours | - |

## 4.3 Who is the focus of the ‘treatment’? Who is selected for it or self-selects into it?

A final essential part of understanding the treatment is to understand who is selected for the treatment, why and how. In other words, it is crucial to understand the selection dynamics of the intervention itself and account for these in how the evaluation is designed. This is vital as in most development interventions, the selection of participants for a programme or the recipients of an intervention is rarely done randomly and almost always purposive. Selection is either done by the implementing agency in favour of participants who have particular characteristics or are located in particular regions or who meet certain criteria to be a part of the project; or happens through self-selection where participants of a certain profile and type choose to join or be a part of an intervention voluntarily. **The central premise of selection effects of the intervention is that the group receiving the intervention is systematically different from those not receiving it, making it imperative for the evaluation to take this into account.**

Understanding selection, and its effects, is also critical for impact evaluation in the field of sustainability standards. Often, a clear understanding of the ‘factual’, as detailed in the previous section, should bring to light details of who participates in standard systems, who standards themselves target, how participants are selected or select themselves (as often entering a scheme or getting certified is a voluntary act) and how the targeted intervention group differs, if at all, from a generic or random group in the same study context. Selection can take place in different ways and the table below draws from the three DIPI study contexts to describe the nature of participant selection that needed to be accounted for in the evaluation design.

| **Selection in the intervention of sustainability standards: examples from the DIPI cases** | |
|---|---|
| DIPI study of BCI cotton project in India | The focus of the intervention was Adoni Mandal where the implementation partner was tasked with bringing 10,000 cotton farmers under the BCI scheme over a five year period. The mandal has a variety of soil types (black soil, red soil, mixed soil), with black soils known to be the most conducive for cotton production. Given the implementation partner had never worked in Adoni before but had worked with the BCI standard before, a choice was made to focus the efforts of the first three years of the project on villages in Adoni Mandal with predominantly black soil profiles. The thinking was that such a choice would generate ‘a better demonstration effect within the mandal, consequent to which more mixed soil and red soil areas can be added to the project in subsequent years.’ Given the intervention was adopting a ‘saturation approach’ (aiming to cover the entire mandal over a period of time), the focus on regions with one soil type at the start does not mean that other soil types would not be included but just that they would be included at a later stage. This means that the evaluation design needs to account for this pre-selection of the focus of the intervention on black-soil areas. |
| DIPI study of Fairtrade and UTZ coffee certification in Kenya | The focus of the intervention in the short-term is the certification of coffee farmers’ producer organisations to the Fairtrade and UTZ Standards and the associated package of interventions thereafter. However, the researchers noted that selection of organisations to get certified was highly likely at this level as there are incentives for the implementation partner (the coffee marketing agency) to select organisations with higher potential to obtain the certification in comparison with any random group of coffee farmers in the same region. For example, a choice might be made to focus on better organised farmers’ groups, a larger number of farmers, those with more aggregate production, farmers in a better position to meet standards, farmers groups producing coffee of a certain quality and so on. In addition to potential selection by the implementation partner, one also needs to account for self-selection in this case as decisions by farmers’ groups to join certification schemes are mostly voluntary (although local support organisations can strongly influence certification choices) and taken on the basis of a mix of political, market and practical considerations of whether the group will meet the standard or not. It is fair to say then that there are certain types of farmers’ groups that are more likely to get certified or adopt certification. |
| DIPI study of RA/SAN and 4C coffee certification in Indonesia | The focus of the intervention in the short term is the certification of coffee farmers in Semendo region of Indonesia to the base 4C Coffee Standard and then subsequently to the more demanding RA/SAN Standard. The intervention began in 2012 with the implementation partner, a large local coffee processing company, establishing a buying station in the region and initiating farmers into the 4C Code. The implementation partner led the formation of farmer groups, carried out socialization programmes to introduce certification, asked farmers whether they wished to get involved and if so, conducted audits to determine if groups could meet certification requirements and then paid for the cost of obtaining certification. After initiating farmers in the region into the 4C Code, a group of eligible farmers were then ‘upgraded’ to meet the RA/SAN standard, again by the implementing agency. In this case, there was a deliberate selection made by the implementation partner to focus on getting coffee farmers in a particular region that supplied its buying station certified and not just any coffee farmer in the region. Also, we note that some ‘eligibility’ criteria were adopted to target farmers for the base 4C certification and also the advanced RA/SAN certification. |

As the DIPI cases illustrate, selection effects from sustainability standards can be quite strong as most standard systems and their implementation partners focus certification on particular types of entities that do differ qualitatively from the average entity. Even when selection is not deliberately made by the system or partners, self-selection is a reality of how sustainability standards are adopted and can often go unnoticed by the system itself. Understanding the selection dynamics of an intervention strengthens our understanding of the factual and helps establish the 'impact on what' question that evaluations grapple with.

Standards implement a **'package of interventions'** making it critical to understand and define 'treatment' clearly for evaluation purposes. This includes clarity on **who standards target for their interventions, who implements the interventions and how.**

# 5. Designing the counterfactual: concepts and challenges

The core difference between impact evaluation and other forms of evaluation is attribution analysis - our ability to attribute some part of the observable change to the specific policy, programme or intervention in question. Put another way, impact evaluation is centrally about a 'with versus without' analysis: what happened with the programme (a factual record) compared to what would have happened in the absence of the programme (a counterfactual record). Understanding what happened without the programme is difficult in reality and so necessitates the construction of a reality in which the programme did not take place or exist – a counterfactual. The counterfactual is a comparison between what actually happened and what would have happened in the absence of the intervention (White, 2006).

For most impact evaluators, the construction of a relevant and robust counterfactual is the central challenge to address. Questions of what constitutes a suitable and credible counterfactual, and how it is to be found in the complex real world where development interventions transpire, has informed much of the debate around social science and policy evaluation for decades. Some of the main learnings from this debate have informed evaluation work in the field of standards and are useful to recapture here[4].

- The limits of 'before/after' or 'pre-/post-' approaches where evaluation involves taking mean values of target outcome variables for the intervention groups before the intervention took place and then again after it had taken place. Although useful to capture broad patterns of change in the target group, the main limitation of the approach is in its analytical inability to attribute change to the intervention being studied given the range of factors that could have caused the observed changes. It is simplistic in its ability to attribute impact and is not considered a reliable and credible method.
- The difficulty of randomised treatment given the complex implementation dynamics of development initiatives and difficulty of assigning the 'intervention' to target and control groups from the start. In addition, even where possible, randomised control trials often struggle with spill-over effects, differential attrition rates between treatment and control groups and unobservable behaviour effects on the treatment group.
- The rise of quasi-experimentalist approaches as dominant designs for impact evaluation research, given the limitation of randomised control trials.
- The consensus on the validity of 'mixed method' approaches and theory-based evaluation designs that combine a range of qualitative and quantitative techniques to data capture and analysis that bring different strengths to the design in their ability to capture context, intervention complexity, causal inference and maximise learning potential for study participants.
- Finally, the debate on attribution vs contribution - are we trying to attribute an impact to the intervention or evaluate the contribution the intervention is making to the impact? Attribution involves a causal claim about the intervention as the cause of the impact, and measurement of how much of the impact can be linked to the intervention. This contrasts with contribution, which makes a causal claim about whether and how an intervention has contributed to an observed impact.

The questions and debates above are well acknowledged in the field of standards' evaluation and are informing design and method choices for robust impact evaluations. In this section, we unpack some core concepts in counterfactual thinking and design in the context of evaluating the impacts of standards. This paper focuses on a discussion of traditional statistical counterfactual designs as this is the approach that the three DIPI evaluations followed (given their focus on the measurement of attributable impact). **Counterfactual thinking can and is increasingly going beyond these traditional routes into new and innovative approaches. For a good, practitioner-oriented overview of commonly used quasi-experimental and non-experimental impact evaluation designs with associated levels of robustness, see Annex 3**. We would also recommend reading Bamberger et al. (2009) for those interested in exploring alternative counterfactual approaches.

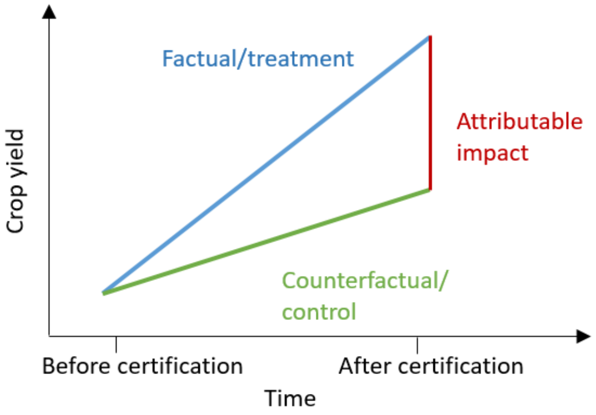
[4] For a fuller consideration of the ongoing debates in impact evaluation that are relevant and useful for work in this field, see Stern *et. al*, (2012). For a specific summary of the challenges of constructing counterfactuals in certification evaluation, see Blackman and Riviera (2010).

## 5.1 The Attribution Problem

As stated earlier, the core difference between impact evaluation and other forms of evaluation is attribution analysis (otherwise called causal or causality inference), which is our ability to attribute some part of the observable change to the specific policy, programme or intervention in question. This is often regarded as the central question of most evaluations. This involves first isolating and then estimating the actual contribution of an intervention to the stated outcome and establishing causality from the intervention to the outcome in question. Establishing attribution is key because in reality change is very rarely caused only by the intervention in question (in our case, sustainability standards). Other interventions and contextual factors interacting with the intervention can strengthen or weaken its effects making simple ex-ante and ex-post comparisons erroneous.

As explained by Hearn and Buffardi for example, "if an evaluation demonstrates that there was a significant increase in average agricultural yields in the intervention village when compared to a village with similar characteristics that did not receive the intervention, the impact attributed to the programme would be the difference between agricultural yields in the intervention and non-intervention sites" (Hearn and Buffardi, 2016: 10). The figure below provides a simple depiction of this concept. Impact evaluations use specific methodologies to allow them to isolate and measure attributable impact.

Source: Adapted from Hearn and Buffardi (2016)



Given the 'package' nature of sustainability standards' interventions and multiple implementation actors, it is not hard to see why attribution analysis is particularly challenging in such evaluations. For example, in the figure above, we would want to know what part of what standards' do is resulting in the increased yield. Is it because of the application of the standard or training on best agricultural practices in addition to the standard, or access to finance for better fertilizers to boost production? Is this increase due to the intervention of standards or those of their implementation partners? As we can see, being very clear about what is meant by 'treatment' from an early stage sets a clear scope for the evaluation and aids attribution analysis, ultimately ensuring more meaningful and valid results.

But attribution analysis is a known challenge in the field of sustainability standards evaluation. Successive impact evaluations of agricultural standards across country and product contexts note the frustration and futility of aiming for full and measurable attribution. In their paper on this topic Ton, Vellema and Ge conclude the following:

> "Most impact studies of market-led development strategies tend to focus on outcomes related to the performance of business practices, such as rural incomes or wellbeing, which are difficult to attribute to the actual processes set in motion by the private-sector support. Similar to the support to farmers in certification, these support interventions involve multiple actors and have many intervening factors that influence their performance. This makes it impossible to attribute changes in outcomes to one specific type of activity (treatment), or, even worse, to one specific supporting agency….Monitoring changes in such ultimate outcomes may be possible, but deriving net effects and claiming attribution of changes in these outcomes to a single part of this complex of factors is not."

(Ton, Vellema and Ge, 2016, p45)

The consensus from researchers on how to overcome this focuses on articulating clearer causal pathways in theories of change, asking different research questions (such as on the 'additionality' of standards rather than absolute impact) and finally on methods that focus not purely on attributing impact but rather verifying the role of an intervention.

## 5.2 Selection bias in impact evaluation

In evaluation terminology, 'selection bias' occurs "when the reasons for which an individual participates in a program are correlated with outcomes. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation" (Gertler et. al, 2016, p59). **Put simply, selection bias results when there is a characteristic difference between the treatment group (that receives the intervention) and control group (that does not receive the intervention) that would bias the results of the evaluation**. This is sometimes referred to as 'sample selection bias' (White and Bamberger, 2008), which basically means the treatment and control have not been correctly matched and that the evaluation is effectively comparing apples to oranges. In other words, the evaluation is flawed and results will be biased because the two groups being compared do not share the same characteristics but differ fundamentally. As White explains:

> "Problems occur if the factors affecting whether a group or individual participate in a programme or not, are correlated with the outcomes of interest, since those participating would do better (or worse) than others regardless of the intervention. Hence if there is such a correlation, then a 'naïve impact estimate', which compares average outcomes for programme beneficiaries with those for a sample of non-beneficiaries (the comparison group), will yield a biased estimate of the impact, called selection bias."

(White, 2013, pp31-32)

We illustrate the issue of selection bias with an example below.

### Illustration: selection bias in impact evaluation

"Consider, for example, a vocational training program for unemployed youth. Assume that two years after the program has been launched, an evaluation attempts to estimate its impact on income by comparing the average incomes of a group of youth who chose to enroll in the program versus a group of youth who, despite being eligible, chose not to enroll. Assume that the results show that youth who chose to enroll in the program make twice as much as those who chose not to enroll. How should these results be interpreted? In this case, the counterfactual is estimated based on the incomes of individuals who decided not to enroll in the program. Yet the two groups are likely to be fundamentally different. Those individuals who chose to participate may be highly motivated to improve their livelihoods and may expect a high return to training. In contrast, those who chose not to enroll may be discouraged youth who do not expect to benefit from this type of program. It is likely that these two types would perform quite differently in the labor market and would have different incomes even without the vocational training program.

The same issue arises when admission to a program is based on unobserved preferences of program administrators. Say, for example, that the program administrators base admission and enrollment on an interview. Those individuals who are admitted to the program might be those who the administrators think have a good chance of benefiting from the program. Those who are not admitted might show less motivation at the interview, have lower qualifications, or just lack good interview skills. Again, it is likely that these two groups of young people would have different incomes in the labor market even in absence of a vocational training program.

Thus the group that did not enroll does not provide a good estimate of the counterfactual. If you observe a difference in incomes between the two groups, you will not be able to determine whether it comes from the training program or from the underlying differences in motivation, skills, and other factors that exist between the two groups. The fact that less motivated or less qualified individuals did not enroll in the training program therefore leads to a bias in the program's impact. This bias is called selection bias."

Source: Quoted from Gertler, et.al, 2016, p59.

## 5.3 Standards and the 'missing counterfactual'

In constructing a counterfactual, the first question evaluators seek an answer to is 'where will the counterfactual come from?' This depends on the way the intervention is carried out. If the population in which the treatment is carried out is sufficiently large, then it is possible to select a sufficiently large sample size for the treatment group and control group that ensures statistical validity. But if the population is too large (the entire region or country) or too small (very few organisations or households), finding a counterfactual for the treatment can be very challenging.  The challenge of the 'missing counterfactual' is now well recognised in the sustainability standards arena. The rapid spread and uptake of standards in key geographies, a multi-actor implementation process, the existence of multiple certification (especially in sectors such as coffee, cocoa and tea) and the overlap of standards' activities with those of partner organisations, all make the identification and construction of counterfactuals a veritable challenge. The three DIPI evaluations also faced challenges in relation to counterfactual identification, although in different ways.

| Challenges faced in locating counterfactuals in the three DIPI study contexts | |
|---|---|
| DIPI study of BCI cotton project in India | The BCI project in Adoni aimed to take a 'saturation' approach i.e. reach all cotton-producing households within the area eventually. This created the challenge of there being no 'pure' counterfactual in the area to compare the 'treatment' or 'target' farmers with. |
| DIPI study of Fairtrade and UTZ coffee certification in Kenya | The nature of the intervention in the Fairtrade/UTZ case, with focus on the producer organisation level as the certified entity, meant the counterfactual or control group had to be identified at the produce organisation level. This meant a search for uncertified coffee producing farmers' organisations in the same agro-ecological zone as the target producer organisations (Mount Elgon, Kenya). Potential control organisations were identified from secondary literature and key informant interviews. Although the team expected to make the final choice after the use of statistical methods to find the closest matches, the limited number of candidates meant this was not needed. Instead, macro-level factors dictated decisions, disqualifying two of the candidate control groups. These factors included focus on coffee, placement in the same agro-ecological zone, and distance from each other to control for a potential spillover effect. In addition to finding counterfactuals, retaining them (and target organisations) has proven to be challenging given shifting field realities and decisions to adopt certification (by the controls) and withdraw from certification (by the targets). |
| DIPI study of RA/SAN and 4C coffee certification in Indonesia | Indonesia has witnessed rapid uptake of coffee certification in recent years. However the research team noted that while there had been rapid development of third-party sustainability standards across southern Sumatra over the last five years, there was a stall over the last 12-18 months prior to commencing the research – mainly owing to inroads made by private companies in coffee sustainability programmes.   This had two important implications for this study design – a) It proved difficult to identify a site with impending expansion of a third-party sustainability program that would enable a pre-intervention baseline and the follow-up research design and b) several existing sustainability programs being implemented as a combination of certification/verification along with additional firm-specific interventions meant it would be difficult to (quantitatively) tease out causation related to sustainability standards. |

To understand how the three DIPI impact evaluations were designed to account for this reality of 'missing counterfactuals' and address the attribution and selection bias challenges noted earlier, we enter the domain of experimental and quasi-experimental evaluation designs.

## 5.4 Experimental counterfactual evaluation designs and randomised controlled trials (RCTs)

In an ideal world, the best way to construct a 'pure' counterfactual or control group is to ensure full randomisation in the application of the programme – or the treatment over a large enough population. This concept – of randomly assigning the treatment – is the basic concept of a Randomised Controlled Trial (RCT) or experimentalist control designs[5]. Full randomisation can take place only if a few conditions are met – the study design should be decided ex-ante (before the intervention begins) so that random allocation of treatment is possible and the implementation partner should be on board with such an approach and ensure that randomisation is maintained throughout the programme cycle. In reality, it is very difficult to create the conditions required to undertake RCTs given the manner in which sustainability standards are implemented where often it is impossible to 'manage treatment'.

[5] For a full discussion of the use of RCTs in evaluating development interventions, read White (2013) An introduction to the use of randomised control trials to evaluate development interventions

There is a vibrant debate on the merits and limits of RCTs, often touted as the 'gold standard' in causal evaluation designs (see Scriven 2008; Deaton and Cartwright, 2016). This paper does not advocate or endorse the use of RCTs as the sole approach to robust impact evaluations; much consideration is needed on when and how RCTs can be used in evaluating the impacts of sustainability standards and how results from RCTs should be interpreted. Despite their acknowledged limitations, RCTs can be useful in estimating causality and measuring attributable impact when the intervention context and dynamics allow for it – and when they are combined with other evaluation design elements that address some serious limitations of the RCT approach. This approach was possible in two of the three DIPI evaluations - the case of the DIPI India cotton study of the BCI project in Adoni that was just beginning and in the case of the DIPI Indonesia coffee study, at the level of upgrading the preselected 4C farmers to the RA/SAN standard. This section discusses how these two evaluations designed the RCT in combination with other evaluation tools, how they conceptualised and achieved randomisation and what results the RCTs will yield.

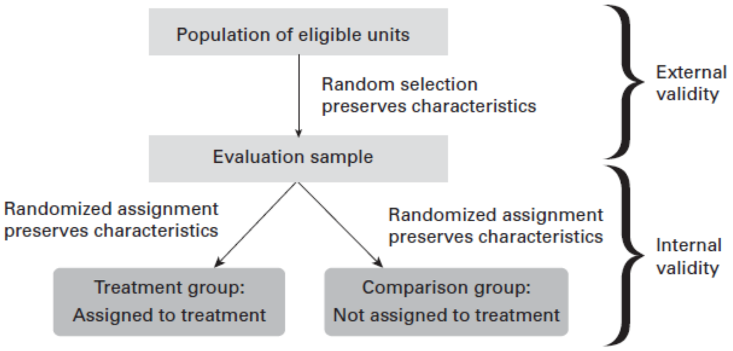
The randomised assignment of treatment is the basic premise underlying a randomised control trial. Under randomised assignment, every eligible unit – an individual, a household, a business, a village, a school or a community – has the same probability of being selected for receiving the treatment in that intervention or programme. **It is important to note that the randomisation needs to occur at the stage of deciding who participates in the intervention and who does not; not a random selection of those already participating in the intervention and those who are not.** However, this involves an important decision to be taken while designing an RCT – what is the unit of randomised assignment? Is it an individual, a household, a community, a section of a community or an entire village? This is linked closely to how the intervention itself is designed – is the focus of the intervention the farm, the household, the farmers' group or the village? In addition to identifying the appropriate unit for randomisation, an RCT design should also ensure internal and external validity. As Gertler et. al explain:

> "An evaluation is internally valid if it provides an accurate estimation of the counterfactual through a valid comparison group. An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units. The results of the evaluation can then be generalised to the population of eligible units. Internal validity is achieved through randomised assignment of treatment and external validity is achieved through randomised sampling."

(Gertler *et. al,* 2016, p73).

The figure below illustrates this difference:

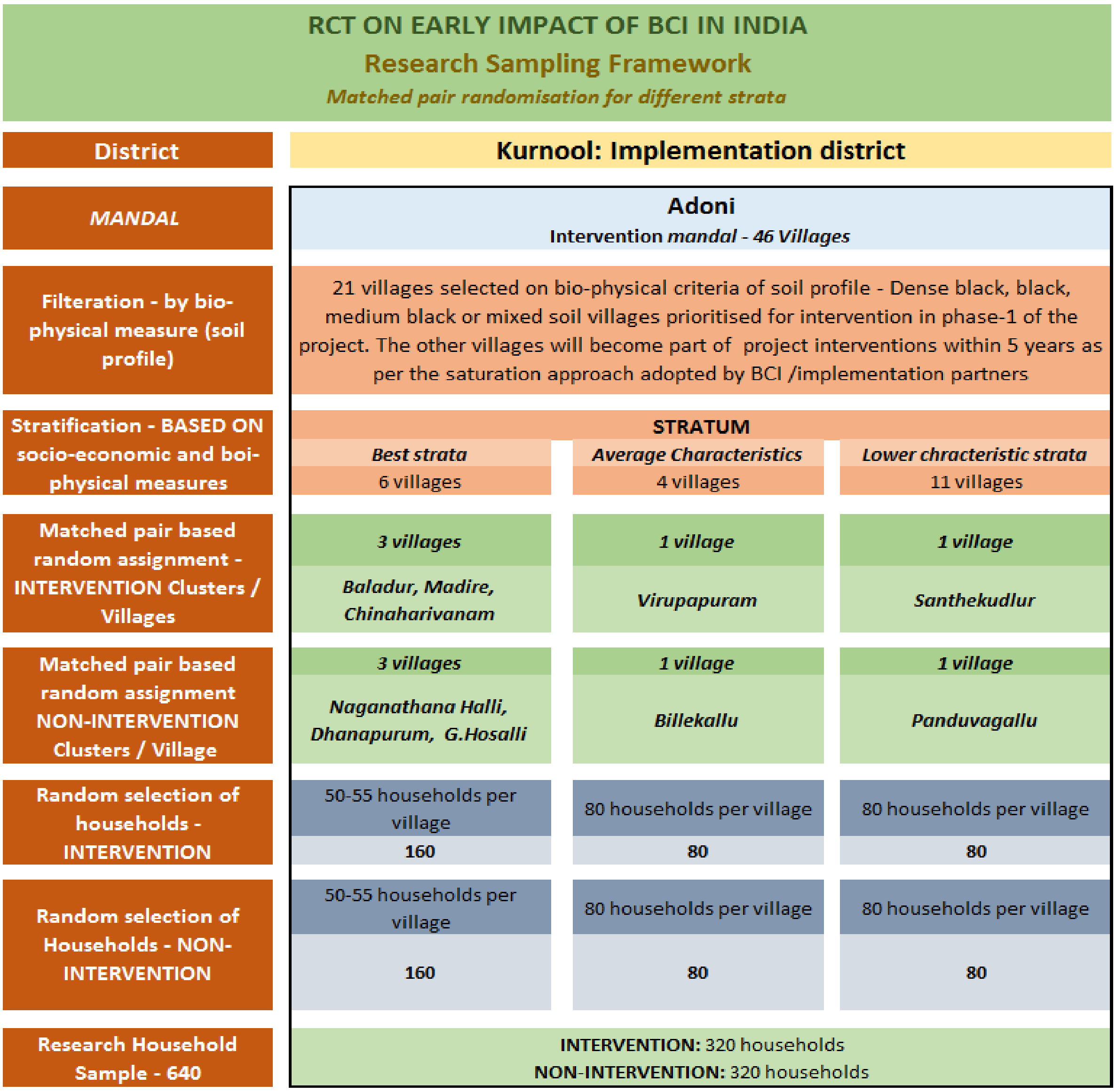
**Figure 4.2 Random Sampling and Randomized Assignment of Treatment**


Population of eligible units
Random selection preserves characteristics
Evaluation sample
Randomized assignment preserves characteristics
Randomized assignment preserves characteristics
Treatment group: Assigned to treatment
Comparison group: Not assigned to treatment
External validity
Internal validity

Source: Gertler et.al, 2016, p73

In the DIPI India study, with Adoni being a new region of implementation for the BCI project and the willingness of the intervention partner to undertake randomised treatment meant that conditions were favourable for designing an RCT type evaluation. Further, a 'saturation approach' to the intervention which meant that the project would cover all farmers in the region over the project period, meant that traditional ethical concerns with RCTs (that the control group stood to gain nothing as they would not be a part of the intervention) would not be a concern here. However, despite the suitability to an RCT design, the heterogeneity of households within the area in bio-physical, economic and social conditions meant that simple randomisation would not be possible. After considering various designs, the evaluation team decided on a cluster RCT approach. This meant that randomisation would be at the level of a cluster, in this case a village, that would then be matched with other villages with similar characteristics for comparisons. This would allow the team to take into account the heterogeneity within the area (by making sure the clusters represented all relevant aspects of the population) and reducing spill-over effects as the intervention would take place at the level of the village.

It is important to mention here that the village was considered to be a cluster and the unit of random assignment, as methodologically it was not possible to randomly assign farmers /households to the intervention within a village, given the saturation approach. The implementation partner wanted to prioritise black or mixed soil areas in the first phase of project implementation (first three years) and hence this bio-physical measure (dense black, medium black, mixed soil) was used as a filter to create the universe for random selection of the clusters /villages. The sampling universe of 21 clusters (so obtained after applying the filter) was divided into 10 best matched pairs (using existing bio-physical and socio-economic parameters) and then from each pair, a treatment and control cluster/ village was randomly assigned. Within each treatment and control village, randomised sampling was used to identify the households that would receive the intervention and those that wouldn't. The figure below explains the key steps in designing the RCT (for a fuller description of the stratification design and sampling framework, read the DIPI India research design document).



**RCT ON EARLY IMPACT OF BCI IN INDIA**

**Research Sampling Framework**

***Matched pair randomisation for different strata***

| District | Kurnool: Implementation district | | |
|---|---|---|---|
| *MANDAL* | Adoni<br>Intervention *mandal - 46 Villages* | | |
| Filteration - by bio-physical measure (soil profile) | 21 villages selected on bio-physical criteria of soil profile - Dense black, black, medium black or mixed soil villages prioritised for intervention in phase-1 of the project. The other villages will become part of project interventions within 5 years as per the saturation approach adopted by BCI /implementation partners | | |
| Stratification - BASED ON socio-economic and boi-physical measures | STRATUM<br>*Best strata*<br>6 villages | *Average Characteristics*<br>4 villages | *Lower chracteristic strata*<br>11 villages |
| Matched pair based random assignment - INTERVENTION Clusters / Villages | *3 villages*<br>*Baladur, Madire, Chinaharivanam* | *1 village*<br>*Virupapuram* | *1 village*<br>*Santhekudlur* |
| Matched pair based random assignment NON-INTERVENTION Clusters / Village | *3 villages*<br>*Naganathana Halli, Dhanapurum, G.Hosalli* | *1 village*<br>*Billekallu* | *1 village*<br>*Panduvagallu* |
| Random selection of households - INTERVENTION | 50-55 households per village<br>160 | 80 households per village<br>80 | 80 households per village<br>80 |
| Random selection of Households - NON-INTERVENTION | 50-55 households per village<br>160 | 80 households per village<br>80 | 80 households per village<br>80 |
| Research Household Sample - 640 | INTERVENTION: 320 households<br>NON-INTERVENTION: 320 households | | |

Source: R. Kumar et. al, 2016, p29

In the Indonesia case, given that the evaluation entered a study context where the 4C intervention had already begun, an RCT approach would not have been possible at the level of trying to assess the impact of the 4C intervention. However, given that the 'upgrading' of the 4C certified farmers to RA/SAN certification was still underway, a randomised approach to that intervention was possible and so an RCT approach was adopted to construct a counterfactual to study the impact of the RA/SAN standard, also following a pipeline approach. The approach followed is detailed below.

As of 2014, the implementing partners were working with 2437 4C verified farmers across 107 farmer groups in 25 villages in three regions and had carried out trainings to help about half of them (50 farmer groups) move to RA/SAN certification. The first steps for the research team was to propose that the implementation partner stagger the remaining roll-out of RA/SAN certification amongst the remaining 57 farmer groups (roughly 1336 uncertified farmers across 20 villages that were already 4C certified) according to random selection generated by the research team. The team randomised the remaining 57 farmer groups (FGs) into 2 groups:

- 29 FGs: treatment (become RA/SAN certified immediately, in the 2015-2016 cycle)
- 28 FGs: control (do not become RA/SAN certified until the 2018-2019 cycle, at soonest (where the farmers are eventually RA/SAN certified is fully at the discretion of the implementing partners)

The research team then stratified the treatment group on two dimensions:

- At the village level, to maximise power (to ensure that treatment occurs in the maximum number of sites)
- According to a measure of livelihood zone. Since randomisation was conducted after the baseline survey has occurred, the researchers were able to use a livelihood measure from the detailed baseline survey as one of the stratification criteria.

The RCT designed for the Indonesia study not only allows a comparison between treatment and control farmers receiving RA/SAN certification but importantly, also enables us to understand the effects of certification in a multiple certification context, i.e. the differential impact of RA/SAN certification with farmers who are already 4C certified. This design therefore allow us to clearly evaluate the impacts of coffee farmers moving from 4C to RA/SAN certification.

## 5.5 Quasi-experimental counterfactual evaluation designs

In cases where random assignment of the treatment is not possible, not permitting an RCT type evaluation, other techniques need to be adopted to generate counterfactuals capable of measuring attributable impact. These are generally referred to as quasi-experimental methods effectively meaning that as the true experiment (randomisation) is not possible, an attempt is made to generate a suitable comparison group with similar characteristics to the treatment that can analytically do the job of a counterfactual. The practical difficulties, costs and ethical concerns with randomisation approaches have also led to the popularity of quasi-experimental approaches in development impact evaluation.

The DIPI evaluations used two such methods that we discuss here. The importance of these methods is that unlike in the BCI cotton study where the treatment was being randomly assigned (through the RCT), in the Kenyan and Indonesian coffee studies, treatment was not or could not be randomly assigned given the nature of the entire intervention. The advantages of the quasi-experimental methods used in these evaluations was that they would allow the research team to measure attributable impact in situations where the criteria used to assign the treatment are not clear or visible (Kenya) or where a control group could be constructed to compare with the treatment group (Indonesia).

The first approach is propensity score matching (PSM) that was adopted in the DIPI Indonesia study. PSM is an approach in which the comparison group is matched to the treatment group on the basis of a set of observed characteristics or by using the "propensity score" (predicted probability of participation given observed characteristics); the closer the propensity score, the better the match[6]. As noted previously, the group of farmers that originally entered the 4C programme in 2012 were not randomly selected and participation was decided by a range of factors and mostly determined by the implementation partner. To detect the effects of the 4C 'treatment' and given randomisation is not possible, propensity score matching was adopted. This involved the identification of a suitable sample of up to 400 farmers in the Semendo region, searching for farmer groups that would have been eligible for the 4C intervention, but were not selected. Using administrative data from the implementation partner in Semendo on the control group for the RCT study, and other available secondary data from the local government and key informants, the team constructed a control group with broadly similar characteristics to the original sample of farmers who received 4C verification.

The DIPI Kenya coffee study posed distinct challenges for the evaluation team on several counts. The intervention (Fairtrade and UTZ certification) takes place at the level of the group, not individual farmers, meaning that the counterfactual also

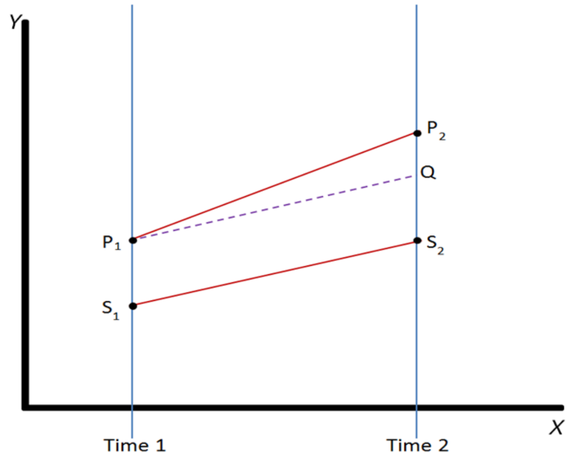
[6] "Propensity score matching (PSM) identifies a group of individuals, households or firms with the same observable characteristics as those participating in the project. It does this by estimating a statistical model of the probability of participating (propensity to participate) using a regression model with participation as the zero-one dependent variable, and a set of observable characteristics, which must be unaffected by the intervention, as the explanatory variables. The coefficients are used to calculate a propensity score, and participants are matched with non-participants based on having similar propensity scores." (White, 2006, p15).

needed to be constructed at that level. This would make randomisation of the treatment (that would have meant random certification of farmers organisations in the region) practically impossible. The target groups for this evaluation were producer organisations that received certification and the control groups were those that did not. The self-selection of farmers' groups into certification and potential selection bias from the implementation partner could have been addressed by having a large sample size (large sample of farmers' groups) but that was also not possible in this case as there were too few groups in the intervention area meeting the necessary criteria to be targets, limiting the statistical power for attribution with quantitative methods alone. Consequently, the control groups (counterfactuals) identified in the Kenya study were two sets of producer organisations (four in total) who did not receive certification – one set that was uncertified and worked with the same implementing partner and the other who were uncertified had no interaction with the same implementing partner. However, the relatively large sample size of farmers selected within each group is large enough to study them as a separate case study.

In addition, all three studies use the difference-in-difference (or double difference) method to aid the analysis of treatment versus control. As noted earlier, the control groups identified in the Kenya study were two sets of producer organisations who did not receive certification. However, such a selection still does not help overcome the possible bias in the findings that would result by comparing the end-line results between treatment and control in cases where the starting point/baseline was not the same. In the Kenya case, this would mean finding a way to account for the differences between certified and non-certified producer organisations at the baseline stage and then look at differences at the end-line stage. A difference-in-difference approach helps overcome this problem[7] as is explained in the note and diagram below, drawn from the DIPI Kenya baseline report.

In a training program, if the goal is to estimate the impact of training on yields, we can use the graphic representation below, where the vertical axis (Y) represents the level of average yields, and the horizontal axis (X) represents time. The yield evolution of the target group is represented by line P, while for the control group it is represented by line S. The yield level is measured for both target and control group at Time 1 (baseline) before either group has received the training, represented by the points P1 and S1. The target group then receives the training and the yield level is measured again for both groups after the training at Time 2 (end line), represented by the points P2 and S2. Not all of the difference between the target and control groups at Time 2 (that is, P2 minus S2) can be explained as being the effect of training on yield, given that a difference already existed between the target and control group at Time 1. If the target group did not receive training, the yield growth of the target group would follow the dotted line Q, which is parallel to the line S. DID can overcome selection bias and generate an unbiased estimator of the impact of training, which is equal to:

$$DID = (P2 - S2) - (P1 - S1) = (P2 - P1) - (S2 - S1)$$



Source: Bennett et. al, 2016

[7] "The difference between the outcome in the treatment group (project area) and comparison group is a single difference estimate. The validity of this estimate as an estimate of project impact requires that the treatment and control groups had the same values of the outcome prior to the intervention. If this is not so then the single difference estimate will be biased. If the treatment group already had superior outcomes prior to the intervention then their better performance post intervention cannot all be attributed to that intervention. The double difference – which is the difference in the change, or, equivalently, the change in the difference – allows for this possibility. Double differencing removes time invariant differences in factors influencing the outcome between the project and comparison groups. However, the validity of the double difference estimate still relies upon the assumption that external determinants of the outcome were the same for treatment and comparison groups during the course of the intervention." (White, 2006, p10).

We can now appreciate the full nature of the counterfactual evaluation design developed in each of the three DIPI evaluations in light of this discussion.

**Counterfactual evaluation design in the three DIPI evaluations**

| | Treatment Group | Control Group | Counterfactual design elements |
|---|---|---|---|
| **DIPI study of BCI cotton project in India** | Households randomly selected within identified treatment villages (across different strata) to receive the 'intervention' - being brought under the BCI Standard over the next 2 years | Households from matched villages and strata as the treatment households, not receiving the intervention in the next 2 years | Randomised control trial with a clustered RCT and a matched pair randomisation approach. Difference-in-difference analysis. |
| **DIPI study of Fairtrade and UTZ coffee certification in Kenya** | Farmers that are members of either one of two groups that receive Fairtrade / UTZ certification | Farmers belonging to one of two sets of control groups. A – producer organisations with similar characteristics and working with the same marketing agent but not receiving certification. B- producer organisations in the same region not working with the same marketing agent and not receiving certification. | Two sets of control groups with difference-in-difference approach to analysis. |
| **DIPI study of RA/SAN and 4C coffee certification in Indonesia** | Intervention 1: Farmers that receive the 4C standard intervention | 1. A group of farmers in the same region that have not received 4C intervention and that do not work with the implementation partner | 1. Propensity Score Matching with difference-in-difference analysis |
| | Intervention 2: Farmers that are 4C verified and then receive additional RA/SAN certification | 2. A group of farmers that receive 4C verification but that do not receive the RA/SAN certification | 2. A randomised control trial with a subset of 4C verified households. |

## 5.6 Other methodological challenges

### Small N[8] problem

Most impact evaluations adopting statistical counterfactual approaches are conducted in contexts where the intervention is being carried out over a large enough population (N) i.e. with a sufficiently high number of units to allow the evaluation to draw statistical relevant sample size (n) groups for treatment and control. These are called 'large n' impact evaluations that involve statistical tests between the treatment and control groups. However, in reality, many interventions many not allow for 'large n' approaches as the units assignment of the intervention are very small in number (such as a programme targeted at a small number of schools or villages) i.e. a small 'N' (population) also results in a small 'n' (sample size). In such cases, studies need to innovate and adopt 'small n' approaches to impact evaluations. The fundamental difference between large and small

[8] Note that in statistics, a capital 'N' refers to the population size whereas a small 'n' refers to the sample size from any given population

‘n’ approaches is that in a ‘large n’ evaluation, causation is established through statistical means but in a ‘small n’ evaluation, given statistical comparison is not possible, causation is established ‘bottom up’ – based on the strength of available evidence, strength of argument and absence or ruling out of alternate explanations for causation (White and Phillips, 2012). There are many approaches to addressing the small n problem[9] but here we detail how the DIPI evaluation in Kenya that faced this challenge addressed it.

The DIPI case study in Kenya focusing on Fairtrade and UTZ certification of farmers organisations is a classic ‘small N’ problem that we encounter with sustainability standards. This is because the unit of assignment (of certification) is the producer organisation and often there are very few such organisations in the same region getting certified to form the sample size for the treatment and very few or no similar organisations in the region to treat as a counterfactual. This makes statistical tests of significance between treatment and control groups at the level of the producer organisation impossible in such cases. The solution involved a combination of methods to ensure statistical validity of the findings and added tools to strengthen causal inference of impact. At the level of the intervention (unit of assignment), the study adopted a purposive sample of 6 producer organisations (2 treatment and 4 controls) meaning n=6 in this case. However, within each selected treatment and control producer organisation, the sample size of farms is large enough to detect effects at farm level within the producer organisation, account for the heterogeneity of farm types and allow for cross-analysis across the 6 producer organisations. The study randomly selected 166 farms within each sample producer organisation to ensure that the sample size was large enough to detect change at the household level within each producer organisation. In addition, a large sample size of farms within each producer organisation also allows the study to compare results between treatment and control producer organisations for particular sub-groups of farms (disaggregated analysis based on certain characteristics such as farm size, farmer age and education etc.).

It is important to note that this still means that the study cannot statistically establish the difference in impact between treatment and control at the level of the producer organisation. Given that the unit of assignment of the treatment is not the farm but the producer organisation, additional tools were needed to understand the difference between certified and non-certified producer organisations and attribute change at the farm level to the intervention. This comes through the contribution analysis framework that the research team developed along with another research tool (the producer organisation survey) which will be used to understand differences between the treatment and control producer organisations.

## Spillover effects

A ‘spillover’ happens when the intervention affects a non-participant thereby adulterating the difference between ‘treatment’ and ‘control’ groups that lies at the heart of impact evaluation. Spillover effects are very common in development interventions as they do not take place in controlled, clinical settings, but the real world with constant interaction between people, villages, and communities. Spillover effects are also a common challenge in the standards’ world as parts of the treatment, such as training farmers on better agricultural practices or building better links to market, may not strictly remain within the group of certified producer organisations and farmers but naturally spill over into surrounding regions.

All the three DIPI case study interventions faced the challenge of spillover effects that had to be accounted for in the evaluation design. In the case of the DIPI study in Kenya, spillovers were at the producer organisation level) and the household level. The study used qualitative methods to detect spillover effect. For producer organisation-level spillover effect, they interviewed nearby non-certified producer organisations to contrast with the treatment producer organisations. For household-level spillover effect, the study interviewed non-certified households who live the in the same community of certified households. In the Indonesia study with RA/SAN and 4C certification, the study checks for potential spillover effects by querying control group farmers (in the RCT) in terms of how well informed they are about the main points of the RA/SAN training. The team also hypothesized that spillover effects would be minimal as geographic and social spillovers are likely to be strongest within farmer groups, and the information contained in the training is relatively complex and hard to convey comprehensively. In a similar vein, the DIPI case study of BCI cotton India faced a high probability of spillover given the intervention was implemented in the same region (mandal). However, given the RCT approach and the fact that intervention would be rolled out in treatment villages (and not the control villages), the team felt this would minimise spillover given the distance between villages and reduce potential contamination.

[9] For a full exploration of small n approaches in development impact evaluations, see 3IE’s working paper by White and Philip (2012).

The key concept in impact evaluation is **attribution analysis** that ideally requires a **comparison of the factual with a counterfactual**. Standards systems must design and integrate **counterfactual thinking** into their evaluations.

# 6. Mixed method approaches for evaluating standards: what are they and how are they used?

Mixed methods are gaining currency as a preferred approach to impact evaluation in many sectors and are becoming the norm in sustainability standards research. “Mixed methods approaches integrate social science disciplines that have a predominantly quantitative and predominantly qualitative approach to theory, data collection, data analysis and interpretation” with the aim of strengthening the reliability of data, validity of findings and understanding of the context and effects of interventions (Bamberger, 2012, p1). As the term suggests, mixed methods approaches involve combining various evaluation methods and tools of data collection and analysis to generate a full picture of the intervention and analyse cause and effect relationships. They also allow evaluators to test the causal links and assumptions embedded within Theories of Change more thoroughly. But how are methods mixed in reality? What mix of methods is effective and are there specific mixes that are more effective in sustainability standards research?

Although many evaluators now routinely use a variety of methods, “what distinguishes mixed-method evaluation is the intentional or planned use of diverse methods for particular mixed-method purposes using particular mixed-method designs” (Greene, 2008). Most commonly, methods of data collection are combined to make an evaluation mixed method, but it is also possible to combine conceptual frameworks, hypothesis development, data analysis, or frameworks for the interpretation of the evaluation findings (Bamberger, 2016). Mixed method is more than combining methods in a study – it is a purposive and selective mix of specific methods and tools – for theory building, data collection, analysis and causal interpretation – that is undertaken for specific evaluation objectives. Mixed method approaches originated as a solution to the age-old ‘qualitative-quantitative’ debate in the social sciences and although obvious, it’s worth restating the main benefits of such an approach from a methodological standpoint.

According to Greene (2008, pp.255-56), the main benefits of adopting such approaches are to strengthen triangulation of findings and ensure comprehensiveness of evaluation findings, support research design and development and a diversity of data capture techniques that complement each other and can adapt to different study context and finally, that mixed methods allows for the integration of a diversity of evaluation values – such as ensuring robustness, maximising learning, empowering programme participants and so on. In addition to these general benefits, Bamberger (2012) also provides a useful summary of the practical and operational benefits of using mixed methods[10].

All three DIPI impact evaluations use a mixed methods approach to data collection and analysis. The table below details the various methods and tools ‘mixed’ in each case as described by the researchers and the reasons for choice of particular tools.

[10] For a fuller description of the full range of quantitative and qualitative data collection tools commonly used and mixed in mixed method approaches, see Michael Bamberger’s excellent guidance note on the use of Mixed Methods in Impact Evaluation (Bamberger, 2012). This includes a useful section for research managers on how to resource and manage mixed method evaluations. See especially box 2 titled ‘Operational benefits from the use of mixed methods’ on page 5.

| Details of 'mixed method' evaluation approach in all three DIPI cases | | |
|---|---|---|
| | The 'mix' of methods and tools (qual = qualitative; quant = quantitative) | Practical benefits of the mixed methods approach for each study context |
| DIPI study of BCI cotton project in India | Method: Combining a theory-based evaluation approach with a randomised control trial.<br>Tools:<br>• Structured farmer household survey as the basis for data needed for the RCT (quant)<br>• Farmer focus group discussions (qual)<br>• Purposively selected blind household panel interviews and case studies (qual)<br>• Key informant interviews (qual)<br>• Validation workshop[11] (qual) | Use of theory of change improve causal inference which is a known weakness in an RCT type evaluation. The RCT allows for a strong estimate of attributable impact based on differences between control and treatment.<br>Data collection and perspective-gathering with other supply chain actors (ginners etc) given survey focusses at household level<br>In-depth insight into the change at household level over the project implementation to add to survey data<br>Understanding of the local market dynamics important for particular pathways in the theory of change |
| DIPI study of Fairtrade and UTZ coffee certification in Kenya | Method: Quantitative analysis using a double difference method combined with a strong contribution framework for causal inference.<br>Tools:<br>• Key informant interviews (qual)<br>• Participatory rural appraisals (qual)<br>• Semi-structured interviews (qual)<br>• Farm-household survey (quant)<br>• Producer organisation survey (quant)<br>• Second round of interviews (qual)<br>• Focus group discussions with farmers and producer organisation management (qual) | The study was designed such that the quantitative tools and approach were aimed at measuring the change between baseline and endline while the qualitative approach was aimed at establishing causation to the intervention.<br>The study used a qualitative approach at the scoping phase, a quantitative approach at the data collection phase followed by a qualitative approach at the analysis and interpretation phase. |
| DIPI study of RA/SAN and 4C coffee certification in Indonesia | Method: Quantitative analysis through propensity score matching (first treatment) and randomised control trial (second treatment) combined with livelihood mapping and village case studies.<br>Tools:<br>• Household survey (quant)<br>• Livelihood mapping exercise at the scoping phase to identify nature of heterogeneity in treatment sample and allow for sub-group analysis of results based on different socio-economic strata (qual)<br>• Village-level case studies using ethnographic methods (qual)<br>• Farmer perceptions survey and survey of sustainability field agents (quant)<br>• Key stakeholder interviews (qual) | The village case studies were done in four villages where a number of farmers are already participating in sustainability programs. They helped identify the relative position of participating households within society, assess the reach of schemes into the broader community, and identify how the poorest households interact with sustainability programs.<br>The surveys will be undertaken in 2016, building on the baseline findings of 2015 as a means of 'checking the pulse' of how the sustainability programmes are being implemented. |

In addition to understanding the benefits of mixed method approaches in theory, it is also useful to reflect on their implementation in practice. The table below shares insights on the strengths and challenges of implementing mixed methods approaches from the DIPI baseline experience.

**Mixed method approach in standards' evaluation: learnings from the DIPI baselines**

| Strengths | Challenges |
|---|---|
| Allows the evaluation to meet needs of statistically valid and measurable impact data along with strong causal inference – can help understand what changed, how much and why. | Is more resource-intensive than a design that is only qualitative or only quantitative. Usually more expensive, time-consuming for research team (to design and implement), research participants (farmers, local stakeholders) and research managers. |
| Allows for a variety of data collection and analytical tools to research the complexity of sustainability standards and the complex contexts in which they operate. | Integration of results to draw firm conclusions can be a challenge if the different methods yield different results on the same indicator. Triangulation improves validity of findings but methods can also throw up contradictory findings. Also, mixed method impact evaluation reports often report 'quantitative results' and 'qualitative results' rather than integrated results – research managers should look out for this. |
| Allows to incorporate a range of values in the study – from robustness and statistical validity to learning and participant empowerment. | Bringing together researchers from different disciplinary backgrounds and evaluation philosophies (with a predisposed preference for quantitative or qualitative approaches) can subject the evaluation to the debates of which approach is more valid. |
| Useful in triangulation and explanation – but also points to where their might be a flaw in intervention implementation (or the quality or fidelity of implementation) and therefore useful for improving the project | A more complex research design usually makes findings complex and interpretation of those findings difficult within commissioning organisations. More effort is needed from monitoring and evaluation teams to translate findings into valid impact claims and help interpret findings for internal learning. |
| Useful way to make use of existing data for evaluation without undertaking new primary data collection | Usually involves more intense research management as the evaluation design is more complex and usually takes more time and money than designs that are only quantitative or only qualitative. |

[11] For more insights on the utility of validation workshops, read this ISEAL blog: https://www.isealalliance.org/online-community/blogs/on-the-virtues-of-a-validation-workshop

Mixed method designs are gaining currency in sustainability standards research. It involves combining **different evaluation methods**, data collection tools and analyses to **generate a full picture** of the intervention and study cause and effect relationships.

# 7. Conclusion: practical tips for standards systems and researchers

We undertake impact evaluations for a range of reasons, not least because ISEAL's Impacts Code requires ISEAL member systems to commission or undergo regular evaluations. Although challenges abound, both in conducting evaluations and drawing conclusions from them, there continue to be strong arguments for standards systems to commit to good quality evaluation. As one source puts it, "impact evaluations are complex but worthwhile exercises" (Gertler et.al, 2016, p319).

The three impact evaluations commissioned by ISEAL as part of the DIPI project also took on the challenge of designing and undertaking complex evaluations that went a step further than what an individual standards system might do as they looked at multiple certification scenarios. Through this paper, we hope to have shed light not only on the nature of challenges faced, but also on how identifying them early on can help design evaluations that can control and overcome them.

We conclude by sharing a few practical recommendations emerging from ISEAL's experience with the baseline stage of the DIPI evaluations. As noted in the introduction, although comprehensive methodological learning is possible only after the end line stage in 2019, baseline stages of evaluation often yield rich insights, some of which are forgotten with the focus on results at the end line stage. Many of the recommendations are not easy to link to individual evaluation issues discussed in this paper and are therefore best read as a package – much like our interventions!

## 7.1 On theory-based evaluation:

**Embed evaluation in the theory of change:** One of the main reasons it is recommended for systems to develop theories of change is precisely so the theory can be put to test in evaluations. There are strong advantages to adopting theory-based approaches in evaluating standards systems and so much more effort is needed in embedding evaluations within these frameworks than is currently the case.

**Understand the theory of change:** In cases where systems are commissioning research to external research teams, effort is needed to ensure the research team fully understands the standards system's generic theory of change thoroughly and correctly interprets it. It is likely that researchers will have comments on the theory of change through the course of the evaluation (and this is to be welcomed) but teams must have a strong fundamental understanding of how the system in question articulates its causal pathways.

**Pick parts of the theory of change to focus on:** Theories of change are usually highly complex and elaborate frameworks as they reflect the entirety of what a standard system does and seeks to change in the short and long-term. No one evaluation can do justice to a system's full theory of change, making choices of which pathways to focus on essential. We strongly recommend that systems and researchers select the specific pathways or causal chains that will be the focus of a specific impact evaluation. Ideally this should be done at the stage of developing the terms of reference for the study or in very early stages of the evaluation.

**Generic theory of change or local theory of change:** In the case that the impact evaluation is focusing on a unique project or context in which the standards system is operating, consider developing a customised causal pathway or theory of change for the project that becomes the framework for the study (or see if the researchers can do this as part of developing the study design). This will aid causal inference and ensure results are locally valid and relevant, although it will make generalisations difficult.

**Consider 'time' in your theory of change:** Change to deliver impact (long-term change) happens slowly and through an iterative process. Most theories of change are not time-bound but impact evaluations happen at a specific point in time in the intervention and change cycle. This makes accounting for time an important point to consider for standards and their researchers. What is the right time to do an impact evaluation? What level of change does the system expect to have made through the course of an impact evaluation? How much time difference is needed between baseline and endline stages of an impact evaluation?

**Flesh out all parts of your system's work:** Sustainability standards are unique market-based tools but often these pathways are under-developed in existing theories of change. More attention can be paid to detailing the market interaction and connections that lead to sustainability outcomes and impacts. This will ensure that the evaluation also captures change at that level and not just at the certified entity/production level.

## 7.2 On understanding the intervention and defining 'treatment':

**Understand the intervention package and choose which part to evaluate:** Given sustainability standards are usually a 'package of interventions', choices of which specific parts of the package will be the focus of evaluation are useful - ideally at the stage of developing the study terms of reference.

**Understand selection dynamics of the intervention:** Specific attention should be paid at the research design stage to understand any selection criteria and effects in the implementation of the intervention. This requires effort and time from the system in question – their monitoring and evaluation, local programme staff and specifically, inputs from the local implementation partners.

**Speak to implementation partners:** to be clear about what happens and who does what locally. Detailed consultation is needed with field partners to fully grasp the specific nature of the intervention and understand 'who is doing what' as part of implementation. Development of an intervention matrix or clear layout of the intervention landscape is highly recommended as part of the research design. Additional time should be built into the scoping phase of evaluations for this.

**Define 'treatment' clearly:** The evaluation research design should clearly define what is being considered the 'treatment' for that evaluation. Ideally, this should be based on discussion between the research team and system or standard in question.

## 7.3 On counterfactual thinking and designs:

**Think ahead for impact evaluations:** If the standard is being implemented in a new country or context and it is highly likely that you will want to conduct an impact evaluation of its work there in the future, think ahead to design an appropriate evaluation and consider where an RCT approach could be adopted or how you could implement the intervention in a way that would aid impact evaluation at a later stage.

**Account for the missing counterfactual:** Much more thought is required from commissioning organisations when conceptualising an impact evaluation on what the possible counterfactuals could be in the given implementation context. This helps define clear (and relevant) research questions (for example a shift from questions such as 'how much change' to 'additional change') and expectations of measurable impact data from the evaluation.

**Minimise selection bias in the evaluation design:** Strengthen the rigour in selecting the treatment and groups from the 'intervention population' to minimise selection bias in the evaluation and ensure sampling frameworks deliver internal and external validity for the results.

**Get innovative in constructing counterfactuals:** With the reality of missing counterfactuals in the standards' world, more creativity and innovation is needed in constructing valid counterfactuals. The field of development impact evaluation is rapidly advancing and many more ideas and methods are now in use than earlier to aid robust impact evaluations that could include quantitative and qualitative elements.

## 7.4 On adopting mixed method approaches:

**Experience of mixing methods:** As choosing 'mixed method' designs can be difficult, research managers could, as part of tendering processes in commissioned evaluations, ask or require that teams that propose mixed method designs if they have experience of implementing similar mixed method approaches in that study context i.e. do research teams know that the mix works? While this is not always possible (given the encouragement to innovate), more dialogue is needed while selecting particular designs on how methods and tools will be mixed and why.

**A research team committed to the mixing of methods:** If the evaluation seeks to adopt a truly mixed method approach in design, data collection and analysis, ensure that the full research team is committed to the 'mixed' nature of the evaluation and that the research manager and research lead are committed to the approach in full from the start.

**Mixed method reporting for mixed method designs:** Research managers should encourage the team to broaden the range of presentation and dissemination methods used to ensure that the full richness of mixed method data is captured, as often data visualisation and presentation is biased towards quantitative types of data leaving behind qualitative data for narrative representation that is often not read. This often requires dialogue and integration in the way results are reported to avoid scenarios that report 'quantitative results' and 'qualitative results' separately. A move and commitment towards 'open data' by both standards systems and research teams will enable this and be a huge positive step in this direction.

## 7.5 On methodological reporting

Irrespective of what design, approach and tools impact evaluations use, rigour in methodological design must be accompanied with rigour in methodological reporting. This requires research teams providing all details on the design, tools and data analysis approach that a particular evaluation is based on. With attempts to make evaluation reports more reader-friendly and accessible, methodological reporting is often cut short with many a report just referring to the bare bones of the evaluation design but missing essential information such as on research site selection criteria and logic, sample size, variables of analysis, basic tests and diagnostic statistics, details of RCT sampling methodology or PSM analyses and so on. Such detail is essential to help readers assess the rigour levels of a study and also determine if individual studies can be included in systematic reviews and other meta reviews that often use methodological criteria as a filter to include or exclude studies.

There are strong arguments for **standards systems to commit to good quality impact evaluation**. We hope this paper helps identify and address key conceptual issues that will improve the rigour of evaluation efforts in this field.

# References:

Alkin, Marvin C., and Christina A. Christie. "An evaluation theory tree." *Evaluation roots: Tracing theorists' views and influences* (2004): 12-65.

Bamberger, Michael, Fred Carden and Jim Rugh, *Summary of Session 713 Think Tank: Alternatives to the Conventional Counterfactual*, American Evaluation Association: Orlando (2009) Retrieved from: http://www.alnap.org/resource/8222.aspx

Bamberger, Michael, "Introduction to mixed methods in impact evaluation." *Impact Evaluation Notes 3* (2012): 1-38.

Bamberger, Michael and White, Howard. "Using Strong Evaluation Designs in Developing Countries: Experience and Challenges." Journal of Multidisciplinary Evaluation, 4; 8; October 2007.

Befani, Barbara, and John Mayne. "Process Tracing and Contribution Analysis: A combined approach to generative causal inference for impact evaluation." *IDS bulletin* 45.6 (2014): 17-36.

Bennett, Mica, Carlos de los Rios, Matthew Himmel, Lydia Wairegi, *Impacts of Certification on Organized Small Coffee Farmers in Kenya*, Committee on Sustainability Assessment (2016). Retrieved from: http://www.isealalliance.org/online-community/resources/iseal-dipi-project-three-commissioned-impact-evaluations-baseline-full-reports-and-

Blackman, Allen, and Jorge E. Rivera, "The evidence base for environmental and socioeconomic impacts of 'sustainable' certification." *Resources for the Future Discussion Paper* 10-17 (2010). Retrieved from: http://www.rff.org/research/publications/evidence-base-environmental-and-socioeconomic-impacts-sustainable-0

Crosse, William, and Deanna Newsom, and Elizabeth Kennedy "Recommendations for Improving Research on Certification Impacts" in Barry, M., et al. *Toward sustainability: the roles and limitations of certification, Final Report.* Prepared by the Steering Committe of the State-of-Knowledge Assessments of Standards and Certification: Washington, DC (2012): A-169 – A-180

Deaton and Cartwright, "Understanding and misunderstanding RCTs", National Bureau of Economic Research, August 2016, accessed at http://www.princeton.edu/~deaton/downloads/Deaton_Cartwright_RCTs_with_ABSTRACT_August_25.pdf

Djimeu, Eric W., and Deo-Gracias Houndolo. "Power calculation for causal inference in social science: sample size and minimum detectable effect determination." *Journal of Development Effectiveness* 8.4 (2016): 508-527.

Gertler, Paul J., et al. *Impact Evaluation in Practice, Second Edition*, Inter-American Development Bank and World Bank: Washington DC (2016). Retieved from: https://openknowledge.worldbank.org/handle/10986/25030

Greene, Jennifer C. "Is mixed methods social inquiry a distinctive methodology?" *Journal of mixed methods research* 2.1 (2008): 7-22.

Hearn, Simon and Anne L. Buffardi, *What is impact? A Methods Lab publication*. Overseas Development Institute: London (2016) Retrieved from: https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/10302.pdf

Leeuw, Frans L., and Jos Vaessen. *Impact evaluations and development: NONIE guidance on impact evaluation*. Network of networks on impact evaluation, 2009.

Milder, Jeffrey C., et al. "Measuring impacts of certification on biodiversity at multiple scales: experience from the SAN/Rainforest Alliance system and priorities for the future." *Policy Matters* Issue 21. CEESP and IUCN: Gland, Switzerland (2016): 11-24

Nelson, Valerie, and Adrienne Martin. "Exploring issues of rigour and utility in Fairtrade impact assessment." *Food* Chain 4.1 (2014): 14-33.

R. Kumar et. al, *Evaluation of the early impacts of the Better Cotton Initiative on smallholders cotton producers in Kurnool district, India: Research design*, Natural Resources Institute (2016) Retrieved from: http://www.isealalliance.org/sites/default/files/private/ISEAL%20DIPI-ResearchDesign_India.pdf

Rogers, Patricia J., “Introduction to impact evaluation”, *Impact evaluation notes 3* (2012) Retrieved from https://www.interaction.org/sites/default/files/1%20-%20Introduction%20to%20Impact%20Evaluation.pdf

Scriven, Michael. “A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research” Journal of Multidisciplinary Evaluation, 5; 9; March 2008.

Stern, Elliot, et al. *Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development*. Department for International Development: London (2012) Retrieved from: http://www.alnap.org/resource/8196.aspx

Ton, Giel. "The mixing of methods: A three-step process for improving rigour in impact evaluations." *Evaluation* 18.1 (2012): 5-25.

Ton, Giel, Sietze Vellema, and Lan Ge. "The triviality of measuring ultimate outcomes: Acknowledging the span of direct influence." *IDS Bulletin* 45.6 (2014): 37-48.

Ton, Giel, Sietze Vellema, and Marieke De Ruyter De Wildt. "Development impacts of value chain interventions: how to collect credible evidence and draw valid conclusions in impact evaluations?" *Journal on chain and network science* 11.1 (2011): 69-84.

Westhorp, Gill. "Realist impact evaluation: an introduction." *Overseas Development Institute: London* (2014): 1-12.

White, Howard, *Impact evaluation: the experience of the Independent Evaluation Group of the World Bank*. World Bank: Washington DC (2006). Retrieved from: http://lnweb90.worldbank.org/oed/oeddoclib.nsf/b57456d58aba40e585256ad400736404/35bc420995bf58f8852571e00068c6bd/$FILE/impact_evaluation.pdf

White, Howard. "Theory-based impact evaluation: principles and practice." *Journal of development effectiveness* 1.3 (2009): 271-284.

White, Howard, “An introduction to the use of randomised control trials to evaluate development interventions.” *Journal of Development Effectiveness*, 5:1 (2013): 30-49

White, Howard, and Michael Bamberger. "Introduction: impact evaluation in official development agencies." *IDS bulletin* 39.1 (2008) Retrieved from: https://opendocs.ids.ac.uk/opendocs/handle/123456789/8243

White, Howard, and Daniel Phillips. *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework*. International Initiative for Impact Evaluation: New Dehli (2012). Retrieved from: http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf

# Annexes

## *Annex 1: Detailed research questions of all three baseline studies*

### Better Cotton Initiative, Cotton, India

- To what extent has the process of becoming or being certified under BCI sustainability standards had an impact (positive or negative, expected or unexpected) upon smallholders (farmers and households) in Kurnool district? What are the economic (yield, productivity, incomes, food security) and social (child labour, farm workers, no discrimination in wages for women) impacts?
- To what extent do we see an improvement in environment variables connected with cotton production (uptake of fertiliser use, reduction in pesticide use, efficient water use, soil health, habitat /biodiversity)?
- To what extent can Producers Unit and /or Farmer Producer Company 'empower' cotton farmers and households – both economically and socially?
- Can we see an increase in Better Cotton availability and uptake in the district /beyond? How can this be strengthened? What are the relative benefits and costs of meeting BCI standards and achieving certification for intended beneficiaries and supply chain actors?

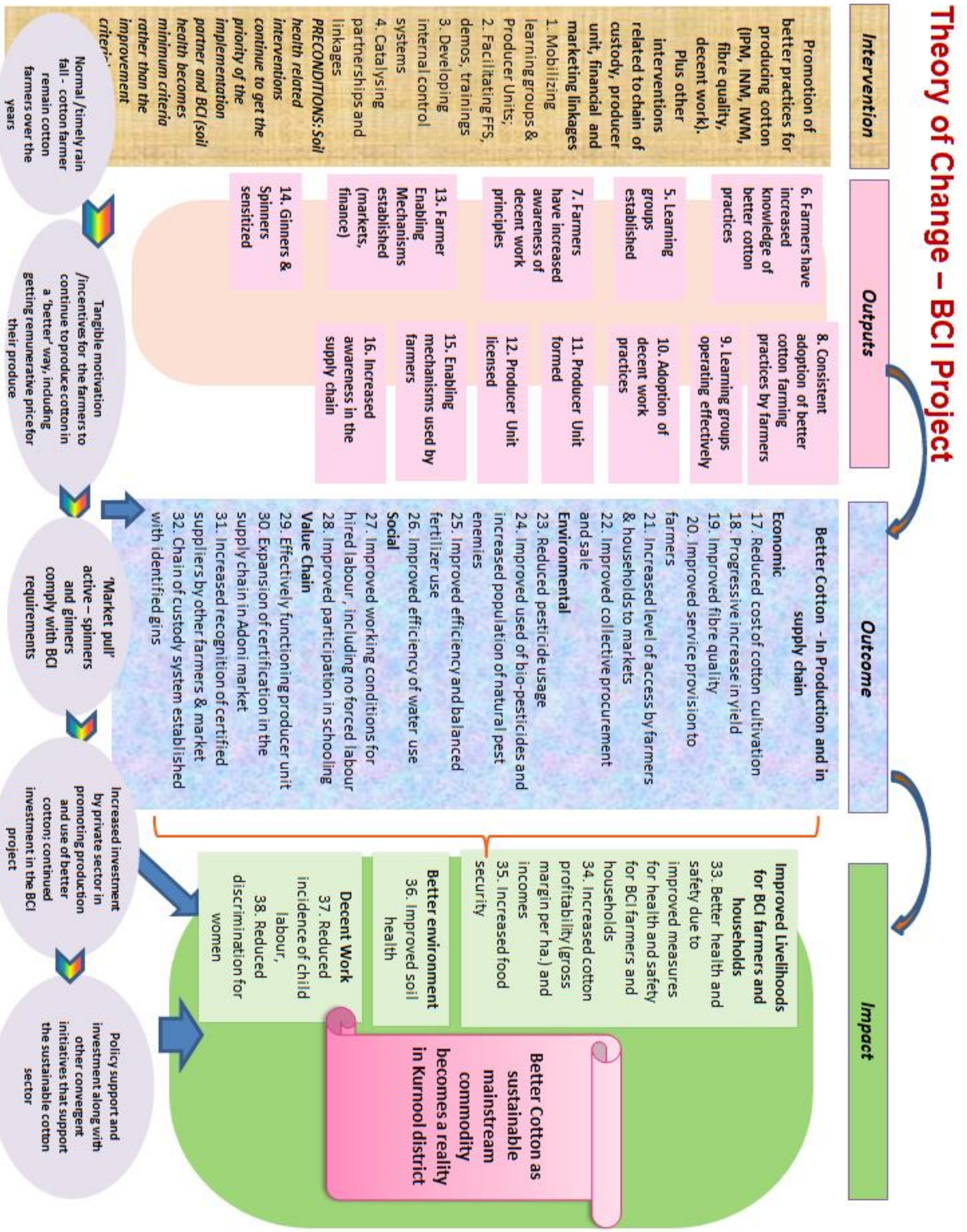
### Fairtrade and UTZ, Coffee, Kenya

- What are the changes that occur at the farm, household, and cooperative levels leading up to certification to the combined Fairtrade and UTZ standards and again after three years of certification?
- Do different types of farmers, such as those with different initial assets, poverty levels, or gender, experience differing changes in outcomes over time and what is the degree of difference?
- Can any observed changes in farm or PO performance be attributed to the combined Fairtrade and UTZ standard systems?
- What is the added value that Fairtrade and UTZ standards systems bring to POs, farms, and households, beyond training? This will include but not be limited to examining the extent to which farmers and PO managers feel satisfied with the experience of certification (in terms of challenges and cost-benefit perceptions).
- What contextual factors significantly influence the effect of Fairtrade and UTZ standards systems on PO, farm, and household changes in performance? The factors to test for influence are: the market orientation of the program, Kenyan and global coffee prices, the PO management and structure, livelihood and poverty context, cultural context, and project implementation experience.
- What are the reasons that different types of farmers (for example, those with different initial assets, poverty levels, or gender) experienced different changes in outcomes, if any such differences are identified in the quantitative analysis?

### 4C and Rainforest Alliance, Coffee, Indonesia

- What is the annual reach and market presence of the standard systems in southern Sumatra?
- What types of producers and producer groups are engaging with the standards systems? Are they reaching smallholders and marginalised farmers?
- Are producer groups and producers making progress along the outcome pathways identified in the conceptual framework?
- Do we see improvements in human well-being at the household level, particularly for small holders and marginalised producers?

*Annex 2: Theory of Change, contribution analysis frameworks for the three DIPI evaluations.*



Theory of Change – BCI Project

Intervention

Promotion of better practices for producing cotton (IPM, INM, IWM, fibre quality, decent work). Plus other interventions related to chain of custody, producer unit, financial and marketing linkages

1. Mobilizing learning groups & Producer Units;
2. Facilitating FFS, demos, trainings
3. Developing internal control systems
4. Catalysing partnerships and linkages

*PRECONDITIONS: Soil health related interventions continue to get the priority of the implementation partner and BCI (soil health becomes minimum criteria rather than the improvement criteri*

Outputs

6. Farmers have increased knowledge of better cotton practices

5. Learning groups established

7. Farmers have increased awareness of decent work principles

13. Farmer Enabling Mechanisms established (markets, finance)

14. Ginners & Spinners sensitized

8. Consistent adoption of better cotton farming practices by farmers

9. Learning groups operating effectively

10. Adoption of decent work practices

11. Producer Unit formed

12. Producer Unit licensed

15. Enabling mechanisms used by farmers

16. Increased awareness in the supply chain

Outcome

Better Cotton - In Production and in supply chain

Economic
17. Reduced cost of cotton cultivation
18. Progressive increase in yield
19. Improved fibre quality
20. Improved service provision to farmers
21. Increased level of access by farmers & households to markets
22. Improved collective procurement and sale

Environmental
23. Reduced pesticide usage
24. Improved used of bio-pesticides and increased population of natural pest enemies
25. Improved efficiency and balanced fertilizer use
26. Improved efficiency of water use

Social
27. Improved working conditions for hired labour, including no forced labour
28. Improved participation in schooling

Value Chain
29. Effectively functioning producer unit
30. Expansion of certification in the supply chain in Adoni market
31. Increased recognition of certified suppliers by other farmers & market
32. Chain of custody system established with identified gins

Impact

Improved Livelihoods for BCI farmers and households
33. Better health and safety due to improved measures for health and safety for BCI farmers and households
34. Increased cotton profitability (gross margin per ha.) and incomes
35. Increased food security

Better environment
36. Improved soil health

Decent Work
37. Reduced incidence of child labour,
38. Reduced discrimination for women

Better Cotton as sustainable mainstream commodity becomes a reality in Kurnool district

Normal /timely rain fall - cotton farmer remain cotton farmers over the years

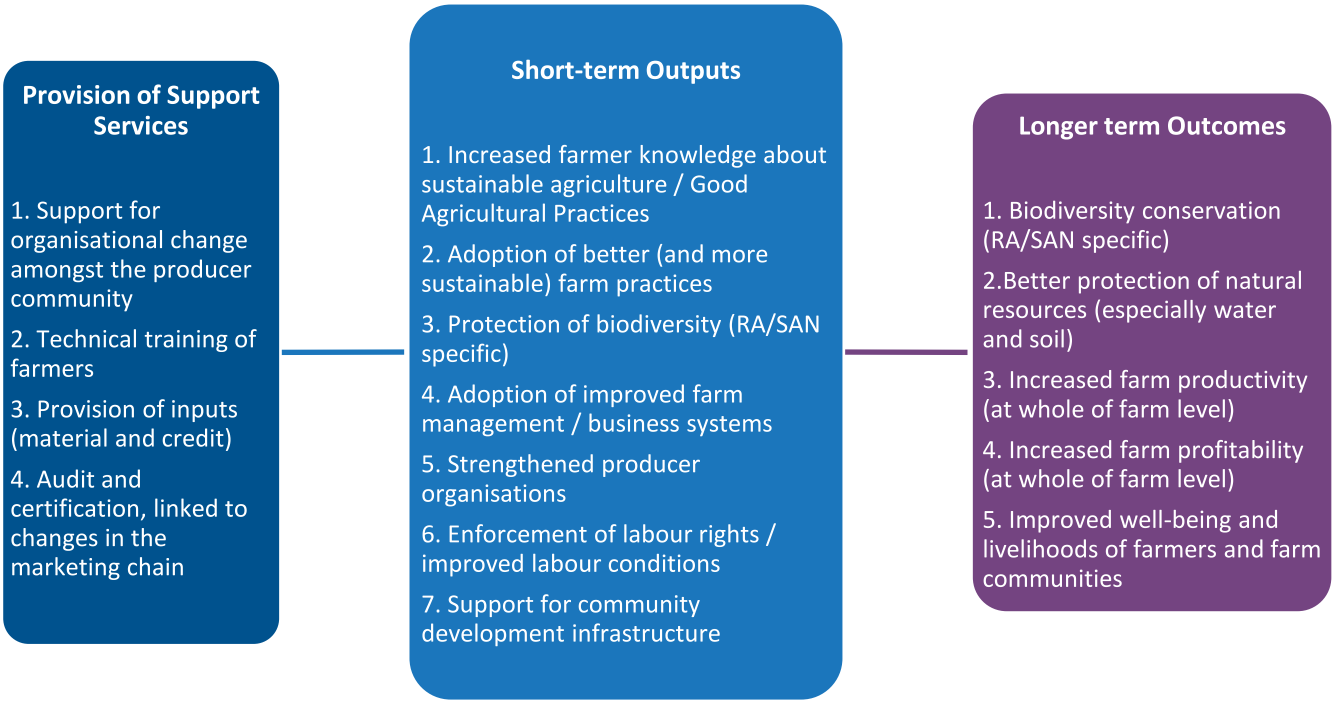
Tangible motivation /incentives for the farmers to continue to produce cotton in a 'better' way, including getting remunerative price for their produce

'Market pull' active – spinners and ginners comply with BCI requirements

Increased investment by private sector in promoting production and use of better cotton; continued investment in the BCI project

Policy support and investment along with other convergent initiatives that support the sustainable cotton sector

Theory of Change for 4C + RA/SAN Coffee Certification in Semendo, Indonesia



**Provision of Support Services**

1. Support for organisational change amongst the producer community
2. Technical training of farmers
3. Provision of inputs (material and credit)
4. Audit and certification, linked to changes in the marketing chain

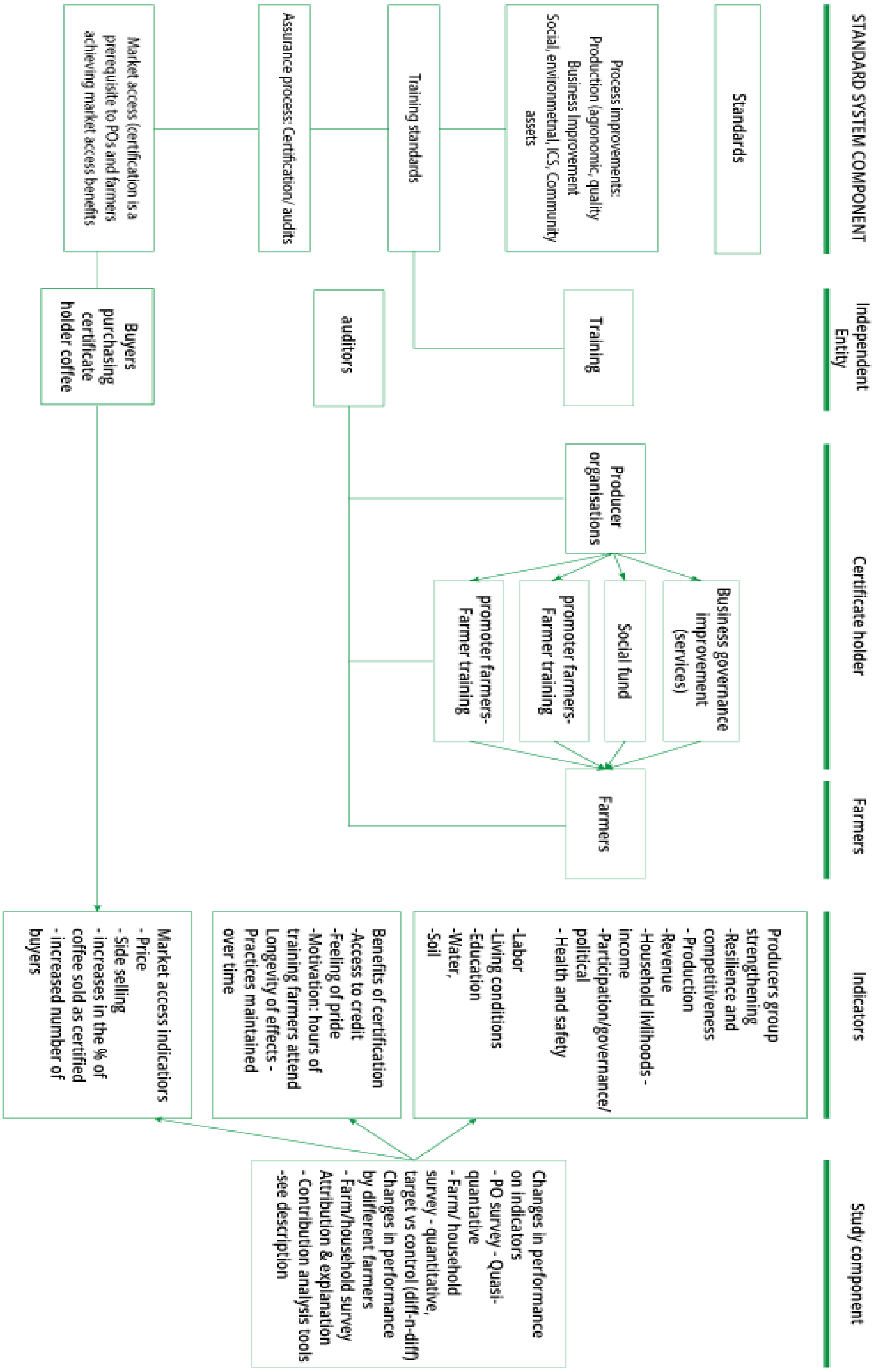
**Short-term Outputs**

1. Increased farmer knowledge about sustainable agriculture / Good Agricultural Practices
2. Adoption of better (and more sustainable) farm practices
3. Protection of biodiversity (RA/SAN specific)
4. Adoption of improved farm management / business systems
5. Strengthened producer organisations
6. Enforcement of labour rights / improved labour conditions
7. Support for community development infrastructure

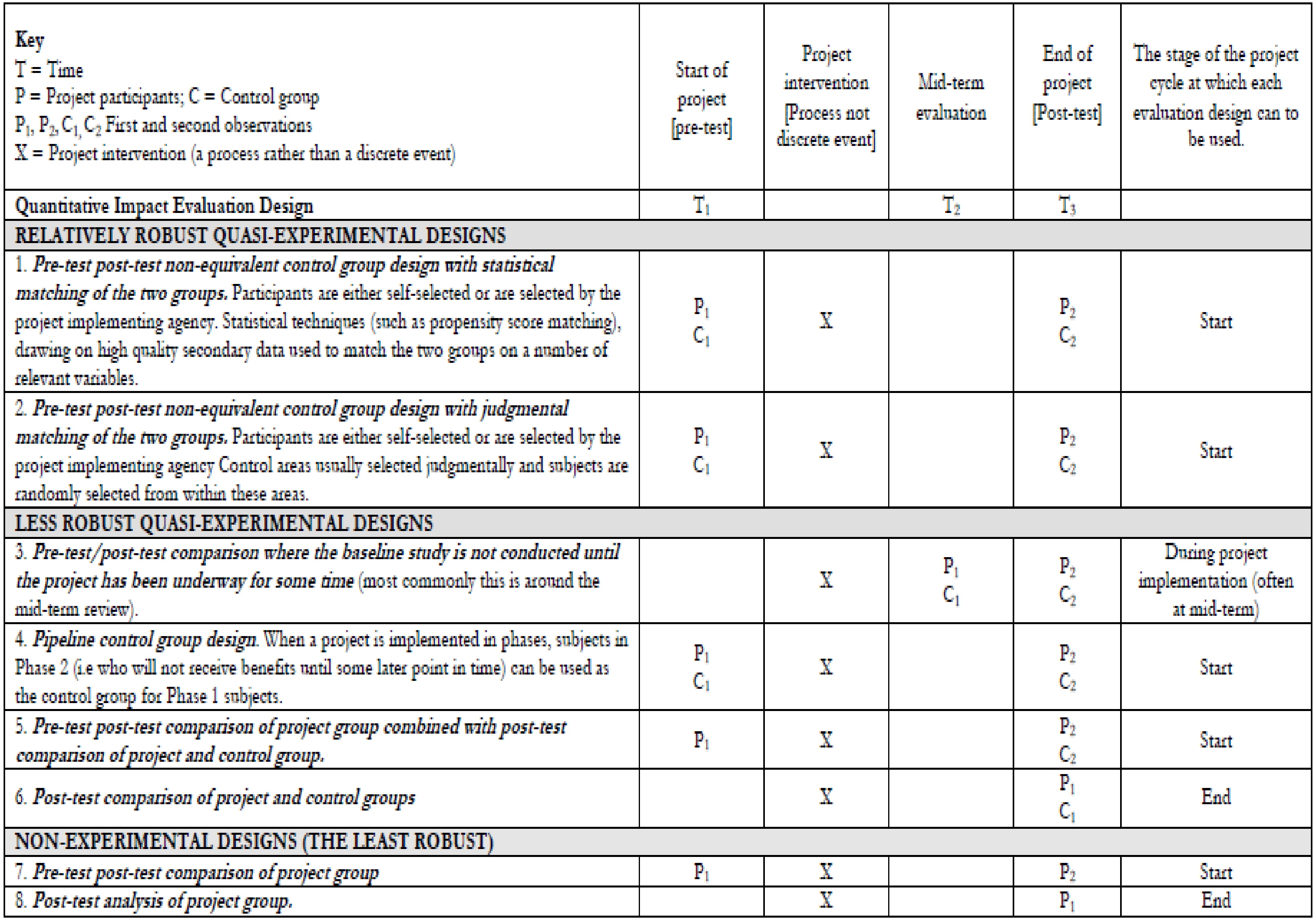
**Longer term Outcomes**

1. Biodiversity conservation (RA/SAN specific)
2.Better protection of natural resources (especially water and soil)
3. Increased farm productivity (at whole of farm level)
4. Increased farm profitability (at whole of farm level)
5. Improved well-being and livelihoods of farmers and farm communities

Theory of Change for UTZ and Fairtrade Coffee Certification in Kenya

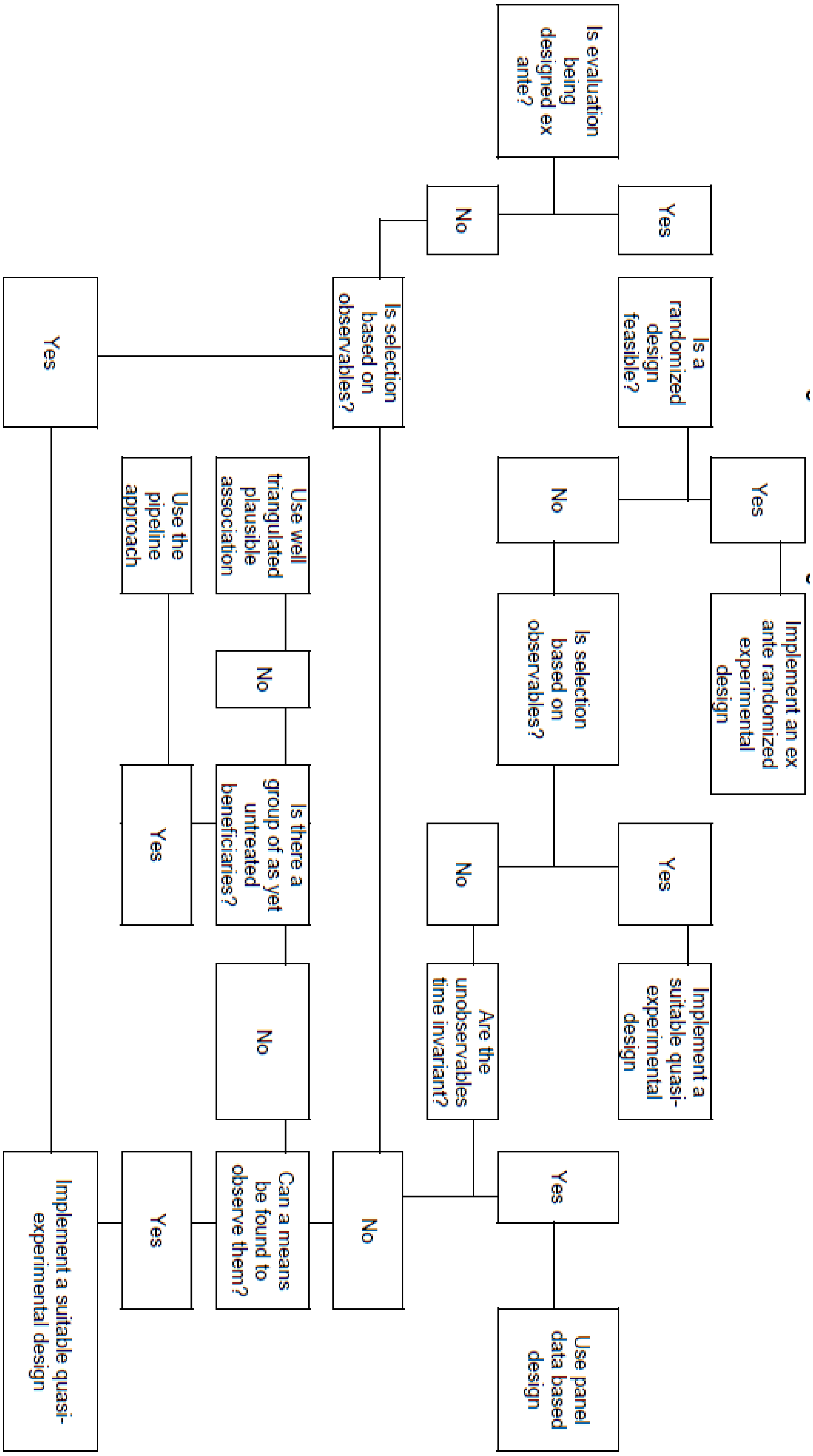

STANDARD SYSTEM COMPONENT
Independent Entity
Certificate holder
Farmers
Indicators
Study component
Standards
Process improvements: Production (agronomic, quality Business Improvement Social, environmetnal, ICS, Community assets
Training standards
Assurance process: Certification/ audits
Market access (certification is a prerequisite to POs and farmers achieving market access benefits
Training
auditors
Buyers purchasing certificate holder coffee
Producer organisations
Business governance improvement (services)
Social fund
promoter farmers- Farmer training
promoter farmers- Farmer training
Farmers
Producers group strengthening -Resilience and competitiveness - Production -Revenue -Household livlihoods - income -Participation/governance/ political - Health and safety -Labor -Living conditions -Education -Water, -Soil
Benefits of certification -Access to credit -Feeling of pride -Motivation: hours of training farmers attend Longevity of effects - Practices maintained over time
Market access indicators - Price - Side selling - increases in the % of coffee sold as certified - increased number of buyers
Changes in performance on indicators - PO survey - Quasi-quantative - Farm/ household survey - quantitative, target vs control (diff-n-diff) Changes in performance by different farmers - Farm/household survey Attribution & explanation - Contribution analysis tools -see description

*ANNEX 3: Commonly used quasi-experimental and nonexperimental impact evaluation designs (reproduced from Bamberger and White, 2007: 68).*



| **Key**<br>T = Time<br>P = Project participants; C = Control group<br>$P_1$, $P_2$, $C_1$, $C_2$ First and second observations<br>X = Project intervention (a process rather than a discrete event) | Start of project [pre-test] | Project intervention [Process not discrete event] | Mid-term evaluation | End of project [Post-test] | The stage of the project cycle at which each evaluation design can to be used. |
|---|---|---|---|---|---|
| Quantitative Impact Evaluation Design | $T_1$ | | $T_2$ | $T_3$ | |
| RELATIVELY ROBUST QUASI-EXPERIMENTAL DESIGNS | | | | | |
| 1. ***Pre-test post-test non-equivalent control group design with statistical matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency. Statistical techniques (such as propensity score matching), drawing on high quality secondary data used to match the two groups on a number of relevant variables. | $P_1$<br>$C_1$ | X | | $P_2$<br>$C_2$ | Start |
| 2. ***Pre-test post-test non-equivalent control group design with judgmental matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency Control areas usually selected judgmentally and subjects are randomly selected from within these areas. | $P_1$<br>$C_1$ | X | | $P_2$<br>$C_2$ | Start |
| LESS ROBUST QUASI-EXPERIMENTAL DESIGNS | | | | | |
| 3. ***Pre-test/post-test comparison where the baseline study is not conducted until the project has been underway for some time*** (most commonly this is around the mid-term review). | | X | $P_1$<br>$C_1$ | $P_2$<br>$C_2$ | During project implementation (often at mid-term) |
| 4. ***Pipeline control group design***. When a project is implemented in phases, subjects in Phase 2 (i.e who will not receive benefits until some later point in time) can be used as the control group for Phase 1 subjects. | $P_1$<br>$C_1$ | X | | $P_2$<br>$C_2$ | Start |
| 5. ***Pre-test post-test comparison of project group combined with post-test comparison of project and control group.*** | $P_1$ | X | | $P_2$<br>$C_2$ | Start |
| 6. ***Post-test comparison of project and control groups*** | | X | | $P_1$<br>$C_1$ | End |
| NON-EXPERIMENTAL DESIGNS (THE LEAST ROBUST) | | | | | |
| 7. ***Pre-test post-test comparison of project group*** | $P_1$ | X | | $P_2$ | Start |
| 8. ***Post-test analysis of project group.*** | | X | | $P_1$ | End |

*Annex 3: Decision tree to select evaluation design to control for selection bias in impact evaluations (Source: White, 2013)*


Is evaluation being designed ex ante?
Yes
No
Is a randomized design feasible?
Yes
Implement an ex ante randomized experimental design
No
Is selection based on observables?
Yes
Implement a suitable quasi-experimental design
No
Are the unobservables time invariant?
Yes
Use panel data based design
No
Is selection based on observables?
Yes
Implement a suitable quasi-experimental design
No
Can a means be found to observe them?
Yes
No
Is there a group of as yet untreated beneficiaries?
Yes
Use the pipeline approach
No
Use well triangulated plausible association

This paper is written by Vidya Rangan, ISEAL Alliance with inputs from Kristin Komives and Marta Maireles.
For comments and feedback, please write to vidya@isealalliance.org

FORDFOUNDATION
*Working with Visionaries on the Frontlines of Social Change Worldwide*

**ISEAL Alliance**
Wenlock Studios
50-52 Wharf Road
London N1 7EU
United Kingdom

+44 (0)20 3246 0066
info@isealalliance.org
twitter.com/isealalliance
**www.iseal.org**

**Photography credits**
Cover: Cotton, USA © BCI
Page 7: (from left to right) Telésfero Gamonal holding coffee bean, Peru © Rainforest Alliance | Senegal Cotton Producer © Stefan Lechner, Fairtrade Africa | Coffee farmer, Brazil © Rainforest Alliance
Page 8: Coffee, Guatemala © UTZ CERTIFIED
Page 12: Cotton worker harvesting, China © Better Cotton Initiative
Page 16: SONOMORO Coffee, Peru 2015 © Santiago Engelhardt for Fairtrade International
Page 22: (from left to right) Gerardo Goicochea, Peru © Rainforest Alliance | Carrying cotton in Mozambique © Better Cotton Initiative | Drying coffee beans in Vietnam © UTZ Certified
Page 23: Cotton, USA © BCI
Page 34: Coffee, Kenya © Giuseppe Cipriani. UTZ
Page 38: Female gin workers in India © Better Cotton Initiative
Page 42: Selection of coffee beans © David Macharia for Fairtrade International