THE CAMPBELL COLLABORATION

Protocol for a Systematic Review: Effects of Certification Systems for Agricultural Commodity Production on Socio-economic Outcomes in Low and Middle-Income Countries

Carlos Oya, Deborah Johnston, Evans Muchiri, Florian Schaefer, Dafni Skalidou, Kelly Dickson, Claire Stansfield

Submitted to the Coordinating Group of:

□	Crime and Justice
□	Education
	□ Disability
☒	International Development
	□Nutrition
□	Social Welfare
□	Other:

Plans to co-register:

☒	No		
□	Yes	□ Cochrane	□ Other
□	Maybe		

Date Submitted: 19 December 2014
Date Revision Submitted:
Approval Date:
Publication Date: 1 July 2015

Table of contents

Background ... 3

How the interventions might work ... 8

Why is it important to do this review? ... 13

objectives of the review ... 16

methodology ... 18

data inclusion and coding ... 29

Data synthesis and analysis ... 39

References ... 43

Review Authors ... 50

Roles and responsibilities ... 52

Potential conflicts of interest ... 52

Funding ... 53

Request support ... 53

Preliminary timeframe ... 53

Plans for updating the review ... 53

Author Declaration ... 54

1. Research design and analytical method assessment ... 55

2. Provisional set of electronic search terms ... 55

3. Critical appraisal of included studies of effects ... 57

4. overall risk of bias assessment ... 66

5. Risk of bias tool for studies examining the barriers and facilitators ... 67

6. outline of coding tool ... 69

7. Advisory group ... 71

BACKGROUND

Introduction

The role of international trade in reducing poverty and increasing welfare in low- and middle-income countries (LMICs) remains an issue of controversy and debate (Winters, 2003; McCulloch *et al.,* 2001). Open economies may perform better in the long term, Winters (2002) argues, but in the short term trade liberalisation can have adverse effects on the most vulnerable actors in the economy and some risk getting trapped in poverty. This is likely to happen to agricultural producers in developing countries, as Nicholls and Opal (2004) highlight, where deficient microeconomic conditions (poor market information, limited access to markets and credit, lack of ability to adapt rapidly to market changes, among others) are coupled with chronic macroeconomic failures, such as the lack of infrastructure and investment, heavy dependence on only few primary commodities and corruption. Primary commodity producers are often particularly vulnerable to price volatility and inadequate and asymmetric price transmission mechanisms.

In addition, international markets for agricultural commodities are increasingly demanding in terms of quality and production conditions, whether related to social or environmental sustainability (Henson & Humphrey, 2010). For example, demand trends in consuming countries have included the emergence of the 'specialty coffee market', which is becoming ever more important as traceability and other specific features become valued by consumers (Daviron & Ponte, 2005). Partly as a result of these broad world demand trends in agricultural trade, a wide range of voluntary private standards (or codes of conduct) have emerged in the past few decades to complement public standards to deal with the trade in agricultural commodities, typically monitored through private audits and third-party certification (Barrientos *et al.,* 2003; Schuster & Maertens, 2015). Such voluntary private standards can be classified either as own company standards, which affect only the workings and supply chain of a single company, or collective standards at both national and international levels, which are available to any number of actors as long as they can fulfil the requirements set by the standard (Henson & Humphrey, 2010). We focus in particular on certification schemes for agricultural commodity production, by which we mean collective standards, subject to third-party certification and auditing processes. Usually these standards should or tend to conform to internationally recognised guidelines such as ISO/IEC 17065:2012[1]. A broad definition provided by ISO/IEC states that 'the overall aim of certifying products, processes or services is to give confidence to all interested parties that a product, process or service fulfils specified requirements'[2].

[1] Which replaced ISO/IEC Guide 65:1996 (http://www.iso.org/iso/home.html)

[2] https://www.iso.org/obp/ui/#iso:std:46568:en

Since such standards increasingly determine the terms of integration of agricultural producers in LMICs into global supply chains (Gibbon & Ponte, 2005), an important debate has emerged about the effectiveness of certification in raising the welfare of direct producers and workers. Given the wide range of certification schemes, certified products and countries involved, it is perhaps not surprising that impact evaluations have found different results. Many studies tend to report mixed findings with some positive and other negative elements, or cases where effects are only marginal (Nelson & Martin, 2013). Some have even found that certification schemes may actually undermine the incomes of the poorest farmers (Henson & Jaffee, 2008), some reported positive impacts for some certification types, but not others (Chiputwa, Spielman & Qaim, 2014), others found effects only for richer farmers (Hansen & Trifković, 2014), while still others showed how certification schemes can help raise rural incomes and reduce poverty (Maertens & Swinnon, 2009). In the case of Fair Trade[3] standards the evidence from primary studies is not conclusive either and the quality of studies measuring effectiveness is uneven and uncertain, as a number of studies and meta-reviews show (Ruben, 2013; FTEPR, 2014; Valkila & Nygren, 2009, Terstappen *et al.,* 2012, International Trade Centre, 2011; Nelson & Pound, 2009; Nelson & Martin, 2013). Many of these studies and existing meta-reviews acknowledge the fact that mixed results may be consistent with the inherent complexities within these different types of interventions, the presence of many factors outside the control of interventions, such as the specifics of commodities and value chains considered, the different standards in question, as well as the diversity of implementation contexts, even within the same certification scheme.

This debate is likely to become increasingly relevant as the sales of agricultural commodities through market channels that require these kinds of certification expand rapidly. For example, in the case of Fairtrade, there are now more than 30,000 certified products worldwide (Fairtrade International, 2014) and the UK market has grown to over 4,500 Fairtrade certified products (http://20years.fairtrade.org.uk/). In the UK in 2013 alone 'sales of Fairtrade products exceeded an estimated value of £1.7bn, a 12% increase on 2012' (Fairtrade Foundation, 2014, p. 11). Part of the growth of certified products like Fairtrade has to do with the fact that large suppliers and retailers have embraced the branding opportunities involved, as in the case of Nestlé, which launched its own Fair Trade coffee in 2005, and Sara Lee/Douwe Egberts' growing association with UTZ Certified coffee for the European market (Tropical Commodity Coalition, 2012).

Description of Certification Schemes and their Interventions

Most certification schemes (CS hereafter) for agricultural commodity production have their roots in ideas about ethical trading in Europe and the US going back at least to the 1980s (Blowfield, 1999; Barratt-Brown, 1993). With supply chains lengthening as a result of the

[3] We distinguish between Fair Trade, a broad movement for ethical trade including various organisations, and Fairtrade, the organisation in charge of setting up standards through FLOCert.

spread of global value chains, consumers – and some firms –began to question the pay and working conditions of the workers and producers in LMICs. Ethical trade seemed to offer an alternative and by the late 1990s voluntary private standards were firmly established in a number of sectors (Barrientos, 2000). More recently, food standards, aimed primarily at quality assurance, have become vital to food exports from LMICs (Hansen & Trifković, 2014).

Generally, certification schemes aim to improve on the effects of free trade by offering better trading conditions, supporting smallholder producer organisations to gain better market access, assisting to enhance product quality, designing specific interventions or incentives to raise productivity, or a combination of these aspects. A challenge for any study of certification of agricultural commodities is that standards tend 'to vary in terms of their reach and objectives' and 'there are also major differences regarding the scope of the offering of certified commodities and products' (von Hagen *et al.*, 2010: 1). They encompass a wide range of different goals and of different methods of achieving those goals. An important differentiation has to be made between the act of licensing itself and direct interventions that precede or follow the licensing process. While the act of certification itself is not a development intervention per se, the introduction of codified standards, often, but not always, in the form of a consumer label following an auditing process, may induce behavioural changes in farmers, resulting for example in specific investments that benefit production conditions and open access to better market opportunities, without any direct intervention at farm level by the certifying body. But most certification schemes do require direct interventions at the level of the farm, the producer group or the workers' group. In short, different certification schemes are best understood as bundles of interventions, guided by a variety of theories of change.

As a result, certification schemes differ greatly in the populations they target, in the outcomes they seek to certify, in the implementation models, and in the audit and certification process itself. Besides, levels of compliance and requirements for improvement over time will also vary between certification schemes. There are certification schemes which operate primarily to enhance the quality and sustainable farming practices achieved by agricultural producers to ensure their products qualify for better market niches, more amenable to sustainable income generation, as is the case of some MPS certificates for flowers, or GlobalG.A.P. (previously EurepG.A.P.) for horticultural produce, which, for example, do not result in consumer labels. Other schemes more directly seek to establish ethical trading conditions by offering alternative markets with higher prices and/or floor prices that cover the costs of production. Amongst these certification schemes, one influential set are Ethical Trading Initiative (ETI) schemes and particularly Fairtrade, which aim to address the adverse effects of international trade by offering better trading conditions to, and securing the rights of, small agricultural producers, workers and their communities and helping them to organise to achieve these goals (Dragusanu *et al.*, 2014). Fair trade-type schemes, unlike other CS more concerned about the quality and characteristics of the product, were primarily designed to directly affect socio-economic outcomes and the empowerment of agricultural producers and workers through different direct interventions, originally as an alternative to perceived 'unfair' free market channels (Barratt-Brown, 1993). Fairtrade, for instance, operates through a set of standardised and audited interventions (floor prices, provision of a premium, credit-availability, assistance to access the market, support to small producer organisations –SPOs – such as small farmers' cooperatives and/or workers' organisations, and so forth) which are conditional on a number of requirements related to democratic processes, participation, transparency and the adherence to environmental and labour standards (see Fairtrade ToCs for a more detailed account of the various pathways to impact considered in relation to their different aims and interventions)[4].

It is worth mentioning that sometimes CS coexist alongside *additional* interventions by NGOs that adhere to the CS social and environmental sustainability standards, as is the case of OXFAM and the Fairtrade certification or TechnoServe (see section on study design for a more detailed discussion). Therefore, as well as the direct interventions being implemented by CS themselves, there is often some form of external support (by NGOs, donor agencies, buyers) which may have been leveraged because the producers or groups of producers have obtained a certification. This can affect the interpretation of findings, so the review will have to consider these additional factors as part of the moderator analysis, as long as these instances are actually reported by included studies.

The range of mediating factors is obviously wide and variegated. However, an important aspect, given that the focus is on agricultural commodities, are the market conditions and value chain characteristics for each particular commodity that may be subject to a range of standards set and monitored by various CS. Therefore, some CS may be more or less effective in reaching some socio-economic outcomes depending on the nature of the value chain and also on the dynamics of agricultural commodity markets, particularly for interventions that focus on prices and premium for producers. Commodity market cycles are likely to have an impact on producers and their workers thereby affecting the interpretation of findings about effects of CS depending on the period of time considered, even in the case of longitudinal studies.

Overall, in any case, by implementing the various bundles of interventions, CS are *expected* to produce positive outcomes that improve the wellbeing of beneficiaries in terms of higher and more stable incomes, better services to improve businesses, as well as education, health and other aspects of human welfare, and decent working conditions for wage workers. These

[4] Available at: http://www.fairtrade.net/fileadmin/user_upload/content/2009/resources/140112_Theory_of_Change_and_Indicators_Public.pdf

interventions are expected to directly and indirectly empower marginalised agricultural producers, workers and their communities.

A further complication is that CS increasingly expand their set of standards to qualify for a wider range of markets, products and consumers, and to compete with other certification schemes. A quick scoping survey of CS shows that overlaps may be significant and the wording of standards and codes of conduct are often strikingly similar despite very different histories and modus operandi[5]. Each certifying organisation may operate different certifications at the same time, depending on the standards applied and the target group (whether small or large farmers, workers, individual producers or organisations). Therefore it is not possible to assign a particular type of certification to a single particular scheme. Many of these schemes, for example apply conventional decent work ILO labour standards as part of their commitment to ethical trade, or share emphasis on 'sustainable farming methods'. In other words, while CS may be very different in some respects they may also overlap substantially on some of their standards. It is therefore necessary to distinguish between a certification scheme (Fairtrade, MPS, Utz Certified, Rainforest Alliance, and so forth) and a standard (more broadly social or environmental standards, and, more specifically, a living wage, the prohibition of certain chemicals, democracy in producer organisations, and so forth).

Moreover, overlaps may also happen at the beneficiary level, when producer organisations or individual producers/employers may receive more than one certification making attribution particularly difficult especially when studies do not report the timing of different certifications and the extent of compliance for each of the certifications (see for instance Woubie *et al.*, 2015 on the implications of double certification).

A related challenge in any review of studies of the effects of certification on socio-economic outcomes is that a particular type of certification can be provided by a variety of certifying bodies/organisations, which may fall under the broad category of voluntary 'social sustainability standards' and conform to broad internationally recognised guidelines such as ISO/IEC 17065:2012, which replaced ISO/IEC Guide 65:1996[6]. For instance, Fair Trade certification may be provided by the Fairtrade International (FLO), or alternative trade organisations within the WFTO, such as CTM Altromercato. Indeed, the Fair Trade network has evolved significantly in the past three decades and has given rise to a variety of organisations that may share a similar ethos and objectives but may differ in terms of focus, outreach, interventions and auditing processes (Jaffee & Henson, 2004; ProForest, 2005; Muradian & Pelupessy, 2005; Kolk, 2005). There can also be various levels of certification by

[5] See, for instance for MPS, http://www.my-mps.com/en/certificates-producer, for Fairtrade http://www.fairtrade.net/our-standards.html, for Utz Certified https://www.utzcertified.org/aboututzcertified and for organic with ethical trade http://www.sacert.org/farming/ethicaltrade
[6] See http://www.iso.org/iso/home.html for details.

the same certifying body as in the case of MPS, depending on what particular standards are applied. There is therefore a multiplicity of standards and certifications that often overlap and compete with one another (von Hagen *et al.*, 2010). The types of CS that are the focus of this review are described in detail in section 3.1.2.

A systematic review could in theory be conducted on every single intervention, which could happen under different CS, as in the case of labour standards interventions that are common to most schemes subscribing to ethical trade standards. However, the reality is that most CS operate with bundles of interventions and most studies will report on the certifications and not on single interventions. Also, seemingly similar interventions may be structured and implemented differently in different places and at different times, encompassing different intervention components. For example, technical assistance and capacity building for better farming practices or to improve organisational performance, may be implemented with a variety of intervention components. This makes the analysis of the causal chain particularly complicated because endpoint outcomes may be attributable to a bundle of interventions without sufficient evidence on which particular intervention component is more effective. For example, in the case of MPS-SQ is the certification more effective because of the enforcement of labour standards or because of the quality standards imposed and their spill-over effects on other intermediate outcomes? In the case of Fair Trade-type interventions, if effects on endpoint outcomes are considered positive, can they be attributed to the setting of a price premium or to the adequate investment of a premium or simply to a balanced combination of both in addition to other direct support to producers' organisations? Most impact evaluations will find it difficult to disentangle the specific effects of these different interventions under the same scheme. At the same time, most CS will consider that what matters is the specific mix of interventions, for instance including various forms of capacity building for producers, and not any one intervention in particular. However, studies may report relevant information that may give insights into the key causal mechanisms either through quantitative or qualitative evidence and this will be used to assess Review Question 2 (see the section on objectives for questions addressed in this review).

HOW THE INTERVENTIONS MIGHT WORK

A major challenge for this review will be that different certification schemes that aim to improve the welfare of agricultural producers and workers in agriculture differ in their model of intervention and in their theory of change (ToC). For example, Fair Trade schemes focus on prices, market access and organisational empowerment, while MPS is mainly about sustainable quality and social standards, and UTZ Certified, while similar to Fair Trade schemes in terms of the broad aims, works in terms of improvements in farming practices and quality rather than price mechanisms**.** Moreover, each certification scheme may also incorporate different grades of certification, as in the case of MPS for flowers, or the different standards applied by Fairtrade to SPOs (Small Producer Organisations) or HLOs (Hired Labour Organisations, that is, large-scale plantations).

Given the wide variety of certification schemes, their intended outcomes and methods of intervention there is no *single* theory of change that is valid for *all* types of certification schemes. There have been attempts by researchers to develop a ToC valid for more than one CS (Nelson & Martin, 2011; Nelson & Martin, 2013). Indeed in 2009, as reported by these authors, 'sustainability standards had yet to articulate their own theories of change [...] although this situation has now changed as a result of the ISEAL Impacts Code and with contributions from this research project', including subsequent studies that drew on NRI reports to develop ToC for impact assessments. Some certification schemes have also recently produced an explicit theory of change, but readers may benefit from consulting the ToC developed, for example, by Utz Certified and Fairtrade as indicative examples[7]. ISEAL, as an umbrella organisation for a range of sustainability standards have also produced a ToC and routinely publish drafts and discussions of ToC for individual member organisations[8].

Drawing from these initial attempts at developing ToC to evaluate impact evidence on sustainable standards, as well as on the recent ToC developed by CS themselves, we have produced a simplified synthetic ToC that summarises the key linkages in the causal chain between *types* of interventions, intermediate outcomes and endpoint outcomes, bearing in mind the focus of this review on socio-economic outcomes, and in line with some earlier attempts such as Nelson and Martin (2011). Some of the organisational ToC mentioned above, particularly the one developed in 2013 by Fairtrade, may be more complex and multifaceted than the synthetic ToC we propose here. This is because CS like Fairtrade also focus on actions and advocacy among consumers to expand the market for Fairtrade certified products and generally the values of Fair Trade. They also include environmental standards as part of the broad canvass of sustainable outcomes. The focus of this review is, however, on the role of standards and interventions that more directly affect the wellbeing of producers and workers involved in the production of certified commodities. The aim is not to evaluate the work of all these different CS, rather to evaluate and synthesise the existing evidence on socio-economic outcomes associated with interventions under CS as defined in this review. This is necessary to keep the review manageable and allow a consistent framework that can be applied to a wider range of CS.

Below are illustrations of how different types of interventions, which are used by different certification schemes, may affect intended outcomes, and therefore the assumed causal chains, which will be analysed in this review.

[7] See, for example, Fairtrade ToCs, published in 2013, http://www.fairtrade.net/fileadmin/user_upload/content/2009/resources/140112_Theory_of_Change_and_Indicators_Public.pdf

[8] See http://www.isealalliance.org/tag/theory-of-change

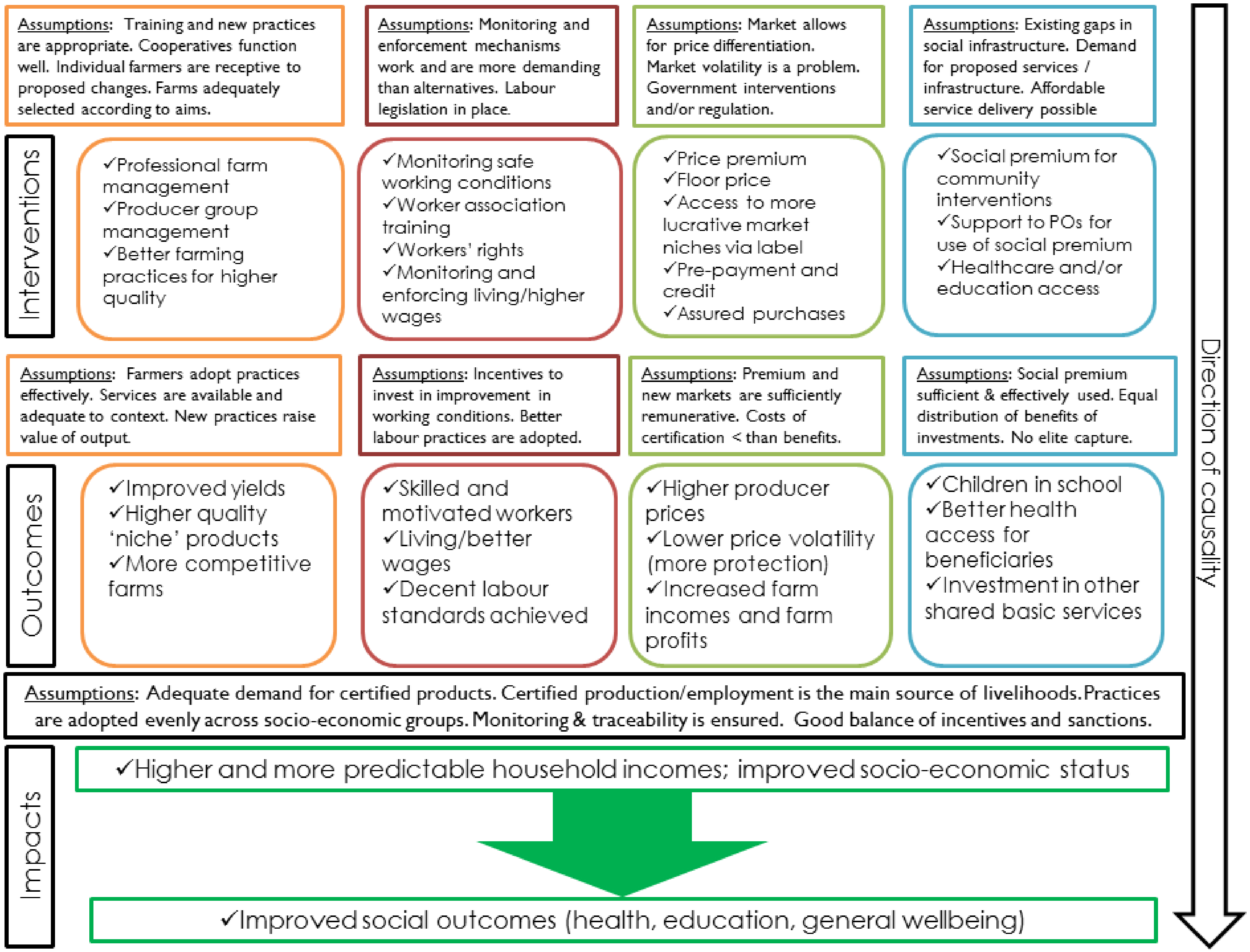
1. **Price and contract interventions**. Interventions to offer price premium or floor prices are expected to → contribute to higher and more stable producer prices, which can → result in higher net profits for agricultural producers, assuming they are not offset by high certification costs. In addition, both pre-payment, credit and longer-term contracts can → improve income stability and reduce vulnerability to shocks. These effects can → result in higher income and consumption at household level.
2. **Market access interventions**. Access to alternative and/or additional market certification schemes can → contribute to: higher prices; better contracts; strengthened market power and negotiation capacities of producer organisations and ultimately → to their members' empowerment.
3. **Product quality**. Better farming practices (including in some cases environmental standards developed through organic farming practices that may lead to better remunerated markets) and associated technical assistance (capacity building) are expected to → lead to higher agricultural incomes and strengthened market power of beneficiaries thereby → raising their capacity to invest in their own production.
4. **Monitoring of producer organisation practices and technical assistance (capacity building) to producer organisations and individual agricultural producers**. Democracy standards (Fairtrade) and other organisational improvements can →result in strengthened organisations in terms of their legitimacy, participation and capacity to negotiate, which → can lead to members' empowerment and access to better services.
5. **Premium.** The premium, as an additional sum of money paid on top of the minimum price → can be invested (by farmers and workers' organisations) in a variety of assets/infrastructure (for example, social, environmental and economic developmental projects), leading to possible positive outcomes → better education and health access/outcomes; incomes if economic infrastructure/assets improve; empowerment via strengthened beneficiary organisations; better working conditions.
6. **Labour standards**. Their implementation can directly → impact workers' wellbeing through of living/better wages, and better working conditions, especially when health and safety conditions improve and affect workers' health. Outcomes should be reviewed for workers employed by all types of agricultural producers from smallholder to large scale organisations.

In Figure 1 we present a simplified synthetic theory of change, which captures the overall logic of interventions under certification schemes. This is synthesised from multiple theories of change from some of the most prominent CS types. The synthetic theory of change was developed to be broad enough to be able to capture all intervention methods we are going to encounter under various certification schemes. It summarises potential causal pathways to impact and key assumptions for four different broad categories of intervention, namely interventions around farm practices, on prices, markets and purchasing agreements, on labour standards and through premiums. Some schemes focus on one or more of the interventions mentioned, while others focus only on one. This synthetic theory of change illustrates the difficulties inherent in aggregating results on effectiveness over a heterogeneous body of schemes and interventions.

A key aspect of any theory of change is a listing of the assumptions that must hold at each step along the causal chain for interventions to have their desired effect. If assumptions do not hold effects may be diminished, skewed, or entirely absent. In the worst case there may even be unintended adverse effects on producers or workers. However, assumptions also differ in their importance for different interventions and thereby certification types. For instance in some cases farmers' pre-existing capacities and therefore self-selection into the scheme (as in quality-oriented schemes) are more important than others (such as Fair Trade schemes for example, see section on study design for more details). In other cases assumptions about distribution of benefits among members of a group matter more when beneficiaries are targeted in groups (as with the premium in Fairtrade certification of SPOs for instance) than when they are targeted individually. The distribution of benefits may also not be equal between workers and employers, where large employers are targeted, or there may be differences between different types of workers. The distribution of personal protective gear may for instance benefit the most at risk workers (for example, sprayers), but may have very little impact on workers who are less at right (for example, those in packaging).

Finally, it is important to note that this review does not take the ToC shown below as the definitive ToC that will guide the final analysis of findings. This simplified ToC may omit some other possible pathways to impact that may not conform to the chains illustrated in the diagram. However, it provides what we think are the most important ones, based on knowledge of literature and other ToC that have been developed so far. In any case, the results of the integrated synthesis for Review questions 1 and 2 will be used to update and reconsider some of the linkages, assumptions and pathways to impact anticipated in the ToC developed for this protocol. It is also hoped that the resulting ToC will be of use to organisations in the process of revising or developing their own ToC, given that a ToC of certification standards is still work in progress.

Figure 1: Simplified synthetic theory of change



Assumptions: Training and new practices are appropriate. Cooperatives function well. Individual farmers are receptive to proposed changes. Farms adequately selected according to aims.

Assumptions: Monitoring and enforcement mechanisms work and are more demanding than alternatives. Labour legislation in place.

Assumptions: Market allows for price differentiation. Market volatility is a problem. Government interventions and/or regulation.

Assumptions: Existing gaps in social infrastructure. Demand for proposed services / infrastructure. Affordable service delivery possible

✓Professional farm management
✓Producer group management
✓Better farming practices for higher quality

✓Monitoring safe working conditions
✓Worker association training
✓Workers' rights
✓Monitoring and enforcing living/higher wages

✓Price premium
✓Floor price
✓Access to more lucrative market niches via label
✓Pre-payment and credit
✓Assured purchases

✓Social premium for community interventions
✓Support to POs for use of social premium
✓Healthcare and/or education access

Assumptions: Farmers adopt practices effectively. Services are available and adequate to context. New practices raise value of output.

Assumptions: Incentives to invest in improvement in working conditions. Better labour practices are adopted.

Assumptions: Premium and new markets are sufficiently remunerative. Costs of certification < than benefits.

Assumptions: Social premium sufficient & effectively used. Equal distribution of benefits of investments. No elite capture.

✓Improved yields
✓Higher quality 'niche' products
✓More competitive farms

✓Skilled and motivated workers
✓Living/better wages
✓Decent labour standards achieved

✓Higher producer prices
✓Lower price volatility (more protection)
✓Increased farm incomes and farm profits

✓Children in school
✓Better health access for beneficiaries
✓Investment in other shared basic services

Assumptions: Adequate demand for certified products. Certified production/employment is the main source of livelihoods. Practices are adopted evenly across socio-economic groups. Monitoring & traceability is ensured. Good balance of incentives and sanctions.

✓Higher and more predictable household incomes; improved socio-economic status

✓Improved social outcomes (health, education, general wellbeing)

WHY IS IT IMPORTANT TO DO THIS REVIEW?

This systematic review will address the extent to which, and under what conditions, interventions under various certification schemes for agricultural commodity production result in higher socio-economic welfare for agricultural producers and workers in low- and middle-income countries (LMICs) – questions about which there is an ongoing and as yet unsettled debate.

As briefly noted above, the current evidence base for the overall impact of interventions resulting from certification schemes for agricultural commodity production on agricultural producers and workers is generally mixed in terms of the reported results, including a range of studies that report either quite positive or negative results. There have, however, been some attempts to systematically review the evidence. A study by the International Trade Centre (2011), one of a four part review series on certification schemes, for instance seeks to present the overall findings of the relevant literature using systematic review methods. Unfortunately, the study uses vote counting, rather than a meta-analytic method that takes effect sizes into account, to synthesise the evidence and no information on effect sizes is presented. While a quality appraisal was undertaken, the results of this exercise for individual studies are not shared with the reader in any detail. The search methods used by the study also cast doubts on how comprehensive its literature coverage is. Searching seems to have been limited almost exclusively to two databases containing only academic journals.

Similarly, a review by Blackman and Rivera (2010) also uses systematic review methods to synthesise the available evidence on sustainability standards. Sadly, this review suffers from very similar issues as the study by the International Trade Centre, namely the reliance on a simple vote counting method, a lack of detail on quality appraisal and an unconvincing search strategy. In short, the existing reviews of the evidence suffer from serious shortcomings that make them unsuitable for research or policy use and the need for a high-quality systematic review using more sophisticated methods of searching and synthesis remains. There have also been many studies that have mapped the various codes of conduct, especially for wage workers, and the way these incorporate issues of gender and how they operate, but these tend to be focused on the nature, process and actors in these schemes rather than on their impact (see Barrientos *et al.*, 2003 for a seminal study of this kind of mapping).

The situation is not much different considering only the literature on Fair Trade interventions, for which more reviews are available. Partly as a result of the rapid increase in sales of Fair Trade products (Krier, 2007; Raynolds, 2000), the number of studies assessing the impact of Fair Trade has substantially increased from 2000[9]. Nevertheless, very few efforts have been made so far to synthesise this body of research. In an attempt to compile existing studies on the impact of Fairtrade, a literature meta-review was commissioned by the Fairtrade Foundation to map and analyse the impact of Fairtrade certification, including 80 academic and development agency reports of which only 23 provided evidence of economic impacts from 33 different separate case studies of Fairtrade certified producers (Nelson & Pound, 2009), while a similar compilation was

[9] Possibly this is a result of criticism regarding the lack of studies. See for instance Ronchis (2002) and Weitzmans (2006).

conducted by Vagneron and Roquigny (2011). Further, Terstappen *et al.* (2012) undertook a systematic scoping review on the social dimensions of Fairtrade, focusing on gender, health, labour and equity in particular. Overall, the three reviews present an account of the existing research, identify some methodological issues (Terstappen *et al.*, 2012; Nelson & Pound, 2009), and make future research recommendations (Terstappen *et al.*, 2012; Vagneron & Roquigny, 2011). Chan and Pound (2009) and Nelson and Martin (2013) have also extended the literature reviews on Fair Trade (including within them previous non-systematic meta-reviews like Nelson & Pound, 2009) to other CS, even if the additional number of studies was limited. Despite their possible influence on selected CS and policymakers (DFID), none of these reviews, however, provides an audit trail of the searching and synthesis process, nor do they systematically assess the quality of the studies they include. Moreover, they do not attempt a statistical meta-analysis of effect sizes or a rigorous and exhaustive synthesis of the qualitative evidence. In this sense, they may not be directly policy actionable.

Efforts have recently been made to increase both the quantity and quality of the evidence on the impact of Fairtrade in particular. However, as reported by Terstappen *et al.* (2012), FTEPR (2014) and Ruben (2013) the main bulk of studies is still characterised by evaluation designs vulnerable to validity threats, while the description of data collection and analysis tends to be poor, preventing assessments of the quality of the evidence. Moreover, there is a bias towards giving more attention to independent agricultural producers as opposed to wage workers (International Trade Centre 2011: 19). Therefore, the need for a systematic review with an inclusive framework, which identifies this expanding body of literature and critically appraises its quality, is clear and timely. Moreover, given the variety of potential mediating factors as well as the various methodological and contextual moderators to consider, this review will endeavour to systematically collect as much information as possible on contexts of implementation and particularities of interventions to be included in the coding and moderator analyses of the effectiveness analysis.

Moreover, each intervention may have differing effectiveness for different groups of rural inhabitants, particularly between rural inhabitants who focus on the production of certified products and those who are mostly dependent on wage labour. The existing evidence focuses much more attention on producers compared to wage workers, especially in the case of CS such as Fairtrade, Utz Certified and others in which smallholder farmers are a core constituency. A lot of research lacks either a baseline or other data on seasonal hired labour inputs and wages (Nelson *et al.*, 2002; Barrientos, 2003; Greenberg, 2004). Partly, this is the result of a wider research gap around rural wage labour, and the prevalence of wage labour in export commodity production is generally vastly underestimated (see for instance the 2013 World Development Report). This is particularly unfortunate, as a lot of research shows that farm workers, rather than farmers, are usually amongst the poorest of the poor (Barrett *et al.*, 2001; Sender, 2003; Hurst *et al.*, 2005; Jayne *et al.,* 2010). It has thus been argued that evidence of effects on wage workers under different schemes is especially limited, and some organisations, such as Fairtrade International, recognise that standards and auditing procedures need review in this respect, as exemplified by a recent Fairtrade International call for evaluation studies and evidence on the impact of smallholder certification on wage workers (http://www.fairtrade.net/vacancies.html on 28th August 2014; see also FTEPR (2014) on the issue of wage workers in Fairtrade certified smallholder farms and an acknowledgement of this gap by Fairtrade International 2015 report). There are of course other CS, which focus more directly on wage employment conditions and labour standards, as is the case of some ETI certifications and the well-known SA800 code established by Social Accountability International (http://www.sa-intl.org/).

The results of this review will be immediately relevant and hopefully actionable to both policy and practice, since they will provide guidance to certifying organisations, such as those who are members of the ISEAL Alliance, sectoral codes of conduct (such as MPS) and broadly ethical trading partners, as to the most effective elements of their interventions. Certifications are also becoming increasingly important to successful entry into global value chains, and are therefore receiving more and more attention in development policy circles. In addition, some of these CS, for example Fair Trade schemes, also receive public funding from government agencies aiming to improve rural livelihoods (for example, DFID) and organisations that provide financial or technical support to such certification efforts can also benefit from this comprehensive effectiveness review. The results will of course also be of direct interest to corporations engaged in buying agricultural produce from LMICs, and can contribute to debates around corporate social responsibility (Mezzadri, 2014). Stakeholders will also be interested in learning about any evidence on (negative or positive) *unintended* effects when studies report these. Moreover, we hope that the results will contribute to ongoing academic debates around the effectiveness of agricultural certification schemes and can help guide future research into areas where the evidence is either weak or ambiguous. Consumer groups or associations will also be interested, as they can gain knowledge to better inform their campaigns and priorities. Lastly, we hope the review be of use for agricultural producer organisations and workers' organisations, which invest resources in the certification processes of their members, as well as for individual agricultural producers who also invest in certification to achieve positive outcomes.

Given the inherent complexity of interventions associated with CS, the variety of implementation contexts and the specificities of different CS and their conditions facing their ultimate beneficiaries, this review may not settle the existing debates about the extent to which, and under what conditions, interventions under various certification schemes for agricultural commodity production result in higher socio-economic welfare for agricultural producers and workers in low- and middle-income countries (LMICs). However, it is hoped that the systematic nature of this review and the detailed data extraction it will entail will provide sufficiently rich information that may be policy actionable, and particularly help researchers and evaluators improve methods and implementation of impact assessments/evaluations of CS interventions in the future.

OBJECTIVES OF THE REVIEW

The primary objective of the review is to evaluate and synthesise evidence on the effects of certification schemes for sustainable agricultural commodity production on key socio-economic outcomes at the level of the individual producer and/or worker. As stated in the previous sections, the main aim is *not* to evaluate the work of all these different CS in relation to all their objectives as standard-setting organisations, but rather to evaluate the existing evidence on socio-economic outcomes associated with interventions under different CS. Although there is an increasing number of impact studies and non-systematic reviews of evidence on certification schemes in agriculture, both independent academic and commissioned research, the evidence base for the effects of such interventions on the economic and social welfare of their beneficiaries appears to remain limited, and – given the inherent difficulty and expense of conducting good impact evaluations – is likely to be characterised by high risk of bias.

An up-to-date systematic review is necessary to assess the quality of this growing evidence base, and synthesise the most important and reliable findings, which may help direct policy to the most effective uses and direct research towards areas where knowledge about the effects of such certification schemes on socio-economic outcomes is most limited. Further research is especially vital in areas where certification schemes may be shown to have had negative impacts. Based on the main review question outlined below, this systematic review will synthesise outcomes along the causal chain, making a distinction between *intermediate* outcomes such as price levels, farm profits, wages, better farming practices for higher output quality and productivity, or the provision of community infrastructure and services, and *endpoint* outcomes, including measures of household welfare such as household income, health and education outcomes. A related objective of the synthesis is to explore and discuss the heterogeneity of interventions and outcomes and the diversity of moderators that may affect the effectiveness of certifications schemes and their variation. As we explain in greater detail in the method section below, CS are a complex set of interventions bundles applied in a wide variety of circumstances. Rather than a seeking a single answer as to 'do they work', we are interested in knowing what works where, for whom and under what circumstances. To give a satisfactory answer to these questions, we must combine meta-analytic methods with a detailed examination of qualitative material and process documentation.

Accordingly, the review will seek to answer the following questions:

Primary Review Question (Review Question 1):

1. What are the effects of certification schemes for sustainable agricultural production, and their associated interventions, in terms of *endpoint* socio-economic outcomes for household/individual wellbeing, such as income (incl farm income), consumption, assets, working conditions, education, health (including nutrition and food security), empowerment, as well as primary *intermediate* outcomes (farm incomes in target crops, net returns to farm incomes in target crops, price levels and their volatility, wages and non-wage conditions, investments in community infrastructure, and so forth – see section 3.1.4) in low- and middle-income countries?

Subsidiary Review Question (Review Question 2):

2. Under what circumstances and why do certification schemes for agricultural commodities have the *intended* and/or *unintended* effects? What are the barriers and facilitators to such certification's *intended* and/or *unintended* effects?

This systematic review will report on *both* intermediate and endpoint outcomes, since many CS are primarily focused and interested in these intermediate outcomes, which may often be only *one* of many contributors to the ultimate or endpoint outcomes (Ton *et al.*, 2014). The synthesis proposed in this protocol will take this into account by considering different theories of change embedded in different certification schemes (see background section 1.3 and discussion of ToC)) and the limitations of available methods in establishing clear causal attribution on outcome effects to particular certification schemes and their interventions. Endpoint outcomes may be hard to attribute to CS even in the best-implemented impact evaluations but this does not make the discussion of effects on endpoint outcomes trivial, as Ton *et al.* (2014) would suggest, as long as sufficient account is taken of contextual and methodological differences through moderator analysis (in the effectiveness review) and through a narrative synthesis of relevant qualitative evidence. Indeed evidence on endpoint outcomes is policy relevant and should not be ignored even if CS do not always pretend to have a direct impact on these outcomes.

The subsidiary review question is important for a number of reasons. First, as stated above, this review will try to synthesise and evaluate evidence on what works where, for whom and under what circumstances. Second, there is an abundance of qualitative and mixed-method research in impact evaluations of CS, which can provide valuable evidence for the subsidiary review question, even if it cannot be used to address the primary review question. Third, while the ToC of most CS is explicit about the expected positive outcomes, there seems to be a gap in understanding *unintended* outcomes, whether negative or positive, and the circumstances in which these arise. While the effectiveness review can pick up key unintended outcomes based on counterfactual methods, the qualitative synthesis can add to this by bringing relevant evidence on implementation particularities, on process constraints and perspectives from both beneficiaries and implementers on unintended outcomes and the balance between intended and unintended outcomes in any given intervention. Fourth, the effects of CS are likely to be differentiated by type of scheme, intervention and context, and their benefits and costs unevenly distributed among stakeholders. Therefore, it is important to find and assess evidence relevant to these questions. The information collected in the review of qualitative evidence, around the four aspects mentioned above, will be also instrumental to conduct moderator analysis in the effectiveness review. Therefore, the qualitative analysis proposed under Review Question 2 will also feed into the effectiveness review for RQ1 via moderator analysis. However, it is not the aim of this review to restrict qualitative analysis to those cases or studies that are eligible for the quantitative effectiveness analysis. There is much to learn from qualitative and mixed-method studies that may not use experimental or non-experimental designs in order to shed light on barriers and facilitators as well as evidence of process and perspectives from both beneficiaries and implementers.

METHODOLOGY

The key principles for selection are noted here in detail. Studies will be included in the review if they meet the following selection criteria.

Criteria for including and excluding studies

Types of participants

The review will include studies on *agricultural producers* and *wage workers* living in low- and middle-income countries, as defined by the World Bank at the time the intervention was carried out. The target group may include individuals, households or producers' and workers' organisations. Depending on the availability of data in the included studies, the review will examine whether findings differ according to gender, age, socio-economic status, location, type of production (smallholder vs plantation), type of product, types of certification scheme, and length of participation in the supply chain of the relevant agricultural certification schemes.

The review will exclude studies that report on the impact of agricultural certification schemes on consumers only.

Types of interventions

The review will include studies on the effects of farm-level interventions in the production of agricultural commodities under certification schemes that have clearly defined socio-economic goals and third party auditing, even if socio-economic improvements are not the explicit primary aim of the certification scheme. The certification schemes, such as interventions that follow the Fair Trade principles, as defined by the World Fair Trade Organisation (WFTO), as well as other for examples under the social sustainability umbrella, must aim directly and explicitly to improve the wellbeing of beneficiaries.

Interventions that simply aim at advocating the objectives and activities of, for example, Fair Trade or other forms of ethical trade will be excluded, as they are designed to raise awareness among consumers without directly affecting the welfare of beneficiary agricultural producers and workers. Interventions and certification for the use of environmentally friendly production processes or environmental sustainability will also be excluded unless (intended or unintended) socio-economic outcomes are reported in studies and/or the certification includes ethical trade standards in addition to environmental standards. There are certification schemes, like Rainforest Alliance, that have environmental sustainability as a primary outcome, but also have explicit objectives in relation to improvements in labour standards. Therefore, studies that include evidence of the impacts of Rainforest Alliance, or similar schemes, on their intended labour standards will be included. Generally, organic standards focus on environmental sustainability and organic production practices, but there is substantial diversity especially if 'organic by-default' is included in the group, and some organic certifications also incorporate social sustainability (or ethical trade) standards that are directly relevant to socio-economic outcomes (Bennett & Franzel, 2013). Unintended effects of organic certification may also affect net returns to production and producers' wellbeing when productivity is negatively affected, so their inclusion can be of interest to this review. Indeed there are some studies that essentially report on socio-economic outcomes associated with *organic* certification, and these will be considered as unintended outcomes or intended depending on whether the organic certification includes ethical trade or other explicit criteria relating to the socio-economic wellbeing of producers and/or wage workers (for example, Ayuya *et al.*, 2015; Bolwig *et al.*, 2009; Bennett & Franzel, 2013). We are aware that there is a growing appreciation of the intertwined nature of social and environmental change processes, and the examples mentioned above attest to this reality. However, it is also true that certification schemes may aim to achieve environmental outcomes in their own right and with no necessary link with socio-economic outcomes. Other previous (non-systematic) literature reviews (Chan & Pound, 2009; Nelson & Pound, 2009) have noted the difficulties in comparing and aggregating impact findings from studies focused on ethical trading and those dealing with environmentally-driven standards. In this sense, a full appreciation of the effects of certification schemes on environmental outcomes would warrant a systematic review alone. Therefore, our consideration of environmental outcomes is subordinated to the main focus of this review on socio-economic outcomes and interventions with socio-economic objectives.

To give the review meaningful boundaries, CS that are not third-party certifications, such as certifications internal to particular corporations (for example, Nestle's AAA standard), will be excluded.

All interventions associated with CS included in this systematic review should have one or more of the following components:

1. Price and contract interventions that guarantee a floor price to agricultural producers and/or offer a price premium and/or provide credit and/or pre-payment and longer term contracts, compared to 'conventional' non-certified market channels.
2. Market access interventions that facilitate access to alternative, niche, specialty, and/or additional markets for agricultural producers, including labels that signal quality or traceability premiums, as in the 'specialty coffee' or flower markets, which are expected to directly benefit farmers through higher prices.
3. Provision of technical assistance and various forms of capacity building to individual agricultural producers for better farming practices that are designed to increase the quality and productivity of their commodities, partly designed to meet more demanding market standards, which aim to result in higher incomes and better market access.
4. Interventions that provide technical or organisational assistance and generally direct capacity building to agricultural producers organisations or workers' organisations. Such interventions may include capacity building of farmers or workers for production, or improvements in quality, and marketing improvements, as well as support for more effective self-organisation and monitoring of discrimination against vulnerable social groups as determined by local context (but typically women, especially if widowed, divorced or separated, children, youth, ethnic, caste or religious minorities).
5. Social or economic premium interventions that pay a premium for social or economic development projects which can be invested to improve production, marketing and/or

community services and infrastructure under the assumption of widely shared benefits at community level.

6. Labour standards interventions that set standards for living wages and improved working conditions. Such interventions include the monitoring of workers' rights and labour standards violations, and educational activities on workers' rights and labour standards.

Types of comparison and study design

For the primary research question, study design refers to the method of selecting participants and to the way in which group equivalence is ensured between treated and comparison groups. To answer Review Question 1 the review will *normally* include studies that compare agricultural producers or wage workers receiving a relevant intervention with a control group that receives no intervention. However, there is potentially significant heterogeneity both across CS and even within the same CS when the implementation models vary (Chan & Pound, 2009). In this regard, studies may also compare several different CS at once, and there may not be an untreated ('pure') control group. Such studies will be included as the comparisons are highly policy relevant. The key issue is that a comparison between different CS interventions may also provide a sensible counterfactual scenario, by comparing CS with the next best, or a similar, alternative. Such a comparison may be preferable to losing policy-relevant information in case there is no 'pure' control group. Besides, in the context of agricultural production it is usually very difficult to find a 'pure' control group in which other kinds of interventions or unobservables may not be present. In such cases there will also be additional efforts to account for the possibility of belonging to multiple schemes, via coding and subgroup and moderator analysis, in the context of Review Question 1. In addition, such information on direct comparisons is important to ascertaining relative effectiveness and can give information on barriers and facilitators, so these studies will also be considered in the synthesis for Review Question 2.

Comparison may be in terms of before/after (that is, a time before the introduction of certification), and/or cross-sectional, (that is, a group of non-participants or a location where certification has not yet been introduced). But before after/after studies will only be included if they have adequate controls for confounding, otherwise a causal attribution of effects to the intervention is not possible. Individuals will be associated with outcomes of certification where there are groups of agricultural producers or workers, producers' organisations or trade unions, or geographic areas when these correspond to locations dominated by, or with very strong presence of, certifying organisations.

In sum, study designs, whether for intermediate or endpoint outcomes, should ideally control for both observable and unobservable systematic differences between the certified and the control group, construct the counterfactual in a way that best simulates randomisation, account for spill-overs and drop-outs and explore heterogeneity of impact across sub-groups of participants. Therefore, in order to comply with best practice in systematic reviews, this synthesis of effects will include studies that have the methodological strength to deal with above mentioned challenges. Hence, studies eligible for inclusion to answer Review Question 1 are: experimental designs (where randomised assignment to the intervention is made at cluster level), which are unusual in the literature on certification schemes, and quasi-experimental designs, including controlled before and after (CBA) studies with contemporaneous data collection and with two or more control and intervention sites, as well as ex post observational studies with non-treated comparison groups and adequate control for confounding. Studies will be deemed to have adequate controls for confounding if they use statistical matching to equate the compared groups and/or employ multivariate statistical controls in outcome equations.

Studies will draw on a variety of statistical analysis methods to create valid comparisons, such as regression discontinuity designs (RDD), difference-in-difference analysis (DID), instrumental variable estimation (IV) and Two or Three Stage Least Square (2SLS/3SLS), and interrupted time series studies (ITSs).

Results obtained from single group studies, whatever the study design, will never be analysed together with results from controlled experimental or quasi-experimental studies. Studies that do not control for confounding using these methods, such as those based on inter-temporal comparison groups (pre-test post-test with no non-intervention comparison group), will be excluded from the effectiveness review.

For the subsidiary research question (Review question #2), we will include qualitative evidence which examines 'how' and 'for whom' certification works, by (a) paying attention to direct and indirect linkages between interventions and outcomes; (b) understanding mediating factors; and (c) explaining heterogeneous distributional outcomes (for example, gender and socio-economic status) (Mallett *et al.*, 2012: 453). Studies eligible to answer this review question are: independent academic qualitative research on CS, eligible studies under the primary review question, commissioned research on CS interventions with low-risk-of bias qualitative evidence (that is, not simply advocacy reports), process evaluations obtained on the interventions evaluated in the effects review, before-after comparative studies, which contain rich information about intervention implementation and 'pathways to impact', even though they may not include a comparison group; other qualitative studies, ethnographies and other types of studies that present evidence on the outcomes of certification interventions as long as they meet the conditions below. Examples of eligible studies include:

- non-experimental studies that examine the direct and indirect impacts of agricultural certification schemes, such as of Fairtrade in this case, using qualitative and mixed-methods like interviews (Ronchi, 2002) or participatory action research and surveys (Bacon, 2010).

The studies considered above for the subsidiary question should:

1. Report on CS interventions on both their processes and outcomes.
2. Contain primary evidence.
3. Provide evidence on either intended or unintended effects and of the causal mechanisms, particularly on the key assumptions detailed in the synthetic Theory of Change.
4. Report at least some information on all of the following: the research question, procedures for collecting data, sampling and recruitment, and at least two sample characteristics.

A variety of research designs and methods are in principle capable of fulfilling the above criteria. These include, but are not limited to qualitative comparative case study research, life histories, rapid appraisals, participatory assessments, participant observation and broadly ethnographic methods.

We will not limit the studies used for the subsidiary question to those that provide information on the same context or country as the eligible studies for the primary review question. This might substantially reduce the number of eligible studies and therefore limit the quantity of relevant information collected to address the subsidiary question. The quality appraisal criteria for qualitative studies will be the main selection criteria in this case.

Types of outcome measures

The review will include studies that contain data on outcomes related to the synthetic theory of change. Outcomes may be intermediate or endpoint, intended or unintended. The focus of the review is on the endpoint outcomes for wellbeing and empowerment of beneficiaries and the conditions of their activities. The review will however also include studies that report on both primary and secondary outcomes, as defined below:

Primary outcomes, divided by *endpoint* and *intermediate* outcomes, include:

1. Household income or consumption or other measure of socio-economic status (monetary measures of total household income or consumption, asset or wealth index, as used in Demographic Health Surveys) (*endpoint outcome*).
2. Health and education of adults and children (years of schooling, literacy, current enrolment status, work days lost due to illness, infant mortality rate) (*endpoint outcome*).
3. Gender equity in the outcomes above (*endpoint outcome*).
4. Producers' and workers' empowerment (*endpoint outcome*). At this stage it is not yet clear whether studies produce consistent measures of ‘empowerment’ and whether some of them overlap with outcomes mentioned above. There is a rich literature on ‘women's empowerment’ that may help operationalise this set of outcomes. Kabeer (2001: 81) broadly defines it as “expansion in the range of potential choices available to women”. However, there is a wide range of measures that attempt to capture effects of an intervention on empowerment. Indeed this can be the case in the context of diverse CS in LMICs. Some measures or understandings of ‘empowerment’ may be in the form of concrete outcomes such as the co-ownership of processing/trading businesses as in the case of Kuapa Kokoo in Ghana and Divine Chocolate (Doherty & Tranchell, 2005), while some may be reported as subjective assessments (perceptions) of greater capacity to control and/or influence, change or participate in a value chain (for producers) or perceptions of greater capacity to engage in collective action for better working conditions in the case of wage workers. Empowerment measurements can also be organised around the notions how interventions affect the ‘existence of choice’, ‘use of choice’ and ‘achievement of choice’ of producers and

workers (Alsop & Heinsohn, 2005). During the data extraction process, once studies have been included, different measurements of empowerment will be considered.

5. Gross or net returns to certified production (*intermediate outcome* as all other outcomes below), measured as gross/net farm profits or as farm income associated with target crop depending on how reported by studies.
6. Quality of commodities (measured in terms of grades or quality premium specific to commodities and which normally result in higher prices/returns).
7. Productivity of commodities (yield, that is, output per land unit).
8. Price levels (for certified commodity and as farm-gate prices, that is, those effectively received by certified producers).
9. Price volatility (for certified commodity). Actual year-on-year historical volatility in standard deviation units or CV.
10. Wages (nominal and/or real, daily equivalent or other time unit). This outcome is of course part of household income, or contributes to it as an intermediate outcome, but may be reported separately as labour standards are a core component of many CS in ethical trading so it should be assessed separately.
11. Non-wage labour conditions (health and safety: number of work-related injuries, access to health care, type of heath care available; benefits and entitlements: sick pay, paid holidays, maternity and paternity leave, free or subsidized food, clothing or shelter, freedom of association, and so forth).
12. Organisational empowerment of producers' and workers' organisations (that is, empowerment as a collective group and not just at individual level), which requires a consideration of the challenges in measuring empowerment as noted above (in order to operationalise, studies may report various measures of enhanced capacity to benefit from value chain or engage in collective action; this can take the form of direct participation in market institution decision-making bodies or on concrete facts about successful collective bargaining).
13. Investments in services and infrastructure, funded by social or economic premium, as advances or direct transfers from certifying organisations. The indicators can take the form of counts of infrastructure or service units created (health posts, housing for teachers/pupils, km of roads, processing plant, warehouse, and so forth).

Secondary outcomes include *both* endpoint outcomes (that are related to empowerment or equity) and intermediate outcomes, as follows:

1. Unintended outcomes (may be positive or negative/adverse); Unintended effects of certification, which can affect the above endpoint outcomes, such as effects on production costs (certification costs), debt, and workload, and local market conditions (that is, local prices, access to local markets) will also be included.

2. Environmental outcomes (either as operational outcomes such as adoption of organic methods or knowledge about environmentally friendly practices or more ‘endpoint’ type outcomes if they affect reported socio-economic outcomes.

To be eligible for inclusion in the review, studies that report on the secondary outcome must also report on at least one of the primary outcomes.

Approach

The specification of the inclusion/exclusion criteria outlined here will be thoroughly tested by conducting a pilot search and screening exercise, which will clarify what types of study are likely to be included in the systematic review. Pilot searches will be conducted by one researcher who could draw on the assistance of a search specialist to ensure that final searches are as exhaustive as possible. Pilot screenings will be conducted independently by research assistants, who are going to conduct the final screening. Comparative screening reports will be used to identify and correct discrepancies. Pilot screenings will be repeated until screening consistency is ensured. The inclusion/exclusion criteria will guide the initial pilot search and, in line with best practice in systematic review methodology, will then be re-applied by research assistants to the sets of studies found by the search process to determine the final set of studies that will be analysed. The research assistants will work under close supervision by review team members. Since transparency is considered to be an effective way to maintain the rigour of systematic reviews, even when including qualitative and mixed-methods research, in addition to quantitative studies (Snilsveit, 2012), specific difficulties that will arise when conducting the review shall be dealt with transparently and consistently.

Study language

The review will only consider results from studies published in English, Spanish, French, German and Portuguese, which are the languages most likely to be used in the literature on CS, given that they are spoken in the biggest consumer markets for certified agricultural commodities.

Summary table

A summary of selection criteria for the primary and secondary research question is shown in tables 1 and 2 below.

Table 1: Inclusion/Exclusion Criteria for Primary Research Question

Parameters	Inclusion Criteria	Exclusion Criteria
Location	Low- and middle-income countries	High-income countries
Language	English, Spanish, French, German, Portuguese	Other languages
Timeframe	1990 until present	
Population	Individual agricultural producers Agricultural wage workers	Consumers Non-agricultural producers Non-agricultural wage workers
Interventions	Studies concerned with the effects of farm level interventions (including processes and audits towards certification) in the production	Studies that are concerned with the effects of farm level interventions of own-

	of agricultural commodities under certification schemes that have clearly defined socio-economic goals and third party auditing	company standards monitored through private audit Studies that are concerned with the effects of interventions in the production of non-agricultural commodities Studies that are concerned with the effects of farm level interventions that have *only* environmental sustainability or environmental-friendly production goals, unless they report on substantial unintended socio-economic outcomes Studies concerned with the promotion of ethical trade generally
Outcomes	Studies will include information on primary (and possibly secondary) outcomes, as specified above in section 3.1.4.	Studies include no information on primary outcomes
Study Type	Primary research that uses quantitative experimental or quasi-experimental methods (see Study Design Criteria for more detail)	Book reviews, literature reviews, policy documents, qualitative research, quantitative methods that are not experimental or quasi-experimental

Table 2: Inclusion/Exclusion Criteria for Subsidiary Research Question

Parameters	**Inclusion Criteria**	**Exclusion Criteria**
Location	Low- and middle-income countries	High-income countries
Language	English, Spanish, French, German, Portuguese	Other languages
Timeframe	1990 until present	
Population	Individual agricultural producers Agricultural wage workers Producers’ and workers’ organisations (in the case of organisational empowerment)	Consumers Non-agricultural producers Non-agricultural wage workers
Interventions	Studies concerned with the effects of farm level interventions (including processes and audits towards certification) in the production of agricultural commodities under certification schemes that have clearly defined socio-economic goals and third party auditing	Studies that are concerned with the effects of farm level interventions of own-company standards monitored through private audit Studies that are concerned with the effects of interventions in the production of non-agricultural commodities Studies that are concerned with the effects of farm level interventions that have *only* environmental sustainability or environmental-friendly production goals, unless they report on substantial unintended socio-economic outcomes

		Studies concerned with the promotion of ethical trade generally
Outcomes	Studies that include relevant information on barriers and facilitators to both intended or unintended outcomes	Studies include no relevant information on barriers and facilitatorsor on unintended outcomes
Study Type	For the interventions evaluated in the effectiveness review, background project documentation, project completion reports and process evaluations. More generally on relevant interventions, primary research that uses quantitative, mixed-methods, ethnographic and other types of qualitative studies (see Study Design Criteria for more detail)	Book reviews, literature reviews, policy documents and advocacy reports without primary research or without reporting on methods and research process Generally, studies that are not based on primary research

Search strategy

We will search for studies, both published and unpublished, that report on the effects of certification schemes (CS) for agricultural commodities and their associated interventions, as well as for studies that examine the circumstances under and the reason(s) for which such interventions have intended or unintended effects (barriers and facilitators to CS effectiveness). Additionally, we will conduct targeted searches to include process evaluations and background project/programme documentation related to the interventions evaluated in the effects review. Searches will be restricted to studies published from 1990 and onwards[10], without language restrictions at this stage (see below). In accordance with guidelines by Hammerstrøm *et al.* (2010) for Campbell Systematic Reviews we will work closely with an information specialist/librarian to devise and quality-assure our general search strategy, and to ensure that it is as exhaustive as possible. We will use the EPPI-Reviewer bibliographic software (Thomas *et al.*, 2010) to manage retrieved references. All references will be downloaded along with the necessary fields (that is, abstract, article identifier, index terms/thesaurus) and imported to the EPPI-Reviewer. If the reference source is not supported by exportation facilities, relevant references will be imported manually. Duplicates will be removed automatically with EPPI-Reviewer, and, where this fails, manually during the screening process. All the searches will be documented and a detailed record of the type of search (that is, electronic, hand-searching, and so forth), specific search strategy, number of references retrieved, date of search and search source will be provided in the final review report in order to provide a transparent trail for replication and validation.

Electronic Searches

To ensure we conduct the most comprehensive search possible, we will search multiple databases, as suggested by Hammerstrøm *et al.* (2010), including general social science-related bibliographic

[10] Our rationale for adopting this cut-off point is the following: although some of the certification schemes covered in this review have existed long before 1990, it is probably the Rio Declaration (1992) that for the first time recognises at a global level the need "to promote sustainable consumption and production" (Potts *et al.*, 2014:19). We therefore believe that it is very unlikely for any study evaluating the effects of such schemes to have been conducted before 1990.

databases, subject specific databases covering agriculture and international trade/economics, systematic review databases, and national and regional databases. We will cover the following databases:

- AgEcon
- Africa Wide
- CAB Abstracts
- International Bibliography of the Social Sciences (IBSS)
- Social Sciences Citation Index (SSCI) / Web of Science
- Econlit
- US National Agricultural Library
- JOLIS
- British Library for Development Studies (BLDS)
- IDEAS repec
- 3ie systematic reviews and impact evaluations database
- The Campbell Library
- AGRICOLA
- Labordoc
- SCIELO

We will also search grey literature databases, as well as websites of research institutions, organisations related to CSs for agricultural commodities, funders and donors. We will cover the following:

- Networked Digital Library of Theses and Dissertations
- ProQuest dissertation database
- Best Evidence Encyclopaedia (BEE)
- ELDIS/Institute of Development Studies (IDS)
- ESRC (Economic and Social Research Council)
- World Bank
- IFPRI
- R4D, DFID
- ISEAL Alliance
- COSA, Committee on Sustainability Assessment
- World Fair Trade Organisation
- Fairtrade Foundation
- Center for Fair and Alternative Trade
- Fairtrade International
- Fair Trade Resource Network
- Fair Trade Institute
- European Fair Trade Association
- Fair Trade USA
- Traidcraft

- Oxfam
- MPS (Fair flowers fair plants)
- Soil Association certification (ethical trading)
- Utz Certified
- Rainforest Alliance
- GlobalG.A.P.
- TWIN
- AGRIS
- FAO Catalogue Online
- CGIAR
- Agra.org
- USAID

Finally, we will cover relevant databases of studies in French, German and in Spanish, even if they are not indexed in English-language databases. This will require targeted selection of relevant databases, especially for Latin America (such as SCIELO).

In order to produce a comprehensive list of keywords related to the review's inclusion criteria (PICOs), we will combine brainstorming and pearl-harvesting (collecting keywords from studies that meet the inclusion criteria) as suggested by Sandieson (2006). Additionally, we will study the thesaurus of each database and customise our general strategy accordingly, including the appropriate controlled vocabulary for each database. The following basic search strategy will be adapted to each database, combining text terms with indexing terms using Boolean (AND /OR) and Proximity (NEAR/WITHIN/ADJ) operators: '[certification terms] AND [population terms] '.

A provisional set of electronic search terms is provided in Appendix 2.

All customised search strategies will be piloted in order to assess their relevance and precision and to identify the most optimal set of search terms. We will prioritise high sensitivity of the search terms over precision in order to avoid omitting relevant studies, which do not report sufficient information in their title or abstract. Reviewers will be over-inclusive at the first screening stage of titles and abstracts. Potentially relevant abstracts will be double screened by two independent reviewers to determine which papers should be retrieved and reviewed at full text. Two reviewers will then independently assess full-text studies for inclusion, and possible disagreements will be arbitrated by a third reviewer. When possible, in order to ensure the review is as inclusive and up-to-date as possible, search strategies will be saved in the database system and we will update our searches during the synthesis phase, limited to the period from the last search and onwards, to include any additional relevant record indexed in the meantime. Additionally, we will set up alerts for relevant authors and thesaurus terms.

Other searches

We will use snowballing on a continuous basis while "the study unfolds", as recommended by Greenhalgh and Peacock (2005:1064). In order to locate eligible studies that could have escaped the electronic search we will screen the references of included studies and of the existing literature and systematic reviews. Additionally, we will use special citation tracking databases such as the Social Science Citation Index (SSCI)/Web of Science, Google Scholar and Scopus to forward track all included studies and selected key papers in order to identify articles that have subsequently cited those papers (ibid). We will conduct hand searches of the books and journals that do not appear in the indexed search results, as well as of recent issues of indexed journals of interest that have not been indexed yet. Moreover, we will make use of our existing knowledge of the literature and we will contact and consult our advisory group (see Appendix 7), key researchers, relevant academic networks and organisations working in the field of CS in order to identify additional eligible studies, including unpublished papers or on-going research. Last but not least, we will be "alert to serendipitous discoveries", that is, finding a relevant study when looking for something else (Greenhalgh & Peacock, 2005:1065).

DATA INCLUSION AND CODING

Study design inclusion criteria

The review will adopt a theory-based, mixed-methods approach and will include a broad range of evidence from both quantitative and qualitative research (Snilstveit 2012). As argued in the background section of this protocol, there are inherent challenges in a systematic review that includes a range of different interventions and a variety of intermediate and endpoint outcomes. Some studies will report on some outcomes (for example, incomes, health outcomes, and so forth) and refer to some interventions (for example, premium in a Fairtrade scheme) under various CS, and other studies will focus on other outcomes (for example, wages, gross or net return to farming, empowerment) and different interventions (technical assistance to SPOs, training in farming methods). The breadth of the review and the heterogeneity of possible documents may give rise to a wide range of study designs.

In order to assess the effects of the agricultural CS, the review will include studies using experimental and quasi-experimental designs, as detailed in section 3.1.3 above. In order to investigate under which circumstances interventions resulting from certification work and for whom, the review will include qualitative, quantitative or mixed methods studies which collect and analyse primary data from beneficiaries, extension agents or experts, as explained in section 3.1.3. Additionally, we will draw on background programme/project documentation, project completion reports and process evaluations, whenever available and use these to provide background information relevant to included quantitative studies. As discussed in section 1.4, the diversity of CS interventions is also mediated by the variety of forms of implementation (even within a single CS) and the diversity of implementation contexts. For this reason the systematic data collection and extraction on the interventions will be given high priority and incorporated in the coding and moderator analyses. Advocacy research that does not incorporate reliable and relevant factual evidence, and/or does not report methods and study characteristics will be excluded in order to ensure the independence of the literature included, as explained in section 3.1.3. The study design criteria are tailored to the requirements of each core review question.

Examples of eligible primary studies

Examples of eligible studies that have been identified through an initial scoping search are the following in the case of Fairtrade certification:

- Quasi-experimental studies that measure the effect of agricultural certification schemes on agricultural producers and their families using multiple regression models (such as Becchetti and Costantino (2006), who compare three different treatments on a variety of dependent variables related to wellbeing with a control group using Tobit models with relevant control variables, and jointly estimate a treatment equation to control for selection bias arising due to self-selection; Becchetti and Michetti (2010), who similarly use logit models with selected control variables and a simultaneously estimated treatment function to arrive at effect sizes free of selection bias) and Propensity Score Matching (PSM) techniques (for example, Ruben and Zuniga (2011), who compare farmers under three different CS with an 'untreated' control group selected through random sampling in each sub-group and matched on a vector of variables describing household and farm characteristics; Ruben *et al.* (2009), which uses PSM to construct comparisons groups receiving no intervention across a variety of locations and crops to evaluate direct and indirect impacts of certification, but is an example of a badly reported study as it fails to give details of matching techniques or balancing assessments, and does not report any details on sample selection).

The review will thus include studies that control for confounding factors with a comparison, that is, which compare agricultural producers or wage workers receiving one or more relevant intervention, with a control group that receives either no intervention or a different type of intervention. The latter is a likely scenario given the increasing prevalence of multiple certifications for a single SPO holding (see below). Comparison may be in terms of before/after (that is, a time before the introduction of certification), and/or cross-sectional (that is, a group of non-participants or a location where certification has not yet been introduced). Generally, study designs that use different, though comparable research locations, may not always account for all relevant confounding factors, since producers in different locations may be affected by different (unobserved) factors. However, in some cases the only feasible way of assigning the 'treatment' to groups of dispersed producers in areas where certified farmers organisations contain thousands of members is the comparison between 'areas' with certification and comparable 'areas' without certification (as discussed in FTEPR (2014)). As mentioned above, another particularity to bear in mind is that there may also be comparisons between different certification schemes for a given commodity or in a given context or between a specific CS and an alternative development intervention (see Chiputwa *et al.* (2015) on Fairtrade, Utz Certified and organic; Parrish *et al.* (2005) for a comparison of the effects of Fairtrade certification against TechnoServe business development). These study designs may be included as long as there is at least a comparable 'control group', which may be in some cases a group treated with a different certification as pointed out in section 3.1.3. An example identified by early scoping:

- Quasi-experimental studies that compare the effect of different agricultural certification schemes on agricultural producers and their families against uncertified control groups using PSM to control for confounding. For example Chiputwa *et al.* (2015) use PSM with three treatments to compare self-selected groups of farmers in three different certification schemes and one control group.

Data extraction and management

Abstract and full-text screening

The screening of studies for inclusion and the subsequent data extraction (coding) will proceed in three distinct stages. Initially, the titles and abstracts of studies identified during the search process will be screened for relevance. Thereafter, studies found to be of relevance to the review will be downloaded in full text and screened against the sets of inclusion criteria set out in this protocol. Lastly, studies selected for inclusion will be coded according to a detailed coding manual.

The first screening for relevance on titles and abstracts will be done by research assistants working under the oversight of team members, against clearly defined exclusion criteria. The code sheet for this stage will be piloted and we will test for agreement amongst coders. Screening on title and abstract will not be double-coded, but pilot screening until consistency is reached among coders will ensure uniformity. A coding manual for screening on title and abstract will be put in place to support the coders, while ongoing communication with the rest of the review team will ensure that screening consistency is maintained. Moreover, coders are instructed to be over-inclusive at this stage and will work under close supervision by more senior team members. All exclusions will be logged in an electronic form in EPPI Reviewer 4. Studies will generally not be disregarded on methodological grounds at this stage, as a serious assessment of study quality cannot normally be based on abstracts only. Exceptions are studies where the abstract clearly indicates an unsuitable study type or design, for example, book reviews, literature reviews with no primary evidence or advocacy/policy documents, as listed in section 3.1. Reviewers, when in doubt, will choose to include studies at this stage, so as not to lose relevant evidence.

The reports selected for further screening based on their titles and abstracts will then be downloaded as full text into a database to be evaluated in greater detail against the inclusion criteria set out above. We will undertake independent, double-coding at this stage, and the coding sheet will be trialled and refined. A database will be created in EPPI Reviewer 4. Reviewers will complete an electronic form for each study reviewed. The forms will be retained both to ensure transparency and to allow for the creation (and possibly analysis) of an excluded studies table later in the review. The broad range of interventions relevant to the review questions pursued necessitate the inclusion of an equally wide variety of studies in terms of method. Different inclusion criteria will be deployed for both quantitative and qualitative studies.

Studies may be excluded from the review on grounds of study design at this stage (see section above). For each study an inclusion/exclusion checklist will be filled in and retained. These checklists will be more detailed than is perhaps common, as this review is the first systematic review following Campbell criteria to target this particular area of the literature.

Data extraction and coding

All studies that meet the inclusion criteria based on full-text review will then be coded against a detailed code book. Study quality assessment and effect size calculation (where appropriate) will be undertaken after detailed coding. Studies will be independently coded by two reviewers using the EPPI-Reviewer software. We will test for agreement between coders statistically using Cohen's kappa statistic on a random sample of studies. The coding will also be checked and overseen by the PI, who will moderate any disagreements between coders We will develop a detailed coding manual during the course of the review, which has the advantage of allowing us to employ an iterative process. A sample of studies selected for inclusion will be tested against the initial version of the coding manual and lessons will be drawn to optimise the amount of information extracted from studies. Please see Appendix 6 for an outline of our coding manual.

Given that we use different sets of inclusion criteria to assess studies that answer the different review questions, a variety of study designs will be included in the review, not all of which can be coded with the same instruments. In particular, qualitative studies, which are used to address Review Question 2 will require detailed content coding. At minimum, information such as authors, publication date and type, population (including geographical location, crop type, production system and size, gender, age), time period under study, intervention type, intervention process and context, study design, study quality, data collection method, and outcomes (intermediate and endpoint) will be collected.

The coding sheet is designed to capture information for use in moderator analysis. Following Lipsey (2009) we will use extrinsic, methodological and substantive moderators. Extrinsic moderators are ones that relate to aspects of the study that are not directly related to research finding, such as they type of publication it appeared in. Methodological moderators capture the choice of method made by study authors, such as experimental and quasi-experimental study designs. Lastly, substantive moderators relate to the characteristics of the intervention and the population under investigation, such as the gender of participants for instance.

In particular our coding sheet will record information on the following moderator variables:

- **Extrinsic:**
    - Publication type: peer-reviewed or not
    - Funding (for example, whether study commissioned by CS)
- **Methodological:**
    - Research design
    - Characteristics of the control groups
    - Sample size
    - Statistical analysis methods
    - Risk of bias
- **Substantive:**
    - Intervention type
    - Type of CS or type of standards (ethical trading, quality, environmental, and so forth)

- Geographical region / location
- Commodity
- Length of exposure to the intervention
- External assistance: in many cases SPOs are assisted by NGOs
- Implementation type: whether through NGO, private company, directly by CS
- Multiple certification
- Production system: large-scale plantation, smallholder production
- Gender
- Year/period of study

More moderators will no doubt be found over the course of the review, especially as the evidence for Review Question 2 will yield insights into barriers and facilitators, suggesting new contextual moderators.

Measure of effect sizes

For studies that speak to Review Question 1, and are – at least in principle – suitable for inclusion in a meta-analysis, we will also collect information necessary for calculating effect sizes such as the outcome variable estimated, the way the outcome value was calculated, sample size (for example, treatment and control group sizes and attrition rates), sample variance, sampling method, group allocation mechanism, the estimated effect, and associated confidence interval (generally set for $\alpha=0.05$) or the standard error of the estimated effect.

Studies for which an effect size (or several, for different outcome variables) can be calculated, will either be comparing different group means (for example, treatment and control) or will be based on measures of association, as in various forms of regression analysis. Where studies compare group means we will calculate either standardised mean differences (SMDs, for continuous data) or odds ratios (ORs, for dichotomous data). For SMDs, we will generally calculate (corrected) Hedges' *g*, given that sample sizes may be small. Where studies report measures of association, we will extract regression coefficients or similar effects sizes. Group comparison measures can be converted into associative measures and vice versa to make comparisons across effects more convenient (using methods detailed in Ellis (2010) or Borenstein *et al.* (2009)). We will convert all comparison measures to SMDs for the sake of synthesis. For each effect size we will also calculate standard errors and use them to construct a 95 per cent confidence interval to give a measure of the precision of the estimate. In all cluster-allocated studies, we will check whether standard errors have been appropriately widened to correct for unit of analysis errors. Uncorrected unit of analysis errors lead to artificially low p-values and narrower confidence intervals and would increase the weight given to a study during meta-analysis. Where unit of analysis errors are present and have not been corrected, we will perform the necessary correction ourselves. Such errors are most prevalent in cluster-allocated RCTs.

We will perform unit of analysis error correction for any outcome measure that requires it. For instance, for associational measures the adjustment method requires the intra-class correlation co-efficient (ICC) where available to estimate the design effect. The 'design effect' is given as deff=1+ICC(m-1), where m is the mean cluster size. The 'effective sample size' ESS=total sample size/design effect.

Standard errors associated for SMD in cluster studies may be inflated by the square root of the design effect unless where authors have implicitly adjusted for the clustering.

In many cases studies do not report the information necessary to calculate effect sizes and associated standard errors. In such cases we will seek the necessary information directly from the corresponding author(s) of the study in question. Where the corresponding author is unable or unwilling to provide the necessary information, we will calculate response ratios, which are simply the quotient of the change in outcome between the treatment group and the control group.

Quantitative effect estimates will be synthesised using inverse-variance random effects model meta-analysis in Stata 13.0 (StataCorp, 2013). Using the random effects model allows the true effects in each study to vary according to some distribution. That is, we do not assume that the variance component is constant across studies, which given the heterogeneity of contexts would be an unreasonable assumption. Full details are given in the synthesis section below.

Dependence of effect sizes

Dependent effect sizes may occur when a single paper reports the findings of more than one study, when more than one paper reports the findings of a single study, when studies include more than one intervention groups compared to one control group, or when studies report outcomes at more than one point in time. They can also occur where studies report multiple specifications, multiple outcome constructs (or groups thereof) or report results for different subgroups of participants. The Campbell Guidance (Becker *et al.*, 2007) will be followed to ensure that only effect sizes that are statistically independent are included in any one synthesis either of a meta-analysis or in the qualitative synthesis. Where a single study is reported in multiple papers, only the paper providing the most relevant data about the study will be chosen as the 'main' paper while other papers will be considered 'secondary reports' where additional information may be drawn about the study. Where a single paper reports the results of more than one study, this paper will separated into the number of studies reported, where they will be coded and analysed separately. A number of factors will be considered to identify papers that could be linked or separated in this way with source of funding, authorship, and interventions considered in the initial coding process will be used.

We will only include one effect size estimate per study in any given synthesis. And only studies that report similar outcomes will be synthesized together. In selecting or calculating the most relevant effect size estimate we closely follow the approach taken by Waddington *et al.* (2014). Where multiple outcome estimates are reported for the same outcome due to different specifications, we will select the specification that is likely to have the lowest risk of bias. Where outcomes are measured during a time of follow-up, the we will construct a synthetic average effect size prior to synthesis. Where estimates are reported separately for different subgroups of participants we will report data on subgroups separately.

We distinguish on the one hand between certification schemes, which are bundles of interventions under a clear and unique label (Fairtrade, Utz Certified, and so forth), and interventions, which are specific measures required by a certification schemes (for example, the payment of premium). On the other hand we distinguish between studies, which are distinct *research projects* and reports, which are write-ups (either partial or complete) of such research projects. So there may be multiple reports of a single study, either because of different publication formats (for example, a working paper followed by a journal article), or because they report on different outcome measures or aspects of the study. A similar situation arises when multiple reports draw on the same dataset. Conversely, a report may contain data on outcome measures from several studies.

Our coding scheme is designed to capture both instances by recording information on the certification scheme, the intervention, the country and area, and the timeframe of the intervention. Based on this information each report will be given unique report- and study-level identification codes to allow for easy matching of duplicates and the identification of reports which violate independence assumptions. Coding in this manner will also aid the synthesis of results by allowing us to synthesize according to specific interventions (rather than intervention types) or programmes.

A further complication arises when a single intervention in a given location may be covered by multiple reports. While our coding scheme allows us to identify such instances easily, a decision must then be taken on how to arrive at an overall effect size for that specific intervention. Multiple measures of the same outcome within one study will not be synthesised as only independent findings will be allowed for meta-analysis purposes. The paper that reports more relevant data will be chosen as the 'main' paper and the others will be considered 'secondary reports' where additional information will be drawn to complement the 'main' paper. We will seek to use the effect size that is expected to be measured with the highest level of precision, which we will assess by looking at sample size and study design. All such decisions will be clearly documented. Should this not be possible, we will, following Baird *et al.* (2013), we will take a simple average of reported effect sizes.

Risk of bias assessment for included studies

Included study types

This review will synthesise a wide range of studies with different study designs, within the boundaries explained in section 3.1.3. This section discusses some of the methodological challenges that can inform the risk of bias assessment for included studies. We expect included studies to deal with the four key problems below in various ways, which will require consideration of possible risk of bias.

Measuring the effects of CS involves a series of methodological challenges and therefore appropriate study designs are required in order to disentangle and attribute effects on welfare outcomes. First, the presence of a wide range of confounding factors (that is, coexistence of several CS and/or other supporting organisations, like NGOs, as well as producer organisations) causes attribution errors and seriously undermines the validity of measured effects. For example, the

Fairtrade certification for smallholder crops always operates through small producer organisations (SPOs) and very often coexists with organic certification (Barham & Weber, 2012), which means that study designs need to recognise and distinguish between the effects of affiliation to a producer organisation, adoption of certified organic farming and participation in Fairtrade, as Becchetti, Conzo and Pisani (2011) highlight. Moreover, aid agencies and NGOs that support SPOs (through training, access to credit, and so forth) often facilitate agricultural certifications as a value chain up-grading strategy (Beuchelt & Zeller, 2011), thus introducing additional interventions that may not be part of the certification process as an additional confounding factor. For example, an NGO like OXFAM may support a particular smallholder cooperative with a number of interventions that are separate from Fairtrade-related interventions, but which coexist and possibly interact, making impact attributions even more challenging.

Second, self-selection is often present in various instances of CS. A certification scheme that opens a lucrative market because of the imposition of strict quality criteria without any substantial direct intervention, apart from the certifying and auditing processes themselves, may induce some farmers to adopt practices that enhance quality and therefore place producers in a more competitive position within a value chain. In these cases self-selection occurs almost by definition as these standards are voluntary and farmers either opt for certification or not. Personal preferences or social pressure may influence producers to adopt a certification associated with a specific farming system (that is, organic, shade-grown, and so forth) or to join an already certified producer organisation, hence selecting themselves into the "treatment" group. Any form of self-selection into CS related interventions, whether at individual or collective level, poses serious threats to validity, since subsequent higher performance of certified producers compared to non-certified ones may be driven by ex-ante differences between the two groups, which are correlated both with certification and good performance (Becchetti, Castriota & Solferino, 2011). Nonetheless, self-selection into CS should not simply be assumed. In fact, in some circumstances, certification happens at group level and may be driven by the supplier of the certification, as, for example, in some Fairtrade schemes, where companies that help farmers obtain the certification (for example, Twin Trading) may drive the selection of certain, already existing, producer organisations rather than these organisations selecting themselves. In such cases, many if not most, members of these organisations may be unaware that their leadership is negotiating a certification and may not be aware of having a particular certification as members of the group, especially in organisations (cooperatives) that include thousands of members. In these cases, finding an appropriate ‘control group’ may be particularly hard if an entire geographical area is affected by the certification and there are no uncertified producers in precisely the same area. Finally, in the case of labour standards in ethical trade the certification is usually provided to the employer, so the beneficiaries, (that is, the wage workers in a plantation), do not ‘select-themselves’ but tend to be passive recipients of the potential benefits of the certification process, at least for the initial interventions leading to certification, which is voluntarily adopted by the employer (see for instance Ruben and van Schendel (2008) for banana workers in Fairtrade certified plantations; and FTEPR (2014) for flower plantation workers).

Third, unintended effects, such as spill-overs and drop outs, resulting from intended strategies can always be present in social processes (Balogun & Johnson, 2005) and study designs should control for those when measuring effects. Positive or negative effects can occur without being part of the design of an intervention (Rossi *et al.,* 2004), and affect persons or organisations that did not participate in the intervention (Berk & Rossi, 1999; Roche, 1999). Spill-over effects can undermine the validity of effect estimation, if, for instance, comparisons are made between certified and non-certified producers drawn from the same community, as there are cases where certification effects can 'contaminate' non-certified producers (that is, through increase of demand for hired labour for more labour intensive certification schemes, or investments in community projects that benefit non-certified producers as well). Drop-out effects, on the other hand, can occur due to unintended negative certification effects on participants. Not including ex-participants can lead to impact overestimations if drop-outs are systematic (Alexander-Tedeschi & Karlan, 2010), as negative effects which made participants abandon the certification scheme are being ignored.

Fourth, the benefits of certification, especially when occurring at group level (for example, a producer organisation) may not be equally distributed. A study design that estimates an average treatment effect by comparing across intervened groups and control groups may not be able to capture the unequal distribution of benefits within clusters. We hope, however, that some quantitative studies as well as qualitative studies will be designed to account for potential unequal distribution of benefits, as interventions involving organised groups are sometimes subject to elite capture (Pan & Christiaensen, 2012).

Quality assessment

We will assess the methodological quality of the included studies focusing on their design, implementation and reporting. We will apply screening questions to identify potential sources of bias, and to determine whether a particular bias was controlled for in the conduct and reporting of the study. We will distinguish between studies that report on the effects of CS and use experimental and quasi-experimental designs and studies that examine the circumstances under which such interventions have intended or unintended effects, which can use qualitative, quantitative or mixed methods. Each study will be assessed independently by two reviewers, including a specialist in the methods used by each study (quantitative/qualitative) using a detailed risk of bias assessment tool adapted to the study methods and the context of agricultural CS. Each key dimension of all included studies will be assessed and scored as "high risk of bias", "low risk of bias" or "unclear" if sufficient information is not available to make a clear judgment. In the case of lack of information or uncertainty over the potential for bias, authors will be contacted to clarify aspects of the research. If clarifications are not obtained or uncertainty remains after clarifications, these studies will be considered as "unclear risk of bias". Unresolved disagreements between the two reviewers will be arbitrated by a third reviewer. The results of the assessment of risk of bias along with a detailed explanation, will be reported in the review. Following guidelines by Deeks *et al.* (2011) and Hannes (2011) for quantitative and qualitative research, we will conduct sensitivity analysis according to the risk of bias of each reported outcome to assess how sensitive our findings are to the inclusion of evidence of different quality.

Assessment of risk of bias in included studies of effects

Risk of bias of included studies of effects will be assessed examining the following seven domains of potential bias:

a. Selection bias and baseline confounding
b. Group equivalence: was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding
c. Hawthorne and John Henry effects: was the process of being observed causing motivation bias.
d. Spill-over effect bias: assesses if the study was adequately protected against performance bias
e. Selective outcome reporting bias: assess if the study was free biases arising from outcome reporting
f. Selective analysis reporting bias: assessing whether the study was free from analysis reporting bias
g. Other biases: assesses if the study is free from other biases introduced in the design and conduct of the study

We will combine and adapt to the needs of the review the assessment tool developed by Duvendack *et al.* (2011) (Appendix 1) and the risk of bias signaling questions developed by Jorge Hombrados and Hugh Waddington[11] (Appendix 3), as applied in Vaessen *et al.* (2014). The assessment results will be presented in a 'Risk of bias' table (Appendix 4) where each entry will address a specific domain of potential bias of the study. This table will serve as a transparent decision-rule to present the potential risks of bias for each study against each domain of potential bias. This information will be used in the synthesis to relate findings with the quality of the evidence using the GRADE approach (Deeks *et al.,* 2011).

Assessment of rigour of included studies examining the circumstances under and the reason for which CS are effective

There are no universally accepted standards for assessing the quality of qualitative research, as Vaessen *et al.* (2012) state. In our attempt to assess the quality of studies examining the circumstances under and the reasons for which CS have intended or unintended effects, we will adapt the appraisal tool developed by Waddington *et al.* (2012), based on CASP (2006), to the needs of this review. We will first assess the relevance, clarity of aims and appropriateness of methodology of the studies. Studies which do not satisfy this first filter will be discarded at this stage and we will not continue with the quality appraisal of these studies.

The remaining studies will be assessed against each of the following dimensions:

[11] The signalling questions were developed by Jorge Hombrados and Hugh Waddington, drawing on existing tools, in particular EPOC (n.d.) 'Suggested risk of bias criteria for EPOC reviews'; Coalition for Evidence-Based Policy (2010) 'Checklist for reviewing a randomised controlled trial of a social programme or project, to assess whether it produced valid evidence'; and Deeks *et al.* (2011).

1. Adequacy of research methods and design: clear link to theoretical framework, adequacy of design, adequacy and justification of research site (that is, use of focus groups and/or semi-structured interviews and/or observation, as well as tools for site selection), adequacy of sampling and data collection strategies, soundness of data analysis, triangulation of data, clarity of analysis and conclusions, consideration of conflicts of interest, researcher's bias reflexivity and other ethical issues.
2. Adequacy of reporting: description of context, description of sample and sampling procedures, clarity of data collection process, clarity of methods of analysis.

We will develop the screening questions drawing on the check list included in Appendix 5. Screening questions will be adapted to the CS context and those considered to be of major importance will be weighted applying a factor greater than 1.

DATA SYNTHESIS AND ANALYSIS

The review will synthesise quantitative information on effectiveness to assess the certification scheme interventions aimed at socio-economic outcomes (Review Question 1). Qualitative data and mixed methods studies that do not meet the inclusion criteria for inclusion in the quantitative synthesis will be used to assess Review Question 2 to explore unintended outcomes of interventions and to understand the barriers/facilitators to such certification's effects. All studies will be coded by using a detailed coding sheet for extracting information that describes, amongst other things, the CS method type, target population, type of intervention, scale of intervention, outcomes, the control groups, how the outcomes were measured, location/country, and the timeframe described. Additionally, the coding sheet will be used to extract data that allow us to calculate effect sizes, including sample sizes, means, standard deviations, confidence intervals, and dropout rates for both intervention and control arms. Coding all studies for the CS they cover, the specific intervention(s), the country/area and timeframe of the intervention(s) will allow us to match information pertaining to the same CS and the same intervention.

Review Question 1: Effectiveness review

Effect sizes will be synthesised using random effects meta-analysis and will be reported in detailed forest plots, complete with appropriate confidence intervals. The meta-analysis will be undertaken using Stata 13. Random effects meta-analysis does not assume that there is a single 'true' effect size to be estimated across studies, but rather allows for a variety of underlying effect sizes. In the random effects model the total observed variance can be broken down into within-study variance and (estimates of) between-study variance. Given the real heterogeneity of study contexts and interventions, the random effects model appears appropriate. The random effects model weights each study by the inverse of its variance (which captures sample size) and by an estimate of the between-studies variance component (T^2), thereby producing wider confidence intervals[12]. A fixed effect synthesis would produce artificially narrow confidence intervals (Borenstein *et al.*, 2009).

[12] Using notation from Borenstein *et al.* (2009) T^2 is the estimated value of the real between-study variance τ^2.

Multiple pooled effect sizes

The question of when studies can legitimately be combined to produce pooled effects is a question every systematic review must consider (Petticrew & Roberts, 2006), but it is particularly pertinent in this review due to the wide variety of interventions and certification types under study. The studies we analyse are heterogeneous in terms of methods, location, timing, the types interventions covered, the population studied and the outcome measures used. It is therefore likely that a single overall pooled effect size covering all included studies will be of limited explanatory value, making it less useful for policy makers and other readers. Such a pooled effect, and the spurious precision it implies, would lack a clear interpretation and therefore lend itself to mis- and over-interpretation. Instead, we will use the causal chain given by intermediate and endpoint outcomes to structure our syntheses and conduct separate syntheses for each of the primary outcomes included in the review.

In all cases, only studies that report similar outcomes will be synthesised together. Heterogeneity will be examined through moderator and sensitivity analysis. In order to maintain transparency we will report in a detailed way which outcomes were considered similar and grouped in the same synthesis.

We will assess the reliability of the calculated pooled effects by conducting power calculations based on the weighted mean effect and seeing how many studies have at least a 50 per cent of finding effect of that size (Ellis, 2010).

Assessment of heterogeneity and subgroup analysis

As the pooled effect size by itself is likely to of limited policy value we will thoroughly investigate the expected heterogeneity in our findings through detailed moderator analysis. Initially, we

will assess heterogeneity of the different outcome syntheses statistically, by computing the I^2 statistic, which is a measure of the proportion of observed variance that is due to real differences in effect seizes, that is a measure of the inconsistency across studies. We will also report the estimate of the between-studies variance component T^2 (Borenstein *et al.*, 2009).

Substantive heterogeneity, indicated by high values for I^2, will be further explored using subgroup analysis, or meta-regression, where sufficient data is available. Moreover, given how unreliable I^2 is as an indicator, we will also use subgroup analysis when we have substantive reasons to believe it to be important, even if values for are I^2 low. Particularly useful for readers will be an exploration of heterogeneity across intervention and CS types, which will be explored through subgroup analysis, and the effects of various moderator variables. Where possible, we will investigate three types of moderators as stipulated by Lipsey (2009) to include extrinsic, methodological, and substantive moderators (see section 3.3.2.2 for more details). In particular we will explore whether findings differ by key contextual factors, such as crop/product type, length of exposure to programme, geographical region/location, external assistance received, and so forth

Meta-regression, similar to conventional multiple regression, has the advantage that multiple moderator variables can entered into the model and different specifications can be used to test for the effect of different moderators, while holding all others constant. However, meta-regression requires sufficient data. Sparse cell sizes can yield unreliable estimates, therefore the constraining the use of meta-regression should not enough data be available. Whether or not we employ meta-regression or rely on ANOVA-type methods (that is, pairwise comparisons) will depend on the amount of data the review finds for each outcome.

Sensitivity analysis

We will conduct a full sensitivity analysis for each separate synthesis, especially by testing for the effects of study quality by analysing effect sizes according to different categories of risk of bias. We will also test for the effects of excluding outliers, but no outlier will be excluded on statistical reasons alone.

As detailed in our search strategy, we will attempt to minimise the possible influence of publication bias on pooled effect sizes. However, we are aware that studies may nonetheless be missing because of existing biases acting against the write up, reporting and publication of unfavourable findings. We will address publication bias by conducting an extensive search, and also using statistical approaches that assess the funnel plots (Egger, 1997; Palmer *et al.*, 2008). We will also try to assess the pooled effect sizes of published and unpublished studies separately and test whether differences are significant (Ellis, 2010).

Review Question 2: Barriers, facilitators and beyond

We will assess under what circumstances and why interventions have intended and unintended effects and what the barriers and facilitators of those effects are. As stated in the section on Review Question 1 above, the point of this review is not to provide a single answer, along the lines of: ‘do certification schemes for agricultural commodities work?’ Rather we seek to understand the complex and interrelated set of questions around what works when, for whom, for which products, in what kind of market environment, and whether benefits are equally shared amongst different socio-economic groups.

We will begin by looking for and synthesising evidence of barriers and facilitators to effectiveness along the causal chains suggested by the different theories of change of the various interventions we analyse. However, given the field of literature we are analysing, stopping here would mean disregarding large quantities of qualitative studies not directly related to questions of effectiveness. As we have discussed above, we will also include evidence on the unintended consequences, and barriers and facilitators of effectiveness for interventions that have not been included in response to Review Question 1. As a result, our narrative synthesis will have three parts: one devoted to evidence on the interventions included in Review Question 1; one detailing other evidence; and finally, an overview of both sets of evidence.

We will employ thematic narrative synthesis methods (Thomas *et al.*, 2008) of studies containing qualitative data to support a fuller understanding of the theory of change underlying CS and the potential barriers and facilitators of implementation and engagement in CS. Detailed evidence tables will be prepared to describe: the methodological quality of each study; details of the CS examined; study site/population; and full findings. Two reviewers will read and re-read data contained within the evidence tables, apply thematic codes and to capture the content of the data, and then group and organise codes into higher-order themes. These themes will be used to generate narrative that addresses issues relevant to the delivery and process of CS in low- and middle-income countries.

Integrated synthesis

To be useful to a wide readership the quantitative results of systematic review must be carefully interpreted. This means not only providing readers with substantive interpretations of mean effect sizes, but also placing these findings in the complex networks of qualitative understanding that exist around the subject. As we are conducting separate syntheses of evidence for the different outcomes under Review Question 1, we will integrate the results of our reviews of questions one and two for each outcome to deliver an overall assessment of what works when and for whom, and also to point to the gaps in existing knowledge. The integrated synthesis will be structured around the synthetic TOC developed in the course of the review and will also seek to address the validity of the different TOCs of different intervention types. Any resulting implications for the validity of the different TOCs will be reported in order to facilitate future improvements.

In addition to this we will also use information from Review Question 2 not tied to an intervention under Review Question 1 to try and draw further lessons for the design and implementation of certification schemes. It is expected that a lot of the information gathered under Review Question 2 will not be directly matched to included data for review question one, as there may not be many papers that many our inclusion criteria for Review Question 1. However to exclude studies from Review Question 2 simply because they cannot be matched to *included* studies under Review Question 1 would lead to an unacceptable loss of information.

Our findings will be integrated into a summary of findings table to elaborate to readers how results were used to form conclusions. The GRADE tool will be applied in our synthesis to enable transparent and structured interpretation of results (Guyatt, 2011).

REFERENCES

Alexander-Tedeschi, G. & Karlan, D. (2007). Cross sectional impact analysis: bias from dropouts. *Mimeo*. The Financial Access Initiative, New York.

Alsop, R. & Heinsohn, N. (2005). Measuring Empowerment in Practice: Structuring Analysis and Framing Indicators. *World Bank Policy Research Working Paper* n. 3510. Washington DC: World Bank.

Ayuya, O. I., Gido, E. O., Bett, H. K., Lagat, J. K., Kahi, A. K., & Bauer, S. (2015). Effect of certified organic production systems on poverty among smallholder farmers: empirical evidence from Kenya. *World Development*, 67, 27-37.

Bacon, C. M. (2010). A spot of coffee in crisis Nicaraguan smallholder cooperatives, fair trade networks, and gendered empowerment. *Latin American Perspectives*, 37(2), 50-71.

Balogun, J., & Johnson, G. (2005). From intended strategies to unintended outcomes: The impact of change recipient sensemaking. *Organization Studies*, 26(11), 1573-1601.

Barham, B. L., & Weber, J. G. (2012). The economic sustainability of certified coffee: recent evidence from Mexico and Peru. *World Development*, 40(6), 1269-1279.

Barratt Brown, M. (1993). *Fair Trade*. London: Zed Books.

Barrett, C. B., Reardon, T., & Webb, P. (2001). Nonfarm income diversification and household livelihood strategies in rural Africa: concepts, dynamics, and policy implications. *Food Policy*, 26(4), 315–331.

Barrientos, S. (2000). Ethical trade and globalisation: assessing the implications for development. *Journal of International Development*, 12(4), 559-570.

Barrientos, S. (2003). *Labour Impact Assessment: Challenges and Opportunities of a Learning Approach*. Presented at the EDIAIS Conference University Of Manchester. Retrieved from http://www.sed.manchester.ac.uk/research/iarc/ediais/pdf/EINJuly04.pdf

Barrientos, S., Dolan, C., & Tallontire, A. (2003). A gendered value chain approach to codes of conduct in African horticulture. *World Development*, 31(9), 1511–1526.

Becchetti, L., Castriota, S., & Solferino, N. (2011). Development projects and life satisfaction: An impact study on Fair Trade handicraft producers. *Journal of Happiness Studies*, 12(1), 115-138.

Becchetti, L., & Costantino, M. (2006). *The effects of Fair Trade on marginalised producers: an impact analysis on Kenyan farmers*. Palma de Mallorca: European Center for Studies on Income Inequality (ECINEQ).

Becchetti, L., Conzo, P., & Pisani, F. (2011). Virtuous interactions in removing exclusion: the link between foreign market access and access to education. *Journal of Development Studies*, 47(9), 1431-1454.

Becchetti, L., & Michetti, M. (2010). When Fair Trade generates social capital by creating room for manoeuvre for pro-poor policies. *African Journal of Business Management*, 4(14), 2903-2914.

Bennett, M., & Franzel, S. (2013). Can organic and resource-conserving agriculture improve livelihoods? A synthesis. *International Journal of Agricultural Sustainability*, 11(3), 193-215.

Bert, R.A., & Rossi, P. H. (1999). *Thinking about program evaluation*. Washington D.C.: Sage Publications

Beuchelt, T. D., & Zeller, M. (2011). Profits and poverty: certification's troubled link for Nicaragua's organic and Fairtrade coffee producers. *Ecological Economics*, 70(7), 1316-1324.

Blackman, A., & Rivera, J. (2010). *The evidence base for environmental and socioeconomic impacts of "sustainable" certification*. Washington D.C.: Resources for the Future.

Blowfield, M. (1999). Ethical trade: a review of developments and issues. *Third World Quarterly*, 20(4), 753- 770.

Bolwig, S., Gibbon, P., & Jones, S. (2009). The economics of smallholder organic contract farming in tropical Africa. *World Development*, 37(6), 1094-1104.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

Chan, M.K. & Pound, B. (2009). *Final report: literature review of sustainability standards and their poverty impact*. NRI, London.

Chiputwa, B., Spielman, D. J., & Qaim, M. (2015). Food standards, certification, and poverty among coffee farmers in Uganda. *World Development*, 66, 400-412.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin*, 70(4), 213–220.

Critical Appraisal Skills Programme (CASP). (2013). *10 questions to help you make sense of qualitative research*. Public Health Resource Unit: England. Retrieved from: www.phru.nhs.uk/Doc_Links/Qualitative%20Appraisal%20Tool.pdf

Daviron, B., & Ponte, S. (2005). The coffee paradox: global markets, commodity trade and the elusive promise of development. New York: Zed Books.

Deeks, J.,J., Higgins, J., P., T., Altman, D., G., 2011. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins, J. P. T., & Green, S. (eds) *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]. Retrieved from: handbook.cochrane.org.

Doherty, B., & Tranchell, S. (2005). New thinking in international trade? A case study of The Day Chocolate Company. *Sustainable Development*, 13(3), 166-176.

Dragusanu, R., & Nunn, N. (2014). The impacts of Fair Trade certification: evidence from coffee producers in Costa Rica (Preliminary and Incomplete). *Mimeo*. Harvard University.

Duvendack, M., Palmer-Jones, R., Copestake, J., G., Hooper, L., Loke, Y., & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people*? London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Effective Practice and Organisation of Care Group (EPOC). (n.d.). *Suggested risk of bias criteria for EPOC reviews*. Retrieved from http://epocoslo.cochrane.org/epoc-specific-resources-review-authors.

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629-634.

Ellis, P. D. (2010) *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*, Cambridge: Cambridge University Press.

Fairtrade, Employment and Poverty Reduction team (FTEPR), 2014. *Fairtrade, Employment and Poverty Reduction in Ethiopia and Uganda*. Final Report to DFID, April 2014. London: SOAS, University of London. Retrieved from http://ftepr.org/publications/.

Fairtrade Foundation. (2014). *Annual impact report 2013-2014*. London: Fairtrade Foundation.

Fairtrade International. (2014). *Annual report 2013-2014*. Bonn: Fairtrade International.

Fairtrade International (2015). *Monitoring the scope and benefits of Fairtrade Sixth Edition, 2014*. Bonn : Fairtrade International.

Gibbon, P., & Ponte, S. (2005). *Trading down: Africa, value chains, and the global economy*. Philadelphia PA: Temple University Press.

Greenberg, S. (2004). *Women workers in wine and deciduous fruit global value chains*. Summary report submitted on behalf of Women on Farms Project. Stellenbosch.

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *British Medical Journal*, 331(7524), 1064-1065.

Guyatt, G., Oxman, A., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., deBeer H., Jaeschke R., Rind D., Meerpohl J., Dahm, P., & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction to GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64, 383-394.

Hammerstrøm, K., Wade, A., & Klint Jørgensen, A.-M. (2010). *Searching for studies: a guide to information retrieval for Campbell Systematic Reviews*. Campbell Collaboration: Oslo.

Hannes, K. (2011). Chapter 4: Critical appraisal of qualitative research. In: Noyes, J., Booth, A., Hannes, K., Harden, A., Harris, J., Lewin, S., & Lockwood, C. (eds) *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions*. Version 1 (updated August 2011). Cochrane Collaboration Qualitative Methods Group. Retrieved from: http://cqrmg.cochrane.org/supplemental-handbook-guidance.

Hansen, H., & Trifković, N. (2014). Food Standards are good–for middle-class farmers. *World Development*, 56, 226-242.

Henson, S., & Humphrey, J. (2010). Understanding the complexities of private standards in global agri-food chains as they impact developing countries. *Journal of Development Studies*, 46(9), 1628-1646.

Henson, S., & Jaffee, S. (2008). Understanding developing country strategic responses to the enhancement of food safety standards. *The World Economy*, 31(4), 548-568.

Hurst, P., Termine, P., & Karl, M. (2005). *Agricultural workers and their contribution to sustainable agriculture and rural development*. Geneva: Food and Agriculture Organisation (FAO); International Labour Organisation (ILO); International Union of Food, Agricultural, Hotel, Restaurant, Catering, Tobacco and Allied Workers' Associations (IUF).

International Trade Centre. (2011). *The impacts of private standards on producers in developing countries*. Literature review series on the impacts of private standards: part II. Geneva: ITC.

Jaffee, S., & Henson, S. (2004). *Standards and agro-food exports from developing countries: rebalancing the debate*. Washington, DC: World Bank.

Jayne, T. S., Mather, D., & Mghenyi, E. (2010). Principal challenges confronting smallholder agriculture in sub-Saharan Africa. *World Development*, 38(10), 1384–1398.

Kabeer, N. (2001). Conflicts over credit: re-evaluating the empowerment potential of loans to women in rural Bangladesh. *World Development*, 29(1), 63-84.

Kolk, A. (2005). Corporate social responsibility in the coffee sector: the dynamics of MNC responses and code development. *European Management Journal*, 23(2), 228–236.

Krier, J. M. (2008). *Fair Trade 2007: New facts and figures from an ongoing success story*. Survey prepared on behalf of the Dutch Association of Worldshops (DAWS). Colemborg: DAWS.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In: Cooper, H., Hedges, L. V., & Valentine, J. C. (eds). *The handbook of research synthesis and meta-analysis*, New York: Sage Publications.

Maertens, M., & Swinnen, J. F. (2009). Trade, standards, and poverty: evidence from Senegal. *World Development*, 37(1), 161-178.

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, 4(3), 445–55.

McCulloch, N., Winters, L. A., & Cirera, X. (2001). *Trade liberalization and poverty: a handbook.* London: UK Department for International Development and the Centre for Economic Policy Research.

Mezzadri, Alessandra (2014) Backshoring, local sweatshop regimes and CSR in India. *Competition and Change*, 18(4). 327-344.

Muradian, R., & Pelupessy, W. (2005). Governing the coffee chain: the role of voluntary regulatory systems. *World Development*, 33(12), 2029–2044.

Nelson, V., & Pound, B. (2009). *The last ten years: a comprehensive review of the literature on the impact of Fairtrade*. London: Fairtrade Foundation.

Nicholls, A., & Opal, C. (2004). *Fair trade – market-driven ethical consumption*. London: Sage Publications.

Palmer, T., Peters, J., Sutton, A. J., & Moreno, S. (2008). Contour enhanced funnel plots for meta-analysis. *The Stata Journal*, 8, 242-254.

Pan, L., & Christiaensen, L. (2012). Who is vouching for the input voucher? Decentralized targeting and elite capture in Tanzania. *World Development*, 40(8), 1619-1633.

Parrish, B. D., Luzadis, V. A., & Bentley, W. R. (2005). What Tanzania's coffee farmers can teach the world: a performance-based look at the fair trade–free trade debate. *Sustainable Development*, 13(3), 177-189.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: a practical guide*. Hoboken NJ: Wiley-Blackwell.

Potts, J. Lynch, M., Wilkings, A., Huppé, G., Cunningham, M., & Voora, V. (2014). *The State of Sustainability Initiatives Review 2014*. International Institute for Sustainable Development (IISD) and the International Institute for Environment and Development (IIED).

ProForest. (2005). *Developing a mechanism for palm oil traceability from plantation to end user*. Discussion paper for RT3. Oxford: ProForest.

Raynolds, L. T. (2000). Re-embedding global agriculture: the international organic and fair trade movements. *Agriculture and human values*, 17(3), 297-309.

Ronchi, L. (2002). *The impact of Fair Trade on producers and their organizations: A case study with Coocafé in Costa Rica*. Policy Research Unit. Sussex: University of Sussex.

Rossi, P. H., Lipsey, M. W., & Freeman, H. (2004). *Evaluation: a systematic approach*. Washington, DC: Sage Publications.

Ruben, R. (2013). *Critical review of some recently published Fair Trade impact studies: Deficient study design and little robust evidence*. Radboud University Nijmegen. Retrieved from: http://www.ru.nl/cidin/@883941/pagina/.

Ruben, R., Fort, R., & Zúñiga-Arias, G. (2009). Measuring the impact of Fair Trade on development. *Development in Practice*, 19(6), 777–788.

Ruben, R., & van Schendel, L. (2008). The impact of Fair Trade in banana plantations in Ghana: Income, ownership and livelihoods of banana workers. In: Ruben, R. (ed) *The impact of Fair Trade*. Wageningen, The Netherlands: Wageningen Academic Publishers.

Ruben, R., & Zúñiga, G. (2011). How Standards compete: comparative impact of coffee certification schemes in northern Nicaragua. *Supply Chain Management: An International Journal*, 16(2), 98–109.

Sandieson, R. (2006). Pathfinding in the research forest: The pearl harvesting method for effective information retrieval. *Education and Training in Developmental Disabilities*, 41(4), 401-409.

Sender, J. (2003). Rural poverty and gender: analytical frameworks and policy proposals. In: Ha-Joon Chang (ed). *Rethinking Development Economics*. London: Anthem Press.

Schuster, M., & Maertens, M. (2015). The impact of private food standards on developing countries' export performance: an analysis of asparagus firms in Peru. *World Development*, 66, 208–221.

Snilstveit, B. (2012). Systematic reviews: from 'bare bones' reviews to policy relevance. *Journal of Development Effectiveness*, 4(3), 388-408.

StataCorp (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

Terstappen, V., Hanson, L., & McLaughlin, D. (2013). Gender, health, labor, and inequities: a review of the fair and alternative trade literature. *Agriculture and Human Values*, 30(1), 21-39.

Thomas, J., Brunton, J., & Graziosi, S. (2010). EPPI-Reviewer 4.0: software for research synthesis. EPPICentre software. London: Social Science Research Unit, Institute of Education.

Thomas J., & Harden, A. (2008) Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(45).

Ton, G., Vellema, S., & Ge, L. (2014). The Triviality of Measuring Ultimate Outcomes: Acknowledging the Span of Direct Influence. *IDS Bulletin*, 45(6), 37-48.

Tropical Commodity Coalition. (2012). *Coffee Barometer 2012*. teacoffeecocoa.org. Retrieved on April 28, 2013, from http://issuu.com/spanhuysen/docs/ttc_coffeebarometer.

Vaessen, J., Rivas, A., Duvendack, M., Palmer Jones, R., Leeuw, F. L., Van Gils, G., Lukach, R., Holvoet, N., Bastiaensen, J., Hombrados, J. G., & Waddington, H. (2014). The effects of microcredit on women's control over household spending in developing countries: a systematic review and meta-analysis. *Campbell Systematic Reviews*, 2014(8).

Vagneron, I., & Roquigny, S. (2010). *What do we really know about the impact of fair trade? A synthesis*, Paris: PFCE.

Valkila, J., & Nygren, A. (2009). Impacts of Fair Trade certification on coffee farmers, cooperatives, and laborers in Nicaragua. *Agriculture and Human Values*, 27(3), 321–333.

Von Hagen, O., Manning, S., & Reinecke, J. (2010). Sustainable sourcing in the food industry: global challenges and practices. *Moderne Ernaehrung Heute*, Official Journal of the Food Chemistry Institute of the Association of the German Confectionery Industry, 4, 1-9.

Waddington, H., Snilstveit, B., Hombrados, J., Vojtkova, M., Phillips, D., Davies, P., & White, H. (2014) Farmer Field Schools for Improving Farming Practices and Farmer Outcomes: A Systematic Review. *Campbell Systematic Reviews*, 2014(6).

Weitzman, H. (2006). The bitter cost of "Fair Trade" coffee. *Financial Times*, 8, 50–100. Retrieved from: http://tomweston.net/bittercost.pdf.

Winters, L. A. (2002). Trade liberalisation and poverty: what are the links?. *World Economy*, 25, 1339–1367.

Woubie, A. A., Muradian, R., & Ruben, R. (2015) Impact of multiple certification on smallholder coffee farmers' livelihoods: evidence from southern Ethiopia. In: Ruben, R, & Hoebink, P. (eds). *Coffee certification in East Africa: impact on farms, families and cooperatives*. Wageningen: Wageningen Academic Publishers.

Declarations

REVIEW AUTHORS

Lead review author: The lead author is the person who develops and co-ordinates the review team, discusses and assigns roles for individual members of the review team, liaises with the editorial base and takes responsibility for the on-going updates of the review.

Name:	Carlos Oya (Lead PI)
Title:	Reader in Political Economy of Development
Affiliation:	SOAS, University of London
Address:	Department of Development Studies, SOAS, University of London Thornhaugh St, Russell Square
City, State, Province or County:	London
Postal Code:	WC1H 0XG
Country:	United Kingdom
Phone:	+44 207 898 4566
Email:	co2@soas.ac.uk

Co-author(s):

Name:	Deborah Johnston (PI)
Title:	Reader in Development Economics
Affiliation:	SOAS, University of London
Address:	Department of Economics, SOAS, University of London Thornhaugh St, Russell Square
City, State, Province or County:	London
Postal Code:	WC1H 0XG
Country:	United Kingdom
Phone:	+44 207 898 4494
Email:	dj3@soas.ac.uk

Name:	Dafni Skalidou
Title:	PhD candidate and freelance consultant
Affiliation:	School of International Development/ University of East Anglia
Address:	Norwich Research Park
City, State, Province or County:	Norwich
Postal Code:	NR4 7TJ
Country:	United Kingdom
Phone:	
Email:	D.Skalidou@uea.ac.uk, dskalidou@yahoo.gr

Name:	Florian Schaefer
Title:	PhD candidate and Research Officer
Affiliation:	SOAS, University of London
Address:	Department of Development Studies, SOAS, University of London Thornhaugh St, Russell Square
City, State, Province or County:	London
Postal Code:	WC1H 0XG
Country:	United Kingdom
Phone:	
Email:	fs12@soas.ac.uk

Name:	Evans Muchiri
Title:	Researcher
Affiliation:	Centre for Anthropological Research
Address:	House 10, Research Village, Bunting Road Campus, University of Johannesburg
City, State, Province or County:	Johannesburg
Postal Code:	TBC
Country:	South Africa
Phone:	
Email:	evanmuchiri@gmail.com

Name:	Claire Stansfield
Title:	Information officer
Affiliation:	EPPI-Centre, Social Science Research Unit, Institute of Education

Address:	Institute of Education, University of London 20 Bedford Way
City, State, Province or County:	London
Postal Code:	WC1H 0AL
Country:	United Kingdom
Phone:	+44 (0)20 7612 6816
Email:	c.stansfield@ioe.ac.uk

Name:	Kelly Dickson
Title:	Research Officer
Affiliation:	EPPI-Centre, Social Science Research Unit, Institute of Education
Address:	EPPI-Centre, Social Science Research Unit, Institute of Education
City, State, Province or County:	Institute of Education, University of London
Postal Code:	20 Bedford Way
Country:	London
Phone:	+44 (0)20 7612 6127
Email:	k.dickson@ioe.ac.uk

ROLES AND RESPONSIBILITIES

The protocol was developed by Carlos Oya (CO) and Florian Schaefer (FS) with inputs from Deborah Johnston (DJ), Dafni Skalidou (DS), Evans Muchiri (EM), Claire Stansfield (CS) and Kelly Dickson (KD). CO is the lead principal investigator and DJ is co-PI. DS, EM and FS will conduct the information retrieval, with the help of research assistants. CS will provide expert advice on information retrieval and the process will be coordinated by DS. KD and relevant staff at EPPI-Centre will provide capacity building on relevant software. Screening decisions will be made by DS, EM and FS with oversight from CO and DJ, who will resolve any conflicts. Initial screening for relevance will also be undertaken by research assistants. Qualitative study coding will be undertaken by DS, EM and FS with help from research assistants and inputs from CO and DJ. Critical appraisal will be undertaken by DS, EM and FS, and cross-checked by CO and DJ. EM and FS will calculate effect sizes, with inputs from CO. CO and DJ will lead on final report writing, and will provide oversight and leadership throughout.

POTENTIAL CONFLICTS OF INTEREST

Carlos Oya and Deborah Johnston were investigators in a DFID-funded research project which, among other aims, assessed the effects of Fairtrade certification on wages and work conditions of workers employed by a range of agricultural producers, including smallholder farmers, as the first study that collected data on wage workers employed by smallholder members of Fairtrade-certified organisations (see FTEPR, 2014). A conventional literature review had been conducted for this project. The project ended on 31st March 2014 and the report published on 23rd May 2014. This participation, which was in the form of independent academic research, does not in any way affect the impartiality of the researchers involved. We see this SR as another step towards more independent research in this field. Moreover, any primary studies in which the PIs and other team members have been involved will be coded by *other* team members.

Dafni Skalidou has worked with Fairtrade organisations in Spain and South America in the past, but is no longer professionally related to any of them. She is currently doing her doctoral research on the impact of Fair Trade on cocoa farmers and banana plantation workers in Ghana. Her work is funded by the University of East Anglia and is totally independent from any FT organisation. Dafni has also worked with the 3ie-Systematic Reviews team in the past, however, her working contract was finalised in September 2013.

FUNDING

This systematic review is supported by the International Initiative for Impact Evaluation (3ie), as part of their Systematic Review Grants (Call 6).

REQUEST SUPPORT

No support is requested at this time.

PRELIMINARY TIMEFRAME

The deadline for submission of the final report is being revised following delays during the protocol revision process. The scheduled date for submission of a first draft of the review findings is now the 31st March 2016, with a final deadline scheduled for 30th June 2016. These dates are provisional and subject to final approval.

PLANS FOR UPDATING THE REVIEW

The experimental and quasi-experimental research in this area is very limited and the rate of publication of new high quality studies is likely to be slow. We will keep abreast of the literature in the field and update the review once sufficient high quality studies become available.

AUTHOR DECLARATION

Authors' responsibilities

By completing this form, you accept responsibility for preparing, maintaining and updating the review in accordance with Campbell Collaboration policy. The Campbell Collaboration will provide as much support as possible to assist with the preparation of the review.

A draft review must be submitted to the relevant Coordinating Group within two years of protocol publication. If drafts are not submitted before the agreed deadlines, or if we are unable to contact you for an extended period, the relevant Coordinating Group has the right to de-register the title or transfer the title to alternative authors. The Coordinating Group also has the right to de-register or transfer the title if it does not meet the standards of the Coordinating Group and/or the Campbell Collaboration.

You accept responsibility for maintaining the review in light of new evidence, comments and criticisms, and other developments, and updating the review at least once every five years, or, if requested, transferring responsibility for maintaining the review to others as agreed with the Coordinating Group.

Publication in the Campbell Library

The support of the Coordinating Group in preparing your review is conditional upon your agreement to publish the protocol, finished review, and subsequent updates in the Campbell Library. The Campbell Collaboration places no restrictions on publication of the findings of a Campbell systematic review in a more abbreviated form as a journal article either before or after the publication of the monograph version in *Campbell Systematic Reviews*. Some journals, however, have restrictions that preclude publication of findings that have been, or will be, reported elsewhere and authors considering publication in such a journal should be aware of possible conflict with publication of the monograph version in *Campbell Systematic Reviews*. Publication in a journal after publication or in press status in *Campbell Systematic Reviews* should acknowledge the Campbell version and include a citation to it. Note that systematic reviews published in *Campbell Systematic Reviews* and co-registered with the Cochrane Collaboration may have additional requirements or restrictions for co-publication. Review authors accept responsibility for meeting any co-publication requirements.

I understand the commitment required to undertake a Campbell review, and agree to publish in the Campbell Library. Signed on behalf of the authors:

Carlos Oya

Form completed by:	**Date:**
Carlos Oya	19th December 2014

Appendices

1. RESEARCH DESIGN AND ANALYTICAL METHOD ASSESSMENT

Source: Duvendack *et al.* (2011)

Figure 2: Research design and analytical method assessment

		Statistical Methods of Analysis		
		IV,PSM,2SLS/LIML,DID, RD	Multivariate	Tabulation
Research Design				
RCT		🟥	🟥	🟥
Pipeline		🟥	🟧	🟧
Panel or b/a and w/wo		🟥	🟧	🟨
Either b/a or w/wo		🟥	🟨	🟨
Natural Experiment		🟥		

Legend	Low threat to validity	🟥	High threat to validity	🟨
	Medium threat to validity	🟧		

Note: IV instrumental variables, PSM propensity score matching, 2SLS two-stage least squares, LIML limited information maximum likelihood, DID difference in differences, RD regression discontinuity. Source: Duvendack *et al.* (2011).

2. PROVISIONAL SET OF ELECTRONIC SEARCH TERMS

Example of set of search terms searched in Web of Science, Indexes=SSCI Timespan=1990-2015, field searched: topic.

1. TS=("certification" or "quality standards" or "quality label?ing" or "sustainability standards")
2. TS=((fair* OR ethic* OR alternative OR sustainab* OR responsib* OR specialty OR eco OR ecologic OR ecological OR organic) NEAR/3 (certifi* OR standard* OR label* OR seal* OR scheme* OR trad* OR market* OR "value chain*" OR commodit* OR product*))
3. TS=("fair trade" or fairtrade or fair-trade or transfair or "fair for life" or "Rainforest Alliance" or "Sustainable Agriculture Network" or "UTZ Certified" or "UTZ" or "Global Partnership for Good Agricultural Practice" or "Global GAP" or "GlobalGAP"

or "4C Association" or "Better Cotton Initiative" or "BCI" or "Cotton made in Africa" or Bonsucro or "Ethical Tea Partnership" or Trustea or "International Federation of Organic Agriculture Movements" or IFOAM or "soil association" or "IOAS" or "LEAF" or "Linking Environment and Farming" or "Union for Ethical BioTrade" or "UEBT" or "Roundtable on Sustainable Palm Oil" or "RSPO" "Fair Flowers Fair Plants" or "ProTerra" or "ISO 14001")

4. #3 OR #2 OR #1
5. TS=(Farmer* or farming or agricultur* or horticultur* or grower* or producer* or worker* or labo?rer* or smallholder* or small-holder* or cooperative* or co-operative* or syndicate* or ((trade or labo?r) NEAR union*) or "agricultural sector" or "agricultural trade" or "floriculture" or "crop production" or "agricultural products")
6. TS=(coffee OR cocoa OR tea OR infusion* OR "yerba mate" OR "camomile" OR sugar* OR fruit* OR banana* OR pineapple* OR mango* OR coconut* OR apricot* OR nut* OR cashew* OR "shea butter" OR argan OR rice OR quinoa OR bean* OR chickpea* OR "red kidney" OR lentil* OR soy* OR herb* OR spice* OR "olive oil" OR olive* OR wine OR honey OR cotton OR flower* OR floriculture OR "palm oil" OR (crop* NEAR/2 produc*))
7. #6 OR #5
8. TS=(Afghanistan or Angola or Albania or "American Samoa" or Argentina or Armenia or Armenian or Azerbaijan or Bangladesh or Belarus or Belize or Benin or Bolivia or Bosnia or Herzegovina or Botswana or Brazil or Bulgaria or Burkina Faso or Burkina Fasso or Burundi or Urundi or Cambodia or Cameroon or Cameroons or Cameron or Camerons or Central African Republic or Chad or Chile or China or Colombia or Comoros or Comoro Islands or Comores or Congo or Costa Rica or Cuba or Zaire or Cote d'Ivoire or Ivory Coast or Djibouti or Dominica* or East Timor or East Timur or Timor Leste or Ecuador or Egypt or United Arab Republic or El Salvador or Eritrea or Ethiopia or Fiji or Gabon or Gambia or Gaza or Georgia Republic or Georgian Republic or Ghana or Grenada or Guatemala or Guinea or Guiana or Guyana or Haiti or Honduras or Hungary or India or Indonesia or Iran or Iraq or Kazakhstan or Kenya or Kiribati or Korea or Kosovo or Kyrgyzstan or Kirghizia or Kyrgyz Republic or Kirghiz or Kirgizstan or Lao PDR or Laos or Lebanon or Lesotho or Liberia or Libya or Macedonia or Madagascar or Malagasy Republic or Malawi or Malaysia or Maldives or Marshall Islands or Mali or Mauritania or Mauritius or Agalega Islands or Mexico or Micronesia or Moldova or Moldovia or Moldovian or Mongolia or Montenegro or Morocco or Ifni or Mozambique or Myanmar or Myanma or Burma or Namibia or Nepal or Nicaragua or Niger or Nigeria or Pakistan or Palau or Palestine or Panama or Paraguay or Peru or Philippines or Philipines or Phillipines or Phillippines or Romania or Rwanda or Ruanda or Samoa or Samoan Islands or Sao Tome or Senegal or Serbia or Seychelles or Sierra Leone or Sri Lanka or Solomon Islands or Somalia or South Africa or St Lucia or St Vincent or Grenadines or Sudan or Suriname or Swaziland or Syria or Tajikistan or Tadzhikistan or Tadjikistan or Tadzhik or Tanzania or Thailand or Tonga or Togo or Togolese Republic or Tunisia or Turkey or Turkmenistan or Tuvalu or Uganda or Ukraine or Uruguay or Uzbekistan or Uzbek or Vanuatu or Venezuela or New Hebrides or Vietnam or Viet Nam or West Bank or Yemen or Zambia or Zimbabwe)
9. TS=((developing or "less* developed" or "under developed" or underdeveloped or "middle income" or "low* income" or underserved or "under served" or deprived or poor*) NEAR (countr* or nation? or population? or world or economy or economies))
10. TS=(low NEAR (gdp or gnp or "gross domestic" or "gross national" or GNI))
11. TS=(lmic or lmics or "third world" or lamicountr*)
12. TS=(low NEAR/3 middle NEAR/3 countr*)
13. TS="transitional countr*"
14. #13 OR #12 OR #11 OR #10 OR #9 OR #8

3. CRITICAL APPRAISAL OF INCLUDED STUDIES OF EFFECTS

Source: Waddington *et al.* (2014)

Selection bias and confounding

Mechanism of assignment: was the allocation or identification mechanism able to control for selection bias?

a) For Randomised assignment (RCTs),

Score "YES" if:

- A random component in the sequence generation process is described (for example, referring to a random number table)[13];
- and if the unit of allocation was at group level (geographical/ social/ institutional unit) and allocation was performed on all units at the start of the study;
- or if the unit of allocation was by beneficiary or group and there was some form of centralised allocation mechanism such as an on-site computer system;
- and if the unit of allocation is based on a sufficiently large sample size to equate groups on average.

Score "UNCLEAR" if:

- The paper does not provide details on the randomisation process, or uses a quasi-randomisation process for which it is not clear has generated allocations equivalent to true randomisation.

Score "NO" if:

- The sample size is not sufficient or any failure in the allocation mechanism could affect the randomisation process[14].

b) For discontinuity assignment (Regression Discontinuity Designs)

Score "YES" if:

- Allocation is made based on a pre-determined discontinuity on a continuous variable (regression discontinuity design) and blinded to participants or;

[13] If a quasi-randomised assignment approach is used (for example, alphabetical order), you must be sure that the process truly generates groupings equivalent to random assignment, to score "Yes" on this criteria. In order to assess the validity of the quasi-randomisation process, the most important aspect is whether the assignment process might generate a correlation between participation status and other factors (for example, gender, socio-economic status) determining outcomes; you may consider covariate balance in determining this (see question 2).

[14] If the research has serious concerns with the validity of the randomisation process or the group equivalence completely fails, we recommend to assess the risk of bias of the study using the relevant questions for the appropriate methods of analysis (cross-sectional regressions, difference-in-difference, etc.) rather than the RCTs questions.

- if not blinded, individuals reasonably cannot affect the assignment variable in response to knowledge of the participation decision rule;
- and the sample size immediately at both sides of the cut-off point is sufficiently large to equate groups on average.

Score "UNCLEAR" if:

- The assignment variable is either non-blinded or it is unclear whether participants can affect it in response to knowledge of the allocation mechanism.

Score "NO" if:

- The sample size is not sufficient or;
- there is evidence that participants altered the assignment variable prior to assignment.[15]

c) For assignment based non-randomised programme placement and self-selection (studies using a matching strategy or regression analysis, excluding IV),

Score "YES" if:

- Participants and non-participants are either matched based on all relevant characteristics explaining participation and outcomes, or;
- all relevant characteristics are accounted for.[16][17]

Score "UNCLEAR" if:

- It is not clear whether all relevant characteristics (only relevant time varying characteristics in the case of panel data regressions) are controlled.

Score "NO" if:

- Relevant characteristics are omitted from the analysis.

d) For identification based on an instrumental variable (IV estimation),

Score "YES" if:

- An appropriate instrumental variable is used which is exogenously generated: for example, due to a 'natural' experiment or random allocation.

[15] If the research has serious concerns with the validity of the assignment process or the group equivalence completely fails, we recommend assessing risk of bias of the study using the relevant questions for the appropriate methods of analysis (cross-sectional regressions, difference-in-difference, etc.) rather than the RDDs questions.

[16] Accounting for and matching on all relevant characteristics is usually only feasible when the programme allocation rule is known and there are no errors of targeting. It is unlikely that studies not based on randomisation or regression discontinuity can score "YES" on this criterion.

[17] There are different ways in which covariates can be taken into account. Differences across groups in observable characteristics can be taken into account as covariates in the framework of a regression analysis or can be assessed by testing equality of means between groups. Differences in unobservable characteristics can be taken into account through the use of instrumental variables (see also question 1.d) or proxy variables in the framework of a regression analysis, or using a fixed effects or difference-in-differences model if the only characteristics which are unobserved are time-invariant.

Score "UNCLEAR" if:

- The exogeneity of the instrument is unclear (both externally as well as why the variable should not enter by itself in the outcome equation).

Score "NO" otherwise.

Group equivalence: was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding?

a) For randomised control trials (RCTs) and quasi-RCTs,

Score "YES" if[18]:

- Baseline characteristics of the study and control/comparisons are reported and overall[19] similar based on t-test or ANOVA for equality of means across groups;
- or covariate differences are controlled using multivariate analysis;
- and the attrition rates (losses to follow up) are sufficiently low and similar in treatment and control, or the study assesses that loss to follow up units are random draws from the sample (for example, by examining correlation with determinants of outcomes, in both treatment and comparison groups);
- and problems with cross-overs and drop outs are dealt with using intention-to-treat analysis or in the case of drop outs, by assessing whether the drop outs are random draws from the population;
- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.

Score "UNCLEAR" if:

- Insufficient details are provided on covariate differences or methods of adjustment;
- or insufficient details are provided on cluster controls.

Score "NO" otherwise.

b) For regression discontinuity designs (RDDs),

Score "YES" if:

- The interval for selection of treatment and control group is reasonably small;
- or authors have weighted the matches on their distance to the cut-off point;

[18] Please note that when a), b) or f) score no or large differences in baseline characteristics, we suggest assessing risk of bias considering other study design (Diff-in-Diff, cross-sectional regression, instrumental variables)

[19]Even in the context of RCTs, when randomisation is successful and carried out over sufficiently large assignment units, it is possible that small differences between groups remain for some covariates. In these cases, study authors should use appropriate multivariate methods to correcting for these differences.

- and the mean of the covariates of the individuals immediately at both sides of the cut-off point (selected sample of participants and non-participants) are overall not statistically different based on t-test or ANOVA for equality of means;
- or significant differences have been controlled in multivariate analysis;
- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.

Score "UNCLEAR" if:

- There are covariate differences across individuals at both sides of the discontinuity which have not been controlled for using multivariate analysis, or if insufficient details are provided on controls;
- or if insufficient details are provided on cluster controls.

Score "NO" otherwise.

c) For non-randomised trials using difference-in-differences methods of analysis,

Score "YES" if:

- The authors use a difference-in-differences (or fixed effects) multivariate estimation method;
- the authors control for a comprehensive set of time-varying characteristics[20];
- and the attrition rate is sufficiently low and similar in treatment and control, or the study assesses that drop-outs are random draws from the sample (for example, by examining correlation with determinants of outcomes, in both treatment and comparison groups);
- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.

Score "UNCLEAR" if:

- Insufficient details are provided;
- or if insufficient details are provided on cluster controls.

Score "NO" otherwise.

d) For statistical matching studies including propensity scores (PSM) and covariate matching21,

[20] Knowing allocation rules for the programme – or even whether the non-participants were individuals that refused to participate in the programme, as opposed to individuals that were not given the opportunity to participate in the programme – can help in the assessment of whether the covariates accounted for in the regression capture all the relevant characteristics that explain differences between treatment and comparison.

[21] Matching strategies are sometimes complemented with difference-in-difference regression estimation methods. This combination approach is superior since it only uses in the estimation the common support region of the sample size, reducing the likelihood of existence of time-variant unobservables differences across groups affecting outcome of interest and removing biases arising from time-invariant unobservable characteristics.

Score "YES" if:

- Matching is either on baseline characteristics or time-invariant characteristics which cannot be affected by participation in the programme; and the variables used to match are relevant (for example, demographic and socio-economic factors) to explain both participation and the outcome (so that there can be no evident differences across groups in variables that might explain outcomes);
- in addition, for PSM Rosenbaum's test suggests the results are not sensitive to the existence of hidden bias;
- and, with the exception of Kernel matching, the means of the individual covariates are equated for treatment and comparison groups after matching;
- and, for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate or any appropriate analysis.

Score "UNCLEAR" if:

- Relevant variables are not included in the matching equation, or if matching is based on characteristics collected at endline;
- or if insufficient details are provided on cluster controls.

Score "NO" otherwise.

e) For regression-based studies using cross sectional data (excluding IV)

Score "YES" if:

- The study controls for relevant confounders that may be correlated with both participation and explain outcomes (for example, demographic and socio-economic factors at individual and community level) using multivariate methods with appropriate proxies for unobservable covariates;
- and a Hausman test[22]with an appropriate instrument suggests there is no evidence of endogeneity;
- and none of the covariate controls can be affected by participation;
- and either, only those observations in the region of common support for participants and non-participants in terms of covariates are used, or the distributions of covariates are balanced for the entire sample population across groups;
- and, for cluster-assignment, authors control particularly for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.

[22] The Hausman test explores endogeneity in the framework of regression by comparing whether the OLS and the IV approaches yield significantly different estimations. However, it plays a different role in the different methods of analysis. While in the OLS regression framework the Hausman test mainly explores endogeneity and therefore is related with the validity of the method, in IV approaches it explores whether the author has chosen the best available strategy for addressing causal attribution (since in the absence of endogeneity OLS yields more precise estimators) and therefore is more related with analysis reporting bias.

Score “UNCLEAR” if:

- Relevant confounders are controlled but appropriate proxy variables or statistical tests are not reported;
- or if insufficient details are provided on cluster controls.

Score “NO” otherwise.

f) For instrumental variables approaches,

Score “YES” if:

- The instrumenting equation is significant at the level of F≥10 (or if an F test is not reported, the authors report and assess whether the R-squared (goodness of fit) of the participation equation is sufficient for appropriate identification);
- the identifying instruments are individually significant (p≤0.01); for Heckman models, the identifiers are reported and significant (p≤0.05);
- where at least two instruments are used, the authors report on an over-identifying test (p≤0.05 is required to reject the null hypothesis); and none of the covariate controls can be affected by participation and the study convincingly assesses qualitatively why the instrument only affects the outcome via participation[23];
- and, for cluster-assignment, authors particularly control for external cluster-level factors that might confound the impact of the programme (for example, weather, infrastructure, community fixed effects, and so forth) through multivariate analysis.

Score “UNCLEAR” if:

- Relevant confounders are controlled but appropriate statistical tests are not reported or exogeneity[24] of the instrument is not convincing;
- or if insufficient details are provided on cluster controls (see category f) below).

Score “NO” otherwise.

Hawthorne and John Henry effects: was the process of being observed causing motivation bias?

Score “YES” if either:

- For data collected in the context of a particular intervention trial (randomised or non-randomised assignment), the authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded, or argue convincingly why it is not likely

[23] If the instrument is the random assignment of the treatment, the reviewer should also assess the quality and success of the randomisation procedure in part a).

[24] An instrument is exogenous when it only affects the outcome of interest through affecting participation in the programme. Although when more than one instrument is available, statistical tests provide guidance on exogeneity (see background document), the assessment of exogeneity should be in any case done qualitatively. Indeed, complete exogeneity of the instrument is only feasible using randomised assignment in the context of an RCT with imperfect compliance, or an instrument identified in the context of a natural experiment.

that being monitored in ways that could affect the performance of participants in treatment and comparison groups in different ways.
- The study is based on data collected in the context of a survey, and not associated with a particular intervention trial, or data are collected in the context of a retrospective (ex post) evaluation.

Score “UNCLEAR” if:
- It is not clear whether the authors use an appropriate method to prevent Hawthorne and John Henry Effects (for example, blinding of outcomes and, or enumerators, other methods to ensure consistent monitoring across groups).

Score “NO” otherwise.

4. Spill-overs: was the study adequately protected against performance bias?

Score “YES” if:
- The intervention is unlikely to spill-over to comparisons (for example, participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are unlikely)[25].

Score “UNCLEAR” if:
- Spill-overs are not addressed clearly.

Score “NO” if:
- Allocation was at individual or household level and there are likely spill-overs within households and communities which are not controlled for in the analysis;
- or if allocation at cluster level and there are likely spill-overs to comparison clusters.

5. Selective outcome reporting: was the study free from outcome reporting bias?

Score “YES” if:
- There is no evidence that outcomes were selectively reported (for example, all relevant outcomes in the methods section are reported in the results section).

Score “NO” if:
- Some important outcomes are subsequently omitted from the results or the significance and magnitude of important outcomes was not assessed.

Score “UNCLEAR” otherwise.

[25] Contamination, that is differential receipt of other interventions affecting outcome of interest in the control or comparison group, is potentially an important threat to the correct interpretation of study results and should be addressed via PICO and study coding.

6. Selective analysis reporting: was the study free from analysis reporting bias?

Score “YES” if:

- Authors use ‘common’ methods[26] of estimation and the study does not suggest the existence of biased exploratory research methods[27].

Score “NO” if:

- Authors use uncommon or less rigorous estimation methods such as failure to conduct multivariate analysis for outcomes equations where it is has not been established that covariates are balanced.

See also the following for particular estimation methodologies.

For PSM and covariate matching, score “YES” if:

- Where over 10% of participants fail to be matched, sensitivity analysis is used to re-estimate results using different matching methods (Kernel Matching techniques);
- for matching with replacement, no single observation in the control group is matched with a large number of observations in the treatment group.

Where not reported, score “UNCLEAR”. Otherwise, score “NO”.

For IV (including Heckman) models, score “YES” if:

- The authors test and report the results of a Hausman test for exogeneity ($p \leq 0.05$ is required to reject the null hypothesis of exogeneity);
- the coefficient of the selectivity correction term (Rho) is significantly different from zero ($P<0.05$) (Heckman approach).

Where not reported, score “UNCLEAR”. Otherwise, score “NO”.

For studies using multivariate regression analysis, score “YES” if:

- Authors conduct appropriate specification tests (for example, reporting results of multicollinearity test, testing robustness of results to the inclusion of additional variables, etc).

Where not reported or not convincing, score “UNCLEAR”. Otherwise, Score “NO”.

7. Other: was the study free from other sources of bias?

Important additional sources of bias may include: concerns about blinding of outcome assessors or data analysts; concerns about blinding of beneficiaries so that expectations, rather than the intervention mechanisms, are driving results (detection bias or placebo effects)[28]; concerns about

[26] ‘Common methods’ refers to the use of the most credible method of analysis to address attribution given the data available.

[27] A comprehensive assessment of the existence of ‘data mining’ is not feasible particularly in quasi-experimental designs where most studies do not have protocols and replication seems the only possible mechanism to examine rigorously the existence of data mining.

[28] All interventions may create expectations (placebo effects), which might confound causal mechanisms. In social interventions, which usually require behaviour change from participants, expectations may form an

courtesy bias from outcomes collected through self-reporting; concerns about coherence of results; data on the baseline collected retrospectively; information is collected using an inappropriate instrument (or a different instrument/at different time/after different follow up period in the comparison and treatment groups).

Score "YES" if:

- The reported results do not suggest any other sources of bias.

Score "UNCLEAR" if:

- Other important threats to validity may be present

Score "NO" if:

- It is clear that these threats to validity are present and not controlled for.

important component of the intervention, so that isolating expectation effects from other mechanisms may be less relevant.

4. OVERALL RISK OF BIAS ASSESSMENT

Source: Vaessen *et al*. (2014)

Figure 3: Overall risk of bias assessment

Study	**Design and analysis based assessment: study design, method of analysis)**	**Risk of selection bias and confounding**								**Risk of spill-overs and conta-mination**	**Risk of outcome reporting bias**	**Risk of analysis reporting bias**					**Other risk of bias**
		RCTs	Quasi-experiments														
			PSM CVM	/ IV	OLS	Logit/Probit	Bi-variate	Panel	Other			RCTs	PSM	IV	OLS	Other	

5. RISK OF BIAS TOOL FOR STUDIES EXAMINING THE BARRIERS AND FACILITATORS[29]

1. Is the research aim clearly stated? (Yes/No)

REPORTING:

2. Description of the context? (Yes/No)

3. Description of sampling procedures? (Yes/No)

- *How have the participants and research sites been selected? Were they the most appropriate?*

4. Are sample characteristics sufficiently reported? (sample size, location, and at least one additional characteristic) (Yes/No)

5. Is it clear how the data were collected (eg: *for interviews, is there an indication of how interviews were conducted*? (Yes/No)

6. Methods of recording of data reported? (Yes/No)

7. Methods of analysis explicitly stated? (Yes/No)

METHODOLOGY:

8. Is there a clear link to relevant literature/theoretical framework? (Yes/No)

9. Is the design appropriate to answer the research question? (Yes/No)

- *Has the researcher justified the research design?*

10. Was the sampling strategy appropriate to the aims of the research? (Yes/No)

- *Have the researchers explained how the participants were selected?*
- *Have the researchers explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study?*
- *Have the researchers discussed issues around recruitment? (for example, why some people chose not to take part)*

11. Have the researchers explained why the research locations/sites they selected were the most appropriate to provide access to the type of knowledge sought by the study? (Yes/No)

12. Were the data collected in a way that addressed the research issue? (Yes/No)

- *Were the methods used appropriate and justified?*
- *Did the researcher discuss saturation of data?*

13. Was the data analysis sufficiently rigorous? (Yes/No)

- *Is there a detailed description of the analysis process?*

[29] Developed by Waddington *et al.* (2012), based on CASP (2006)

- *Does the data support the findings?*
- *Is the relationship between the researcher and the participants adequately considered?*
- *To what extent is contradictory data are taken into account?*
- *If the findings are based on quantitative analysis of survey data, are multivariate techniques used to control for potential confounding variables?*

14. Has triangulation been applied? (Yes/No)

- *Data triangulation (location, time and participants)*
- *Investigator triangulation*
- *theory triangulation (several theories)*
- *methodological triangulation*

15. Is the analysis and conclusions clearly presented? (Yes/No)

- *Have the researchers discussed the credibility of their findings? (for example, triangulation, respondent validation, more than one analyst)*
- *Is there adequate discussion of the evidence both for and against the researcher's arguments?*
- *Are the findings explicit?*
- *Are the findings discussed in relation to the original research question?*

16. Was there potential for conflict of interest and if so, was this considered and addressed? (Yes/No)

17. Does the paper discuss ethical considerations related to the research? (Yes/No)

6. OUTLINE OF CODING TOOL

Adapted from Waddington *et al.* (2014)

Table 3: Outline of Coding Tool

Category	Data extracted
General information	Authors Institution Funders of implementation Year of publication Language Primary data?
Publication type	Journal article Working paper Conference paper Project report Other report Book chapter Book
Author affiliation	Funders of research Employee of implementing agency? Employee of another body?
Certification	Certification scheme/label Level (if appropriate) Specific goals/targets Third party auditing? Frequency of audits Multiple certifications?
Interventions	Farm-level interventions Community level interventions Target group Number of people who received intervention Length of intervention (from-to) Additional interventions (not covered by CS)
Context	Country Region Crop
Study method	Quantitative Qualitative Mixed
Study type	RCT quasi-RCT RDD natural experiment DID

	IV ITS PSM 2SLS 3SLS ex post observational studies with control for confounding
Data collection	Period (from-to) Frequency
Control group	Received intervention? Description of intervention Group allocation mechanism
Sampling	Number of clusters (treatment & control) Number of individuals (treatment & control) Attrition (treatment & control)
Spillovers	Geographical separation of treatment and control?
Contamination	Influence of other intervention which differentially affects treatment and comparison groups on relevant outcomes
Study quality	Risk of bias assessment (see section 3.3.3)
Effect sizes	Estimated effect type Differentiated effect estimates? Adjusted or unadjusted analysis
Intermediate effects (where possible with SE)	Net returns to certified production Quality of commodities Productivity of commodities Price levels (for certified commodity) Price volatility (for certified commodity) Wages (nominal and/or real) Non-wage labour Organisational empowerment of producers' and workers' organisations
Intermediate effects measure	List for each effect reported
Endpoint effects (where possible with SE)	Household income or consumption Health and education of adults and children Gender equity in the outcomes above Producers' and workers' empowerment
Endpoint effects measure	List for each effect reported
Qualitative information	Quality of the interventions Diffusion of the intervention Adoption of interventions Interactions with non-certified interventions Uncaptured confounding? Differentiation of effects (socio-economic, gender, age, etc.) Excluded people/groups Distribution of benefits

7. ADVISORY GROUP

1. **Kristin Komives** - Position: Senior M&E Manager; Type of organization; International organization-NGO; Name of organization: **ISEAL Alliance** http://www.isealalliance.org/ This is a key stakeholder, as the global membership association for sustainability standards, which includes several certification bodies as its members. The organization has a strong impact evaluation team which seeks to improve the production, dissemination and use of robust evaluation data on sustainability standards.
2. **Sue Longley** -International Officer for Agriculture and Plantations; Type of organization: Civil Society Organization; Name of organization: IUF - *International Union of Food, Farm and Hotel Workers*.
3. **Robert Hale** - Private Sector Development Adviser; Type of organization: Central Government agency; Name of organization: **Dept for International Development (UKAID).**
4. **Maren Duvendack (University of East Anglia)** – Academic / methodologist - well established scholar in systematic reviews in international development. Type of organization: Academia; Name of organization: University of East Anglia.
5. **Lone Riisgard (DIIS and Roskilde University, Denmark)** – Academic / Area of study – well established scholar in the fields of value chain analysis, private standards/social regulation (certification schemes) impact, workers. Type of organization: Academia; Name of organization: Roskilde University and Danish Institute for International Studies
6. **Sandy Balfour (Divine Chocolate – Kuapa Kokoo; Liberation Foods)** – He has been directly involved in the setting-up and work of two Fairtrade companies (Divine Chocolate – linked to a famous cocoa producer organization in Ghana called Kuapa Kokoo; and Liberation Foods); Type of organization: Private sector – producer organizations.
7. **Jane Njuguna (AGRA)** – M&E officer. Type of organization: Funding body – NGO; Name of organization: AGRA. http://agra-alliance.org/ AGRA is one of the funders of this systematic review and has a strong interest in the findings.
8. **Daniel Phillips (3ie Synthesis and Review Office)** – Type of organization: Think-tank and Commissioning agency – Focal point for this project at 3ie and key resource person for all the requirements and key methodological issues specific to this systematic review.