The challenges of screening and synthesizing qualitative research in a mixed-methods systematic review. The case of the impact of agricultural certification schemes

By Dafni Skalidou[1] and Carlos Oya[2]

ABSTRACT

The number of mixed-methods systematic reviews in international development is growing in recent years. By recognising the value of qualitative research in providing valuable evidence on causal mechanisms, barriers, facilitators and the importance of context, mixed methods systematic reviews go beyond the 'what works' question. However, appropriate methods to screen and synthesise qualitative evidence in these reviews are still in a development phase and the methodological literature dealing with reviewing qualitative evidence in the field of development studies is scarce and under-developed. This paper aims to contribute to this gap by discussing the methodological and practical challenges of including qualitative evidence in a mixed methods systematic review in international development. In particular, this article makes a contribution in terms of offering reviewers and users of systematic reviews a full account of the process of screening and synthesizing a very large volume of heterogeneous qualitative studies. Using as an example a review on the effects of certification schemes for agricultural production (Oya et al, 2017), we report on each reviewing step, describing the problems encountered and solutions found. The paper proposes ways of extracting a large volume of data and integrating the qualitative synthesis with the evidence from the related quantitative effectiveness review.

Keywords: mixed-methods systematic review, methodology, agricultural certification, qualitative synthesis

[1] Universty of East Anglia

[2] SOAS, University of London

1. Introduction

Systematic reviewing is a method that allows 'locating, appraising, and synthesising evidence' and is widely used as an 'aid to evidence based decision making' (Petticrew, 2001: 98). The method aims to tackle two major issues undermining the standard literature review. By following transparent and replicable procedures of finding, appraising and synthesising evidence, it is not susceptible to the authors' (or study commissioners') 'cherry picking' of studies and/or findings within studies to suit their purposes or preferred agenda. Additionally a systematic review separates 'the wheat from the chaff' and by excluding evidence which does not meet explicit quality criteria produces a synthesis of only the trustworthy evidence relevant to the review question (White and Waddington, 2012:352). Employed in medical sciences since the 1970s to synthesise evidence on health-care interventions, systematic reviews have since been used in an increasing range of disciplines outside health care, including natural sciences, education and social sciences (Petticrew 2001; Mallet et al 2012). Despite their relatively recent introduction in the field of development studies (the first systematic reviews in international development following the Campbell Collaboration standards date 2012, although non-Campbell systematic reviews in the field started to emerge in the early 2000s), their use has increased rapidly and considerably over the last years, amid calls for more and better evidence to inform policymaking (White and Waddington, 2012; Mallet et al, 2012).

Originating from the field of medicine, systematic reviews have traditionally focused on synthesising quantitative evidence, typically from randomised control trials (RCTs), to assess the effectiveness of a treatment, i.e. on 'what works'. In the field of international development, where randomisation is not always feasible (or desirable), this has been translated into systematic reviews drawing on evidence produced by both experimental and quasi-experimental study designs able to control for selection bias in order to conclude whether a specific development intervention works or not. Nevertheless, over the last decades it has been increasingly argued across disciplines that different types of questions require different research approaches. Scholars from the field of health and education have highlighted that qualitative studies have a 'distinctive and important contribution' to make in evidence-based policy (Dixon-Woods et al 2000:125), particularly in addressing question regarding the processes by which certain activities are undertaken (Davies, 1999). As a result, the need to move 'beyond effectiveness in evidence synthesis' and no longer consider quantitative methods as the only valid source of evidence has been recognised, and qualitative evidence has established 'a place for itself' in systematic reviewing (Popay 2006; Dixon-Woods and Fitzpatrick, 2001:765). This recognition comes among calls for more mixed methods theory-based evaluations that not only address the effectiveness question (i.e. does the intervention work?) but also questions on causal mechanisms and more specifically the how, why, when and for whom interventions work (Weiss, 1997; Pawson and Tilley, 2004; White, 2009; Snilstveit 2012; Caracelli and Cooksy 2013). Considering the importance of context and addressing the issue of external validity and transferability to other settings also call for a more pluralistic approach to impact evaluation and a broadening of the questions (Cartwright 2010). This is being reflected in an increasing number of systematic reviews that adopt a theory-based and/or realist approach that draws on

both theoretical understandings and empirical evidence in order to elucidate the interplay between 'the context in which the intervention is applied, the mechanisms by which it works and the outcomes which are produced' (Pawson et al 2005:21). Such approaches are becoming common in the medical and health care field (e.g. de Goeij et al 2015; Nicaise et al 2013; O'Campo et al 2011; Wong et al 2010). Calls for transdisciplinary research synthesising knowledge from different academic traditions and contexts (Oliver et al 2017) and systematic reviews and maps combining different types of evidence (McKinnon et al 2016; Pullin et al 2013) have also emerged in environmental studies. In the field of international development, systematic reviews adopting a mixed methods approach drawing on different types of evidence, in order to not only understand what works but also for whom and in what circumstances, have increased their presence over the last years (see Figure 1).

<INSERT FIGURE 1 ABOUT HERE>.

Nevertheless, systematic reviews methods were initially developed to synthesise quantitative evidence from controlled trials. While methodological texts for reviewing qualitative evidence (Atkins et al, 2008; Barnett-Page and Thomas 2009; Dixon-Woods et al 2000) and combining evidence provided by diverse study types (Dixon-Woods et al 2004; Harden and Thomas 2005; Caracelli and Cooksy 2013) are now emerging, mainly in the medical research literature, research in this area is still ongoing (Thomas and Harden, 2008). Related methodological literature dealing with reviewing qualitative evidence and conducting mixed methods systematic reviews in the particular field of development studies is even more scarce and under-developed (Snilstveit 2012). This paper aims to contribute to this gap by discussing the methodological and practical challenges of including qualitative evidence in a mixed methods systematic review in the field of development studies. In particular, this article makes a contribution in terms of offering reviewers and users of systematic reviews a full account of the process of screening and synthesizing a very large volume of heterogeneous qualitative studies. Using as an example a review on the effects of certification schemes for agricultural production (Oya et al, 2017), we report on each reviewing step, describing the problems that we encountered and the ways we dealt with them. Although our focus is methodological, we do not report here on some of the detailed methods used (e.g. precise searching terms or development of coding themes). A detailed account of the methods used to conduct the review is available to the reader in the full technical report (Oya et al. 2017). In this article we focus on particular issues that arose while reviewing qualitative research systematically and in the process of integrating the findings with those of a statistical meta-analysis.

The paper is structured as follows. Section 2 provides a basic background to the review that is used as example, and presents the basic criteria used for inclusion of qualitative studies. Section 3 discusses the challenges of searching and screening a large volume of heterogeneous material, much of which had to be found outside conventional bibliographic databases. Section 4 focuses on the options for quality appraisal, an area for which there is published guidance but still fertile ground to experiment with different combinations of criteria depending on the nature of the qualitative studies found. Section 5 illustrates the hurdles in the process of data extraction and proposes some options for analytically-driven selectivity. Section 6 contains a more detailed

account of the synthesis of qualitative evidence in a review with a large number of studies and provides an account of key decisions along the chain from coding to the narrative synthesis along theory-based thematic blocks. Finally Section 7 concludes and provides a summary of key pointers for researchers willing to integrate a large volume of qualitative material in systematic reviews.

2. Background on the systematic review of agricultural certification

We draw on our experience from reviewing qualitative evidence for a theory-based mixed methods systematic review of the socio-economic effects of certification schemes (CS) for agricultural commodities (Oya et al. 2017). The review was commissioned by 3ie and was conducted according to the Campbell Collaboration standards for systematic reviews. It included a conventional quantitative effectiveness question about the impact of certification on indicators of wellbeing for producers and workers, complemented with a review question on contextual factors, barriers and facilitators which drew on the rich qualitative literature that was found on the patterns and nature of certification systems in agriculture. The focus in this paper is on the qualitative synthesis and the research process to synthesise evidence from qualitative research and integrate it with findings from the quantitative meta-analysis. The review question on barriers and facilitators[3], which is of interest here, was twofold and explored the circumstances under which certification schemes have intended or unintended effects, as well as the factors that act as barriers or facilitators to these effects. To address this question we included studies that met the following inclusion criteria:

1. The research question or objective had to be clearly reported.

2. Data collection methods and, where appropriate, sampling procedures had to be clearly reported.

3. The study had to provide evidence based on primary data collected from CS beneficiaries, facilitators, implementers, extension agents, auditors or experts analysed using qualitative methods.

4. The study had to provide *substantive* evidence on at least one of the key themes of interest (implementation dynamics, distributional dynamics and contextual factors) devised from the Theory of Change (ToC) used in this review

[3] These types of questions are also called 'process' questions as they explore the 'conditions of programme implementation and the mechanisms that mediate between processes and outcomes as a means to understand when and how programmes work (Weiss, 1997:41). See also Moore et al (2014) on process evaluations in the context of health.

These were basic criteria designed to screen a wide range of sources and include only substantive evidence that met the review methodological inclusion criteria from a heterogeneous pool of studies that included a mix of secondary data analysis and primary research ranging from 'quick-and-dirty' advocacy reports to in-depth ethnographic work. By 'substantive' we meant that the material needed to provide either a 'thick description'[4] (i.e. detailed description of the relevant context together with an analysis of how this context affects certification processes) or entire sections devoted to the analysis of certification processes (i.e. studies providing only mentions, or non-analytic descriptions of general or historical context were not included). Overall, 136 qualitative studies were included for synthesising barriers, enablers and other contextual factors. The findings of the qualitative synthesis where then integrated with the results of the statistical meta-analysis which was informed by 43 experimental and quasi-experimental studies. However, the technical report (Oya et al. 2017) reported the quantitative and qualitative syntheses separately first, in order to present the wealth of empirical information obtained from a large pool of studies for both review questions. On the basis of the main findings from each synthesis and a revision of the ToC which informed the research process from the protocol stage we then wrote a narrative integrated synthesis including relevant quantitative and qualitative findings according to the main causal chains stemming from the ToC and the evidence found.

3. Searching & Screening

Review scope

It is debatable whether a systematic qualitative synthesis should include all relevant studies. Discussing meta-ethnography, Doyle (2003:326) argues that 'the sample is purposive rather than exhaustive because the purpose is interpretive explanation and not prediction', as it is the case in meta-analysis. Therefore, aiming for 'conceptual saturation' may be more appropriate as a search strategy for qualitative research, Thomas and Harden (2008) suggest. On the contrary, Barroso et al (2003:154), state that a qualitative meta-synthesis should ideally retrieve all the relevant studies, not only a sample of them and cite Cooper (1998) to support that 'the most important threat to the validity of any research integration effort is to fail to conduct a sufficiently exhaustive search'. Finlayson and Dixon (2008:61), on the other hand, suggest that in qualitative synthesis no method is preferable over the other, and see qualitative synthesis as a 'continuum' of methods from which researchers should choose the approach that best suits the material to be synthesised.

In our case, we developed a common search strategy for both qualitative and quantitative evidence, and aimed at making our search as exhaustive and comprehensive as possible.[5] In the literature on agricultural

[4] See Geertz (1973).

[5] See Thomas and Harden (2008) on why it makes sense to search both quantitative and qualitative studies together before initiating separated screening.

certification schemes it is sometimes impossible to distinguish what is primarily quantitative and qualitative from title and abstract searches. Moreover, any given paper could produce relevant quantitative and qualitative evidence at the same time, especially in the case of mixed-method evaluations. We searched for studies that included qualitative evidence regardless of their design (i.e. qualitative evidence could be provided by purely qualitative studies on CS, quantitative impact evaluations or mixed-methods evaluations) or whether the evidence was linked to programmes included in the effectiveness review. One constraint faced by this review and discussed in the section on synthesis method below, was that theory-based evaluation studies that collected sufficient evidence for process and 'qualitative' questions alongside data for the effectiveness question are scarce in the certification literature. It is also generally hard to find relevant qualitative material from separate studies conducted in the same exact settings as contemporaneous impact evaluations. More importantly, the team regarded qualitative studies and their evidence useful in their own right even when not directly linked to specific impact evaluations and their particular settings. Overall this meant that adopting an 'effectiveness plus' approach (Snilstveit, 2012) and drawing only on additional information from studies included in the effectiveness review, or evidence from different studies but on the same interventions or settings as the quantitative evidence, was not an option that could address the process question in a satisfactory way. Instead, we set out to search and synthesise relevant qualitative evidence on CS, regardless of whether this evidence was in any way linked to the specific programmes reviewed in our meta-analysis. This approach is not new in systematic reviewing, as Snilstveit (2012) underlines providing examples from the medical field (e.g. Thomas et al. 2003; Harden et al 2009). It seems to be more scarce, however, particularly in the field of international development where we are aware only of few other reviews that have followed that path so far (Waddington et al, 2014; Lawry et al, 2014; De Buck et al, 2017). On the contrary, mixed methods reviews narrowing the inclusion of qualitative evidence only to the evidence that is linked to the interventions (or countries) included in the effectiveness review appear to be more common (e.g. Carr-Hill et al 2016; Molina et al 2016; Samii et al 2014; Berg and Denison 2012).

An 'effectiveness plus' mixed-method review has the advantage of reducing the volume of material to screen and especially synthesise. By contrast, including all the relevant existing trustworthy evidence instead of only evidence directly linked to quantitative evaluations is time and resource demanding. This was clearly the case in this review. Despite the high cost involved, however, our experience shows that broader but highly relevant qualitative evidence can be very valuable in illuminating implementation patterns across different contexts, as well as in contributing to our understanding of why the same *type* of intervention can be effective in one context but not in another. This is illustrated by the way 46 qualitative studies included in this review, which contained evidence not directly linked to the quantitative evaluations included in the effectiveness review, allowed a deeper understanding of the crucial role of Producers' Organisations (POs) in the effectiveness of CS more broadly. By synthesising evidence on the governance and implementation dynamics within POs and between POs and buyers, we were able to identify aspects of PO management but also in the characteristics of producers and buyers that could influence certification effectiveness. PO mismanagement and corruption, for instance, was a recurrent issue across different contexts and value chains, which affected producers'

participation in CS and the resulting benefits. On the other hand, transparency in management and transactions and provision of high quality services was identified as a facilitator of CS effectiveness. Externally-imposed POs were more vulnerable to corruption than POs formed on producers' own initiatives and efforts and therefore the latter performed better than the first in terms of certification benefits. PO-buyer relations were also important, with cases of long-lasting, direct and transparent relations with buyers enhancing certification benefits to producers. Should these 46 studies not been included we would have lost all this valuable explanatory evidence on how PO specific characteristics can undermine or boost the effects of CS.

Searching multiple sources

Searching for relevant qualitative evidence is a particularly challenging, time- and resource- consuming task. Qualitative research is more 'widely dispersed than quantitative research' across different databases and journals (Dixon-Woods et al 2000:130; Walters et al, 2006) and is often published in books or theses instead of indexed in electronic databases (Atkins et al 2008) and therefore can be more time consuming to retrieve than quantitative evidence (Lloyd Jones, 2004). In our case, characteristics specific to the certification literature added to the above mentioned challenges. The literature on certification is heterogeneous in its origin (from academic studies to impact evaluations commissioned by certifying bodies or funders of CS), spanning across disciplines (from geography and economy, to gender or environmental studies) and hosted in a vast array of locations, from electronic databases to websites of research institutions, organisations related to standards and certification, funders and donors. It often stands in a grey zone between scholarship and advocacy, where studies commissioned by certification related organisations to be used for marketing purposes abound. A further challenge, not exclusive of the certification literature but one that clearly applies, is the existence of papers with unclear titles and abstracts (or no abstracts at all). This may lead to inappropriate indexing of these papers (Atkins et al 2008) but also to a substantially larger proportion of papers having to be retrieved for full text screening in order to enable inclusion decisions (Lloyd Jones, 2004).

Due to these complexities, it was not possible to rely exclusively on electronic searches of bibliographic databases. Therefore, besides extensively searching multiple electronic databases with different foci (i.e. social science-related bibliographic databases, subject-specific databases covering agriculture and international trade/economics, systematic review databases, and national and regional databases), we also developed a multi-pronged strategy to find relevant 'grey' literature'. This included meticulous targeted searches for reports, students theses and papers not indexed in databases, as well as hand searches of books for relevant chapters. In our initial searches we prioritised higher sensitivity over precision of search terms in order to avoid omitting relevant studies which did not report sufficient information in their title or abstract. Implementing this obviously complex search strategy required highly specialised skills and knowledge on information retrieval and it would not have been possible without the support and advice of two information retrieval specialists, who not only ensured that our search strategy was as exhaustive as possible but also assisted in technical troubleshooting with specific databases. Moreover, the second phase of targeted searches was much

more time consuming than the initial electronic searching of bibliographic databases, an issue that must be considered in advance in order to devise realistic timeframes for this kind of systematic review.

Selecting studies

Screening and selecting studies from a wide range of sources is an established practice in systematic reviewing, as authors seek to identify all the relevant literature and minimise bias from omitting valid evidence (Livoreil et al. 2017). In this process we encountered some well, and other less, documented challenges by scholars with experience in searching and synthesising different types of evidence. First, including sources beyond large databases with export facilities in place resulted in a high number of studies which could not be automatically imported in EPPI-Reviewer 4, our systematic reviewing software, as it is the standard procedure (Bates et al 2007). Instead, studies retrieved from databases or websites with no export facilities, which, given our searching scope, were multiple, had to be screened on title and abstract on the spot. Only the papers eligible for full text screening were then manually imported into the software. This proved to be a more resource intensive process than having all the references automatically imported at once and then screened after duplicate removal.

Second, the screening process required several rounds of screening on partial and full text due to the fact that title and abstract alone often failed to provide sufficient information for exclusion, an issue already highlighted in the health literature (i.e. Evans 2002, Barroso et al 2003, Lloyd Jones 2004). As a result, we undertook a first round of screening on title and abstract only against relevance criteria (i.e. is the title and abstract relevant to the topic of the review), during which we decided to be over-inclusive in order to deal with unclear titles and abstracts. This lead to large number of irrelevant and ineligible studies being included after the first round of screening on title and abstract. Therefore, before proceeding to full text screening we engaged in a partial text review aimed at excluding studies which were irrelevant but could not be excluded without screening, at least partially, the main body of the study. However, during this partial text review we were able to exclude some of the studies that did not meet the first three inclusion criteria (see section 2), mainly those which were not based on primary data or new analysis of existing data. This additional partial text screening also gave the opportunity to separate the included studies according to the type of evidence they contained (quantitative, qualitative, mixed).

We then proceeded to full text screening using two different coding tools containing methodological criteria for quantitative and qualitative evidence. Studies containing mixed evidence were screened against both coding tools. Interestingly, during the full text screening phase there were more disagreements between coders for studies containing qualitative evidence. This is despite the fact that the review relied only on basic methodological criteria for inclusion of qualitative evidence (i.e. research question/objectives and data collection methods should be clearly reported, relevant qualitative evidence should be based on primary data). This is significant of the difficulties in dealing with qualitative evidence where, unlike quantitative studies, there are no clear-cut inclusion criteria for study design and analytical methods. Potential discretion among

coders means that a consistent system of checks and piloting of coding is necessary to avoid disagreements. This in itself adds to the time needed to screen studies. The more disagreements in initial stages, the more piloting and cross-checking was needed until coding decisions finally converged.

The initial full-text review suggested that, although large number of studies contained relevant evidence to the broader field of CS and met all the basic methodological criteria for inclusion, the depth of evidence with explanatory power to address the specific questions on context, barriers and facilitators in which we were interested was much more limited. The result of that first-stage full-text screening was a large number of studies containing qualitative evidence (n=264) but no assurance that they were all suitable for the task of addressing our review question with the required substance. Moreover, trying to synthesise such a large number of studies would have compromised depth of analysis for the sake of breadth, given limited time and resources. We therefore decided to only include studies which contained 'relevant and substantive' evidence on the specific thematic areas of interest, namely implementation dynamics, distributional dynamics and other contextual factors shaping the causal pathways to impact (see section 2 on the definition of "substantive" evidence). This allowed us to exclude a large number of studies which passed the basic methodological criteria and contained relevant evidence, but whose analysis was rather thin and descriptive, findings were not clearly linked to data and overall lacked the ability to explain how, for whom and under what circumstances CS could or could not work.

Discussion of the searching and screening results

The searching and screening process resulted in 136 included studies across 114 individual reports. 20% of the included studies provided qualitative evidence resulting from ethnographic research methods. Our results reaffirm that going the extra mile to search for studies not indexed in electronic databases was indeed necessary in order to avoid missing relevant qualitative evidence. Although we have not tracked the source of our included studies, figure 2 below shows that only 37% of the included studies were journal articles published in peer reviewed journals, and therefore likely to be indexed in bibliographic electronic databases. A substantial 43% of the included studies were in the form of research reports and working papers, which can be easily accessible online, though not always indexed in the main social sciences electronic databases (Figure 2). Finally, the remaining 20% originated from PhD and Master thesis and a small percentage of book chapters. It is unlikely that we could have captured this evidence without targeted searched in databases specialised in theses and dissertations or hand searches.

<INSERT FIGURE 2 ABOUT HERE>

It is important to note here that despite representing a rather small percentage of the total of the included studies, searches for theses and dissertations rewarded us with twenty-four eligible studies which proved to be exceptionally rich sources of trustworthy and insightful primary qualitative evidence, particularly when ethnographic methods were used (e.g. Sen, 2009; Setrini, 2011; Staib, 2012; Naylor, 2014). Having the space

(and obligation) to provide detailed methodological and analytical chapters, these studies commonly met all the methodological criteria for inclusion, and as we shall discuss further on, ticked most of the boxes in the quality appraisal process and provided evidence that could convincingly unpack the 'black box' of the CS processes.

On the contrary, searches in websites of certification-related organizations lead to mostly non-includable studies. Despite being relevant to the topic of the review, these studies often failed to meet the inclusion methodological criteria, including those on relevant and substantial evidence. This does not necessarily mean that these papers did not contain good research. However, being written for a non-academic public, they often failed to report sufficiently on their research design and methods as their readership was not expected to require such details. Unfortunately, poor or absent methodological reporting prevented assessments of the quality of these studies and lead to their exclusion as a rigorous synthesis of the evidence would require these details (e.g. Fairtrade Foundation, 2010; Giovannucci et al, 2008; UTZ, 2016; TWIN, 2013). Another problem with the literature sometimes commissioned and /or conducted by certification systems is that it is often limited to 'fact-finding' information, with no explanatory power to illuminate questions of process and context. Sometimes the operational expediency of this kind of studies lacking enough explanatory depth is related to the constraints on time and resources that commissioning organisations face. It should be highlighted, however, that some studies commissioned by standards bodies and NGOs implementing certification programmes of high research quality and reporting standards have also appeared in recent years (e.g. Waarts et al 2013 &2016). This may indicate a change in the research culture of such organisations and thus make systematic review searches more promising.

Finally, we would like to highlight that research reports, working papers and PhD theses commonly included more details on the data collection and analysis methods than the journal article versions of the same studies. This resulted in journal articles being excluded on the basis of non-substantial evidence, while the linked 'grey' version of the same study was included. It was common, for instance, to include the PhD thesis of an author but to exclude journal articles associated with the thesis. Some examples of excluded articles linked to included PhD thesis are Bacon (2005), Jaffee (2008), Lyon (2007), Shreck (2002), Sen (2014). Part of the problem lies in the space constraints that authors face when submitting to journals, where priority is often accorded to findings and the details of the data analysis, rather than to methodological aspects of data collection.

The above points underline the importance of going off the road to search for studies that although not indexed in main electronic databases, can contain valuable and trustworthy qualitative evidence. However, they also reveal the need to be selective regarding 'grey' literature searches, particularly when dealing with limited resources. Databases and websites addressed to a more general, non-academic public may not yield a satisfactory amount of includable studies due to limitations in their methodological reporting.

4. Quality appraisal

No consensus exists regarding how, or even whether, the quality of qualitative research should be assessed (Campbell et al 2003; Atkins et al 2008; Thomas and Harden, 2008). While some have criticised the 'the quest for permanent or stable criteria' as adopting a positivist approach in social inquiry (Schwandt, 1996:58), others agree that defining what 'good evidence' and 'high quality' research is, is necessary in order to systematically review and synthesise evidence (Popay et al, 1998; Dixon-Woods et al 2004; Hannes 2011). Even when agreeing that qualitative evidence should be appraised, however, the purpose of the process is debated, with qualitative research scholars arguing against excluding studies on the basis of the quality assessment due to the lack of consensus on what is 'good' qualitative evidence (Sandelowski et al. 1997). [6]Further concerns related to quality appraisal of qualitative evidence regard how a single set of criteria can do justice to the diversity of approaches within qualitative research (Dixon-Woods et al 2004) and the risk of excluding good research due to inadequate appraisal criteria (Atkins et al, 2008). Unlike the case of quantitative evidence where it is established that different study types require different appraisal criteria (i.e. a controlled trial needs to be assessed in a different way than a propensity score matching study), qualitative research has been so far regarded by quality assessments as a 'unified field', Dixon-Woods et al (2004:8) underline. Another concern in appraising qualitative (but also quantitative) evidence relates to the challenge of distinguishing between the quality of reporting and the quality of the study design and execution (Dixon-Woods et al 2004).

So how and for what purpose should we assess the quality of qualitative evidence included in mixed methods systematic reviews? Alongside the debate about the extent to which qualitative evidence can or should be assessed and the recognition of the challenges that come with the task, a series of appraising tools, checklists, or list of questions for assessing qualitative evidence has emerged (Spencer et al 2003). One of the most popular tools in qualitative and mixed methods systematic reviews is the Critical Appraisal Skills Programme (CASP) which consists of a series of 10 questions. Originally developed for assessing qualitative evidence in the field of medicine (e.g. Bohren et al, 2015; Munro et al 2007; Campbell et al 2003), the tool has also been adapted for assessing qualitative evidence in development studies systematic reviews (e.g. Waddington et al 2014; Lawry et al 2014; De Buck et al 2017). Comparing how the CASP instrument performs in relation to two other methods of appraising qualitative research, Dixon-Woods et al (2007:45) conclude that the tool encourages judgements on the 'procedural aspects of research', but can be less insightful regarding the depth of analysis

[6] However, there is an increasing tendency in the methodology literature to agree on basic principles of quality that can correspond to the standard validity criteria used for quantitative research. Hannes (2011) mentions the criteria of credibility (qualitative equivalent for internal validity); transferability (external validity); dependability (reliability); and confirmability (objectivity). Reflexivity and plausibility mays also be considered in addition to the core criteria above, especially in cases where the epistemological orientation of the researcher prevents from conforming to the principle of confirmability (Bryman 2012).

of the study or the contribution of knowledge it make in its field. In short, it is a good foundation from which to develop more elaborate and fine-tuned tools that also focus on substance and analysis.

In our case, we developed an adapted version of the CASP (2013) tool to assess the quality of the included qualitative evidence. At the time our systematic review was conducted, the only published mixed methods systematic review in the field of international development, meeting the Campbell Collaboration standards and including all the qualitative evidence and not only the one linked to the effectiveness review was the review on the effects of Farmer Field Schools (FFS) for improving farming practices and farmer outcomes, conducted by Waddington et al (2014). Given the proximity between FFS and agricultural CS as interventions targeting farming populations, and the fact that it was the closest published example of what we were trying to do, we decided to follow Waddington et al (2014) in their selection of qualitative appraisal tool and build on their version of CASP to adapted to our needs. Like Waddington et al (2014) we assess the included studies against the clarity of research question, the existence of a clear descriptions on the context, the sampling procedures, the data collection and analysis methods, and the triangulation of data. While Waddington et al (2014) focused more on the sampling procedures, such as reporting of sample size and location, we were interested in understanding not only how participants were selected into the research but also how *research sites* were selected and in assessing whether research placement could have affected the research findings (i.e. researchers choosing a research site because of previous connections with a certain citification body, NGO involved in a certification programme, certified PO, etc.). This is a critical issue as we consider the selection of research sites as more important for questions of external validity and potential programme and research placement bias than the specific selection of respondents and the size of the samples once sites have been decided. It is also remarkable how many studies of certification tend not to give detailed justification of site selection or give unconvincing reasons, such as sites being suggested by PO leaders (Cramer et al. 2014). We also included a criterion on whether the study reported on the researcher's role/ positionality, while in Waddington et al (2014) this point was addressed by focusing on the authors' conflicts of interest. In the context of qualitative research there may not be explicit conflicts of interest but reflexivity is important given that data analysis processes and interpretation of findings are influenced by researcher's positionality given limited standardisation. Finally, although we have initially included judgements on the extent to which methodological choices were appropriate or not, we found this part of the tool dysfunctional in its application. While checking whether certain methodological elements were included in the research design and reported (e.g. triangulation) was quite straight forward, making judgements about whether the methodological choices were appropriate or not required more time, as well as more experienced coding skills. Given the large amount of included papers and the resources and time constraints of the review, we ended up skipping this stage.

Another difference with Waddington et al (2014) is that we decided a priori not to exclude any studies on the basis of the quality assessment. As described in the previous section, studies that lacked reporting of basic methodological aspects or whose analysis was thin and descriptive without sufficient substantive evidence related to the 'process' question were already excluded during the full text screening stages. Instead, we used

the assessment to provide an account of the quality of evidence used in the synthesis. In that sense, our approach is similar to Atkins et al (2008) and De Buck et al (2017) who applied adapted versions of the CASP instrument to identify the strength and limitations of their evidence base, but not to further exclude studies that already met their basic inclusion criteria. Appraising qualitative evidence without excluding studies on the basis of the assessment was also reported by Lloyd Jones (2004: 276) who comment that this approach allowed 'an understanding of each study on its own terms' and enabled reflections on the ways in which the research methods used by each study shaped 'understandings about the subject of interest'.

We find our experience from assessing qualitative evidence using the CASP tool similar, in the sense that it has enabled a better understanding of how the evidence was produced. As we anticipated, different research approaches performed in different ways in the assessment. Ethnographic studies commonly went into considerable detail in documenting their methodological choices, and although the concept of sampling (in terms of size or standardised criteria of selection) might not apply, relevant information on how the research was conducted was thorough, including issues of reflexivity and triangulation (e.g. Jaffee, 2006; Pollack, 2006; Sutton, 2014). On the contrary, qualitative studies that were rapid assessments focused on gathering perceptions from participants or on understanding the certification process tended to report limited information on methods, usually in the form of a list of methods of data collection with not further information or justification on the selection of research sites, let alone for the sampling of respondents interviewed through focus groups, semi-structured interviews or participatory techniques. Quantitative impact evaluations containing complementary qualitative evidence tended to report even less information on the research procedures that regarded the collection and analysis of qualitative data (e.g. Barham and Weber, 2012; Jena et al, 2012; Subervie and Vagneron, 2013). In many of these cases, qualitative data appeared as ad hoc 'add-ons' for which there was no need to document methods and research process.

Although we did not use the assessment to exclude studies, we did classify the evidence as either high confidence or low confidence. We chose not to produce a synthesised individual quality assessment for each study, but to provide a summary of the quality of the evidence synthesised on the review instead in order to visualise the strengths and limitations of the overall evidence base.

5. Data extraction

Deciding what data to extract for qualitative synthesis is far more challenging that in the case of statistical meta-analysis, where numeric results of experimental or quasi-experimental studies are easily identifiable and extractable (Thomas and Harden, 2008). In qualitative studies, however, the task is more complicated due to varying reporting styles, misrepresentation of primary data or analytical processes as findings, or misuse of quotes and field observations to produce findings (Sandelowski and Barroso, 2002). In short, qualitative data are intrinsically heterogeneous and make standardised procedures of extraction harder to apply. One strategy of dealing with data extraction in qualitative studies is identifying and extracting the 'key concepts' or main

points from each included studies before proceeding with the synthesis (Campbell et al 2003). However, this is not always possible. Thomas and Harden (2008:n.a) describe how while identifying 'data' in the studies was a straightforward task, locating key concepts or 'succinct summaries of finding' was far more challenging, particularly for studies who merely describe and summarise their data, without reaching further analytical depth.

In our case, narrowing the scope of the review to include only 'relevant and substantial' evidence allowed to exclude a large amount of studies with 'descriptive excess' (Sandelowski and Barroso, 2002 citing Lofland and Lofland, 1995) and limited analytical depth. However, our included material drew on different qualitative research approaches and it is true that we had to deal with a variety or reporting styles. Instances of misrepresentations of data as findings or unclear connections between data and claims were also common. To deal with these challenges, we decided to extract data from the entire text of the included studies and not only from the specific study sections in order to avoid missing relevant data due to different reporting styles. At the same time, we opted for being selective in our data extraction and we applied the same logic in including studies to including data from within included studies. This means that we only extracted text from each included study that emanated from primary data and that was both 'relevant and substantial'.

Relevance had to do with our three main areas of interest: the implementation of CS, the distributional dynamics among scheme participants, and contextual barriers and facilitators, especially as these relate to achieving stated goals of CS. For example, there were basic aspects of context like the labour market structures and standards in each country/location; relations and structures in specific value chains both locally and globally; preferences and actions by key value chain actors, which could impact on adoption and effectiveness; the historical trajectories of specific POs and how they could shape internal governance structures and decision making. For this reason, data from qualitative studies that only reported on CS outputs, without providing any insights on implementation or distributional dynamics, or on how the context can shape these outputs, were excluded. Substance had to do with the degree to which the findings within the included studies went beyond simple statements and descriptions and had the power to explain how, why, for whom CS had or did not have certain effects, as well as how contextual factors influenced these effects. Therefore, descriptive data on the impact of CS which were not embedded in the context of the study or were not followed by an analytical discussion were not included for coding. For instance, simple descriptions of the use of social premium, which were commonly found, were excluded from data extraction, unless they were accompanied with some insights on how related decisions on the premium investments were reached, who was able to benefit from these investments (or who was not, and why), how the context played a role in the success or failure of such investments, etc. Additionally, authors' statements or opinions that were not supported by evidence or rich context descriptions were not included for synthesis. Similarly, qualitatively-researched perceptions of CS effects (i.e. farmers' perceptions on the benefits resulting from participation in CS) were not included unless accompanied with factual descriptions or explanations of how or why these perceptions were formed.

Perceptions may be useful in their own right but not so if they do not provide insights into causal mechanisms and key contextual factors.

To illustrate out data extraction strategy we provide two example of included and excluded text extracts. Both extracts are from the same study, which also underlines that qualitative data is not homogenous, but varying qualities of description and analysis can be found within the same study. The following extract from Smith (2010:49) was excluded as despite being relevant in terms of CS implementation (i.e. use of social premium), it did not contain sufficient explanatory power: 'Funds from the Fairtrade Premium had also been used to help combat HIV/AIDS, including co-finance for HIV testing and awareness-raising.' On the contrary the following extract from the same study was included:

'Fairtrade had had another important impact on the social and legal status of Haitians in the Dominican Republic. Funds from the Premium were being used to process passports and working visas, giving them protection from the regular mass expulsions of migrant workers by Dominican authorities. It also reduced the cost of travelling to and from Haiti, as they no longer had to pay 'coyotes' to get them across the border illegally. However, obtaining the right to stay and work in the Dominican Republic did not give automatic rights to workers' children, as Dominican law required children to have a Dominican birth certificate in order to attend school and access other public services. As a result some Haitians were resorting to paying Dominicans to 'adopt' their children so they could obtain a birth certificate and gain full citizenship.' (Smith, 2010: 50)

The above extract was deemed to provide enough substantial relevant evidence on all our key themes of interest for the following reasons: it describes specifically not only how the Fairtrade premium was used (process passports and working visas), but also provides a link with direct effects (provide protection; reduction of travelling costs), and therefore it is was considered relevant for implementation dynamics. Further, it provides insights on how a specific group of workers (illegal immigrants from Haiti) benefited from a certification input and therefore it was relevant for distributional dynamics. Finally, it provides contextual information (expulsions of migrant workers by Dominican authorities) which was deemed important to understand contextual barriers and facilitators that can affect CS effectiveness.

6. Synthesizing qualitative evidence

While methods for synthesising quantitative evidence have become increasingly sophisticated in order to include different study designs in addition to RCT studies synthesis methods for qualitative research have evolved less rapidly (Dixon-Woods et al 2004). Currently, a number of approaches to qualitative synthesis exist, from narrative summaries to grounded theory and meta-ethnography. Each has its own problems and strengths, as Dixon-Woods et al (2004) show, and probably no approach is better than the other, but the

optimal choice depends on the type of the studies to synthesise and the aims of the synthesis (Finlayson and Dixon, 2008).

In our example review we opted for a 'thematic synthesis' approach of the qualitative evidence, as by developed by Thomas and Harden (2008). The approach 'combines and adapts approaches from both meta-ethnography and grounded theory' in order to address process as well as effectiveness questions (Barnett-Page and Thomas, 2009:3). It involves line-by-line coding of the relevant text and then the organisation of these codes into first 'descriptive' and then 'analytical' themes (Thomas and Harden, 2008). The creation of 'analytical themes' that 'go beyond' the findings of the primary studies and generate 'additional concepts, understandings or hypotheses' (ibid: n.a), has similarities to the creation of 'third order interpretations' and the process of translation found in meta-ethnography, while the fact that themes are developed inductively also resonates with grounded theory, Barnett-Page and Thomas (2009) note. This approach to synthesis has been used by Thomas et al (2003) in a systematic review on the barriers and facilitators on healthy easting for children. In the field of international development, Waddington et al (2014) followed the same approach to synthesise qualitative evidence on the barriers and facilitators to the effectiveness of Farmer Field Schools (FFS). The difference between the two reviews is that while Thomas et al (2003) made no use of predetermined themes and allowed the study findings to guide their coding, Waddington et al (2014) combined predetermined themes which resulted from the FFS's ToC with emerging themes from the included studies.

Our approach is similar to Waddington et al (2014) in that we combined predetermined and emerging themes. The approach was therefore driven by our ToC and the previous knowledge of the certification literature by the reviewers involved in the data extraction process. We used the assumptions included in the ToC into core thematic blocks, i.e implementation dynamics; distributional dynamics; other contextual factors, which could be broken down in further sub-themes (e.g. costs of certification, gender dynamics in distribution) more directly linked to critical assumptions in each of the main causal chains. However, we also expected that further sub-themes would emerge in the coding process and like Thomas et al (2003) we kept adding new codes and refining old ones during the coding process, i.e. grounding our revised ToC in the data as they revealed key structures, processes and relations. Additionally, we allowed for every piece of relevant data to be coded under one or multiple codes, a fact which enabled observations of how the different themes interrelate and informed the creation of 'analytical themes' later on.

Coding process

All the included studies were coded using both MS excel and Nvivo. While MS Excel allowed recording basic information on the predetermined themes and managing the bulk of our coding, the flexibility of Nvivo was key in incorporating emerging themes not previously considered. Some of these emerging themes were added as new separate codes, while others were merged with already existing ones, creating new and more refined codes. This allowed us to develop a much more detailed, multi-layered hierarchical tree structure in Nvivo,

while our spreadsheet coding tool remained simpler. Both tools, however, maintained the same basic structure throughout the process.

Table 1 in Annex shows a Matrix illustrating (for a small sub-set of themes) how our initial Excel codes evolved into a restructured, more expanded and detailed 'node'[7] tree in Nvivo. For instance, codes related to certification 'training', 'services' and 'premium' were separated in the initial Excel coding tool. However, during the coding process it became apparent that the three themes were interconnected, with credit, a certification service, often affecting premium payments, for instance. As a result, the three themes got grouped under the term 'Certification inputs' in Nvivo, while at the same time several 'child nodes' emerged under each initially separate theme to capture the variability and complexity of the implementation of CS (e.g. six different child nodes emerged under the 'training' node). As the coding process enlarged our understanding of the reality of CS implementation, we also had to make adjustments in our vocabulary. For example, while initially we had two predetermined codes for the certification premium, financial and social, during the coding process we realised that it was very difficult to isolate the financial premium from the price producers received for certified products, which in its turn was influenced by a number of other factors, such as the efficiency of the cooperative management, the percentage of the total produce sold as certified, the size of the cooperative, etc. As a result the initial theme 'financial premium' was renamed 'price' in the Nvivo tree.

Nvivo also allowed us to easily incorporate and elaborate on new emerging themes which were not initially anticipated, but proved to be significant in illuminating the process question. An example is the role of multi-certification in understanding the CS implementation dynamics, and particularly the combination of organic with social standards. Organic certification in isolation from other certifications was excluded as an intervention from the scope of the review for reasons addressed in the technical report (see Oya et al, 2007:51), but studies covering an eligible certification scheme, commonly Fairtrade, in combination with organic certification were included for synthesis. However, no code for 'organic certification' was included in our initial coding tool, since this was out of the scope of the review. Nevertheless, during the coding process we realised that organic certification was commonly linked to increased costs of production and therefore acted as a barrier to participation also for social standards when those were combined with organic (Milford, 2014; Jaffee, 2006; Abarca-Orozco, 2015). On the other hand, strong demand for organic products also meant that while producers were uncertain to receive the Fairtrade premium, organic markets were more likely to remunerate for the total of the harvest (Valkila, 2009). The Nvivo tree not only allowed us to quickly create a new code and accommodated relevant data, but also to explore the links with other nodes, such as 'costs of certified production', or 'adoption of standards'.

[7] Node is the term that Nvivo uses for 'a collection of references about a specific theme, place, person or other area of interest' (http://help-nv10.qsrinternational.com/desktop/concepts/about_nodes.htm)

Synthesis process

After coding all the primary studies, we proceeded to generation of detailed descriptive themes (Thomas and Harden, 2008). This was done by reviewing the Nvivo coding structure for similarities and differences between the codes, and adjusting its structure by relocating and merging codes accordingly. Then, we summarised the extracted data across studies under each code to produce a descriptive synthesis. This was still close to the extracted original text and did not 'go beyond' the content of the primary studies to generate additional concepts, understandings or hypothesis (ibid).

We then built on the descriptive synthesis to generate 'analytical themes'. This was done by interpreting the descriptive summaries to provide 'new interpretive constructs, explanations or hypotheses' (Thomas and Harden, 2008:n.a) that could inform our process review question. 'Going beyond' the content of the primary studies can be the most controversial part of a qualitative synthesis, as Thomas and Harden underline, since it depends on 'the judgement and insights of the reviewers'. In order to reduce the potential influence of the researcher on the transition from descriptive to analytical themes, each analytical theme was discussed by the three reviewers, who come from different disciplinary backgrounds but have extensive experience in qualitative data analysis and particularly in the specific field of literature covered by this review. The joint work on the synthesis not only contributed to a better calibration of the contributions of different studies to the main core themes but also to the identification of a set of additional sub-themes that provided other contextual information that had not been fully anticipated. As a result, our thematic synthesis does not derive from the interpretation of a single researcher but rather from an iterative dialogue between three researchers.

A key issue that emerged, for instance, from the analytical synthesis was the significance of wealth and resources as a cross-cutting theme across our main areas of interest. In terms of programme implementation, our findings suggest that wealth and resources, and associated sets of power relations, can enhance or prevent entry in the certification process. This is because the adoption of standards required by CS is resource demanding and often depends on the capacity of POs, producers and plantations to bear the extra costs related to certified production (i.e. cost of extra labour and inputs) and marketing (i.e. cost of payment delays). In terms of distributional dynamics, wealth emerged as a crucial factor influencing the ability of producers to benefit from CS, as through the descriptive synthesis it became clear that larger producers, POs and plantation with better access to labour and financial resources were more likely to concentrate benefits. Our analytical synthesis therefore suggested that CS are not generally able to reach and deliver benefits to the smaller and poorer farmers, despite the public discourse of certification related institutions about improving trading conditions for the 'small-scale' and 'economically disadvantaged producers' (i.e. WFTO, 2017; Fairtrade International, 2017) and addressing poverty of 'smallholder' and indigenous farmers (i.e. Utz, 2014; Rainforest-Alliance, 2014).

Integration of findings with quantitative results

As argued in Section X above, an important challenge in this review is the scarcity of 'linked' studies, i.e. of studies that contribute to both the effectiveness and process questions, and particularly the lack of process evaluations. Nonetheless many studies included for the process question contained rich and valuable evidence on barriers and facilitators and more broadly on the contextual factors that matter in the impact of certification on producers and workers. The analytical themes identified in the qualitative synthesis provided an excellent background to go back to the original ToC and the key causal chains contained therein.

In order to integrate the findings of the qualitative synthesis with the evidence from quantitative synthesis, we considered the key causal pathways discussed in the ToC. Specifically we organised the integrated synthesis by moving along the causal chain from intermediate outcomes (producer prices) onto final outcomes (farm and household income) considering the most relevant quantitative and qualitative evidence. We had key barriers and facilitator that affected multiple outcomes but efforts were made to highlight the ways in which the same contextual factors affected specific outcomes. The synthesis was organised around two empirically distinct sets. First, we considered outcomes related to producers' farm profits and revenues, including key components, such as producer prices and yields, both key intermediate outcomes for many CS. We started by summarising the quantitative effects and linked these to a selection of key findings from the qualitative synthesis on implementation and contextual factors that could affect these outcomes. The trick was to have a selection of key findings and to provide a narrative synthesis on how they affected the specific outcomes under consideration. Given how rich and extensive the qualitative synthesis was, the decision over what constituted key barriers and facilitators and which aspects of context mattered most for specific causal chains required an additional exercise of analysis drawing from inputs of two or three reviewers. This analysis also allowed to identify contextual factors that were specific to particular nodes in the causal chain (say, between producer prices and certified farm income) and broader contextual factors that were relevant to various nodes in the causal chain (typically distributional dynamics and PO management and relations cut across different nodes). Second, we focused on labour market outcomes, mainly wages, for which we had a set of clearly defined qualitative themes that were relevant for context and implementation dynamics. In relation to labour standards and specifically wages we had fewer but clearer contextual aspects that could help make sense of the negative effects found in the quantitative synthesis, such as which workers were covered by labour standards and which ones not and the importance of national and local labour market contexts. Then we moved onto key final outcomes for which there was a reasonably good number of studies, i.e. total household income.

Overall, the main empirical substance of the integrated synthesis was concentrated in the first group of outcomes (prices, yields, wages and certified farm income). This was consistent with the consensus in the literature that CS can be more directly linked to these outcomes and that the range of contextual factors is far too wide to establish a clear set of causal links, including barriers and facilitators, in relation to final outcomes. Indeed, it was easier to link some of the most important qualitative themes to that set of outcomes than to

household income, assets and health. The bulk of the qualitative evidence focused on that part of the causal chain. Indeed some of the key insights on implementation dynamics and how benefits were distributed could be clearly linked to the heterogeneity of findings on yields, prices and certified farm income. Studies on labour outcomes also provided sufficient qualitative evidence to help us understand the lack of positive quantitative effects on wages. Some of this qualitative evidence had also been collected for the same quantitative impact evaluations included in the estimation of wage effects, in a rare case of data linking quantitative and qualitative data for the same setting (cf Cramer et al. 2014). It is interesting that there is substantial congruence between the insights from 'linked' evidence in those studies and the contributions from qualitative studies reporting on labour outcomes without quantitative evidence. Therefore, even in the absence of 'linked' quantitative and qualitative data, it is possible to draw on the contextual information provided by 'qualitative-only' studies.

7. Conclusions and recommendations

This article has documented the challenges faced in the process of including a substantial body of qualitative evidence in a mixed-method systematic review of the effects of agricultural certification on the wellbeing of producers and workers. The focus of the paper has been on the various hurdles encountered at every key stage of the systematic review, the options considered to address these hurdles and the solutions devised.

Perhaps one of the most significant challenges was to deal with a very large volume of literature which was hard to screen in early stages, given the lack of detail included in titles and abstracts and the various thematic areas that were relevant to the review. A combination of broad scope and a large and heterogeneous literature generated important challenges during the searching and screening process. In order not to miss out on key contributions in the field, much time had to be devoted to search from multiple sources and to screen in multiple stages.

As argued in this article, criteria for inclusion should be clear-cut but that is not a straightforward task in the case of qualitative research, at least not as clear as with quantitative impact evaluations. By nature, qualitative research is more heterogeneous and does not follow standardised reporting systems. Moreover, it is harder to ascertain how much evidence is relevant from the available studies to decide on whether they can be used for the purposes of synthesis and to combine with quantitative meta-analysis. We had to be flexible while rigorous when screening studies and assessing their substantive contribution to the key thematic components of our qualitative synthesis. This led to a multi-stage screening process, which included full-text review of a very large number of items. The result was a substantial body of literature that was both substantive enough to be included in the review and sufficiently rigorous in terms of reporting on methods. Unfortunately it is possible that some valuable studies were screened out of the review simply because there was only minimal information on methods, which meant that basic inclusion criteria were not met. The experience of our review shows that better reporting is in the interest of authors so their studies can be included in and contribute to systematic reviews.

Qualitative studies were evaluated in terms of their quality and this paper provides a discussion of the main available tools, their usefulness and a series of adaptations that made our assessment of quality more fine-tuned, especially giving due consideration to a range of issues that are important for qualitative researchers, e.g. researchers' positionality, site selection and triangulation. The decisions made during the screening, data extraction and synthesis process were also designed to deal with a largely heterogeneous body of literature spanning the whole range of relevant studies from quick 'fact-finding' commissioned research to in-depth and slow ethnographic work more typical of PhD projects.

The synthesis process was laborious as a result of the volume of studies finally included in the review, but perhaps the most important hurdle was the scarcity of material that could be directly 'linked' to the effectiveness review, i.e to evidence stemming from the quantitative impact evaluations. However, the experience of this review demonstrated that a systematic review of a substantial body of qualitative evidence, even when not directly linked to specific quantitative effects, can shed light on a broad range of relevant issues and provide a sufficiently clear picture of the main contextual factors that matter to understand the how, why and when agricultural certification has positive effects.

Any mixed-methods review that aims to be sufficiently inclusive of qualitative evidence is likely to face challenges in terms of time and resources as it is often hard to anticipate the volume of material that will be found and how laborious the screening and synthesis process will be. Our experience shows that a careful balance must be struck between available resources, the scope of the review, and the explanatory power of the evidence included, but we would encourage future reviewers to secure the resources to include all relevant trustworthy evidence, and not only studies directly linked to the effectiveness review, since there is indeed much to learn from the judicious use of a relatively large volume of qualitative research.

References

Abarca-Orozco, S.J. (2015). Production and marketing innovations in Fair Trade and organic coffee cooperatives in the Cordoba-Huatusco corridor in Veracruz, Mexico. PhD. Iowa State University.

Atkins, S., Lewin, S., Smith, H., Engel, M., Fretheim, A., & Volmink, J. (2008). Conducting a meta-ethnography of qualitative literature: lessons learnt. *BMC medical research methodology*, *8*(1), 21.

Bacon, C.M. (2005). Confronting the coffee crisis: Nicaraguan farmers use of cooperative, Fair Trade and agroecological networks to negotiate livelihoods and sustainability. PhD. University of California, Santa Cruz.

Barham, B.L., and Weber, J.G. (2012). The Economic Sustainability of Certified Coffee: Recent Evidence from Mexico and Peru. *World Development*, 40(6), pp.1269-1279.

Barnett-Page, E. and Thomas, J. (2009). Methods for the synthesis of qualitative research: a critical review. *BMC medical research methodology*, *9* (1), p.59.

Barroso, J., Gollop, C.J., Sandelowski, M., Meynell, J., Pearce, P.F. and Collins, L.J. (2003) The challenges of searching for and retrieving qualitative studies. *Western journal of nursing research*, *25*(2), pp.153-178.

Bates, S., Clapton, J. and Coren, E., 2007. Systematic maps to support the evidence base in social care. *Evidence & Policy: A Journal of Research, Debate and Practice*, *3*(4), pp.539-551

Berg R., C., Denison E. (2012). Interventions to reduce the prevalence of female genital mutilation/cutting in African countries. Campbell Systematic Reviews 2012:9

Bohren, M.A., Vogel, J.P., Hunter, E.C., Lutsiv, O., Makh, S.K., Souza, J.P., Aguiar, C., Coneglian, F.S., Diniz, A.L.A., Tunçalp, Ö. and Javadi, D., Oladapo, O., Khosla, R., Hindin, M., Gülmezoglu, A. (2015). The mistreatment of women during childbirth in health facilities globally: a mixed-methods systematic review. *PLoS medicine*, *12*(6), p.e1001847.

Campbell, R., Pound, P., Pope, C., Britten, N., Pill, R., Morgan, M. and Donovan, J. (2003). Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. *Social science & medicine*, *56*(4), pp.671-684.

Caracelli, V. J., & Cooksy, L. J. (2013). Incorporating Qualitative Evidence in Systematic Reviews: Strategies and Challenges. *New Directions for Evaluation*, *2013*(138), 97-108.Carr-Hill R, Rolleston C, Schendel R. (2016) The effects of school-based decision making on educational outcomes in low- and middle-income contexts: a systematic review. Campbell Systematic Reviews 2016:9

Cartwright, N., 2010. Will this policy work for you? Predicting effectiveness better: how philosophy helps. Presidential Address Philosophy of Science Association. Available from: http://www2.lse.ac.uk/CPNSS/projects/orderProject/documents/Publications/CartwrightPSA.pdf

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

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, 47:2, 108-121

De Buck E, Van Remoortel H, Hannes K, Govender T, Naidoo S, Avau B, Vande veegaete A, Musekiwa A, Vittoria L, Cargo M, Mosler H-J, Vandekerckhove P, Young T. Approaches to promote handwashing and

sanitation behaviour change in low- and middle-income countries: a mixed method systematic review. *Campbell Systematic Reviews* 2017:7.

De Goeij, M.C., Suhrcke, M., Toffolutti, V., van de Mheen, D., Schoenmakers, T.M. and Kunst, A.E., 2015. How economic crises affect alcohol consumption and alcohol-related health problems: a realist systematic review. *Social Science & Medicine*, *131*, pp.131-146.

Dixon-Woods, M. and Fitzpatrick, R., 2001. Qualitative research in systematic reviews: has established a place for itself. *BMJ: British Medical Journal*, *323*(7316), p.765.

Dixon-Woods, M., Fitzpatrick, R., Roberts, K. (2000). Including qualitative research in systematic reviews: opportunities and problems. Journal of Evaluation in Clinical Practice, 7:2, 125-133.

Dixon-Woods, M., Agarwal, S., Young, B., Jones, D. and Sutton, A., 2004. Integrative approaches to qualitative and quantitative evidence. *London: Health Development Agency*, *181*.

Dixon-Woods, M., Sutton, A., Shaw, R., Miller, T., Smith, J., Young, B., Bonas, S., Booth, A. and Jones, D., 2007. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. *Journal of health services research & policy*, *12*(1), pp.42-47.

Doyle, L.H. (2003). Synthesis through meta-ethnography: paradoxes, enhancements, and possibilities. *Qualitative Research*, *3*(3), pp.321-344.

Evans, D., 2002. Database searches for qualitative research. *Journal of the Medical Library Association*, 90(3), p.290

Fairtrade Foundation. (2010). Fairtrade Tea: Early Impacts in Malawi. London: Fairtrade Foundation.

Fairtrade International, (2017). *Journeys to change: Fairtrade theory of change.* [online]. Accessed on January 10, 2017. Available at: http://www.fairtrade.net/fileadmin/user_upload/content/2009/resources/1612-Fairtrade_Theory_of_Change.pdf

Finlayson, K.W., Dixon, A. (2008). Qualitative meta-synthesis: a guide for the novice. Nurse researcher, 15(2), pp. 59-71.

Geertz, C. (1973). Thick description: toward an interpretative theory of culture. In: Geertz, C. *The interpretation of culture: selected essays*. New York: Basic Books.

Giovannucci, D., and Potts, J., with Killian, B., Wunderlich, C., Soto, G., Schuller, S., Pinard, F., Schroeder, K., and Vangeron, I. (2008). *Seeking Sustainability: COSA Preliminary Analysis of Sustainability Initiatives in the Coffee Sector*. Winnipeg, Canada: Committee on Sustainability Assessment.

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 (editors), *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions.* Version 1. Cochrane Collaboration Qualitative Methods Group, 2011. Available from URL http://cqrmg.cochrane.org/supplemental-handbook-guidance

Harden, A., Brunton, G., Fletcher, A. and Oakley, A., 2009. Teenage pregnancy and social disadvantage: systematic review integrating controlled trials and qualitative studies. *Bmj*, *339*, p.b4254.

Harden, A. and Thomas, J., 2005. Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methodology*, *8*(3), pp.257-271.

Jaffee, D.S. (2006). Producing justice: Fair trade coffee, livelihoods and the environment. PhD. The University of Wisconsin - Madison.

Jaffee, D. (2008). 'Better, but not great': the social and environmental benefits and limitations of Fair Trade for indigenous coffee producers in Oaxaca, Mexico. In: Ruben, R., *The Impact of Fair Trade.* Wageningen: Wageningen Academic Publishers, pp.195-220.

Jena, P.R., Chichaibelu, B.B., Stellmacher, T., and Grote, U. (2012). The impact of coffee certification on small-scale producers' livelihoods: a case study from the Jimma Zone, Ethiopia. *Agricultural Economics*, 43(4), pp. 429-440.

Lawry, S, Samii, C, Hall, R, Leopold, A, Hornby, D, Mtero, F. The impact of land property rights interventions on investment and agricultural productivity in developing countries: a systematic review. Campbell Systematic Reviews 2014:1

Livoreil, B., Glanville, J., Haddaway, N.R., Bayliss, H., Bethel, A., Lachapelle, F.F., Robalino, S., Savilaakso, S., Zhou, W., Petrokofsky, G. and Frampton, G., 2017. Systematic searching for environmental evidence using multiple tools and sources. *Environmental Evidence*, *6*(1), p.23

Lloyd Jones, M.L., 2004. Application of systematic review methods to qualitative research: practical issues. *Journal of advanced nursing*, *48*(3), pp.271-278.

Lyon, S. (2007). Maya coffee farmers and fair trade: assessing the benefits and limitations of alternative markets. *Culture & Agriculture*, 29(2), pp.100-112.

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, 445-455 Milford, A.B. (2014). Co-operative or coyote? Producers' choice between intermediary purchasers and Fairtrade and organic co-operatives in Chiapas. *Agriculture and Human Values,* 31(4), pp. 577-591.

McKinnon, M.C., Cheng, S.H., Dupre, S., Edmond, J., Garside, R., Glew, L., Holland, M.B., Levine, E., Masuda, Y.J., Miller, D.C. and Oliveira, I., 2016. What are the effects of nature conservation on human well-being? A systematic map of empirical evidence from developing countries. *Environmental Evidence*, *5*(1), p.8

Molina E, Carella L, Pacheco A, Cruces, G, Gasparini L. Community monitoring interventions to curb corruption and increase access and quality of service delivery in low- and middle-income countries. Campbell Systematic Reviews 2016:8.

Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, Moore L, O'Cathain A, Tinati T, Wight D, Baird J. *Process evaluation of complex interventions: Medical Research Council guidance*. MRC Population Health Science Research Network, London, 2014.

Munro, S.A., Lewin, S.A., Smith, H.J., Engel, M.E., Fretheim, A. and Volmink, J., 2007. Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS medicine*, *4* (7), p.e238.

Naylor, L.B. (2014). Decolonial autonomies: Fair trade, subsistence and the everyday practice of food sovereignty in the highlands of Chiapas. PhD. University of Oregon.

Nicaise, P., Lorant, V. and Dubois, V., 2013. Psychiatric advance directives as a complex and multistage intervention: a realist systematic review. *Health & social care in the community*, *21*(1), pp.1-14.

O'Campo, P., Kirst, M., Tsamis, C., Chambers, C. and Ahmad, F., 2011. Implementing successful intimate partner violence screening programs in health care settings: evidence generated from a realist-informed systematic review. *Social science & medicine*, *72*(6), pp.855-866.

Oliver, S., Garner, P., Heywood, P., Jull, J., Dickson, K., Bangpan, M., Ang, L., Fourman, M. and Garside, R., 2017. Transdisciplinary working to shape systematic reviews and interpret the findings: commentary. *Environmental Evidence*, *6*(1), p.28.

Oya C, Schaefer F, Skalidou D, McCosker C, Langer L. (2017). Effects of certification schemes for agricultural production on socio-economic outcomes in low- and middle- income countries: a systematic review. *Campbell Systematic Reviews* 2017:3

Pawson, R., Greenhalgh, T., Harvey, G., Walshe, K., 2005. Realist review- a new methods of systematic review designed for complex policy interventions. Jounral of Health Services Research and Policy, Vol. 10:1, pp: 21-34

Pawson, R., Tilley, N. (2004). Realist Evaluation. Realist evaluation. *Changes*.

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: myths and misconceptions. BMJ, vol. 322, 98-101.

Pollack, N. (2006). Women's empowerment in Fair Trade coffee co-operatives in Oaxaca, Mexico. Masters. Saint Mary's University.

Popay, J. ed., 2006. Moving beyond effectiveness in evidence synthesis: Methodological issues in the synthesis of diverse sources of evidence. National Institute for Health and Clinical Excellence

Popay, J., Rogers, A., Williams, G. (1998). Rationale and Standards for the Systematic Review of Qualitative Literature in Health Services Research. Qualitative Health Research, 8:3, 341-351.

Pullin, A.S., Bangpan, M., Dalrymple, S., Dickson, K., Haddaway, N.R., Healey, J.R., Hauari, H., Hockley, N., Jones, J.P., Knight, T. and Vigurs, C., 2013. Human well-being impacts of terrestrial protected areas. *Environmental Evidence*, *2*(1), p.19

Rainforest Alliance. (2014). *Rainforest Alliance Certified Cocoa.* [online]. Accessed on January 10, 2017. Available at: http://www.rainforest-alliance.org/articles/rainforest-alliance-certified-cocoa

Samii C, Lisiecki M, Kulkarni P, Paler L, Chavis L. Effects of Decentralized Forest Management (DFM) on Deforestation and Poverty in Low and Middle Income Countries: A Systematic Review Campbell Systematic Reviews 2014:10

Sandelowski, M. and Barroso, J., 2002. Finding the findings in qualitative studies. *Journal of nursing scholarship*, *34*(3), pp.213-219.

Sandelowski, M., Docherty, S. and Emden, C., 1997. Focus on qualitative methods. Qualitative metasynthesis: issues and techniques. *Research in nursing and health*, *20*, pp.365-372.

Schwandt, T.A. (1996) Farewell to criteriology. *Qualitative Inquiry*, *2*(1), pp.58-72.

Sen, D. (2009). From illegal to organic: Fair trade-organic tea production and women's political futures in Darjeeling, India. PhD. Rutgers, The State University of New Jersey.

Sen, D. (2014). Fair trade vs. Swaccha Vyapar: women's activism and transnational justice regimes in Darjeeling, India. *Feminist Studies*, 40(2), pp.444-472.

Setrini, G. (2011). Global niche markets and local development: clientelism and Fairtrade farmer organizations in Paraguay's sugar industry. PhD. Massachusetts Institute of Technology.

Shreck, A. (2002). Just bananas? A Fair Trade alternative for small-scale producers in the Dominican Republic. PhD. Colorado State University.

Smith, S. (2010). Fairtrade Bananas: A Global Assessment of Impact. University of Sussex: Institute of Development Studies.

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

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L., 2003. Quality in qualitative evaluation: a framework for assessing research evidence.

Staib, P.W. (2012). Coffee and the countryside: Small farmers and sustainable development in Las Segovias de Nicaragua. PhD. The University of New Mexico.

Subervie, J., and Vagneron, I. (2013). A Drop of Water in the Indian Ocean? The Impact of GlobalGap Certification on Lychee Farmers in Madagascar. *World Development*, 50, pp. 57-73.

Sutton, S. (2014). Voice, Choice and Governance: The Case of Tanzania's Fairtrade Co-operatives. PhD. Queen Mary, University of London.

Thomas J, Sutcliffe K, Harden A, Oakley A, Oliver S, Rees R, Brunton G, Kavanagh J (2003) Children and Healthy Eating: A systematic review of barriers and facilitators. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

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

TWIN. (2013). Empowering Women Farmers In Agricultural Value Chains. London: TWIN.

UTZ. (2016). *UTZ Impact Report 2016*. Amsterdam, The Netherlands: UTZ.

UTZ. (2014). *How can UTZ address poverty in cocoa farming?*. [online]. Accessed on Janury 10, 2017. Available at:https://www.utz.org/better-business-hub/strengthening-your- reputation/prosperity-for-cocoa-farmers-just-around-the-corner/

Valkila, J. (2009). Fair Trade organic coffee production in Nicaragua: Sustainable development or a poverty trap? *Ecological Economics*, 68, pp.3018-3025.

Waarts, Y., Ge, L., Ton, G., van der Mheen, J. 2013. A touch of cocoa. Baseline study of six UTZ-Solidaridad cocoa projects in Ghana. LEI Wageningen UR

Waarts, Y., Ingram, V., Liderhof, V., Puister-Jansen, L., and van Rijn, F, and Aryeetey, R. (2016). *Impact of UTZ certification on cocoa producers in Ghana, 2011 to 2014*. LEI Wageningen UR.

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

Walters, L.A., Wilczynski, N.L. and Haynes, R.B., 2006. Developing optimal search strategies for retrieving clinically relevant qualitative studies in EMBASE. *Qualitative Health Research*, *16*(1), pp.162-168.

Weiss, C.H., 1997. Theory- based evaluation: Past, present, and future. *New directions for evaluation*, *1997*(76), pp.41-55.

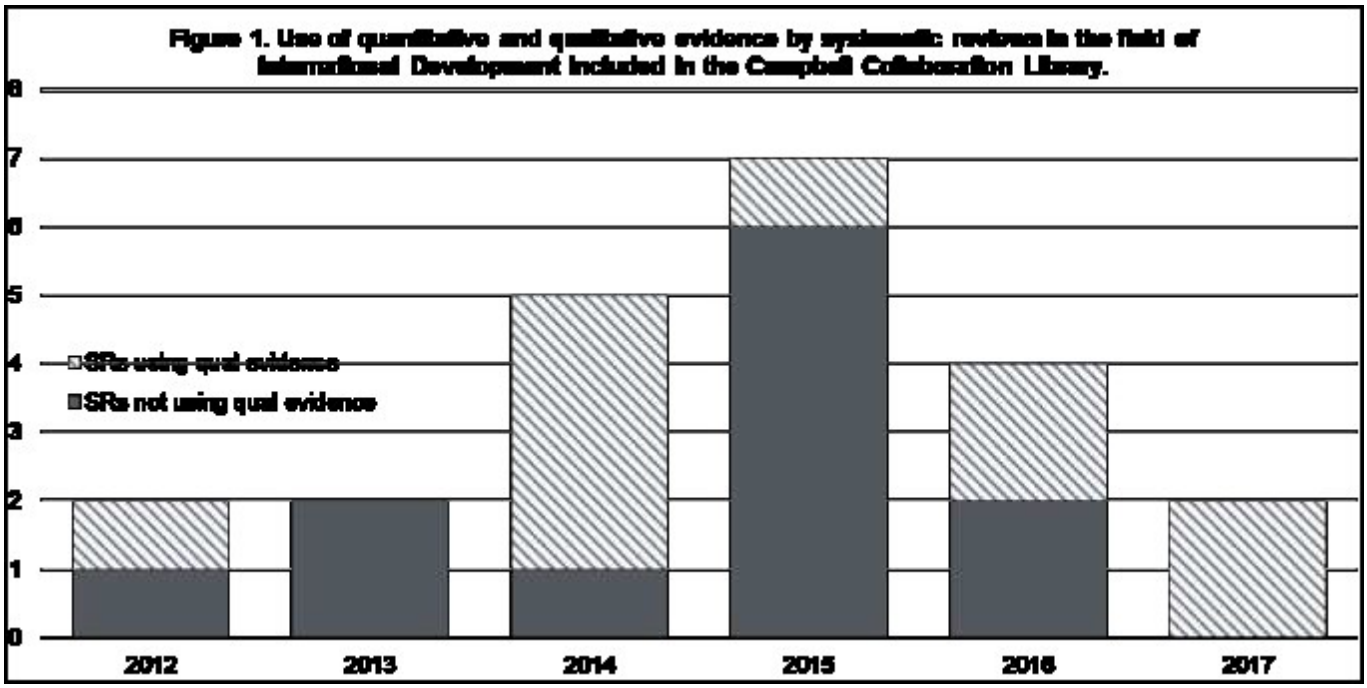
White, H., 2009. Theory-based impact evaluation: principles and practice. *Journal of development effectiveness*, 1(3), pp.271-284.

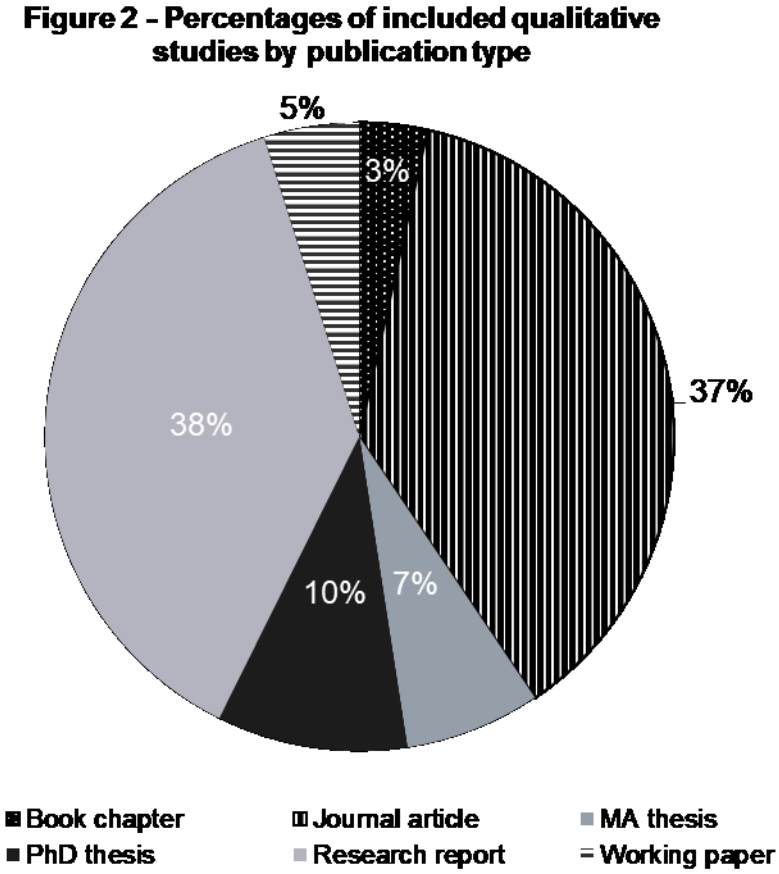
White, H., Waddington, H. (2012). Why do we care about evidence synthesis? An introduction to the special issue on systematic reviews, *Journal of Development Effectiveness*, 4:3, 351-358.

Wong, G., Greenhalgh, T., Pawson, R. 2010. Internet-based medical education: a realist review of what workds, for whom and in what circumstances. BMC Medical Education 2010, 10:12.

World Fair Trade Organization. (2017). *10 Principles of Fair Trade.* [online]. Accessed on January 10, 2017. Available at: http://wfto.com/fair-trade/10-principles-fair-trade

Tables and Figures





Annex

Table 1 – Illustration of tool for synthesis process

	Excel coding tool		Nvivo coding tool			
	Theme	**Description**	**Theme**	**1st level sub-theme**	**2nd level sub-theme**	**3rd level sub-theme**
Implementation Dynamics	Training	Material related to training and new practices (i.e. good agricultural practices)	Certification Inputs	Training & new practices	- Certified markets - Farm management - GAP - Negotiating skills - cooperative/ PO management - Training providers	
	Services	Material related to cetification related services (i.e. distribution of inputs, such as chemiclas and fertilisers; credit services; etc)		- Floor price - Labour standards - Long-term contracts - Minimum wage - Prepayment and credit		
	Financial premium	Material related to financial premium payments & use		Price		
	Social premium	Material related to social premium payments & use		Social Premium		
	Costs	Material related to financial premium use (i.e. amount of money paid to producers on top of the market price)	Costs of certified production	- Certification fees - Paperwork - Quality requirements - Labour requirements	- Household labour	-Female labour - Child labour
					- Wage labour	
	Monitoring	Material related to implementaion costs of certification programmes	Monitoring & Auditing	- External inspection - Internal inspection	- Transparency - Accountability	