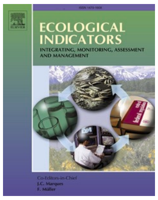
Ecological Indicators 125 (2021) 107490

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Progress and pitfalls: A systematic review of the evidence for agricultural sustainability standards

Rebecca Traldi

University of Maryland College Park, Department of Geographical Sciences, 7251 Preinkert Drive, College Park, MD 20742, United States

ARTICLE INFO

Keywords:
Voluntary sustainability standards
Agriculture
Impact evaluation
Evidence synthesis

ABSTRACT

Over the last decade, there have been increasing calls for robust impact evaluations of voluntary agricultural sustainability standards (VSS's). In response, this study reviews the literature regarding 13 major agricultural standards, asking: where are certified crops being studied? Which sustainability outcomes and indicators are measured? And finally, what does the current evidence base suggest about VSS outcomes? The analysis of 45 peer-reviewed articles suggests a mismatch between what is certified and what is studied. Some crops and standards are over-represented in the literature as compared to their amount of certified production (e.g. coffee and Fairtrade certification), while others are under-represented (cotton, sugar, cocoa, soy, and palm oil, in addition to Organic certification). The review also identifies countries which appear to be under-represented in the literature, including Brazil, Australia, Malaysia, the Ivory Coast, and the United States. When measuring success, economic indicators are the most frequently evaluated, and only 20% of studies analyze economic, social, and environmental indicators simultaneously. When grouped by case, the indicator results tend to be positive on average (51%), followed by no difference (41%) and negative (8%) outcomes. There are no significant differences among sustainability pillars in terms of the average proportion of positive and negative results. These findings should be interpreted carefully, since the evidence base is heavily weighted towards coffee certification (75% of cases analyzed), and impacts are highly context dependent. Finally, the review identifies best practices in conducting robust evaluations, including the importance of addressing sustainability trade-offs and appropriately measuring environmental outcomes. While significant gaps remain, the findings indicate an increase in research credibly measuring VSS impacts.

1. Introduction

Agricultural production is critical for humanity's survival. While agriculture provides an indispensable service, it also results in serious consequences for environmental and social sustainability outcomes. Agriculture is known to be a key driver of negative environmental impacts, including deforestation and subsequent impacts on wildlife habitat and greenhouse gas emissions, nutrient imbalances due to intensive fertilizer application and other practices, and impacts on soil and water resources (Foley et al., 2011, Vitousek et al., 2009). Production can also have a range of implications for human well-being, from changes in livelihoods activities to violations of labor regulations, child labor, and forced labor, among other issues (O'Rourke, 2014, Rasmussen et al., 2018).

A variety of interventions have been implemented to stem these negative externalities and promote resource stewardship and benefits for local communities. One significant category of interventions is voluntary sustainability standards, or VSS's. Voluntary sustainability standards first came to the forefront in the 1980s, with standards like Organic (IFOAM) and the Rainforest Alliance (Potts et al., 2014). Voluntary standards are based on the idea that a combination of positive incentives (e.g. price premiums for producers and/or provision of other services), training and awareness building, clear and consistent criteria for success, and a market-based approach can join forces to boost sustainability (Smith et al., 2019). Most voluntary standards outline requirements related to social, economic, and environmental sustainability, although the specific principles, criteria and indicators vary between the standards, as illustrated by Table 1. There are hundreds of these standards globally, including in forestry and seafood sectors (IISD, 2015). For agriculture, there are about 13 standards that are the most widely adopted and recognized by the international community (Willer et al., 2019). According to recent research, agricultural standards represent about 1.1% of global agricultural area, although their production is not equally distributed among regions (Tayleur et al.,

E-mail address: rtraldi@terpmail.umd.edu.

https://doi.org/10.1016/j.ecolind.2021.107490
Received 22 August 2020; Received in revised form 15 December 2020; Accepted 29 January 2021
Available online 23 February 2021

2017).

While there are slight variations among the standards, they generally operate in the following way: producers voluntarily commit to improving their sustainability practices by adhering to the principles and criteria of the standard (e.g. improving farm efficiency, implementing conservation measures, or ensuring social safeguards). They complete the transition with support from the standard and possibly other stakeholders, depending on the local context. Their operations are then audited by an impartial third party. Finally, their product is "certified", and may be labeled with the appropriate standard seal, depending on the certification (Milder et al., 2015).

Many different stakeholders participate in voluntary standards design and implementation, including non-governmental organizations, private companies, and industry associations. Standards are also a key sustainability tool for many large international brands who source significant amounts of agricultural raw materials. Companies like Coca-Cola, PepsiCo, and many others have committed to sourcing certified crops as part of their sustainability strategies (Smith et al., 2019).

1.1. The need for empirical evaluation of certification's impacts

Over the last decade, there has been increasing interest in

Table 1
Overview of 13 major international agricultural voluntary sustainability standards and their key characteristics, principles, and criteria[1]. There is overlap between sustainability pillars within principles and criteria. Data is from program websites, the ITC Standards Map (2015), Potts et al., 2014, and Willer et al., 2019.

Standard	Crop specific?	Process-based vs. performance-based[2]	Year initiated	Environmental principles and criteria (P&C)	Social P&C	Economic P&C
Better Cotton Initiative	Yes	Process-based	2005	Crop protection, water stewardship, soil health, biodiversity, responsible land use	Decent work conditions	Fiber quality, management systems
Bonsucro (sugarcane)	Yes	Performance-based	2008	Manage biodiversity and ecosystem services, additional biofuel requirements under EU renewable energy directive	Obey the law, respect human rights and labor standards	Production efficiency, continuously improve key areas of the business
Common Code for the Coffee Community (4C)	Yes	Process-based	2006	Biodiversity, energy, soil management, waste management, water management	Work and labor rights, working conditions, gender, health and safety	Profitability and productivity, capacity development, record keeping, market access/ information, quality, traceability
Cotton Made in Africa	Yes	Process-based	2005	Responsible land use, enhance biodiversity, and protect climate and environment; GMO-free cotton, care for water and soil; minimize adverse impacts of crop protection	Responsible business conduct, support smallholder farmers, decent work, respect children’s’ rights and gender equality	Effective management systems; access to high quality inputs and pre-financing; increase productivity and fiber quality; improving living conditions and resilience
Fairtrade	No	Process-based	1997	Agricultural practices e.g. agrochemicals, waste, soil and water, GMOs	Social development, e.g. organizational transparency, worker rights and security, working conditions	Required minimum price and/or price premium (the latter is invested in quality of life improvements), pre-financing
Global Good Agricultural Practices (GAP)	No	Process-based	1997	Waste and pollution management, conservation	Worker health, safety, and welfare, complaints management	Site management, record-keeping, hygiene, recall procedure
Organic cropland (IFOAM)	No	Process-based	1972	Organic ecosystems, crop production	Social justice	Processing and handling
Proterra	No	Process-based	2012	Biodiversity conservation, effective env. management; no GMOs; pollution and waste mgmt.; water mgmt.; GHG and energy; adoption of good ag. practices	Compliance with law; human rights and responsible labor practices; responsible relations with workers & community	Traceability and chain of custody
Rainforest Alliance/ Sustainable Agricultural Network	No	Process-based; some outcome criteria (e.g. specific native vegetation thresholds)	1987	Biodiversity conservation, natural resource conservation	Improved livelihoods and human well-being (e.g. working conditions)	Effective planning and management
Roundtable on Responsible Soy	Yes	Process-based	2006	Environmental responsibility, good agricultural practices	Legal compliance, responsible labor conditions & community relations	Good business practices
Roundtable on Sustainable Biomaterials[3]	Yes	Process-based	2007	GHG emissions, conservation, soil, water, air	Legality, human and labor rights, rural development, local food security, land rights	Planning, monitoring and continuous improvement; use of technology, inputs, and waste management
Roundtable on Sustainable Palm Oil	Yes	Process-based	2004	Protect, conserve, and enhance ecosystems and the environment	Behave ethically and transparently; operate legally; respect human rights; support smallholder inclusion; respect workers' rights and conditions	Optimize productivity, efficiency, positive impacts, and resilience
UTZ[4]	No	Process-based	2002	Soil, waste, water, biodiversity, energy	Labor rights, health and safety, employment conditions, human rights	Price premiums

[1] Certifies several crops, residues, and associated feedstocks, including sugar cane, waste starch from wheat, coconut, *brassica carinata*, jatropha and corn.
[2] UTZ merged with Rainforest Alliance in 2018. Since this is relatively recent, UTZ is treated as a separate standard in this review.
[3] This table is intended to provide an overview, rather than a comprehensive list of all standard criteria. C.A.F.E Practices and Bird Friendly certifications are not covered here. For a more detailed overview regarding the environmental coverage of these standards, see Tayleur et al., 2017.
[4] Process-based standards outline practices that must be implemented, but not specific outcomes; performance-based standards specify actual outcomes to be achieved. Both approaches can exist within one standard (Potts et al., 2014).

understanding the efficacy of voluntary standards in achieving agricultural sustainability objectives. In 2011, Blackman and Rivera emphasized the limited number of robust studies of the impact of certification, arguing that the majority did not rely on research designs which could reasonably indicate causation. This was largely due to inadequate incorporation of a "counterfactual" approach, or a test of what may have happened in the absence of the certification (Blackman and Rivera, 2011). In practice, measuring the counterfactual outcome typically requires comparison between a "treatment" group that has been certified, and a "control" group that has not been certified (Blackman and Rivera, 2011).

Since then, there have been significant developments in assessing certification's outcomes. Notably, there have been two systematic reviews focused on the impacts of agricultural certification in the last five years, published by Oya et al. (2018) and DeFries et al. (2017). The former focuses primarily on socioeconomic outcomes for producers, while the latter looks at outcomes across all pillars of sustainability. These reviews have enabled a clearer picture of the evidence regarding certification's impacts, and set out structured protocols for identifying, reviewing, and selecting studies for inclusion. They also indicated a growing but still imbalanced evidence base, since over 80% of the studies included in DeFries et al. (2017) were focused on coffee certification.

The primary statistical methods utilized in robust impact evaluations of certification include multivariate regression, matched pair comparisons, propensity score matching, instrumental variable approaches, and difference-in-difference methods (DeFries et al., 2017, Ferraro and Hanauer, 2014). Each of these methods necessitates consideration of various assumptions. For example, matching designs require that selection into the certification program occur only due to observed variables (Bolwig et al., 2009). The instrumental variable approach requires identification of an instrument that is correlated with the treatment variable, but not directly correlated with outcomes of interest (Chiputwa and Qaim, 2016). Broadly, these methods fall into the realm of "quasi-experimental" research designs (see e.g. Ferraro, 2009, and Butsic et al., 2017).

There is considerable interest in evaluating the impacts of certification from sustainable development practitioners, evidenced by a growing amount of grey literature. For example, many non-government organizations have published their own reports to distill insights on outcomes of certification (Komives et al., 2018, Petrokofsky and Jennings, 2018). In 2019, an online database called Evidensia was launched with the explicit goal of making certification impact studies widely available (evidensia.eco).

There are also parallels between empirical work evaluating VSS certification, and the evaluation of other sustainability interventions. These evaluations depend on the use of indicators which appropriately measure relevant sustainability outcomes, and are important for a wide range of disciplines from both a theoretical and applied perspective. For example, indicators have been used to assess best management practices and conservation on farmlands (Targetti et al., 2014, Last et al., 2014, Garibaldi et al., 2017, Latruffe et al., 2016), the socioeconomic and environmental impacts of protected areas (Naidoo et al., 2019), and the outcomes of ecosystem management (Breslow et al., 2016, Breslow et al., 2017).

Despite recent developments, substantial gaps remain in our understanding of the impact of agricultural certification. Given the severity of current sustainability challenges, stakeholders may wonder whether changes are needed to the certification model, or question whether and how these interventions should be integrated with other activities at the landscape level (Tscharntke et al., 2015). It's also unclear to what extent general lessons learned can be inferred from study results to date, since evaluations are not representative of all certifications, production systems, or regions globally.

This study responds to these persistent questions regarding the impacts of voluntary sustainability standards for agriculture at the plot, farm, and household level, building upon previous reviews. Specifically, it addresses the questions:

1) Where are certifications being studied, and how does that compare to the extent of certification globally?

2) Which pillars of sustainability are included in evaluation studies, and which indicators are used to measure success?

3) What does the current evidence base suggest regarding outcomes of voluntary agricultural sustainability standards?

The analysis complements the literature on impacts of agricultural VSS's in a few ways. First, over half of the studies assessed here have not been included in previous reviews, due to their recent publication. Since the empirical evaluation of VSS's is a rapidly developing field, we can now ask more nuanced questions regarding research gaps than was previously possible. Second, this analysis provides a new level of detail regarding the distribution of impact evaluations of certification by crop, certification, and country, and identifies specific research gaps and needs. Third, by blending aspects of quantitative and qualitative systematic review, this study synthesizes information about methods to measure success. This enables identification of best practices that can help inform consistent and clear outcome indicators in future impact evaluations. In addition to the literature regarding VSS's, these insights can be applied more broadly to studies measuring the sustainability outcomes of environmental interventions.

2. Materials and methods

There were five main stages of the literature review approach. First, a search was conducted to identify articles to screen for inclusion in the analysis. Second, each article was assessed based on a predefined list of inclusion criteria, based on previous reviews (primarily DeFries et al., 2017; Oya et al., 2018), and an initial list of articles for inclusion was prepared. Third, each of these studies was read and evaluated for key insights, including research design and main findings. Fourth, a detailed indicator table was developed for each study, following DeFries et al., 2017. Finally, findings were synthesized with other data sources to address the research questions and distill key areas of progress and gaps, and descriptive statistics were calculated. More information on each stage of the process is provided below.

This study provides a methodical, replicable, and transparent approach to collecting evidence on VSS outcomes. This is achieved by clearly outlining the search process, inclusion criteria, and evaluation method (Siddaway et al., 2019). A schematic representing the key methodological stages and following PRISMA guidelines (Moher et al., 2009) is shown in Fig. 1. A PRISMA checklist can be found in the Supplementary Materials.

2.1. Identifying articles

Literature searches were conducted through the following sources: 1. Publications selected by recent reviews, 2. All agricultural articles housed in the Evidensia database as of November 2019, and standardized searches through 3. Google Scholar, and 4. Web of Science. The Web of Science search terms included (impact of sustainability certification AND agricultur*), which returned 135 articles as of January 2020. A small number of sources were identified through citations within research articles from the initial search. In total, 240 articles were screened – 135 identified through Web of Science, and 116 identified through other sources listed above. All article searches were conducted only in English, and only English-language articles are included in the analysis. The article list was finalized in May 2020.

2.2. Screening articles

The first round of screening relied primarily on the abstract. Articles which did not include a counterfactual, quasi-experimental research design to assess the impact of certification were eliminated. In cases

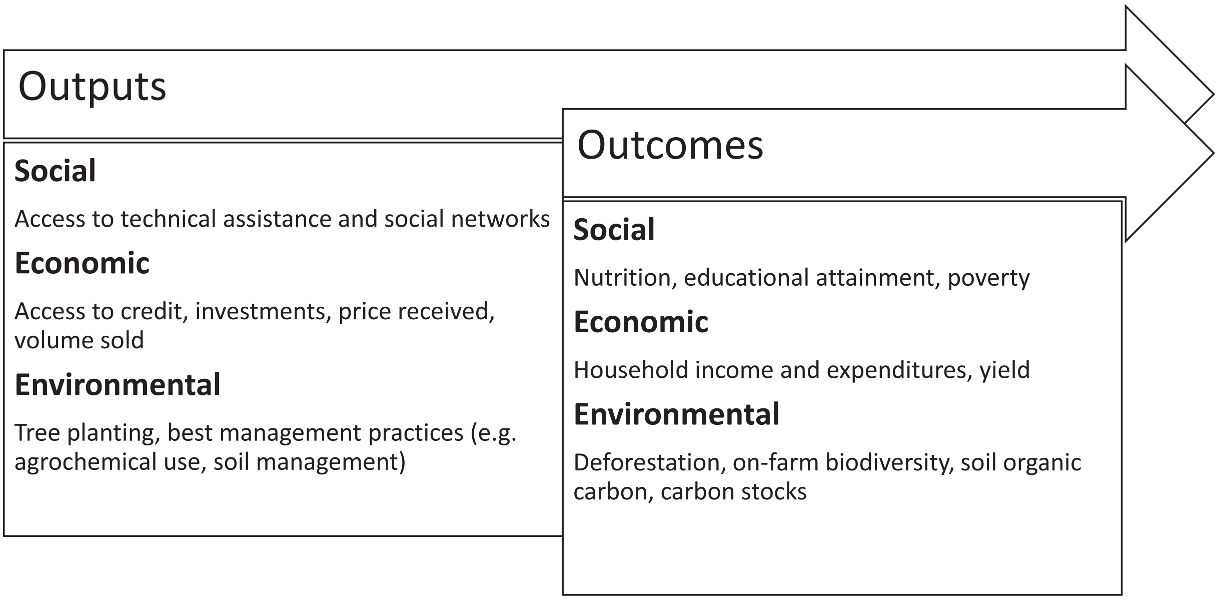

Outputs
Social
Access to technical assistance and social networks
Economic
Access to credit, investments, price received, volume sold
Environmental
Tree planting, best management practices (e.g. agrochemical use, soil management)
Outcomes
Social
Nutrition, educational attainment, poverty
Economic
Household income and expenditures, yield
Environmental
Deforestation, on-farm biodiversity, soil organic carbon, carbon stocks

Fig. 1. Outputs versus sustainability outcomes for voluntary sustainability standards. Standards' outcome goals vary; the figure presents illustrative examples.

where the abstract did not provide sufficient information to ascertain this, the article was read to determine the methodological approach and the extent to which selection bias and group equivalence between treatment and control groups were addressed. Articles that were previously characterized as having high bias (e.g. were noted as "Critical Bias" by Oya et al., 2018's assessment or designated as high bias by Komives et al., 2018) were not included in the analysis. Four articles were identified after the article list was finalized, and thus were not included (Dietz et al., 2018, Dietz et al., 2020, Sellare et al., 2020, Santika et al., 2020; the first two focus on coffee, the latter two on cocoa and palm oil respectively). For an overview of the reviewed articles, see Table 2. Among the included articles, 19 have been included in previous quantitative reviews, and 26 have not.

2.3. Eligibility, analysis, and synthesis

Each study was read and coded for various characteristics, including the region, country, certification and crop studied, the research design and sample size, the pillars of sustainability included in the analysis, and key findings. Some initial characteristics that were appealing to extract from studies were not consistently available – for example, the years since certification was obtained for the study population. At this stage, a few articles initially selected for inclusion were eliminated for various reasons, e.g. methodological approach, resulting in an n size of 45 articles and 66 cases. Several studies include more than one case (multiple certifications in one country, or more than one country of analysis).

For each indicator, the pillar of sustainability and general theme was recorded, as well as the indicator itself, and the results, noting no difference, positive, and negative outcomes (see Table 3). If the significance level of an indicator was uncertain after reviewing a study's results tables, the corresponding author was contacted to clarify. As Fig. 2 illustrates, outputs relate to direct activities resulting from an intervention, while outcomes refer to longer term goals including changes in behavior, social, and environmental conditions (Mascia et al., 2014). Indicators which were contextual in nature or unrelated to a certification outcome were classified as "Other" rather than being incorporated within a sustainability pillar. Overall, 777 indicators were included in the review (see Supplementary Materials for the full indicator table).

To answer the research questions, summary statistics were calculated for key elements of the included studies, including indicator variables. These results were synthesized with other data sources – for example, for question 1, findings were synthesized with insights from Willer et al., 2019 and Tayleur et al., 2017 on the global extent of certification. Finally, indicators grouped by sustainability pillar and by theme (e.g. income, productivity, best management practices) were tested for significant differences using the Kruskal-Wallis analysis of variance in R (R Core Team, 2013).

3. Results

3.1. Where are certifications being studied, and how does that compare to the extent of certification globally?

To answer the first research question, attributes of the studies are compared with information on global certification extent from Willer et al., 2019 and Tayleur et al. (2017 and Tayleur et al. (2018). After determining the proportional coverage of studies for key certifications and crops, several key findings emerge. First, there's a mismatch between what is certified and what is studied. As shown in Tables 4 and 5, certain crops and certifications are under-represented in the literature – cotton, sugar, cocoa, soy, and palm oil are notable examples, as well as Organic certification. Other crops and certifications are over-represented in the literature, including coffee (a point made in previous reviews), as well as Rainforest Alliance, Fairtrade, and UTZ certification. It should be noted that studies on multiple certification are not reflected in Table 5. However, studies of multiple certification occur for the most commonly evaluated standards (e.g. Fairtrade, UTZ, and Rainforest Alliance), so if included would only magnify current trends.

Interestingly, the top five understudied certified crops – cotton, sugar, cocoa, soybeans, and palm oil – are all known to drive significant and urgent sustainability issues. These range from severe labor risks, intensive pesticide and water use, deforestation, and land conversion, to additional human rights and farmer livelihoods issues. Research on these crops appears to be increasing, though, as all of the identified studies which evaluate them were published in 2016 or later. As shown in Table 4, cotton has the most significant discrepancy between its estimated certified area (over 5 million hectares, about 22% of total certified area) and study coverage. Only one study was identified assessing the impact of the Better Cotton Initiative, the predominant cotton certification standard (Zulfiqar and Thapa, 2016). Based on this review, it is possible that two of the most underrepresented certified

Table 2
Studies included in this review, in order of publication year. Some studies include multiple cases or analyses of certification (e.g. multiple crops, countries, or certifications). Combined certification is indicated with a hyphen (e.g. “Fairtrade-Organic”).

Study	Location(s)	Crop	Certification(s)	Methods	Included in previous review[1]?
Arnould et al., 2009	Nicaragua, Peru, and Guatemala	Coffee	Fairtrade	OLS and binomial logistic regression	Yes
Bolwig et al., 2009	Uganda	Coffee	Organic	OLS regression and maximum likelihood estimation	Yes
Fort and Ruben, 2009	Peru	Banana	Fairtrade	Propensity score matching, probit regression	Yes
Ruben et al., 2009	Costa Rica	Coffee, Banana	Fairtrade	Propensity score matching	Yes
Jones and Gibbon, 2011	Uganda	Cocoa	Organic	Instrumental variable and multivariate analysis	No
Ruben and Zuniga, 2011	Nicaragua	Coffee	Fairtrade	Propensity score matching, probit regression	Yes
Weber, 2011	Mexico	Coffee	Fairtrade-Organic	Instrumental variable, probit regression	Yes
Blackman and Naranjo, 2012	Costa Rica	Coffee	Organic	Propensity score matching, probit regression	No
Colen et al., 2012	Senegal	Fruit and vegetables	Global GAP	Panel data analysis and OLS	Yes
Jena et al., 2012[2]	Ethiopia	Coffee	Fairtrade-Organic	Propensity score matching and OLS	Yes
Ruben and Fort, 2012	Peru	Coffee	Fairtrade	Propensity score matching, probit regression	Yes
Kleemann and Abdulai, 2013	Ghana	Pineapple	Organic	Propensity score matching, Endogenous switching regression	No
Rueda and Lambin, 2013	Colombia	Coffee	Rainforest Alliance	Pair-matched case-control	Yes
Takahashi and Todo, 2013[3]	Ethiopia	Coffee	Rainforest Alliance	Propensity score matching, difference-in-difference panel	Yes
Schoonhoven-Speijer and Ruben, 2014	Kenya	Coffee	UTZ	Multiple regression, logistic regression to control for between-group differences	Yes
Takahashi and Todo, 2014	Ethiopia	Coffee	Rainforest Alliance	Probit model/PSM	No
Chiputwa et al., 2015[4]	Uganda	Coffee	Fairtrade, UTZ, Organic	Propensity score matching	Yes
Elbers et al., 2015	Uganda	Coffee	UTZ	Difference-in-difference	Yes
Jena et al., 2015	Nicaragua	Coffee	Fairtrade, Organic	Propensity score matching, endogenous switching regression	No
Rueda et al., 2015	Colombia	Coffee	Rainforest Alliance	Remote sensing analysis, Pair matched case-control	Yes
Cattau et al., 2016	Indonesia	Palm oil	Roundtable on Sustainable Palm Oil	Propensity score matching, analysis of MODIS data	No
Caudill and Rice, 2016[5]	Mexico	Coffee	Bird Friendly	Poisson regression	No
Chiputwa and Qaim, 2016[6]	Uganda	Coffee	Fairtrade, UTZ, Organic	Instrumental variable approach + regression	Yes
Ibanez and Blackman, 2016	Colombia	Coffee	Organic	Matched difference-in-difference	No
Karki et al., 2016	India	Coffee	Fairtrade	Panel data analysis, endogenous switching and quantile regression	No
Qiao et al., 2016	China, Sri Lanka	Tea	Fairtrade-Organic	Propensity score matching	Yes
van Rijsbergen et al., 2016	Kenya	Coffee	Far Trade, UTZ	Matched difference-in-difference	Yes
Zulfiqar and Thapa, 2016	Pakistan	Cotton	Better Cotton Initiative	Propensity score matching and probit regression	No
Haggar et al., 2017	Nicaragua	Coffee	UTZ, Rainforest Alliance, Fairtrade-Organic, Fairtrade, C. A.F.E.	Propensity score matching, multiple regression	No
Jena and Grote 2017	India	Coffee	Fairtrade	Propensity score matching	No
Meemken et al., 2017	Uganda	Coffee	Fairtrade-UTZ	Instrumental variable + cross-sectional model	No
Mitiku et al., 2017	Ethiopia	Coffee	Fairtrade, Organic, Rainforest Alliance, Fairtrade-Organic	Propensity score matching + regression	No
Takahashi and Todo, 2017	Ethiopia	Coffee	Rainforest Alliance	Propensity score matching	No
Akoyi and Maertens, 2018	Uganda	Coffee	Fairtrade-Organic, UTZ-Rainforest Alliance-4C	Instrumental variable + regression	No
Carlson et al., 2018	Indonesia	Palm oil	Roundtable on Sustainable Palm Oil	Propensity score matching, panel models	No
Doanh et al., 2018	Vietnam	Tea	Organic	Propensity score matching + regression	No
Froehlich et al., 2018	Brazil	Various	Organic	Propensity score matching, bounded treatment effect, regression	No
Ingram et al., 2018	Ghana, Ivory Coast	Cocoa	UTZ	Propensity score matching and difference-in-difference, supplemented by focus groups and interviews	No
Meemken and Qaim, 2018	Uganda	Coffee	Fairtrade-UTZ	Entropy balancing + regression	No
Minten et al., 2018	Ethiopia	Coffee	Fairtrade, Organic	Propensity score matching, probit model	No
Mitiku et al., 2018	Ethiopia	Coffee	Rainforest Alliance	OLS with control variables, panel fixed effect models	No
Morgans et al., 2018	Indonesia	Palm oil	Roundtable on Sustainable Palm Oil	Propensity score matching and before and after control impact (BACI) analysis	No
	Uganda	Coffee		Instrumental variable + regression	No

(continued on next page)

Table 2 (*continued*)

Study	Location(s)	Crop	Certification(s)	Methods	Included in previous review[1]?
Vanderhaegen et al., 2018[7]			Fairtrade-Organic, UTZ-Rainforest Alliance-4C		
Filho et al., 2019	Brazil	Strawberry	Organic	Propensity score matching, endogenous switching regression	No
Tran and Goto, 2019	Vietnam	Tea	UTZ	Propensity score matching, probit model	No

[1] Previous reviews considered here include DeFries et al., 2017, and Oya et al., 2018.
[2] Considered high bias by Oya et al., 2018, but included in DeFries et al., 2017.
[3] Same study region as Takahashi and Todo, 2014 and 2017.
[4] Same study regions as Chiputwa and Qaim, 2016, Meemken et al., 2017.
[5] Included in high-level review, but not in the indicator level analysis due to small sample size.
[6] Results for all three certifications are grouped; thus, this is treated as one case when analyzing statistical results.
[7] Same study region as Akoyi and Maertens, 2018.

Table 3
Example of indicator coding.

Source	Pillar	Theme	Country	Certification	Crop	Indicator	Significant difference?	Positive or negative?
Takahashi and Todo, 2017	Environmental	Forest quality	Ethiopia	Rainforest Alliance	Coffee	Forest density classification based on NDVI and survey validation[1]	Yes	Positive
Takahashi and Todo, 2017	Environmental	Forest quality	Ethiopia	Rainforest Alliance	Coffee	Forest classification in buffer areas (as above, but assesses buffer around certified areas, evaluating spillover effects)	Yes	Positive
Zulfiqar and Thapa, 2016	Economic	Income	Pakistan	Better Cotton Initiative	Cotton	Income	Yes	Positive

[1] Relevant characteristics for each density classification are provided, including number of trees and tree species, height ranges, number of strata, and canopy cover.

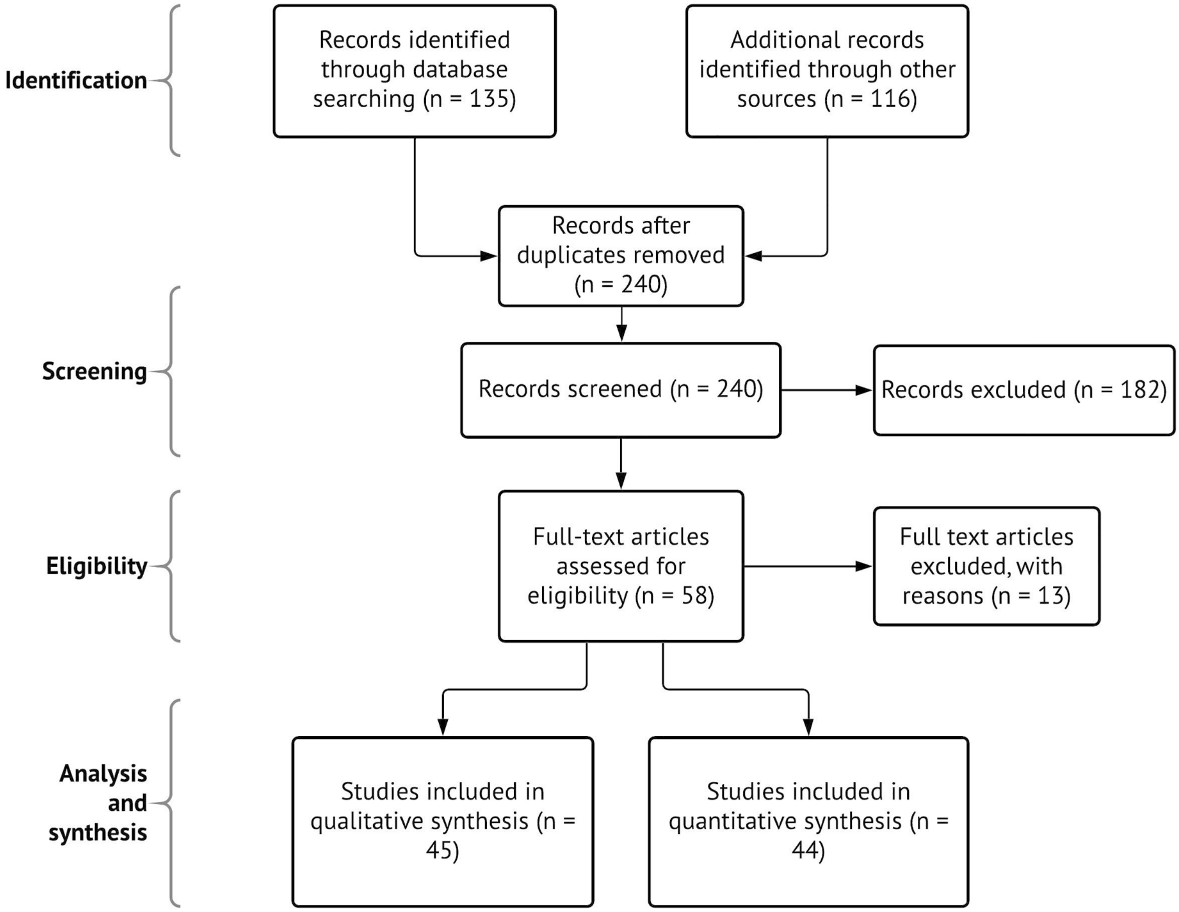

Identification
Records identified through database searching (n = 135)
Additional records identified through other sources (n = 116)
Records after duplicates removed (n = 240)
Screening
Records screened (n = 240)
Records excluded (n = 182)
Eligibility
Full-text articles assessed for eligibility (n = 58)
Full text articles excluded, with reasons (n = 13)
Analysis and synthesis
Studies included in qualitative synthesis (n = 45)
Studies included in quantitative synthesis (n = 44)

Fig. 2. Methodology diagram, following PRISMA guidelines (Moher et al., 2009).

Table 4
Comparison of major crops produced under certification with studies of certifications' impacts, sorted by low to high level of representation. A positive difference indicates the crop is under-represented in the literature. Data sources: Willer et al., 2019, Tayleur et al., 2017 (certified crop area estimates), Author's elaboration.

Major certified crops	Estimated hectares certified (minimum)	Percent of total certified area	Percent of the crop that is certified	Change 2016/2017	Change 2013–2017	Percent of study coverage	Difference between certified area and study coverage
Cotton	5,154,933	22.20%	16.20%	66.80%	172.40%	1.49%	20.71%
Sugar	1,979,979	8.53%	7.60%	88.50%	80.20%	0.00%	8.53%
Cocoa	2,908,640	12.53%	24.80%	22.80%	114.70%	4.48%	8.05%
Soybeans	1,801,269	7.76%	1.50%	−30.20%	−5.90%	0.00%	7.76%
Palm oil	2,537,424	10.93%	11.90%	1.40%	26.10%	4.48%	6.45%
Wheat	1,108,492	4.77%	0.51%			0.00%	4.77%
Other oilseeds	1,002,300	4.32%	0.81%			0.00%	4.32%
Other cereals	549,414	2.37%	0.63%			0.00%	2.37%
Barley	354,881	1.53%	0.72%			0.00%	1.53%
Oats	344,990	1.49%	3.60%			0.00%	1.49%
Pulses	317,446	1.37%	0.40%			0.00%	1.37%
Maize	229,919	0.99%	0.13%			0.00%	0.99%
Fruit & Veg	1,240,463	5.34%	1.03%			4.48%	0.86%
Rice	75,898	0.33%	0.05%			0.00%	0.33%
Root crops	57,928	0.25%	0.10%			0.00%	0.25%
Other	15,433	0.07%	0.17%			0.00%	0.07%
Bananas	340,196	1.46%	6.00%	17.20%	28.60%	2.99%	−1.52%
Tea	668,768	2.88%	16.40%	22.70%	77.30%	5.97%	−3.09%
Coffee	2,533,211	10.91%	23.40%	−8.50%	8.70%	74.63%	−63.72%

[1] There are known challenges in estimating the total certified area at crop and country level, for example due to multiple certification. See Willer et al., 2019 for more information.

Table 5
Comparison of major certifications globally with studies of certifications' impacts, sorted by low to high level of representation. Data sources: Willer et al., 2019, Roundtable on Sustainable Biomaterials, 2018, Author's elaboration.

Area certified for 13 major standards globally	Hectares certified (2017)	Percent of total certified area	Percent of study coverage	Difference between certified area and study coverage
Organic cropland (IFOAM)	69,845,243	71.50%	20.90%	50.56%
Proterra	2,339,259	2.40%	0%	2.39%
Better Cotton Initiative	3,561,000	3.60%	1.50%	2.15%
Global GAP	3,548,194	3.60%	1.49%	2.14%
Cotton Made in Africa	1,619,469	1.70%	0%	1.66%
Roundtable on Responsible Soy	1,259,672	1.30%	0%	1.29%
Bonsucro	1,161,000	1.20%	0%	1.19%
Roundtable on Sustainable Biomaterials	18,100	0.01%	0%	0.01%
Roundtable on Sustainable Palm Oil	3,301,088	3.40%	4.50%	−1.10%
Common Code for the Coffee Community (4C)	1,630,546	1.70%	3.00%	−1.32%
Rainforest Alliance/ Sustainable Agriculture Network	3,458,167	3.60%	11.90%	−8.40%
UTZ	3,376,870	3.50%	13.40%	−9.98%
Fairtrade	2,634,678	2.70%	26.90%	−24.17%

[1] One study looked at Bird Friendly certification, not shown here (<15,000 ha certified). The C.A.F.E. practices standard is also not shown here.

crops – sugar cane and soybeans – have never been assessed in the peer reviewed literature using a robust impact evaluation method. In addition, cotton, sugar cane, and cocoa are all experiencing significant increases in certification, heightening the need for robust impact evaluation. Certified cotton area increased by over 170% from 2013 to 2017, sugar certified area increased by 80%, and cocoa certified area increased by 115% (Willer et al., 2019). From a standards perspective, Organic faces the largest discrepancy between certified area and study coverage. IFOAM organic cropland covered approximately 69 million hectares in 2017, representing 72% of certified area. It is particularly important to evaluate trade-offs when considering Organic certification, since it has been found to produce benefits such as improved soil health and biodiversity, while also negatively impacting productivity and income (Vanderhaegen et al., 2018). Most of the other certifications with the largest gaps in evaluation studies – Proterra, Better Cotton Initiative, Global GAP, Cotton Made in Africa, Roundtable on Responsible Soy, and Bonsucro – conduct their own performance monitoring to measure progress on certified farms, but do not use a counterfactual approach to assess their impact over time.

The analysis also reveals important insights about regional and country-level research coverage and gaps. As illustrated by Table 6, Africa is the most common region of analysis, followed by Latin America (51% and 34% respectively), with Southeast Asia, South Asia, and East Asia representing about 7%, 6%, and 1% of studies. No studies were identified assessing certification in North America or Australia, although certification does take place in these areas. There are 19 countries covered by the included studies – the most common countries are Uganda with 24% of cases, Ethiopia with 16% of cases, and Nicaragua with 13% of cases. Peru, Costa Rica, Colombia, Kenya, and Indonesia each represent about 4% of the reviewed cases. Table 6 presents the top 13 countries that appear to be under-represented in the literature. This list represents a mix of high, middle, and low-income countries on six continents, and includes Brazil, Australia, Malaysia, Indonesia, the United States of America, Canada, Zambia, and the Ivory Coast. There are also several European countries currently under-represented in the literature (Spain, Italy, France, and Germany). These countries likely have significantly different certification compositions. For example, the United States' certified area estimate is largely driven by barley, while other countries like Australia, Brazil and Indonesia have a larger mix of crops. The full country comparison list can be found in the Supplementary Materials. Comparing this list to country-level certified area based on Tayleur et al.'s 2017 analysis of certification extent helps illuminate current research needs. Fig. 3 depicts this information visually on a map at country level. Comparison to Tayleur et al.'s 2018 map of certification extent allows identification of clustered certified regions which appear to have not yet been evaluated. While this does not reflect

Table 6
Abridged comparison of countries with certified production globally with studies of certifications' impacts, sorted by low to high level of representation. Data sources: Tayleur et al., 2017 (certified area estimates), The World Bank (2020), author's elaboration.

Country	Certified area (ha, Tayleur et al., 2017)	Percent of total cert. area	Percent study coverage	Difference between certified area and study coverage	Income level
Brazil	2,386,045.00	15.93%	3%	12.94%	Upper middle-income
Australia	653,733.50	4.36%	0%	4.36%	High income
Malaysia	649,866.00	4.34%	0%	4.34%	Upper middle-income
Ivory Coast	861,866.40	5.75%	1%	4.26%	Lower middle-income
Spain	596,629.00	3.98%	0%	3.98%	High income
Italy	562,749.60	3.76%	0%	3.76%	High income
United States of America	546,591.50	3.65%	0%	3.65%	High income
Zambia	318,680.00	2.13%	0%	2.13%	Lower middle-income
France	309,019.50	2.06%	0%	2.06%	High income
Turkey	307,157.10	2.05%	0%	2.05%	Upper middle-income
Canada	296,972.60	1.98%	0%	1.98%	High income
Germany	268,272.20	1.79%	0%	1.79%	High income
Indonesia	931,536.90	6.22%	4%	1.74%	Lower middle-income

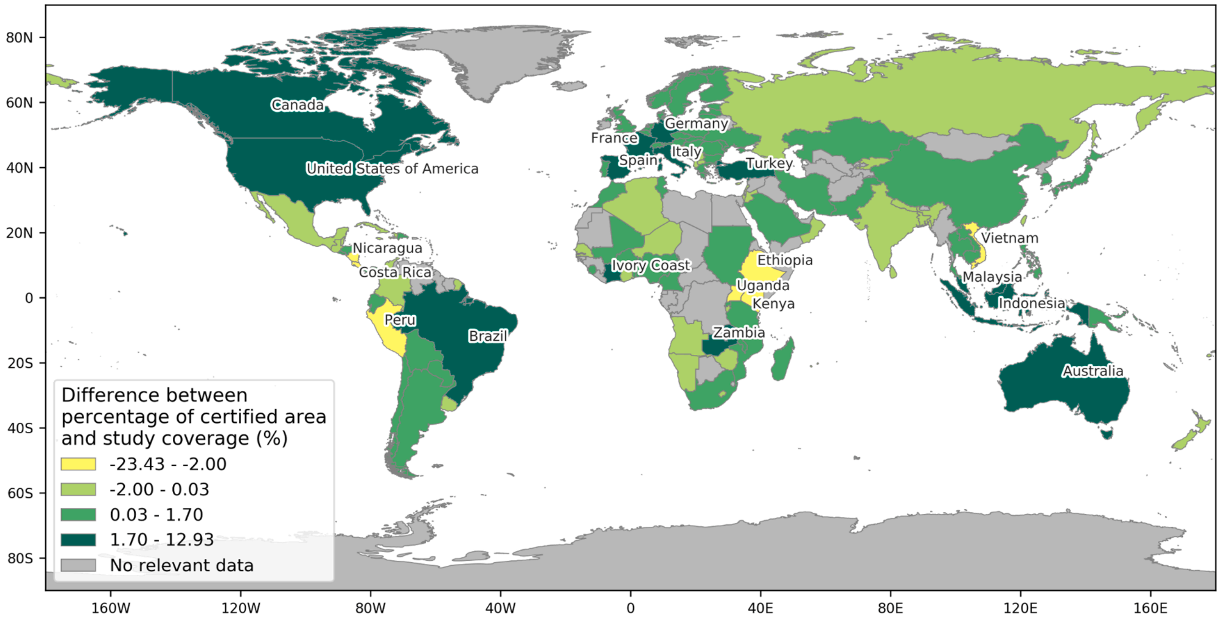


Fig. 3. Country research gaps in evaluations of certified agriculture. Higher values indicate under-represented countries, with the yellow and darkest green classes representing the areas of highest and lowest representation, respectively. Countries falling into these two classes are labeled. Data sources: Tayleur et al., 2017 and author's elaboration. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all certified area, it appears that large areas of certified coffee, sugar, and soy in Brazil, cocoa produced in the Ivory Coast, and oil palm produced in Malaysia have not yet been evaluated in the peer-reviewed literature. Fig. 4 highlights the estimated certified areas for cocoa and coffee in the Ivory Coast and Ghana, as well as several certified crops in Brazil, using 30 km × 30 km grid cells from Tayleur et al., 2018.

3.2. *Which pillars of sustainability are included in evaluation studies, and which indicators are used to measure success?*

The economic pillar is studied the most frequently (84% of the reviewed articles). There is less frequent analysis of social outcomes (43% of articles), as well as environmental outcomes (43% of articles). Only 20% of the studies (nine in total) consider all three pillars, while 30% look at two pillars (typically, economic and social outcomes together).

The analysis provides some important insights regarding environmental indicators. First, the majority of indicators measure "outputs", or direct results from the certification (60% of the reviewed environmental indicators). These are often practice-focused, for example related to agrochemical use, and best management measures related to crops, soil, trees, or water resources. A significant portion of the indicators, about 40%, measure outcomes, or what could be considered "ultimate goals" of certification related to environmental conditions. Outcome indicators focus on topics including forests and trees (e.g. density, diversity, tree cover, forest quality), biodiversity (invertebrate abundance), carbon stocks (tree biomass, soil organic carbon, and total carbon stocks), deforestation, and fire activity (for palm oil). Notably, no direct measurements were employed for water resources, and only a few outcome indicators for soil incorporated direct measurement. For biodiversity, in

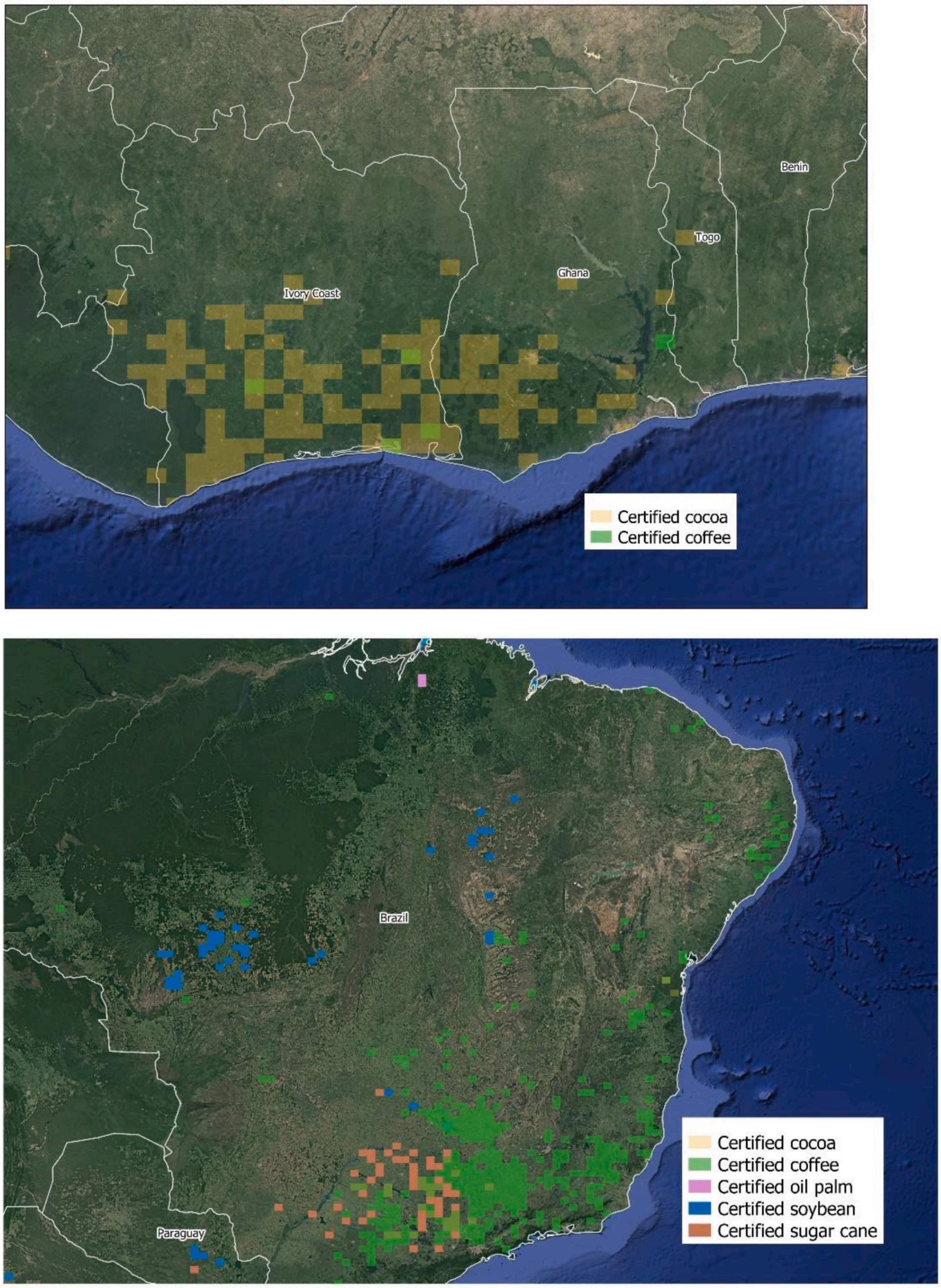


Fig. 4. Spotlight on under-studied countries producing certified agriculture. Certified area for major crops (Tayleur et al., 2018) are displayed for the Ivory Coast and Brazil. Data sources: Tayleur et al., 2017b and author's elaboration. NB: There are known challenges with identifying precise locations of certified area.

addition to invertebrate diversity (Vanderhaegen et al., 2018), one study analyzes orangutan presence (Morgans et al., 2018).

Additionally, the majority of environmental indicators (58%) are measured using survey data, 33% are measured using field observations, and 8% rely on remotely sensed data (e.g. Landsat and MODIS). In terms of the themes covered by environmental indicators, the most common are best management practices (23% of environmental indicators) and agrochemical and input use (20% of indicators). Tree density and diversity is the third-most common theme, followed by carbon storage, soil, and biodiversity (9%, 7%, and 7% respectively). Water conservation is only considered in 2% of the indicators (and only for output indicators), and habitat quality for 3% of the indicators. For additional information on the environmental indicator coding, see the Supplementary Materials.

For social sustainability, about 56% of indicators measure outputs or direct results of the certification, and about 44% measure ultimate outcomes. Some of the most common themes covered by social indicators include 1. perception and satisfaction with 14% of the indicators (e.g. regarding technical assistance, cooperatives, and economic well-being); 2. social networks (12% of indicators), covering topics like participation in farming organizations; 3. nutrition and food security, including the extent of nutrient, vitamin and energy deficiencies (11% of indicators); and 4. gender, e.g. control of resources, assets and decision-making within the household, and participation in agriculture (10% of indicators). Additional social themes include poverty, labor, health, education, and child labor.

The reviewed studies provide useful examples of outcome measurement for themes like education, health, nutrition and food security, and gender. For example, indicators to measure educational outcomes include school attendance and the maximum grade obtained for school-age children, as well as educational expenditures (Arnould et al., 2009, Minten et al., 2018, Meemken et al., 2017). Arnould et al., 2009 use two health indexes to analyze health outcomes – one which measures the extent to which individuals in the household suffered specific illnesses or death, and one measuring receipt of treatment for those illnesses. Chiputwa and Qaim, 2016 and Meemken et al., 2017 both assess nutritional outcomes, primarily by measuring per adult equivalent consumption of calories and micronutrients, using these values to estimate the prevalence and depth of any deficiencies.

In terms of the economic indicators, the measurement of outputs and outcomes is fairly equal, with 48% and 52% of the indicators, respectively. Income is the most common economic sustainability theme, representing 25% of the indicators; it is followed by productivity (12% of indicators), price (9% of indicators), and expenditures (7% of indicators). Income indicators often include overall household income as well as income from the certified crop, and yield is typically measured in volumes per hectare or per tree (for coffee). Several studies also go beyond income to measure the prevalence of poverty among certified and non-certified groups, which can help illustrate systemic challenges to well-being, despite potential improvements in economic output indicators like price and volumes sold (e.g. Mitiku et al., 2017, Jena and Grote, 2017, Akoyi and Maertens, 2018, Vanderhaegen et al., 2018).

3.3. What does the current evidence base suggest regarding outcomes of voluntary agricultural sustainability standards?

What conclusions can we draw from this evidence base regarding outcomes of voluntary agricultural sustainability standards? When grouped by case, the most common results are positive (51%), no difference (41%), and negative (8%). There are no significant differences in terms of the average proportion of positive and negative results when grouped by sustainability pillar. The proportion of no difference results is significantly different between economic and social pillars as measured by the Kruskal-Wallis test, with the social pillar tending to exhibit a higher proportion of no difference results (see Table 7 and Fig. 5). Kruskal-Wallis tests also suggest no significant differences between a selection of sustainability themes, although negative results border on significantly different (see Table 8 and Fig. 6). Sustainability themes with the highest proportion of positive results include environmental outcomes, income, agrochemical and input use, and social networks; conversely, best management practices and perception and satisfaction show high proportions of not significant results, and gender and productivity show the highest amounts of negative results. Negative results for productivity occur primarily for Organic and Fair Trade-Organic certification. While it is possible to further analyze indicator results across certifications and crops, this is likely only advisable when there have been a sufficient amount of studies conducted on the certification or crop in question, to reduce the likelihood of erroneous conclusions.

It is important to treat these results with caution due to the still-limited evidence base. This is particularly true for two reasons – first, that evaluation studies have been largely unanimous in their emphasis on context in driving or enabling certification outcomes, and second, that 75% of the reviewed studies focus on coffee. In addition, participants may perceive the benefits they receive from certification differently than outcome indicators suggest, which can be captured through qualitative methods and direct feedback from participants. One example of this is found in Jena et al.'s 2015 study of Fairtrade certification in Nicaragua, which found no difference between certified and non-certified farms in terms of overall income, although participants reported benefits related to education and health services not reflected in outcome indicators.

Table 7
The average proportion of positive, negative and no difference results when comparing between certified and uncertified participants, grouped by case for each pillar of sustainability. Standard deviations are shown in parentheses. Differences between sustainability pillars were not significant, apart from the proportion of non-significant results (p = 0.017 for Kruskal-Wallis test). The number of indicators per case varies between 1 and 57; the average is 11.

Pillars of Sustainability	Number of cases	Proportion positive	Proportion negative	Proportion no difference
Environmental	26	0.47 (0.36)	0.06 (0.21)	0.47 (0.35)
Social	29	0.33 (0.36)	0.07 (0.16)	0.6 (0.36)
Economic	55	0.51 (0.38)	0.12 (0.27)	0.38 (0.33)
All	62	0.51	0.08	0.41

Contextual factors that have been identified as important influencers of certification's outcomes include: the prevalent poverty and livelihoods conditions (e.g. severity of poverty and dependence on farm income prior to the intervention); market structure (the existence of a price premium and cooperative structure, for example, as well as the contract type); market conditions (e.g. sales to certified markets and global commodity price trends); and the extent to which certification requirements surpass legal requirements (Jena et al., 2015, Minten et al., 2018, Ruben et al., 2009, Qiao et al., 2018, Oya et al., 2018, van Rijsbergen et al., 2016). Certifications operate within in a broader market environment, and there is a risk that severe shocks caused by price volatility can mask the benefits of years of productivity gains, depending on the support systems available to producers (Dave McLaughlin, personal communication). On the environmental side, contextual conditions could include broader environmental trends, for example the regional deforestation level (Rueda et al., 2015, Carlson et al., 2018). Different standards also vary in their requirements, as well as the training, assistance, and other benefits they provide to producers – this is important to keep in mind when attempting to generalize results. For example, Fairtrade certification includes social premium funds that go to the community, to improve services like education, health care, and infrastructure (Karki et al., 2016). Standards also vary in the extent to which they address environmental issues like reduction of greenhouse gas emissions and deforestation (Tayleur et al., 2017, Table 1).

3.4. A qualitative assessment of best practices and novel examples

In addition to addressing the key research questions, this analysis helps identify best practices in assessing certification's impacts. First, there are several studies which provide strong examples of multi-pillar evaluations. The findings indicate that only 20% of studies look at all three pillars of sustainability; of these, five were published prior to 2015, and four were published in 2018 (Fort and Ruben, 2009; Ingram et al., 2018; Minten et al., 2018; Morgans et al., 2018; Ruben et al., 2009; Ruben and Fort, 2012; Ruben and Zuniga, 2011; Schoonhoven-Speijer and Ruben, 2014; Vanderhaegen et al., 2018). Some specific strengths of multi-pillar studies include consideration of trade-offs between environmental and socioeconomic outcomes, as well as data collection regarding farmer's perceived benefits of certification. Another important area of progress has been the consideration of nuanced social indicators, including issues like health, education, gender, and nutrition (see e.g. Chiputwa and Qaim, 2016, Meemken and Qaim, 2018). These studies incorporate data collection approaches from the development field (among others), enabling assessment of intra-household dynamics as well as human well-being outcomes. Finally, there have been significant developments in the assessment of environmental sustainability indicators, particularly using field measurements and remotely sensed imagery (e.g. Haggar et al., 2017, Vanderhaegen et al., 2018, Takahashi and Todo (2013), Takahashi and Todo (2014), Takahashi and Todo (2017), and Cattau et al., 2016). These studies employ clear sampling strategies and analytical approaches to address various sources of bias, and data collection protocols to enable comprehensive and efficient field measurements. Together, these best practices enable authors to identify novel findings related to the outcomes of certification, and it merits

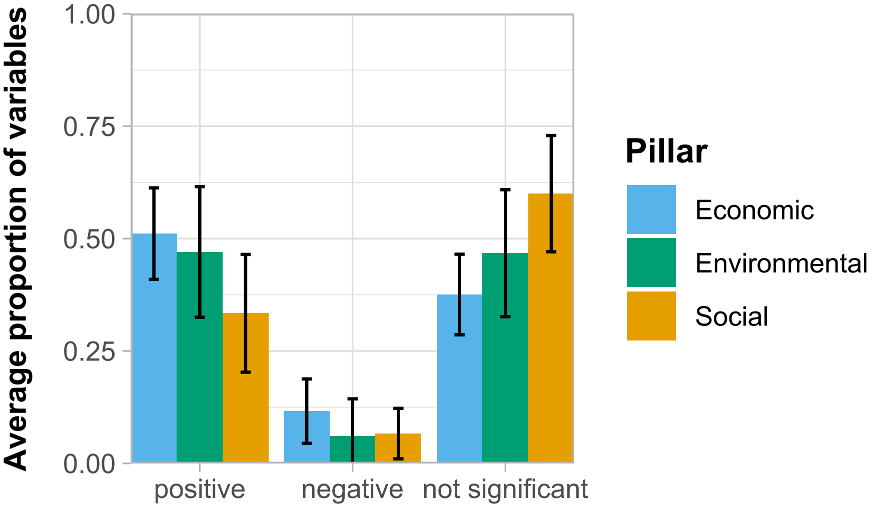


Fig. 5. The average proportion of positive, negative and no difference results for comparison between certified and uncertified participants, when grouped by case, for each pillar of sustainability. Error bars show 95% confidence intervals.

Table 8
The average proportion of positive, negative and no difference results between certified and uncertified participants, grouped by case for a selection of indicator themes. Standard deviations are shown in parentheses. Differences between sustainability themes were not significant as measured by the Kruskal-Wallis test, although negative results bordered on significantly different (p = 0.052).

Indicator theme	Number of cases	Proportion positive	Proportion negative	Proportion no difference
Income	32	0.54 (0.47)	0.08 (0.26)	0.38 (0.44)
Productivity	29	0.37 (0.44)	0.24 (0.41)	0.39 (0.45)
Environmental output	19	0.37 (0.38)	0.09 (0.24)	0.54 (0.40)
Environmental outcome	16	0.63 (0.35)	0.02 (0.08)	0.35 (0.32)
Best management practices	15	0.26 (0.40)	0.08 (0.15)	0.66 (0.40)
Poverty	13	0.38 (0.51)	0.00	0.62 (0.51)
Social network	9	0.41 (0.41)	0.15 (0.34)	0.44 (0.42)
Agrochemical and input use	8	0.50 (0.37)	0.19 (0.37)	0.31 (0.34)
Perception and satisfaction	8	0.16 (0.27)	0.18 (0.35)	0.66 (0.39)
Gender	7	0.18 (0.41)	0.39 (0.45)	0.43 (0.46)

consideration why more studies do not replicate these approaches. These best practices are elaborated upon further in the Discussion section.

3.5. Limitations

Ideally, a review of this type would include a step of cross-checking methodological decisions, including the process for searching for articles, selecting articles for inclusion, and coding indicator results, to ensure inter-rater reliability (Siddaway et al., 2019). However, this is not always possible, as was the case for this review. Here, this is addressed by relying closely upon previous reviews, and providing detailed information on methods. Methodological details are provided in the Supplementary Materials to enable reproducibility, and it assumed that any small divergences in article selection or coding would not greatly affect key trends in the results.

Additional limitations of this review include that it was conducted in the English language, and that it did not consider evaluations which had not been peer reviewed. There is a large amount of grey literature focused on the impacts of VSS's, and some of this literature likely applies a counterfactual approach. Future reviews could aim to broaden the search to additional languages beyond English and consider rigorous grey literature sources in addition to peer-reviewed articles.

4. Discussion

4.1. Relating findings to other recent reviews

Among the papers that have robustly evaluated the impact of certification, there have been mixed findings regarding outcomes. DeFries et al. (2017) found that the majority (58%) of outcome indicators in evaluated studies found no significant difference between certified and non-certified producers, while 34% of indicators represented positive outcomes. When grouped by case, this study finds a slightly more positive trend, as the average proportion of positive indicators was 0.51. Although economic and environmental pillars have a higher average proportion of positive results than social indicators, these differences were not statistically significant. However, the fact that social indicators tended to exhibit no difference between certified and non-certified producers may suggest that sustainability standards are less effective in driving change for social sustainability concerns (see Ingram et al., 2018).

Oya et al. (2018) found that most studies which looked at farm income effects found positive and statistically significant results, although there was variation in individual study-level effects. The authors surmised that this was driven by specific capacity-building activities, farm productivity, and market conditions (e.g. prices received), with the price linkage as key to overall effects on income (Oya et al., 2018). There is also a challenge translating farm income effects into broader household income effects, due to contextual factors, and household dependence on other income sources (Oya et al., 2018, DeFries et al., 2017) – here, we see this reflected in the lower proportion of positive results for poverty indicators as compared to income indicators. A more recent review (Meemken, 2020) also found evidence of positive price and income effects. Overall, this study finds that income indicators tend to be positive; the only theme with a higher average proportion of positive results was the "environmental outcome" category.

4.2. Recent progress – multi-pillar studies, increased consideration of trade-offs, and non-economic benefits to producers

Despite persistent research gaps, significant progress has been made in the evaluated studies, which can present further opportunities for

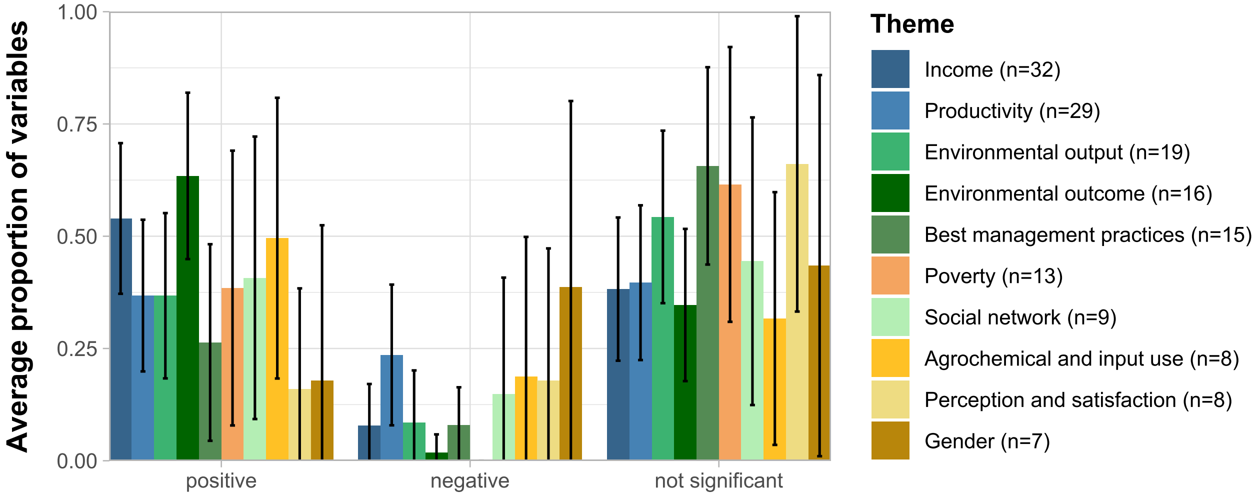


Fig. 6. The average proportion of positive, negative and no difference results between certified and uncertified participants, when grouped by case, for a selection of sustainability themes. Error bars show 95% confidence intervals.

learning and replication. Some of the key areas of progress have been an increase in the number of multi-pillar studies, and the inclusion of important socioeconomic variables beyond price and income. The discussion below highlights a few exceptional studies which provide examples of best practices.

First, in their analysis of Fairtrade and Organic certification in Ethiopia, Minten et al., 2018 explore the receipt of price premiums by certified producers, as well as environmental and social outcomes. In addition to the price premium, outcome indicators include agricultural management practices (e.g. stumping, compost use, and use of agrochemical inputs), as well as school attendance for school-aged children (Minten et al., 2018). The authors also ask farmers about the benefits they see in engaging with cooperatives more broadly. Although this study finds limited evidence of income improvement due to certification, the inclusion of social benefits like school attendance enables the recognition of non-economic benefits that may influence farmers' decisions to participate in certification, in addition to their households' overall well-being (Minten et al., 2018).

Vanderhaegen et al. (2018) provide another useful example of a multi-pillar analysis, in their study of multiple coffee certifications (Fairtrade-Organic and UTZ-Rainforest Alliance-4C) in Uganda. The indicators included in this study are unique among those reviewed, as they span socio-economic issues (e.g. poverty, income, and yield) as well as practice and outcome-based environmental issues. Within the environmental pillar, this study addresses several important themes including biodiversity, carbon storage, and tree density and diversity, with data collected through field observations, in addition to the agricultural best management practices that are typically included in survey-based approaches to evaluate certification (Vanderhaegen et al., 2018). The authors go one step further and conduct a correlation analysis to illuminate trade-offs between sustainability pillars. This approach underpins a key finding – the two multiple certifications have different implications for the level of coffee intensification, with UTZ-RA-4C increasing the likelihood of using agrochemicals (along with practices like use of shade trees and intercropping with legumes), and Fairtrade-Organic reducing the use of inputs and increasing likelihood of soil tillage (Vanderhaegen et al., 2018). The two schemes also have different outcomes for socio-economic performance at farm level and prices, as well as environmental variables like carbon stocks and tree cover density (Vanderhaegen et al., 2018). The findings emphasize that increased agrochemical inputs which tend to occur in UTZ-RA-4C certification lead to yield, labor productivity and income increases, but decreases in invertebrate abundance and diversity (Vanderhaegen et al., 2018). This critical finding regarding trade-offs between yields and ecosystem services would not have been possible with a single pillar approach, which further underlines the importance of applying a multi-pillar methodology to other crops, certifications, and production contexts. This is also important because some studies have suggested that productivity increases are more important than price premiums for increasing overall returns to farmers (Akoyi and Maertens, 2018).

Another important area of progress has been the consideration of nuanced social indicators, including issues like health, education, gender, and nutrition (see Chiputwa and Qaim, 2016, Meemken and Qaim, 2018). In their analysis of Fairtrade-UTZ certification in Uganda, Meemken and Qaim (2018) include indicators like asset ownership for female household members, as well as a 24-hour time recall measuring division of labor, a subjective measure regarding satisfaction with time for leisure activities, and interactions with farmer organizations, extension officers, and training sessions to capture rural services. This approach enables the authors to identify an increase in total household assets for certified female-headed households as compared to non-certified, due to higher coffee revenues. The authors also found evidence that standards impacted the distribution of assets within households, with a positive effect on joint asset ownership within male-headed households (Meemken and Qaim, 2018). Further, indicators related to social networks – for example, organizational participation and measurement, including farmers' organizations – also suggest some positive benefits for certified producers (see van Rijsbergen et al., 2016, Meemken and Qaim, 2018). This lends support to growing evidence that price premiums are just one piece of the puzzle in terms of certification's socioeconomic impacts.

While there have been significant improvements in the exploration of VSS impacts for multiple pillars of sustainability, studies which analyze all three pillars are still the minority in the literature. Conducting evaluations which incorporate social, economic, and environmental components should continue to be a priority.

4.3. Acting on research gaps – representation of different standards, crops, and regions; environmental outcome measurement

The key research gap indicated by this review is an underrepresentation of studies for certain crops, certifications, and regions. This is particularly important since we do not have any reason to believe that findings for coffee – which represents 75% of the cases included – will hold true for other crops. Additionally, several of the main understudied crops – cotton, cocoa, sugar cane, soy, and palm oil – are known to cause severe and urgent sustainability issues.

Why should the high proportion of studies focused on coffee give us pause in generalizing results? There are several ways in which coffee certification substantively differs from certification for other crops. Out of the 13 major agricultural sustainability standards, seven certify coffee. Coffee is shade-tolerant, meaning that it can prosper under full or partial shade (Rainforest Alliance, 2017). For this reason, it is fairly unique among the major certified crops, and shade-grown coffee is known to bring significant environmental and socioeconomic benefits (De Beenhouwer et al., 2013, Jha et al., 2014). The share of shade-grown coffee has been decreasing over time, and several key production countries now rely primarily on full-sun coffee production systems (Jha et al., 2014). Still, many of the major certified crops including cotton, sugar, soybeans, and palm oil do not have an equivalent to shade grown coffee – while production systems may represent differing levels of intensity, agroforestry systems are not common at the global scale. This means that we can expect to see different environmental issues and outcomes for those crops. As one example of this, the Rainforest Alliance standard sets a minimum canopy cover requirement of 40% for their certified coffee; the only other crops with canopy cover minimums are cocoa (30%), clove and vanilla (40%), and pepper (20%) (Rainforest Alliance, 2017). Interestingly, no studies were identified evaluating certified coffee in Brazil, which has one of the highest certified coffee areas (Willer et al., 2019). Further, coffee faces specific market conditions that distinguish it from other crops. While it represents one of the most well-developed markets for sustainably certified agriculture, it also faces challenges, including a mismatch between supply and demand. A 2014 report found that although 40% of global coffee production complied with global standards, only 10% was sold as such (Potts et al., 2014). Finally, we can expect that the willingness to pay a price premium will differ significantly for coffee as compared to a commodity crop like soy or wheat, which may have implications for financial returns to producers.

An additional gap is a lack of direct measurement for environmental outcomes. While 38% of reviewed studies considered environmental indicators, only 22% explicitly considered environmental outcomes. While information on the use of specific agricultural practices is helpful, it is important to complement this with direct measurement as well as the measurement of true outcomes, rather than focusing exclusively on outputs (Milder et al., 2015). Despite this limitation and the overall lower proportion of studies looking at environmental outcomes, there have been several innovative approaches to the topic. The two main methods for measuring environmental outcomes incorporate field observations and direct measurement, and remotely sensed data.

The use of field measurements for environmental outcomes is emblemized by studies from Haggar et al., 2017 and Vanderhaegen et al., 2018. The first study aligns with the Committee for Sustainability Assessment (COSA) method for assessment of coffee sustainability, utilizing indicators that can be measured by trained evaluators (rather than scientific experts), and can be completed in half a day to one day per farm, thus enabling larger sample sizes. Environmental outcomes assessed by this study include tree density and diversity, habitat quality, and carbon storage (Haggar et al., 2017). This balance of rigor and feasibility should continue to be adopted in future studies, including for under-studied crops like cotton, sugar, cocoa, soybeans, and palm oil. Vanderhaegen et al., 2018 collected environmental data in a subsample of 74 coffee fields, using stratified random sampling based on elevation and soil type. These fields were then matched with similar non-certified fields using propensity score matching, and field measurements were completed for half-hectare plots randomly placed throughout fields (Vanderhaegen et al., 2018). Field measurements included diameter at breast height and height for plant species; woody debris, stem and plant counts; litter collection; and soil samples for bulk density and soil organic carbon (Vanderhaegen et al., 2018). This study is also notable in that it included outcome variables related to soil and biodiversity (invertebrate abundance and diversity) (Vanderhaegen et al., 2018).

Takahashi and Todo (2013), Takahashi and Todo (2014), and Takahashi and Todo (2017) and Cattau et al. (2016) provide useful examples of evaluating environmental outcomes using remotely sensed imagery. Takahashi and Todo (2014) look at Rainforest Alliance certification of coffee in Ethiopia, considering outcomes related to forest conservation, deforestation, and forest quality. The study utilizes two years of Landsat 7 imagery (2005 and 2010) and a probit model, supplemented by a household survey to understand the relationship between forest conservation and socioeconomic characteristics. The authors find that certification has a positive effect on forest conservation; conservation was also affected by years of formal education and the total area of agricultural land at the household level (Takahashi and Todo, 2014). They also find evidence that certification had a significant impact on behavior for producers with more limited assets (Takahashi and Todo, 2014). Cattau et al. (2016) evaluate palm oil concessions in Indonesia to determine whether Roundtable on Sustainable Palm Oil certification impacts fire activity. This analysis uses nonparametric matching methods and data from Global Forest Watch and MODIS Active Fire Detections, finding a positive and significant relationship for one out of four outcome variables (density of fire activity on non-peatlands in wet years), with the remaining variables showing no significant difference between certified and non-certified concessions (Cattau et al., 2016).

4.4. Challenges for robust evaluations of certification's impacts

There are clear challenges in designing and implementing robust impact evaluations of certification. One such challenge is the resources needed to implement them (Margoluis et al., 2009). Studies typically rely on field-collected survey data, sometimes in combination with secondary data sources, like government agricultural censuses or remotely sensed imagery. In order to conduct robust statistical analysis, field surveys must cover a sufficient number of farms. Sample design is also complex, with many studies using two or three stage sampling designs, with careful matching between treatment and control groups to account for selection bias. The resource requirements of these studies are significant and the logistics are not uncomplicated, which is characteristic of environmental impact evaluations more broadly (Mascia et al., 2014).

Several additional difficulties of designing robust impact analyses of certification have been identified. These include considerations related to control group selection. For example, studies may include producer communities in close proximity to the intervention group as a control. However, spillover benefits have been found to affect outcomes for nearby communities as well as those receiving certification – this may create an impression of a smaller difference between treatment and control groups, if not identified and appropriately highlighted in results (Komives et al., 2018, Jones and Gibbon, 2011). Some studies have accounted for this by selecting control groups outside of the influence of certification activities, or explicitly considering spillover effects in data collection (see Ingram et al., 2018, Van Rijsbergen et al., 2016). Additionally, differences in timing of certification among certified groups may create issues with identifying trends in the data (Komives et al., 2018).

Another challenge of building the evidence base has been indicator consistency and quality. Milder et al. (2015) emphasized that use of different outcome variables between studies presented a challenge for robust comparative analysis; in addition, studies of environmental changes tended to focus on management practices, rather than environmental outcomes per se (Milder et al., 2015). The authors referenced several programs which sought to define common indicators, including the Committee on Sustainability Assessment (COSA) and the SAI Platform, and advocated for the development of common metrics (Milder et al., 2015, COSA, 2020, SAI Platform, 2014). Indicator consistency involves a delicate balance, as production systems between small and large producers vary significantly, and necessitate different measures of success (Dave McLaughlin, 2020, personal communication).

5. Conclusions

VSS's represent a huge investment in sustainability from the international community. Given the extent of current sustainability problems, it is more important than ever to understand the outcomes of these interventions for local communities and ecosystems. This review articulates several important priorities for the evaluation of voluntary sustainability standards' impacts moving forward. First, future research should build on recent gains in terms of assessing multiple pillars of sustainability. There is evidence that trade-offs can occur between socioeconomic and environmental outcomes – failing to consider these trade-offs will result in an incomplete picture of the benefits and challenges of certification.

Secondly, continued efforts to standardize and align indicators are critical to inform future reviews and meta-analyses of certification. It is important that outcome indicators reflect actual goals of certification, and not contextual factors that cannot reasonably be influenced by the program. While price premiums are a crucial component of benefits to producers, it is also important to robustly evaluate other socioeconomic impacts of certification and producers' perceptions of these impacts, which may influence producers' decisions to participate and stay in the program.

Third, future research should continue to build on progress to measure environmental outcomes. Although environmental outcomes are not frequently evaluated, there have been significant developments in direct measurement and remote sensing methods. Studies should continue to build on these gains to determine whether certifications are achieving their goals. Remote sensing has particular benefits for assessing landscape-level impacts of certification, although there can be challenges in defining suitable proxies for sustainability outcomes at the landscape level and addressing confounding factors (Rueda et al., 2015, Morgans et al., 2018).

This study's findings generally align with DeFries et al. (2017) at the indicator level, but paint a slightly more positive picture, as the average percentage of positive indicators was 51%. However, these trends should be interpreted with caution, and a more balanced evidence base is required to increase confidence in the findings.

Future research should aim to address the research gaps identified within this review, including evaluation of under-represented crops, standards, and geographical areas. Due to the growing acknowledgement that certification alone is not sufficient to achieve sustainability goals, there is also a need to evaluate the adoption and impact of VSS's in combination with other approaches, including landscape-based conservation and policy support for producers. Although growing, the limited evidence base has significant implications for continuous improvement of standards, their potential integration with other policy or program interventions, and their ability to deliver on key sustainability objectives.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author acknowledges Dr. Julie Silva, Dr. George Hurtt, Dr. Alexandra Tyukavina, Dr. Michael Paolisso, and David McLaughlin for their invaluable feedback and support on the manuscript; the author also thanks two anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2021.107490.

References

Akoyi, K.T., Maertens, M., 2018. Walk the talk: private sustainability standards in the Ugandan coffee sector. J. Develop. Studies 54 (10), 1792–1818.

Arnould, E.J., Plastina, A., Ball, D., 2009. Does fair trade deliver on its core value proposition? effects on income, educational attainment, and health in three countries. J. Public Policy Market. 28 (2), 186–201.

Blackman, A., Naranjo, M.A., 2012. Does eco-certification have environmental benefits? organic coffee in Costa Rica. Ecol. Econ. 83, 58–66.

Blackman, A., Rivera, J., 2011. Producer-level benefits of sustainability certification. Conserv. Biol.: J. Soc. Conserv. Biol. 25 (6), 1176–1185. https://doi.org/10.1111/j.1523-1739.2011.01774.x.

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

Butsic, V., Lewis, D.J., Radeloff, V.C., Baumann, M., Kuemmerle, T., 2017. Quasi-experimental methods enable stronger inferences from observational data in ecology. Basic Appl. Ecol. 19, 1–10.

Breslow, S.J., Allen, M., Holstein, D., Sojka, B., Barnea, R., Basurto, X., Carothers, C., Charnley, S., Coulthard, S., Dolšak, N., Donatuto, J., García-Quijano, C., Hicks, C.C., Levine, A., Mascia, M.B., Norman, K., Poe, M., Satterfield, T., St. Martin, K., Levin, P. S., 2017. Evaluating indicators of human well-being for ecosystem-based management. Ecosyst. Health Sustainability 3 (12), 1–18.

Breslow, S.J., Sojka, B., Barnea, R., Basurto, X., Carothers, C., Charnley, S., Coulthard, S., Dolšak, N., Donatuto, J., García-Quijano, C., Hicks, C.C., Levine, A., Mascia, M.B., Norman, K., Poe, M., Satterfield, T., Martin, K.S., Levin, P.S., 2016. Conceptualizing and operationalizing human wellbeing for ecosystem assessment and management. Environ. Sci. Policy 66, 250–259.

Carlson, K.M., Heilmayr, R., Gibbs, H.K., Noojipady, P., Burns, D.N., Morton, D.C., Walker, N.F., Paoli, G.D., Kremen, C., 2018. Effect of oil palm sustainability certification on deforestation and fire in Indonesia. Proc. Natl. Acad. Sci. USA 115 (1), 121–126.

Cattau, M.E., Marlier, M.E., DeFries, R., 2016. Effectiveness of Roundtable on Sustainable Palm Oil (RSPO) for reducing fires on oil palm concessions in Indonesia from 2012 to 2015. Environ. Res. Lett. 11 (10), 105007.

Caudill, S.A., Rice, R.A., 2016. Do Bird Friendly® Coffee Criteria Benefit Mammals? Assessment of Mammal Diversity in Chiapas, Mexico. PLOS ONE, 11(11), e0165662. https://doi.org/10.1371/journal.pone.0165662.

Chiputwa, B., Qaim, M., 2016. Sustainability standards, gender, and nutrition among smallholder farmers in Uganda. J. Develop. Stud. 52 (9), 1241–1257.

Chiputwa, B., Spielman, D.J., Qaim, M., 2015. Food standards, certification, and poverty among coffee farmers in Uganda. World Dev. 66, 400–412.

Colen, L., Maertens, M., Swinnen, J., 2012. Private Standards, Trade and Poverty: GlobalGAP and Horticultural Employment in Senegal. World Econ., 35(8), 1073–1088. https://doi.org/10.1111/j.1467-9701.2012.01463.x.

COSA: Committee on Sustainability Assessment. Indicator Library: Master List. (2020). https://thecosa.org/master-list/.

De Beenhouwer, M., Aerts, R., Honnay, O., 2013. A global meta-analysis of the biodiversity and ecosystem service benefits of coffee and cacao agroforestry. Agric. Ecosyst. Environ. 175, 1–7.

DeFries, R.S., Fanzo, J., Mondal, P., Remans, R., Wood, S.A., 2017. Is voluntary certification of tropical agricultural commodities achieving sustainability goals for small-scale producers? a review of the evidence. Environ. Res. Lett. 12 (3), 033001.

Dietz, T., Grabs, J., Chong, A.E., 2018. Mainstreamed voluntary sustainability standards and their effectiveness: evidence from the Honduran coffee sector. Regul. Governance. https://doi.org/10.1111/rego.12239.

Dietz, T., Estrella Chong, A., Grabs, J., Kilian, B., 2020. How effective is multiple certification in improving the economic conditions of smallholder farmers? evidence from an impact evaluation in colombia's coffee belt. J. Develop. Stud. 56 (6), 1141–1160.

Doanh, N.K., Thuong, N.T.T., Heo, Y., 2018. Impact of conversion to organic tea cultivation on household income in the mountainous areas of Northern Vietnam. Sustainability 10 (12), 4475. https://doi.org/10.3390/su10124475.

Elbers, W., van Rijsbergen, B., Bagamba, F., Hoebink, P., 2014. Chapter 2 The impact of Utz certification on smallholder farmers in Uganda. In Coffee certification in East Africa: impact on farms, families and cooperatives (Vol. 1–0, pp. 53–82). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-805-6_2.

Ferraro, P.J., 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Eval. 2009 (122), 75–84. https://doi.org/10.1002/ev.297.

Resende Filho, M.A., Andow, D.A., Carneiro, R.G., Lorena, D.R., Sujii, E.R., Alves, R.T., 2019. Economic and productivity incentives to produce organically in Brazil: evidence from strawberry production in the Federal District. Renew. Agric. Food Syst. 34 (2), 155–168.

Ferraro, P.J., Hanauer, M.M., 2014. Advances in measuring the environmental and social impacts of environmental programs. Annu. Rev. Environ. Resour. 39 (1), 495–517.

Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., Balzer, C., Bennett, E.M., Carpenter, S.R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., Tilman, D., Zaks, D.P.M., 2011. Solutions for a cultivated planet. Nature 478 (7369), 337–342.

Fort, R., Ruben, R., 2009. The impact of Fairtrade on banana producers in northern Peru. In 2009 Conference, August 16-22, 2009, Beijing, China (No. 50964; 2009 Conference, August 16-22, 2009, Beijing, China). International Association of Agricultural Economists. https://ideas.repec.org/p/ags/iaae09/50964.html.

Froehlich, A.G., Melo, A.S.S.A., Sampaio, B., 2018. Comparing the profitability of organic and conventional production in family farming: empirical evidence from Brazil. Ecol. Econ. 150 (C), 307–314.

Garibaldi, L.A., Gemmill-Herren, B., D'Annolfo, R., Graeub, B.E., Cunningham, S.A., Breeze, T.D., 2017. Farming approaches for greater biodiversity, livelihoods, and food security. Trends Ecol. Evol. 32 (1), 68–80. https://doi.org/10.1016/j.tree.2016.10.001.

Haggar, J., Soto, G., Casanoves, F., Virginio, E. de M., 2017. Environmental-economic benefits and trade-offs on sustainably certified coffee farms. Ecological Indicators, 79, 330–337. https://doi.org/10.1016/j.ecolind.2017.04.023.

Ibanez, M., Blackman, A., 2016. Is eco-certification a win-win for developing country agriculture? organic coffee certification in Colombia. World Dev. 82, 14–27. https://doi.org/10.1016/j.worlddev.2016.01.004.

IISD: International Institute for Sustainable Development, 2015. Voluntary Sustainability Standards. International Institute for Sustainable Development, IISD https://www.iisd.org/topic/standards.

Ingram, V., Van Rijn, F., Waarts, Y., Gilhuis, H., 2018. The impacts of cocoa sustainability initiatives in West Africa. Sustainability 10 (11), 4249. https://doi.org/10.3390/su10114249.

ITC (International Trade Centre). Standards Map. (2015). http://standardsmap.org/identify2.aspx.

Jena, P.R., Chichaibelu, B.B., Stellmacher, T., Grote, U., 2012. The impact of coffee certification on small-scale producers' livelihoods: a case study from the Jimma Zone, Ethiopia. Agric. Econ. 43 (4), 429–440. https://doi.org/10.1111/j.1574-0862.2012.00594.x.

Jena, P., Stellmacher, T., Grote, U., 2015. Can coffee certification schemes increase incomes of smallholder farmers? Evidence from Jinotega, Nicaragua. Environ. Dev. Sustain. 19 https://doi.org/10.1007/s10668-015-9732-0.

Jena, P., Grote, U., 2017. Fairtrade Certification and Livelihood Impacts on Small-scale Coffee Producers in a Tribal Community of India. Appl. Econ. Perspect. Policy 39 (1), 87–110. https://doi.org/10.1093/aepp/ppw006.

Jha, S., Bacon, C.M., Philpott, S.M., Ernesto Méndez, V., Läderach, P., Rice, R.A., 2014. Shade Coffee: update on a disappearing refuge for biodiversity. Bioscience 64 (5), 416–428. https://doi.org/10.1093/biosci/biu038.

Jones, S., Gibbon, P., 2011. Developing agricultural markets in Sub-Saharan Africa: organic cocoa in Rural Uganda. J. Develop. Stud. 47 (10), 1595–1618. https://doi.org/10.1080/00220388.2011.579107.

Karki, S.K., Jena, P.R., Grote, U., 2016. Fairtrade Certification and Livelihoods: A Panel Data Analysis of Coffee-growing Households in India. https://doi.org/10.15488/1712.

Kleemann, L., Abdulai, A., 2013. Organic certification, agro-ecological practices and return on investment: Evidence from pineapple producers in Ghana. Ecol. Econ. 93, 330–341. https://doi.org/10.1016/j.ecolecon.2013.06.017.

Komives, K., Arton, A., Baker, E., Kennedy, E., Longo, C., Newsom, D., Pfaff, A., Romero, C., 2018. Conservation impacts of voluntary sustainability standards: How has our understanding changed since the 2012 publication of "Toward sustainability: The roles and limitations of certification"? Washington DC: Meridian Institute. Available at merid.org/content/projects/supply_chain_sustainability_research_fund.

Last, L., Arndorfer, M., Balázs, K., Dennis, P., Dyman, T., Fjellstad, W., Friedel, J.K., Herzog, F., Jeanneret, P., Lüscher, G., Moreno, G., Kwikiriza, N., Gomiero, T., Paoletti, M.G., Pointereau, P., Sarthou, J.-P., Stoyanova, S., Wolfrum, S., Kölliker, R., 2014. Indicators for the on-farm assessment of crop cultivar and livestock breed diversity: a survey-based participatory approach. Biodivers. Conserv. 23 (12), 3051–3071. https://doi.org/10.1007/s10531-014-0763-x.

Latruffe, L., Diazabakana, A., Bockstaller, C., Desjeux, Y., Finn, J., Kelly, E., Ryan, M., Uthes, S., 2016. Measurement of sustainability in agriculture: a review of indicators. Stud. Agric. Econ. 118 (3), 123–130. https://doi.org/10.7896/j.1624.

Margoluis, R., Stem, C., Salafsky, N., Brown, M., 2009. Design alternatives for evaluating the impact of conservation projects. New Direct. Eval. 2009 (122), 85–96. https://doi.org/10.1002/ev.298.

Mascia, M.B., Pailler, S., Thieme, M.L., Rowe, A., Bottrill, M.C., Danielsen, F., Geldmann, J., Naidoo, R., Pullin, A.S., Burgess, N.D., 2014. Commonalities and complementarities among approaches to conservation monitoring and evaluation. Biol. Conserv. 169, 258–267. https://doi.org/10.1016/j.biocon.2013.11.017.

Meemken, E.-M., Qaim, M., 2018. Can private food standards promote gender equality in the small farm sector? J. Rural Stud. 58, 39–51. https://doi.org/10.1016/j.jrurstud.2017.12.030.

Meemken, E.-M., Spielman, D.J., Qaim, M., 2017. Trading off nutrition and education? a panel data analysis of the dissimilar welfare effects of Organic and Fairtrade standards. Food Policy 71, 74–85. https://doi.org/10.1016/j.foodpol.2017.07.010.

Meemken, Eva-Marie, 2020. Do smallholder farmers benefit from sustainability standards? a systematic review and meta-analysis. Global Food Security 26 (September), 100373. https://doi.org/10.1016/j.gfs.2020.100373.

Milder, J.C., Arbuthnot, M., Blackman, A., Brooks, S.E., Giovannucci, D., Gross, L., Kennedy, E.T., Komives, K., Lambin, E.F., Lee, A., Meyer, D., Newton, P., Phalan, B., Schroth, G., Semroc, B., Rikxoort, H.V., Zrust, M., 2015. An agenda for assessing and improving conservation impacts of sustainability standards in tropical agriculture. Conserv. Biol. 29 (2), 309–320. https://doi.org/10.1111/cobi.12411.

Minten, B., Dereje, M., Engida, E., Tamru, S., 2018. Tracking the quality premium of certified coffee: evidence from Ethiopia. World Dev. 101 (C), 119–132.

Mitiku, F., De Mey, Y., Nyssen, J., Maertens, M., 2017. Do private sustainability standards contribute to income growth and poverty alleviation? a comparison of different coffee certification schemes in Ethiopia. Sustainability 9 (2), 246. https://doi.org/10.3390/su9020246.

Mitiku, F., Nyssen, J., Maertens, M., 2018. Certification of semi-forest coffee as a land-sharing strategy in Ethiopia. Ecol. Econ. 145, 194–204. https://doi.org/10.1016/j.ecolecon.2017.09.008.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group, 2009. *P*referred *r*eporting *i*tems for *s*ystematic reviews and *m*eta-*a*nalyses: the PRISMA statement. PLoS Med 6 (7), e1000097. https://doi.org/10.1371/journal.pmed1000097.

Morgans, C.L., Meijaard, E., Santika, T., Law, E., Budiharta, S., Ancrenaz, M., Wilson, K. A., 2018. Evaluating the effectiveness of palm oil certification in delivering multiple sustainability objectives. Environ. Res. Lett. 13 (6), 064032 https://doi.org/10.1088/1748-9326/aac6f4.

Naidoo, R., Gerkey, D., Hole, D., Pfaff, A., Ellis, A.M., Golden, C.D., Herrera, D., Johnson, K., Mulligan, M., Ricketts, T.H., Fisher, B., 2019. Evaluating the impacts of protected areas on human well-being across the developing world. Sci. Adv. 5 (4), eaav3006. https://doi.org/10.1126/sciadv.aav3006.

O'Rourke, D., 2014. The science of sustainable supply chains. Science 344 (6188), 1124–1127. https://doi.org/10.1126/science.1248526.

Oya, C., Schaefer, F., Skalidou, D., 2018. The effectiveness of agricultural certification in developing countries: a systematic review. World Dev. 112, 282–312. https://doi.org/10.1016/j.worlddev.2018.08.001.

Petrokofsky, G., Jennings, S., 2018. The effectiveness of standards in driving adoption of sustainability practices: A State of Knowledge Review. 3Keel and ISEAL Alliance.

Potts, J., Lynch, M., Wilkings, A., Huppe, G., Cunningham, M., Voora, V., 2014. The state of sustainability initiatives review 2014: Sustainability and transparency. International Institute for Sustainable Development and the International Institute for Environment and Development.

Qiao, Y., Martin, F., Cook, S., He, X., Halberg, N., Scott, S., Pan, X., 2018. Certified organic agriculture as an alternative livelihood strategy for small-scale farmers in china: a case study in Wanzai County, Jiangxi Province. Ecol. Econ. 145 (C), 301–307.

Rainforest Alliance. (2017). Rainforest Alliance – Sustainable Agriculture Standard Version 1.2. Red de Agricultural Sostenible, A.C. (Sustainable Agriculture Network). https://www.rainforest-alliance.org/business/wp-content/uploads/2017/11/03_rainforest-alliance-sustainable-agriculture-standard_en.pdf.

Rasmussen, L.V., Coolsaet, B., Martin, A., Mertz, O., Pascual, U., Corbera, E., Dawson, N., Fisher, J.A., Franks, P., Ryan, C.M., 2018. Social-ecological outcomes of agricultural intensification. Nat. Sustainability 1 (6), 275–282. https://doi.org/10.1038/s41893-018-0070-8.

Qiao, Y., Halberg, N., Vaheesan, S., Scott, S., 2016. Assessing the Social and Economic Benefits of Organic and Fair Trade Tea Production for Small-Scale Farmers in Asia: A Comparative Case Study of China and Sri Lanka. Renew. Agr. Food Syst. 313, 246–257. https://doi.org/10.1017/S1742170515000162.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Roundtable on Sustainable Biomaterials. 2018. "Understanding Our Footprint: Outcomes of RSB's Monitoring & Evaluation System in 2018." Roundtable on Sustainable Biomaterials. https://rsb.org/wp-content/uploads/2019/06/RSB-Monitoring-Evaluation-System-Report-2018-Draft-2.pdf.

Ruben, R., Fort, R., Zúñiga-Arias, G., 2009. Measuring the impact of Fairtrade on development. Development in Practice 19 (6), 777–788. https://doi.org/10.1080/09614520903027049.

Ruben, R., Fort, R., 2012. The impact of fairtrade certification for coffee farmers in Peru. World Dev. 40 (3), 570–582.

Ruben, R., Zuniga, G., 2011. How standards compete: comparative impact of coffee certification schemes in Northern Nicaragua. Supply Chain Management: Int. J. 16 (2), 98–109. https://doi.org/10.1108/13598541111115356.

Rueda, X., Lambin, E., 2013. Responding to globalization: impacts of certification on colombian small-scale coffee growers. Ecol. Soc. 18 (3) https://doi.org/10.5751/ES-05595-180321.

Rueda, X., Thomas, N.E., Lambin, E.F., 2015. Eco-certification and coffee cultivation enhance tree cover and forest connectivity in the Colombian coffee landscapes. Reg. Environ. Change 15 (1), 25–33. https://doi.org/10.1007/s10113-014-0607-y.

SAI Platform. (2014). Sustainability Performance Assessment Version 2.0. https://saiplatform.org/uploads/SPA%20Guidelines%202%200.pdf.

Santika, Truly, Wilson, Kerrie A., Law, Elizabeth A., St, Freya A.V., John, Kimberly M., Carlson, Holly Gibbs, Morgans, Courtney L., Ancrenaz, Marc, Meijaard, Erik, Struebig, Matthew J., 2020. Impact of palm oil sustainability certification on village well-being and poverty in Indonesia. Nat. Sustainability 1–11. https://doi.org/10.1038/s41893-020-00630-1.

Schoonhoven-Speijer, M., & Ruben, R. (2014). Chapter 5 Maintaining sustainable livelihoods: effects of Utz certification on market access, risk reduction and livelihood strategies of Kenyan coffee farmers. In Coffee certification in East Africa: impact on farms, families and cooperatives (Vol. 1–0, pp. 149–174). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-805-6_5.

Sellare, Jorge, Meemken, Eva-Marie, Qaim, Matin, 2020. Fairtrade, agrochemical input use, and effects on human health and the environment. Ecol. Econ. 176 (October), 106718 https://doi.org/10.1016/j.ecolecon.2020.106718.

Siddaway, A.P., Wood, A.M., Hedges, L.V., 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annu. Rev. Psychol. 70 (1), 747–770. https://doi.org/10.1146/annurev-psych-010418-102803.

Smith, W.K., Nelson, E., Johnson, J.A., Polasky, S., Milder, J.C., Gerber, J.S., West, P.C., Siebert, S., Brauman, K.A., Carlson, K.M., Arbuthnot, M., Rozza, J.P., Pennington, D.

N., 2019. Voluntary sustainability standards could significantly reduce detrimental impacts of global agriculture. Proc. Natl. Acad. Sci. 116 (6), 2130–2137. https://doi.org/10.1073/pnas.1707812116.

Takahashi, R., Todo, Y., 2013. The impact of a shade coffee certification program on forest conservation: a case study from a wild coffee forest in Ethiopia. J. Environ. Manage. 130, 48–54. https://doi.org/10.1016/j.jenvman.2013.08.025.

Takahashi, R., Todo, Y., 2014. The impact of a shade coffee certification program on forest conservation using remote sensing and household data. Environ. Impact Assess. Rev. 44, 76–81. https://doi.org/10.1016/j.eiar.2013.10.002.

Takahashi, R., Todo, Y., 2017. Coffee certification and forest quality: evidence from a wild coffee forest in Ethiopia. World Dev. 92, 158–166. https://doi.org/10.1016/j.worlddev.2016.12.001.

Targetti, S., Herzog, F., Geijzendorffer, I.R., Wolfrum, S., Arndorfer, M., Balàzs, K., Choisis, J.P., Dennis, P., Eiter, S., Fjellstad, W., Friedel, J.K., Jeanneret, P., Jongman, R.H.G., Kainz, M., Luescher, G., Moreno, G., Zanetti, T., Sarthou, J.P., Stoyanova, S., Viaggi, D., 2014. Estimating the cost of different strategies for measuring farmland biodiversity: Evidence from a Europe-wide field evaluation. Ecol. Ind. 45, 434–443. https://doi.org/10.1016/j.ecolind.2014.04.050.

Tayleur, C., Balmford, A., Buchanan, G.M., Butchart, S.H.M., Ducharme, H., Green, R.E., Milder, J.C., Sanderson, F.J., Thomas, D.H.L., Vickery, J., Phalan, B., 2017. Global coverage of agricultural sustainability standards, and their role in conserving biodiversity: certification standards and biodiversity. Conservation Lett. 10 (5), 610–618. https://doi.org/10.1111/conl.12314.

Tayleur, Catherine, Juliet Vickery, Stuart Butchart, Christine Corlet Walker, Graeme Buchanan, Fiona Sanderson, Jeffrey Milder, et al. 2017b. "GIS Data for: Where Are Commodity Crops Certified, and What Does It Mean for Conservation and Poverty Alleviation?" 2 (October). https://doi.org/10.17632/mpdf6ytswm.2. License: https://creativecommons.org/licenses/by-nc/3.0/.

Tayleur, C., Balmford, A., Buchanan, G.M., Butchart, S.H.M., Corlet Walker, C., Ducharme, H., Green, R.E., Milder, J.C., Sanderson, F.J., Thomas, D.H.L., Tracewski, L., Vickery, J., Phalan, B., 2018. Where are commodity crops certified, and what does it mean for conservation and poverty alleviation? Biol. Conserv. 217, 36–46. https://doi.org/10.1016/j.biocon.2017.09.024.

The World Bank. World Bank Country and Lending Groups. (2020). https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

Tran, D., Goto, D., 2019. Impacts of sustainability certification on farm income: Evidence from small-scale specialty green tea farmers in Vietnam. Food Policy 83, 70–82. https://doi.org/10.1016/j.foodpol.2018.11.006.

Tscharntke, T., Milder, J.C., Schroth, G., Clough, Y., DeClerck, F., Waldron, A., Rice, R., Ghazoul, J., 2015. Conserving biodiversity through certification of tropical agroforestry crops at local and landscape scales. Conservation Letters 8 (1), 14–23. https://doi.org/10.1111/conl.12110.

Vanderhaegen, K., Akoyi, K.T., Dekoninck, W., Jocqué, R., Muys, B., Verbist, B., Maertens, M., 2018. Do private coffee standards 'walk the talk' in improving socio-economic and environmental sustainability? Global Environ. Change 51, 1–9. https://doi.org/10.1016/j.gloenvcha.2018.04.014.

van Rijsbergen, B., Elbers, W., Ruben, R., Njuguna, S.N., 2016. The ambivalent impact of coffee certification on farmers' welfare: a matched panel approach for cooperatives in central Kenya. World Dev. 77, 277–292. https://doi.org/10.1016/j.worlddev.2015.08.021.

Vitousek, P.M., Naylor, R., Crews, T., David, M.B., Drinkwater, L.E., Holland, E., Johnes, P.J., Katzenberger, J., Martinelli, L.A., Matson, P.A., Nziguheba, G., Ojima, D., Palm, C.A., Robertson, G.P., Sanchez, P.A., Townsend, A.R., Zhang, F.S., 2009. Nutrient imbalances in agricultural development. Science 324 (5934), 1519–1520. https://doi.org/10.1126/science.1170261.

Weber, J.G., 2011. How much more do growers receive for Fairtrade-organic coffee? Food Policy 36 (5), 678–685. https://doi.org/10.1016/j.foodpol.2011.05.007.

Willer, H., Sampson, G., Voora, V., Dang, D., Lernoud, J., 2019. State of Sustainable Markets 2019: Statistics and Emerging Trends. International Trade Center. https://www.deslibris.ca/ID/10102592.

Zulfiqar, F., Thapa, G.B., 2016. Is 'Better cotton' better than conventional cotton in terms of input use efficiency and financial performance? Land Use Policy 52, 136–143. https://doi.org/10.1016/j.landusepol.2015.12.013.