The Triviality of Measuring Ultimate Outcomes: Acknowledging the Span of Direct Influence

Giel Ton, Sietze Vellema and Lan Ge

Abstract Sustainability standards and certification schemes have been promoted as a market-driven instrument for realising development impacts and receive public funding. As a result, companies, NGOs and supporting donors and governments want to know if these ambitions have been fulfilled. Their tendency is to commission household surveys to assess net effects of certification in areas such as poverty, productivity and food security. This article argues that, rather than trying to measure precise net effects on farmer income, the focus should be on detailed measurement of more immediate outcomes in terms of knowledge and implementation of good agricultural practices. Contribution analysis is proposed as an overall approach to verify the theory of change, combining survey-based net-effect measurement of these immediate and intermediate outcomes with less precise, lean monitoring of indicators to verify the contributory role of these outcomes that are outside the span of direct influence, such as household income and poverty alleviation.

1 Introduction

The private sector and market-led strategies have become increasingly central to development policy and practice. Moreover, non-governmental organisations (NGOs) are teaming up with companies or private–public partnerships. This shift from public to private-led development strategies is based on changing expectations of the role of trade versus aid for poverty alleviation. In many donor countries, this policy is increasingly based on the assumption that the private sector is more effective in reaching development goals than development aid through governments or NGOs. Accordingly, donor agencies have begun to re-allocate public resources to companies and private–public partnerships. From a public perspective, the obvious question for impact evaluation is how to demonstrate this assumed effectiveness.

In general, donor agencies prefer precise measurements of net effects in relation to the Millennium Development Goals (MDGs), with income generation and poverty reduction as main objectives (DCED 2010; DGIS 2011). This often translates into survey-based research designs, including baseline studies, randomised sampling, and comparison groups. This article challenges the exclusive emphasis on precise measurement of income effects in quasi-experimental evaluation designs. Net effects, especially those related to business performance and income, are influenced by a wide range of intervening factors that are impossible to control for under real-world conditions. This makes it difficult to attribute effects to the actual interventions and provides little information on the effectiveness of developmental activities. Based on our experiences with impact studies of certification-induced training programmes for farmers (Ton, Vellema and de Ruyter de Wildt 2011; Ton 2012; Waarts *et al*. 2012; Waarts *et al*. 2013a, 2013b, 2013c), we argue that there are good reasons to limit this dominant focus on measuring net effects in ultimate outcomes, and propose to shift attention to the domain of immediate and intermediate outcomes.

This article uses the example of certification to discuss the methodological challenges for impact evaluation of market-led development interventions. Sustainability standards and the related certification schemes, implemented in tropical commodity chains such as cocoa and

IDS Bulletin Volume 45 Number 6 November 2014
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

coffee, aim to enhance environmental sustainability, social justice and economic viability. Multinational firms and global NGOs partner in defining and implementing these sustainability standards (Vellema and van Wijk 2014). Government and donor agencies are motivated to support such endeavours because they believe that implementation of these standards is instrumental to achieving development goals. Standards systems aim to enhance their public accountability, but also to shift attention to their impact on more intermediate outcomes. We describe recent advances in these efforts by certification schemes and illustrate the challenges in impact evaluation of these types of interventions.

The challenge addressed by this article is to find ways to get credible data on outcomes that are still attributable to the support interventions that are related with certification, and to do so in a way that allows comparison between different possible support modalities that may lead to the same type of outcomes. We propose to measure and compare the effectiveness of activities foremost on the increase in knowledge (immediate outcomes) and improved business practices (intermediate outcomes). Further, we aim to verify the contribution of these intermediate outcomes to business performance (ultimate outcomes) and development impact. This entails a combined use of the realist notion of verifying and refining programme theories (Pawson and Tilley 1997; Rogers 2009; Ton 2012; Vellema *et al*. 2013) and a mix of methods to collect evidence that bolsters the 'contribution story' (Mayne 2001, 2012). Data collection in an impact evaluation along the lines of contribution analysis uses multiple strands of evidence to verify, support or challenge the key assumptions in the intervention logic. 'The research builds a compelling case with evidence from which it is reasonable to conclude with confidence that the intervention has made a contribution and why' (Mayne 2012). Contribution analysis combines the precise measurement of the outcomes and the analysis of the causal processes set in motion within the span of direct influence of an intervention, with the monitoring of outcomes and influencing factors outside the span of direct influence. In addition to survey-based research, a mix of methods is used to enable cross-case comparative analysis, as well as for finding out from stakeholders and expert panels how relevant the intervention is compared to alternative strategies (benchmarking).

The article first contextualises the challenge addressed. It reviews some important initiatives to improve reporting on the impacts of standards-setting and certification in the cocoa and chocolate industries. Second, we reflect on our experiences with the design and implementation of survey-based impact evaluation in cocoa production in Côte d'Ivoire, Finally, we discuss the implications of our findings for future impact evaluations of private-sector support programmes. We propose to limit rigorous measurements of net effects to outcomes and processes 'within the realm of the programme' and to use a mix of methods to collect information to verify the assumption that these business practices (intermediate outcomes) are contributing factors that together generate a change in business performance and development impact.

2 Setting: impact evaluation of certification

Development impacts are generally framed in terms of the triple P (RSCE 2009): People-Planet-Profit. Accordingly, texts accompanying certification schemes, such as UTZ Certified, Rainforest Alliance, Fairtrade or Organic, suggest contributions to environmental sustainability (reflected in benign farming practices and conservation of forests, natural resources and biodiversity), social justice (reflected particularly in labour rights, improved working conditions and inclusion of marginalised groups) and economic fairness (reflected mainly in business opportunities for smallholder farmers, improved rural incomes and living conditions, and vitality of a sector). In addition to these developmental goals, standard systems have more internal objectives related to the logistics and verification of quality and quantity of transactions in the value chain, reliable and cost-efficient sourcing models, and traceability in the chain of custody.

The objective of fostering sustainability in the supply chain through certification is aligned with concerns for corporate social responsibility on the part of leading companies involved in global trade and processing of tropical commodities. Likewise, governments and public donor agencies support certification because they themselves are committed to sustainability as a public goal, and consider market-led intervention strategies as an effective vehicle to achieve this. As a consequence, NGOs, businesses and governments in the field of certification and sustainability in cocoa need to report on their achievements, both to account

for public funding and to convince consumers of the benefits of paying an additional price for the products with a certificate.

2.1 Public accountability requirements

In 2010, a group of cocoa-processing companies, retailers, chocolate manufacturers, NGOs and Dutch ministries signed a letter of intent to support the revitalisation of cocoa production in West Africa to enhance the consumption of sustainably produced and certified chocolate in the Netherlands (Chocolate Working Group 2010). The Netherlands is the world's largest importer of cocoa and is home to large processing facilities as well as the offices of several voluntary standards bodies that govern the certification process, which explains the private and public interest in such a partnership in this country. The endeavour is linked to the Dutch Ministry of Economic Affairs' policy concerning International Cocoa Agreements and to public–private partnerships working on the Roundtable for a Sustainable Cocoa Economy, the World Cocoa Foundation and the Dutch Sustainable Trade Initiative (IDH). Stakeholders in the partnership have agreed to source only certified cocoa in 2025, as a joint commitment to enhance sustainability.

The assumption underlying this partnership is that an increase in market share of certified chocolate would lead to an increase in sustainability of the cocoa supply. The letter of intent places a strong emphasis on measuring the market share of certified chocolate products in the Dutch market as a proxy-indicator of impact. The following year, the Netherlands Ministry of Foreign Affairs gradually increased evaluation requirements for public funding of private-sector support programmes and required them to report on impact on poverty and food security. In the 'Protocol on Evaluability and Attainment of Results' (DGIS 2011), it demanded a monitoring and evaluation plan that included baseline, progress and end-of-project measurements and the use of control groups for robust net-effect measurements. The protocol suggested measuring and reporting on the impact of private-sector development support on nutritional status and household income of the beneficiary population. In addition to these mandatory impact areas, the protocol suggested, among several other things, to measure productivity of land use, input efficiency, access to training and finance, and quality of the business environment. This tendency to increase the requirements on private-sector recipients of development aid, in order to better elucidate the effectiveness of their interventions, is not specific to the Netherlands, but is a generalised trend among all OECD donors (DCED 2010).

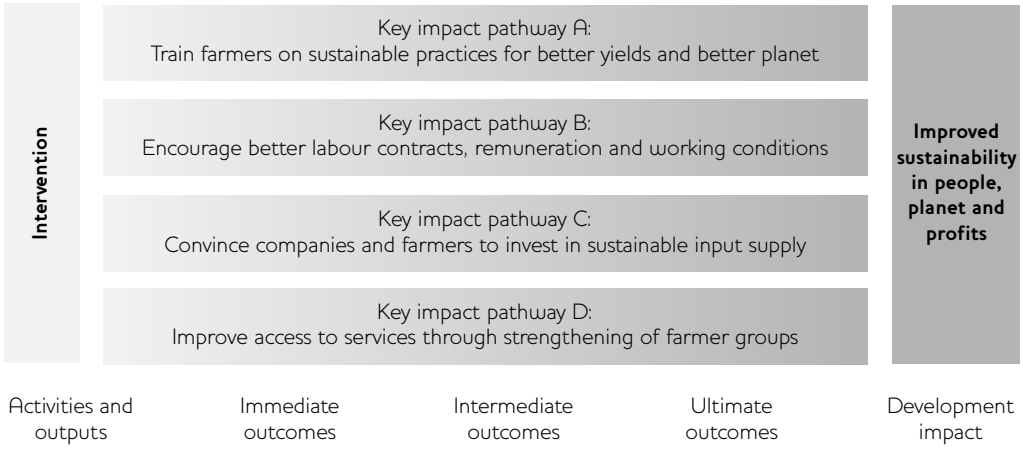
2.2 Harmonised indicators and rigorous measurement

The International Social and Environmental Accreditation and Labelling Alliance (ISEAL Alliance) set out to improve the quality of impact evaluation of certification and to respond to accountability requirements with credible evaluation research. ISEAL aims to introduce minimum quality requirements for monitoring and evaluation by standards systems and certification schemes (ISEAL Alliance 2014) and to advance towards harmonising outcome indicators between sustainability systems (ISEAL Alliance 2013).

ISEAL requires standards systems and certification schemes to ensure the quality of performance-monitoring data and of outcome and impact evaluations to guarantee transparency of the sustainability claims communicated to consumers (ISEAL Alliance 2014). The scheme owner must ensure that at least some of these are independent impact evaluations. Harmonisation of indicators used to track outcomes and impact would permit benchmarking and comparison between standards systems.

Where donor communities emphasised reporting on sustainable economic development and poverty reduction (DCED 2010), discussions within ISEAL shifted attention to the measurement of more tangible outcome areas within the sphere of control of these voluntary standards organisations. This means evaluating such aspects as the adoption of conservation practices, yields, sales practices, satisfaction with crop profitability, perceptions about changes in natural resources, etc. Several of these common indicators are still in the process of being fine-tuned, for example comparative measurements to obtain knowledge on and adoption of good agricultural practices in specific crops (Rigby *et al*. 2001; Russillo and Pintér 2009; El Hage 2012), assessment of the capabilities of farmer groups to engage in marketing and value-chain coordination (Donovan and Stoian 2012; Ton *et al*. 2014), and the use of a common multi-dimensional poverty index.

Figure 1 **Stylised representation of the theory of change developed with UTZ Certified composed of distinct impact pathways**



Source Authors' own elaboration, based on discussions with UTZ Certified. A more detailed version of their theory of change is published in UTZ Certified (2014: 11–14)

As early as 2007, the demand for better impact evaluation of voluntary sustainability standards had led to an international, multi-stakeholder initiative to improve the quality of these measurements: the Committee on Sustainability Assessment (COSA, http://thecosa.org/). COSA's main efforts have been to develop, pilot and implement metrics and indicators for measuring sustainability outcomes over time. COSA emphasises the need for times-series data and comparison groups and the use of econometric methods to limit selection bias. They propose to gather information on the different aspects of sustainability by using lists of questions, which are converted into dummy variables, and through Principal Component Analysis converted into factor scores representing the relative position of the respondents in relation to various aspects of sustainability (COSA 2013). However, even these sophisticated quasi-experimental designs, or designs that deviate from random assignments of treatments, struggle with increasing numbers of observable and unobservable factors that influence the ultimate outcomes in farm performance that are only indirectly influenced by activities in the field.

2.3 Refined theories of change

Methodological challenges related to the above initiatives on impact evaluation encouraged a discussion between practitioners and researchers about feasible ways to register and attribute impact. Next to being a way to be publicly accountable for their role in reaching the MDGs, impact evaluation needed to be instrumental for gathering information that could help to improve the intervention strategy itself (Nelson and Martin 2012). ISEAL initiated and supported a series of consultations to develop and modify an Impact Code (ISEAL Alliance 2014), specifying how voluntary standards organisations should show the outcomes and impact of certification and standards in a credible way, while respecting the logistic and budgetary constraints of the implementing partners. The discussions in ISEAL stimulated a growing interest in learning how to construct and refine theories of change, and to select those performance indicators that would help to verify the key assumptions in their intervention logic (Rogers 2008; White 2009; Ton 2012).

In 2011, we helped several voluntary standards organisations and their implementing partners to define their theory of change. For this, we used Mayne's framework (2001), which differentiates the main activities and outputs per stakeholder group, the immediate outcomes in knowledge of these stakeholders, the intermediate outcomes in improved practices of the stakeholders and the ultimate outcomes in performance indicators related to these modified (business) practices. Based on a detailed 'cloud' of outcome areas, derived from their programme documents and mission statements, we developed a stylised representation (Figure 1), in which we identified several different impact pathways.

The stylised representation resulted from our exercise with UTZ Certified, and was later further refined and modified by UTZ in several of their communications, for example the 2014 Impact Report (UTZ Certified 2014). The impact logic of the support of UTZ Certified assumes several pathways that are expected to lead to poverty reduction. For example, compliance with the prescribed agricultural practices and the provision of extension services is expected to increase the efficiency of cocoa production and consequently result in higher and more stable household incomes. Moreover, support to farmer organisations in managing an internal control system is expected to enhance their capacity to negotiate prices and/or to obtain better access to credit and agri-inputs. Reliable access to output markets and predictable incomes makes it possible for cocoa farmers to invest in their farms and offer better remuneration and working conditions to farm labourers.

This identification of various pathways proved useful to the key evaluation questions in commissioned research on impact, and helped to find appropriate outcome areas needing to be monitored. Each pathway embodies a specific sequence of causal steps and configurations of influencing factors, and each will have a specific result chain to graphically represent this causal logic. These result chains embody the assumptions about causal relationships between the main activities per stakeholder group, the main outputs of these activities, and the outcomes and development impacts.

2.4 The evaluation challenge in certification

Even when certification bodies communicate impressive differences in yields and income between certified farmers and comparison groups (UTZ Certified 2014), the attribution claims reported in the more rigorous studies are more modest, as selection biases and the influence of confounding factors cannot be entirely ruled out. Whereas it would be ideal, from the perspective of donors and standard-setting bodies, that impact studies measure exact net effects on poverty and environment, in reality this becomes difficult or impossible, as these outcomes are in fact influenced only marginally by the certification-related activities in the field. The efficacy of the intervention becomes increasingly dependent on activities of other actors or factors.

Only truly experimental designs, such as randomised controlled trials (RCTs), are sufficiently robust to handle the influence of many observable and unobservable confounding factors, if they are based on random assignment of fairly uniform treatments. However, random assignment of the support is, from the perspective of the implementers of this support and the traders that need to sell the certified products logistically highly undesirable. Certification efforts need a crucial mass of farmers in a geographically constrained area and an internal control system that builds on the locally available organisational social capital. The context will vary and activities with farmers (treatments) tend to be fairly heterogeneous, as they are adapted to cope with these contextual differences.

Designs of evaluations, therefore, will need to use quasi-experimental, observational studies, in which a group of beneficiaries is compared with a group of farmers that did not receive support. Several survey-based research designs are available to cope with differences in context and characteristics between these two groups (Shadish, Cook and Campbell 2002; Khandker, Koolwal and Samad 2009). However, the econometric methods to find 'comparable' treatment and comparison households in these quasi-experimental designs depend on the limited information on key characteristics that is available. Even so, they are contingent on normative decisions about what to include or exclude as a variable in the matching model. Net effects on ultimate outcomes measured with quasi-experimental designs always have, therefore, a high level of inaccuracy and are subject to the positive or negative biases of enumerators, data analysts and inferring researchers.

Nevertheless, in the field of certification, reporting on net-income effects tends to be the prime focus of impact evaluations. Information on the mean and variations in income tend to be the anchors for statistical power calculations and the determination of minimum sample sizes. To calculate household income, fairly detailed quantitative information about crop revenue and input costs are needed. In the context of diversified farm systems with multiple crops, most common in smallholder agriculture, the disaggregation of labour time and input costs for the target crop is notoriously difficult. This results in ambiguity in the constructs that are

used as proxies for income effects (Ton *et al*. 2010), for example net income including self-consumption of production, cash income, agricultural cash income, and income from target crops. The estimates given by the farmer of market prices, input costs and labour time spent is prone to recall bias, which results in unreliable income estimates when used in calculations.

Recent systematic reviews of the effectiveness of certification point to inconclusive results and, therefore, limited usefulness of net-income estimates in studies on the impact of certification (Blackman and Rivera 2010; Alvarez and von Hagen 2011, 2012; Blackmore *et al*. 2012; Crosse, Newsom and Kennedy 2012; SCSSC 2012). Except when price premiums are an important component of the intervention, as with Fairtrade and Organic certification, the positive or negative changes in farmers' income are only very remotely related to the support and services provided to comply with the certification requirements. A wide range of factors determines income (yield, input use, costs, etc.). In most certification-related interventions, a change in yield or farm income will be influenced by a set of agricultural practices that are being promoted, such as the use of improved varieties, different handling protocols for plants and products, or new or enhanced soil conservation measures. However, even more important are the factors over which the intervention has no control at all, such as site-specific weather patterns that define yields, changes in market prices in response to site-specific trade dynamics, competition between buyers, or changes in crop patterns or off-farm income due to employment-generating activities and seasonal out-migration. Multi-year agronomic research could provide more convincing evidence to support the assumed impact of improved agricultural practices on yields and income than the estimates derived from information collected through household surveys.

Our aim has been to link the above methodological considerations with ongoing discussions on impact assessments within ISEAL and the practitioners' domain of standards-setting and certification (e.g. Vellema 2010; Vellema and Ton 2012). Our premise is that impact evaluation can better focus on the measurement of the change in knowledge on and implementation of agricultural practices, and 'reason through' the likely effect of these practices on yields and household income. We have illustrated the evaluation challenge by describing one of our experiences in the design and implementation of an evaluation study on certification of cocoa farmers in Côte d'Ivoire.

3 Case study: evaluation of training for certification in Côte d'Ivoire

As researchers of Wageningen UR, we were contracted by a number of organisations to co-design survey-based impact evaluations of cocoa certification initiatives in Ghana, Côte d'Ivoire and Indonesia. Cocoa certification schemes typically require a codified set of good practices related to cocoa production and farm management, and include third-party auditing to confirm that the requirements are met or will be met within a specified time frame. In designing the methodology for these evaluation assignments, we attempted to identify the outcomes and processes in field-level certification initiatives that could be attributable to certification initiatives. In most cases, the main intervention was related to training of farmers in good agricultural practices – a mandatory requirement by certification schemes such as UTZ Certified – and the organisation of internal control systems.

3.1 Multiple actors, treatments and contexts

UTZ Certified is a certification scheme whose aim is to support sustainable farming worldwide. Its mission is to create a world in which sustainable farming is the norm – a world in which farmers implement good agricultural practices (GAPs) to manage their farms profitably with respect for people and planet, industry invests in and rewards sustainable production, and consumers can enjoy and trust the products they buy. In 2007, UTZ Certified launched its cocoa programme with founding members Cargill, Ecom, Heinz, Mars, Nestlé and Ahold and the not-for-profit organisations Solidaridad, Oxfam Novib and World Wildlife Fund (WWF), with the first pilots in Côte d'Ivoire starting in 2008. In 2012, UTZ Certified and one of its implementing partners, the Dutch NGO Solidaridad, commissioned LEI-Wageningen UR to design and implement an impact evaluation of their cocoa programme in Côte d'Ivoire.

Implementation of the cocoa programme involved a heterogeneous group of actors, each with different objectives, roles and responsibilities. NGOs, traders, private partnerships, governments and international public organisations were

Table 1 Example of the variation in treatments in support programmes targeting cocoa farmers in Côte d'Ivoire

Type of treatment	Number	Explanation
Trader-specific CSR programmes	8	Trading companies have their own CSR programme related to certification
Intervention activities per trader	2 to 5	Traders may provide training on business skills, organising demo plots, providing gifts, inputs, nursery and seedling supply, etc.
Phases in certification schemes	6	Farmers can be in starting phase of the programme or have been in the certification programme for up to five years
Training on different topics	>10	Topics including a variety of elements covering production methods, use of Personal Protection Equipment (PPE), waste management, etc.

Source Authors' own elaboration, based on field data reported by Ingram *et al.* (2014).

partners in implementing the certification programme. All cocoa farmers in Côte d'Ivoire covered by UTZ Certified were organised as producer groups, which were generally cooperatives of varying sizes. Most producer groups were linked to particular traders, who assisted them in attaining certification. These traders could target more than one certification scheme (for example Rainforest Alliance Certification, Fairtrade). As a result, half of the 86 UTZ Certified producer groups had multiple certifications with overlapping requirements. Approximately 21 per cent of the farmers participating in the UTZ programme were also Rainforest Alliance certified, and 2 per cent were both UTZ and Fairtrade certified. This resulted in a high level of diversity in the actual 'treatment' that the beneficiaries received (Table 2). The modalities of training differed from centralised sessions with professional agronomists to more intensive and participatory methods of knowledge exchange, such as field demonstrations in so-called Farmer Field Schools. Training was sometimes combined with additional support such as the supply of agro-inputs or credit.

This complexity of the interventions, which was detected at baseline, complicated the design of the impact study. The initial approach, which was to assess impact through a (matched) comparison of the 'UTZ-programme group' (the 'treatment group') and the 'non-UTZ group' (the 'control group'), needed to be modified due to the diversity of treatments in the 'certified group'.

Furthermore, both prior to and during the UTZ Certification programme, other activities had taken place relating to sustainable cocoa production, which had addressed the same type of practices that were promoted through the programme. Obviously, this history of support from multiple sources in different agro-ecological zones made it even more challenging to find treatment and comparison groups that would allow us to attribute changes in outcomes to the UTZ programme activities.

The variations in treatments were further compounded by differences in agro-ecological conditions in which the smallholder farmers operated. The statistical analysis of baseline indicators showed a very high level of variability (standard deviations of more than 75 per cent of the mean) in indicators such as productivity and yield or net income. We used a regression analysis with more than 20 explanatory variables to detect the differences in outcomes that could be related to differences in characteristics between the group of farmers that were included in certification-related activities (treatments), and those that were not. The heterogeneity in treatments and farmer characteristics meant that few treatments had a significant correlation because of the dominant impact of being situated in specific agro-ecological zones having different production potentials. This pointed to the need to, at least, match treatment and comparison groups according to their agro-ecological zone.

3.2 Sample sizes for rigorous measurements

Variability in treatments and contexts has substantial implications for the sample size required to detect a significant difference in income by comparing treatment and comparison groups. Table 2 presents an estimate of the minimum sample size that would be needed using

Table 2 Minimum sample size calculations based on estimated effect sizes and baseline standard deviations

Outcome indicators		Trained group in same agro-ecological zone at baseline				Plausible net-effect to be captured in the research (a)	Minimal sample size needed for the hypothetical difference between the two groups to be statistically significant ($p<0.05$) using a two sample T-test*
Category	**Indicator**	**Actual sample size (N)**	**Mean**	**Standard deviation**	**Variability (b)**	**%**	**Total N**
Immediate outcomes	Knowledge score	436	0.246	0.110	0.45	30	74
Intermediate outcomes	Implementation score	436	0.241	0.054	0.22	20	42
Ultimate outcomes	Yield (kg/ha)	406	531	416	0.78	10	1914
	Net income (USD/ha)	326	712	666	0.93	10	2718

Source Authors' own elaboration, based on Ingram *et al.* (2014).

Note *Sample sizes were estimated using the statistical package Stata 13 (*StataCorp. 2013. Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP) with the command: power two means *1 1+a*, power(0.8) sd(b), where *a* is the plausible net-effect, *b* is the variability of the indicator (calculated as the standard deviation divided by the mean).

a simple T-test to detect the expected differences in means between a trained and an untrained group. Although we used regressions to derive impact estimates, these minimum standard sizes were indicative of the minimum size of the household survey needed to detect significant differences between the different groups and reject the null hypothesis of no impact.

The measurement of net effects in knowledge increase and implementation of farming practices appear to be constrained by the logistics and budget of an impact evaluation (Table 2). The minimum sample sizes needed to detect the (plausible) effects on yield and income between two groups were prohibitively high. Why should we collect very precise data on income and yields, with long interviews and burdensome data cleaning, when this would anyhow not give strong evidence of net effects? We decided that, for net-effect estimates, we could better restrict ourselves to the precise measurement of immediate and intermediate outcome indicators. With a reliable instrument to assess knowledge levels and farmer practices, we could verify the key assumption that training induced by certification is instrumental in changing farmer practices. The positive impact of these practices for yield and income would have to be verified with other methods, not household surveys.

Anticipating this, we had piloted an instrument during the baseline study to assess the level of knowledge on GAPs in cocoa and the implementation of these practices by the farmers as immediate and intermediate outcome indicators. The farmers' knowledge level was estimated as a 'knowledge score' derived from their answers to a range of multiple-choice questions on GAPs as required by UTZ Certified. Farmers' implementation of GAPs was similarly measured as an 'implementation score', based on

their answers to questions about the practices that they had implemented on their fields. The valuation of the 'correctness' of the various answer categories for each of these questions was determined in consultation with agronomic specialists from UTZ Certified and local research institutes.

We reflected on the practices that were being promoted in consultations with groups of farmers and cocoa experts. Farmers and researchers may disagree on which practices are considered to be 'good agricultural practices', as they might use different criteria to judge some of these GAPs. Regular consultations with a multi-disciplinary expert panel to validate the local appropriate set(s) of GAPs in cocoa is a cost-effective way to capitalise on the available experiences and evidence and 'reason through' the contribution of these modified practices to yields. To feed this discussion with farmers and experts, data were collected through household surveys on the reasons for implementing the practices or not. Special attention was paid to those practices that multiple training programmes have tried to convince farmers to introduce, but where implementation was low.

4 Discussion

The evaluation challenge for certification, described in this article, applies to a wider range of development interventions. Most impact studies of market-led development strategies tend to focus on outcomes related to the performance of business practices, such as rural incomes or wellbeing, which are difficult to attribute to the actual processes set in motion by the private-sector support. Similar to the support to farmers in certification, these support interventions involve multiple actors and have many intervening factors that influence their performance. This makes it impossible to attribute changes in outcomes to one specific type of activity (treatment), or, even worse, to one specific supporting agency.

Because outcomes can be quite diverse, they may be difficult to simply compare between treatment and comparison groups. For example, the enhanced social network that results from certification-related activities may provide access to additional sources of credit, or, when the social network of a farmer is extended, this may stimulate the migration of their children to gain more promising livelihoods.

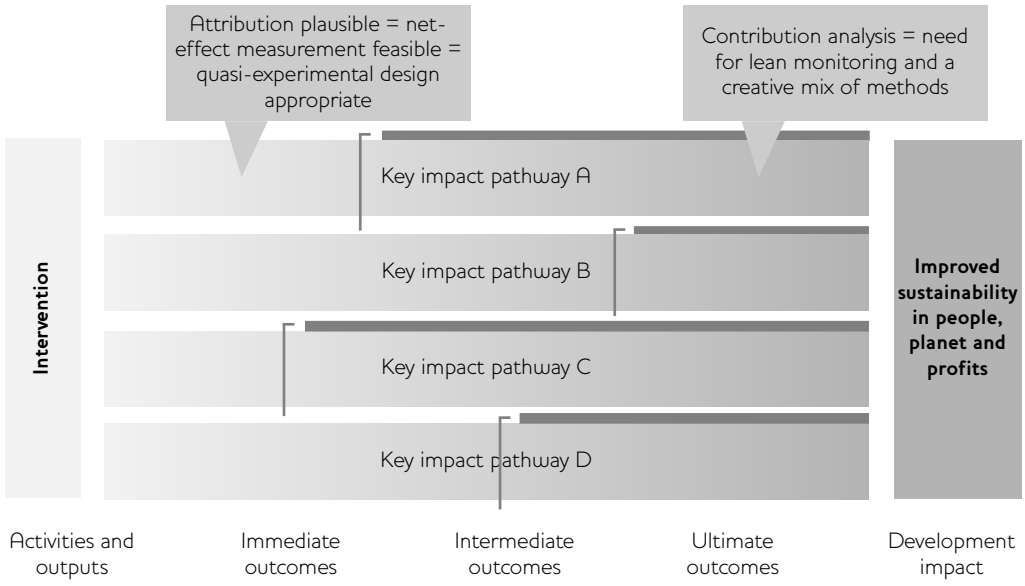
We showed that training in GAPs is, at most, one of the contributory factors (Mayne 2001; Stern *et al*. 2012) to ultimate outcomes, as are household income or yields. Monitoring changes in such ultimate outcomes may be possible, but deriving net effects and claiming attribution of changes in these outcomes to a single part of this complex of factors is not. Instead of attributing net effects using farmer or business surveys, other methods to verify the role of an intervention are needed. A research approach is needed that examines whether the type, amount and timing of support is right, instead of focusing on the causal effect of only one treatment – just as it takes matches, fuel and oxygen to start a fire (Mackie 1965).

One of the possible ways to handle the evaluation challenges in private sector support, which emerged out of our collaborative research experience with certification organisations, is the importance of better explaining the theories of change. In this work, we identified impact pathways and discussed where to draw boundaries around the direct span of influence of the intervention in each of these pathways to impact. This exercise helped unravel plausible causal processes triggered by interventions and to identify multiple intervening factors that influenced outcomes; these factors, or combinations of factors, are often more essential to effecting changes and at the same time largely unpredictable.

It is evident that a creative mix of methods is needed to collect the multiple strands of evidence to ascertain the contribution story (Mayne 2001, 2012). We expect that large-scale surveys of household income and expenditures will be less useful for reflecting on how certification improves sustainability and alleviates poverty than more qualitative approaches. It is also important to analyse how interventions contribute to wider development processes. Expert panels that reflect on sector dynamics and identify the strengths and weaknesses of various interventions can assess the achievement of intended ultimate outcomes; this can help to estimate the added value of the support.

The implication of the above is that, when moving towards ultimate outcomes, it is important to recognise where interdependencies of factors and actors become too dominant to make net-effect measurement feasible. This suggests a shift from

Figure 2 **Different impact evaluation designs within and outside the span of direct influence of an intervention**



Source Authors' own elaboration.

'impossible' quantitative-attribution-oriented research on ultimate outcomes to a precise identification of proxy-indicators for key immediate and intermediate outcome areas that are still (plausibly) attributable to the intervention (Figure 2). As illustrated in Figure 2, each pathway in the intervention logic will have a different boundary of the span of direct influence – a boundary for which net-effect measurements become impossible. This boundary will be a result of causal-theoretical logic, and a function of budgetary and methodological constraints (Bamberger *et al*. 2004). Quasi-experimental research designs may be appropriate when assessing net effects within this bounded span of influence. They will need to focus on those immediate and intermediate outcome indicators within the span of direct influence, where a change (+ or -) is still indicative for performance of the intervention. Only on those outcomes can these designs provide the 'credible counterfactual' (Ruben, Fort and Zuniga-Arias 2009; Alvarez and von Hagen 2011) and generate meaningful and informative net-effect estimates.

Beyond this boundary of the span of direct influence, monitoring information on ultimate outcomes (poverty, income, yields) may still be informative, but not for establishing net effects attributable to the intervention. Therefore, rough indicative values are sufficient; for example, to compare the participant group with others, exact measurements to calculate absolute values in net-effect calculations are not needed. Instead of collecting detailed data on these outcomes in household surveys, there is a need for lean proxy-indicators that help to map the relative poverty position of a household, such as the Progress-out-of-Poverty indicator (PPI). While the PPI is not appropriate nor intended for net-effect calculations (Chen and Schreiner 2009), this simple questionnaire takes very little time for both the respondent and enumerator. Being a common indicator, it may provide useful information to compare the targeting of the support in relation to alternative interventions that have similar goals.

5 Conclusions

The article builds the case to reconcile precise measurements of immediate and intermediate outcomes in business practices that are considered to be within the span of direct influence of an intervention, combined with other methods to verify the causal assumptions of contribution of these practices to development impacts, following the logic of 'contribution analysis'. We propose to refine the intervention logic in distinct impact pathways, to identify key assumptions and key outcome areas, and draw the boundary of the span of direct influence of the intervention. We propose contribution analysis as

an overarching approach that combines precise survey-based net-effect measurement of immediate and intermediate outcomes, with less precise, lean monitoring of indicators and the use of a creative mix of methods to verify the contributory role of these outcomes in household income and poverty alleviation. Using similar indicators in key outcome areas would create enhanced opportunities for systematic cross-case analysis and for benchmarking and learning.

References

Alvarez, G. and von Hagen, O. (2012) *When Do Private Standards Work?*, Literature Review Series on the Impacts of Private Standards, Part IV, Geneva: International Trade Centre

Alvarez, G. and von Hagen, O. (2011) *The Impacts of Private Standards on Producers in Developing Countries*, Literature Review Series on the Impacts of Private Standards, Part II, Geneva: International Trade Centre

Bamberger, M.; Rugh, J.; Church, M. and Fort, L. (2004) 'Shoestring Evaluation: Designing Impact Evaluations Under Budget, Time and Data Constraints', *American Journal of Evaluation* 25: 5

Blackman, A. and Rivera, J. (2010) *The Evidence Base for Environmental and Socio-Economic Impacts of 'Sustainable' Certification*, Washington DC: Resources for the Future

Blackmore, E.; Keeley, J.; Pyburn, R.; Mangnus, E.; Chen, L. and Yuhui, Q. (2012) 'Assessing the Benefits of Sustainability Certification for Small-Scale Farmers in Asia', in J. Mayers (ed.), *IIED Natural Resource Issues*, London: International Institute for Environment and Development

Chen, S. and Schreiner, M. (2009) *Progress out of Poverty Index: A Simple Poverty Scorecard for Indonesia*, Washington DC: Progress out of Poverty – Grameen Foundation

Chocolate Working Group (2010) *Letter of Intent: Sustainable Cocoa Consumption and Cocoa Production*, The Hague: Sustainable Trade Initiative – Ministry of Agriculture Nature and Food Quality

COSA (2013) *The COSA Measuring Sustainability Report: Coffee and Cocoa in 12 Countries*, Philadelphia: Committee on Sustainability Assessment

Crosse, W.; Newsom, D. and Kennedy, E. (2012) 'Appendix 1: Recommendations for Improving Research on Certification Impacts', *Steering Committee of the State-of-Knowledge Assessment of Standards and Certification (2012). Toward Sustainability: The Roles and Limitations of Certification*, Washington DC: RESOLVE Inc.

DCED (2010) *The DCED Standard for Measuring Achievements in Private Sector Development: Control Points and Compliance Criteria, Version V*, Cambridge: Donor Committee for Enterprise Development

DGIS (2011) *Protocol Resultaatsbereiking en Evalueerbaarheid in PSD 2011 e.v.*, The Hague: DGIS-DDE-IOB

Donovan, J. and Stoian, D. (2012) *Capitals: A Tool for Assessing the Poverty Impacts of Value Chain Development*, Technical Series, Technical Bulletin 55, Turrialba: CATIE

El Hage, N. (2012) *Guidelines for Sustainability Assessment in Food and Agriculture*, Rome: FAO

Ingram, V; Waarts, Y.; Ge, L.; van der Vugt, S. and Puister-Jansen, L. (2014) *Impact of UTZ Certification of Cocoa in Ivory Coast: Assessment Framework and Baseline*, The Hague: LEI Wageningen UR

ISEAL Alliance (2014) *Code of Good Practice for Assessing the Impacts of Social and Environmental Standards Systems: Revision 2.0*, London: ISEAL Alliance

ISEAL Alliance (2013) *Demonstrating and Improving Poverty Impacts: ISEAL Common Core Indicators*, London: ISEAL Alliance

Khandker, S.; Koolwal, B. and Samad, H. (2009) *Handbook on Impact Evaluation*, Washington DC: World Bank

Mackie, J.L. (1965) 'Causes and Conditions', *American Philosophical Quarterly* 2.4: 245–64

Mayne, J. (2012) 'Contribution Analysis: Coming of Age?', *Evaluation* 18: 270–80

Mayne, J. (2001) 'Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly', *Canadian Journal of Program Evaluation* 16: 1–24

Nelson, V. and Martin, A. (2012) 'The Impact of Fairtrade: Evidence, Shaping Factors, and Future Pathways', *Food Chain* 2: 42–63

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage Publications

Rigby, D.; Woodhouse, P.; Young, T. and Burton, M. (2001) 'Constructing a Farm Level Indicator of Sustainable Agricultural Practice', *Ecological Economics* 39: 463–78

Rogers, P.J. (2009) Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation, *Journal of Development Effectiveness* 1: 217–26

Rogers, P.J. (2008) 'Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions', *Evaluation* 14: 29–48

RSCE (2009) *Round Table on a Sustainable World Cocoa Economy – RSCE2/7 Draft Principles for a Sustainable Cocoa Economy*, Roundtable for a Sustainable Cocoa Economy, http://tinyurl.com/lfyjls9 (accessed 24 April 2014)

Ruben, R.; Fort, R. and Zuniga-Arias, G. (2009) 'Measuring the Impact of Fair Trade on Development', *Development in Practice* 19: 777–88

Russillo, A. and Pintér, L. (2009) *Linking Farm-Level Measurement Systems to Environmental Sustainability Outcomes: Challenges and Ways Forward*, Winnipeg: International Institute for Sustainable Development (IISD)

SCSSC (2012) *Toward Sustainability: The Roles and Limitations of Certification*, Washington DC: Steering Committee of the State-of-Knowledge Assessment of Standards and Certification (SCSSC) – RESOLVE Inc

Shadish, W.R.; Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston MA: Houghton Mifflin Co.

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, London: Department for International Development

Ton, G. (2012) 'The Mixing of Methods: A Three-Step Process for Improving Rigour in Impact Evaluations', *Evaluation* 18: 5–25

Ton, G.; Vellema, S.R. and de Ruyter de Wildt, M. (2011) 'Development Impacts of Value Chain Interventions: How to Collect Credible Evidence and Draw Valid Conclusions in Impact Evaluations?' *Journal on Chain and Network Studies* 11: 69–84

Ton, G.; Flores, L.; Monasterios, R. and Yana, E. (2014) 'Capabilities and Performance in Collective Marketing: The Importance of Learning to Cope with Agency Dilemmas', in R. Christy, C. da Silva, N. Mhlanga, E. Mabaya and K. Tihanyi (eds), *Innovative Institutions, Public Policies and Private Strategies for Agro-Enterprise Development*, Hackensack NJ: World Scientific Publishing Co. Inc.

Ton, G.; Taylor, S.; Vlaming, J. and Hiller, S. (2010) 'Evaluating Poverty Impacts of Bottom-of-the-Pyramid Irrigation Technology Supply: IDE's Rolling Baseline Approach to Household Income Impact Assessment', in P. Kandachar, I. de Jongh and M. Halme (eds), *International Conference on Impact of Base of the Pyramid Ventures*, Delft: Faculteit van het Industrieel Ontwerpen, TU Delft

UTZ Certified (2014) *UTZ Certified Impact Report January 2014: Combining Results from 24 External Impact Studies and Data from UTZ Certified*, Amsterdam: UTZ Certified

Vellema, S. (2010) 'Measuring Impacts: About Lean Approaches, Mechanisms, and Critical Control Points', unpublished lecture presented at ISEAL Alliance Conference, London, 22 June 2010

Vellema, S. and Ton, G. (2012) 'Is Measuring Ultimate Impacts the Way Forward?', paper presented at ISEAL Conference 'Who's Who and What's On in Research about the Poverty Reduction Impacts of Sustainability Standards', London, 27–28 November 2012

Vellema, S. and van Wijk, J. (2014) 'Partnerships Intervening in Global Food Chains: The Emergence of Co-Creation in Standard-Setting and Certification', *Journal of Cleaner Production*, in press – corrected proof, www.sciencedirect.com/science/article/pii/S0959652614003230 (accessed 12 April 2014)

Vellema, S.; Ton, G.; de Roo, N. and van Wijk, J. (2013) 'Value Chains, Partnerships and Development: Using Case Studies to Refine Programme Theories', *Evaluation* 19: 304–20

Waarts, Y.; Ge, L. and Ton, G. (2013a) *From Training to Practice: Mid-Term Evaluation of the UTZ-Solidaridad Smallholder Tea Programme in Kenya*, The Hague: LEI Wageningen UR

Waarts, Y.; Ge, L. and Ton, G. (2013b) *From Training to Practice: Mid-Term Evaluation of the UTZ-Solidaridad Smallholder Tea Programme in Malawi*, The Hague: LEI Wageningen UR

Waarts, Y.; Ge, L.; Ton, G. and Van der Mheen, J. (2013c) *A Touch of Cocoa: Baseline Study of Six UTZ-Solidaridad Cocoa Projects in Ghana*, The Hague: LEI Wageningen UR

Waarts, Y.; Ge, L.; Ton, G. and Jansen, D. (2012) *Sustainable Tea Production in Kenya: Impact Assessment of Rainforest Alliance and Farmer Field School Training*, The Hague: LEI Wageningen UR

White, H. (2009) 'Theory-Based Impact Evaluation: Principles and Practice', *Journal of Development Effectiveness* 1: 271–84