Credible evidence on complex change processes: key challenges in impact evaluation on agricultural value chains.

Ton, Giel and Vellema, Sietze and DeRuyterDeWildt, Marieke

LEI Wageningen UR, LEI Wageningen UR, LEI Wageningen UR

1 June 2011

CREDIBLE EVIDENCE: ANTICIPATING VALIDITY THREATS IN IMPACT EVALUATIONS OF AGRICULTURAL VALUE CHAIN SUPPORT

Giel Ton & Sietze Vellema & Marieke de Ruijter de Wildt

-LEI Wageningen UR-

March 2010

Working Paper, presented at the WICANEM 2010.
Before quotation, please contact the authors: giel.ton@wur.nl

ABSTRACT

Although a growing field of policy intervention, the effectiveness of public-private value chain support is regularly questioned in the policy realm. Partly resulting from stronger pressures on aid money to show its worth, convincing evidence is asked for the effect on poverty alleviation. However, impact evaluations of interventions are challenging: outcome indicators are often multi-dimensional, impact is generated in dynamic and open systems and the external validity of conclusions are often limited, due to contextual particularities. Therefore, there is a strong case for theory-based evaluation where logic models indicate how the intervention is expected to influence the incentives for people's behaviour. The key assumptions inherent in these casual models can be tested through observation and measurement of specific outcome indicators, using mixed methods in triangulation. The mix of methods will have to anticipate the major threats to validity to the type of evaluative conclusion that the evaluation is expected to generate .Following the work of Shadish, Cook and Campbell (2002), validity threats relate to: 1) statistical conclusion validity; 2) internal validity; 3) construct validity; and, 4) external validity. The authors propose the combined use of data-set observations and causal-process observations in a comparative case-study design, based on critical realist concept of context-mechanism-outcome configurations. The use of a realist method to describe and analyze intervention pilots, facilitates the exchange of experiences between development agencies with evidence-based research. Its defined generalisation domain may prevent uncritical embracement of good practices. Certain value chain upgrading strategies may be viable and effective in a range of situations but are not the panacea, the standard solution, for creating market access; they all involve specific institutional arrangements that 'fire' specific mechanisms and incentives that depend on the institutional environment and social capital of stakeholders involved.

CREDIBLE EVIDENCE: ANTICIPATING VALIDITY THREATS IN IMPACT EVALUATIONS OF AGRICULTURAL VALUE CHAIN SUPPORT

1. Introduction

Value chain development has emerged as an area of donor interventions for poverty reduction in developing countries. The World Development Report 2008 (World Bank, 2007) has put it as a centrepiece to agricultural policy in developing countries. Value chain support focuses on capacities and capabilities of value chain actors and the enabling policies and institutions that facilitate change processes that benefit the poor. The performance of a specific value chain can be enhanced by, for example, increasing the scale of operations, improving service provision to producers, developing capacities to comply with (buyer-driven) quality requirements or addressing the process of value creation and value distribution. Value chain development is a container concept that has strong parallels with policy approaches such as ‘private sector development’ (Donor Committee for Enterprise Development), ‘making markets work for the poor’ (DFID), ‘growing inclusive markets’ (UNDP), and ‘opportunities for the majority’ (IADB).

Although a growing field of policy intervention, the effectiveness of public-private chain support is regularly questioned in the policy realm. A commonly heard criticism, for example, is that value chain support picks ‘winners’, focusing on a relatively small group of entrepreneurial poor and hence has a limited impact on average poverty levels (Humphrey and Navas-Aleman 2009). Partly a result of stronger public pressures on aid money to show its worth, convincing evidence is asked for the effect on poverty alleviation (OECD 2008; SDC 2009).

These calls for credible evidence have led to more stringent accountability requirements for agencies to defend the logic and demonstrate the impact of these interventions (Tanburn, 2008). However, generating convincing evidence on the link between development ‘output’ and donor supported intervention ‘inputs’ is not easy. Many ‘traditional’ research designs for evaluating impact prove impractical or inappropriate for analyzing value chain interventions. Value chains are complex, multi-layered and open socio-technical systems that are influenced by a myriad of intervening actors, and are continuously shaped and reshaped to adapt to changing conditions. Attribution of impacts of interventions in this dynamic ‘cloud’ of complex and intertwined sets of institutional arrangements is difficult, but necessary to answer legitimate questions on relevance, effectiveness and replicability of value chain development support (Roche and Roche, 1999, DAC, 2008). Decision makers on value chain support need comparable information on policies that work to choose effective instruments from the available policy menu.

One of the promising initiatives to generate credible and comparable information on value chain interventions originates from the Donor Committee for Enterprise Development (DCED) (2008). The initiative proposes minimum standards for reporting on private sector development, in which monitoring income changes and calculating attribution to program interventions is a required practice. When implemented, this would be a great leap forward towards developing a body of evidence on value chain support. The conclusions and policy recommendations derived from evaluative research need to be supported by data and information that has been collected and analyzed in a credible way. To meet the standard, lean research designs are needed that can face the most common threats to validity (Bamberger et al., 2006, Creevey and Woller, 2006, Shadish et al., 2002).

In this paper we add to the discussions on the design of impact evaluations tools and methods. We present a framework to evaluate the methodological design to assess change and impacts in value chain configurations based on validity threats to the evaluative conclusion. We propose a multi method research strategy to collect and analyze information that can stand up to scrutiny (Brady et al., 2006). The paper consists of three sections. First, we briefly discuss the basic evaluation question in impact assessments and illustrate the different threats to validity when concluding on these questions. Second, we elaborate three major methodological areas that are specific for value chain development and pose methodological challenges: measuring outcome patterns; attribution of impact in open systems; and the limits of the generalization domain as a result of social embeddedness. In the final section, we reflect on the applicability of our approach to assess and design impact evaluation methods that generate valid information on replicable principles and practices of value chain support. We stress the importance of linking ex-post impact evaluation processes with ex-ante constructions of plausible impact theories and credible outcome measurement methods.

2. Impact evaluation research: the quest for valid conclusions

There are many different reasons for doing an evaluation. Three types of evaluation are distinguished: evaluations that primarily look for accountability, for knowledge, or for development (Chelimsky and Shadish, 1997). Accountability evaluations look at the value of public expenditures, focusing on issues of costs and efficiency; knowledge evaluations aim for insights into public problems, policies, programs and processes, critiquing old methods in order to develop new ones; and, development evaluations seek to strengthen institutions and agencies in a particular evaluative area. The first two types are largely summative in nature, while the third type is largely formative. Although there is an overlap in tools and processes, these three types of evaluation are underpinned by different purposes and the evaluation methods will differ accordingly. In this paper, we focus on the first two types of evaluations that are especially interested in the evaluation of the processes that generate outcomes: impact evaluation.

Impact evaluation has three basic questions for which information and evidence is collected:

- Does it work? What positive and negative changes did the intervention generate in the performance of the value chain?

- How does it work? What components of the support did work, for whom, and under what conditions?

- Will it work elsewhere? What components might work for whom under what conditions?

Though each evaluation assignment will have a different emphasis, these three questions are in varying wordings asked by the stakeholders commissioning the evaluation and are intimately related. The first question is a quest for evidence and especially relevant when public or private investments have alternatives and need an indication of the extent to which their support contributed to stated objectives. Pawson and Tilley (1997) argue that this first question is far too dominant in evaluation research whilst the second question is more productive in providing guidance to stakeholders and in generating useful policy recommendations. Likewise, Ravallion, chief evaluator at the World Bank, points to the dominance of methods that limit themselves to show that policies work or not, without generating additional information on how they work and could work in other settings. He opposes, specifically, the dominance of econometric impact assessment methods that only compare average values of indicators between treated and control groups. According to Ravallion, the audience of most impact assessments, policy makers, do indeed rarely bother about the outcomes of statistically rigorous randomized impact assessments: "They also want to answer questions like: Does the intervention work the way it was intended? What types of people gain, and what types lose? What proportion of the participants benefit? What happens when the program is scaled up? How might it be designed differently to enhance impact" (Ravallion, 2009). The third evaluation question is often the main motivation for an evaluation. Often, an impact evaluation is commissioned to asses the possibilities to replicate it in other contexts, or upscale the intervention from 'pilot' to 'mainstream'. This third question is most directly related to the policy recommendations of an evaluation and is consequently the most read part and most vulnerable to critique.

These three questions in impact evaluation require different sets of methods to generate and analyse information. They need different kinds of information, or, at least, with different 'depth and detail'. Whereas the first question may treat the intervention as a one-package 'black box', the second question explicitly opens the black box to know how incentives are created and perceived during the intervention. Answers to the second question need to be based on more detailed information about contextual factors that influence the outcomes of the intervention in specific (groups of) persons and details on the reasons of persons to react (or not) to the incentives offered through the intervention. The third question, interpreting the data and conclusions of the first two questions, intends to formulate generalized inferences and explores the possibilities to extrapolate findings to other contexts.

Shadish et al. (2002) indicate that no generalised causal inference has absolute validity, there will always be some specific conditions that limit the generalisation domain of the conclusion. They stress the need to design procedures that (partially) control some of the limitations of the research methods used that may weaken the validity claims of causal inferences. They distinguish four dimensions of validity that have to be convincingly addressed in the design of evaluation research:

- Statistical conclusion validity: the way inferences about correlations are made in data-set observations. It emphasises the need to comply with proven methods to estimate association or correlation between variables.
- Internal validity: the way causality is attributed in the evaluation. This refers to the logic behind the observed correlations and explains why and how interventions contribute to the observed change.
- Construct validity: the way that generalisations are made from the categories used in the evaluation to broader units of representation. It stresses the importance of precise definitions and concepts.
- External validity: the way that the findings are generalizable to other persons, times and contexts. This requires to be precise about conditions and requirements that define the generalization domain.

S*tatistical conclusion validity* is key if the research method involves statistical analysis of data-sets. In data analysis, we often compare between groups of respondents and calculate averages or other measures of comparison in the sample population. We then use several tests to conclude on the probability or 'significance' of a correlation between their characteristics and the outcomes. Conclusions have to be supported by tests on assumptions of correlation and for example, indicate probability intervals for means and effect sizes of the factors in a regression. Just producing an output table that indicates 'significant' relations is insufficient. All statistical tests have assumptions and pre-conditions related with the data, like the 'normal distribution of the data' or the 'homogeneity of variance of the different groups'. Taking statistical conclusion validity seriously, we need to be explicit about such assumptions, and design methods to check these assumptions.

Internal validity is intimately related to the argumentations to support a causal inference. It is important to be clear how the evaluative research makes the link between an intervention (cause) and specific outcomes in the value chain. There are three basic conditions that define causality: the cause needs to be active before the effect is produced; the cause must be related to the effect produced; and alternative explanations of the effect must be discarded. In value chain development, it is unlikely that there is just one cause of the change. The effect of interventions is usually a result of a constellation of positive and negative factors active in a particular context, in which each individual factor in that constellation is a so-called *inus condition:* in itself *insufficient* to explain the outcomes of a support intervention, but a *non-redundant* part of a wider constellation of factors that is *unnecessary* but *sufficient* to produce the outcome (Mackie, 1965). Hence, we will have to make plausible that the value chain support was indeed necessary for producing the outcomes that we observe. Non-observables, characteristics or factors that are 'hidden' and not registered in the data-set, may be part of the constellation of factors and with the potential to provide alternative explanations of the observed effects. To support an evaluative conclusion on the effectiveness of a value chain support intervention, the importance (non-redundancy) of the latter in this' cloud' of causal factors has to be made plausible. The research tools need to be designed in a way that they generate sufficient information to do so.

The quest for replicable models underscores the importance of *construct validity*. The evaluators need to be explicit about the way they generalise the concepts and constructs that they find in the evaluation. If they conclude something about the effectiveness of a certain intervention in the chain, e.g. “investments in cooling tanks is effective in linking dairy producers to markets”, we immediately face several threats to construct validity: is ‘dairy producers’ a good construct, or do we need to make distinctions in small and bigger dairy farmers, diversified farms or specialized farms? Does the inference hold for all types of investment support that facilitate cooling tanks in this specific case, or do we need to make distinctions in grants and credit schemes, or farmer-driven and government-driven schemes? Is it valid for all markets, or only for the urban fresh milk markets and not for cheese and yoghurt markets? To face threats to construct validity, we need to be precise about the concepts and constructs used and design our research methods accordingly.

Even more challenging are the threats to *external validity*. Even when we come to the conclusion that in a specific context the intervention was a key factor with positive results, this will not necessarily hold in other settings. Hence, we need to argument why, and to what extent, the findings can be generalized and remain valid for other contexts and conditions. Like all four types of validity, but stronger than in the earlier three, absolute validity does not exist. All ‘best practices’ and lessons learnt can be questioned by indicating a ‘peculiarity’ related to the context; conclusions will be limited to conditions such as consumer price margins, civil peace or existing trust levels that limit opportunistic behaviour. Hence, we need to design tools that collect information to respond to the most obvious or relevant questioning of the validity of our policy recommendations; we have to generate sufficient information to describe and defend our 'generalisation domain' (Chen, 1994).

Few evaluations in international development systematically address issues of validity; the field of value chain support is no exception to this (Zandniapur et al., 2004, Humphrey and Navas-Aleman, 2009). "While many evaluations refer to threats to validity in their initial design, it is much less common to find any systematic assessment of validity in the presentation of findings and conclusions. Often the only reference to validity is a brief note stating that given the budget, time, data (or sometimes political) constraints under which the evaluation was conducted, the findings should be treated with some caution" (Bamberger, 2007).

3. Methodological challenges

We will now discuss the validity challenges in three core methodological areas that relate with the evaluation questions: Does it work? How does it work? Will it work elsewhere? Our first concern is the problem of measuring outcome patterns. Performance indicators vary between relative simple indicators to complex constructs that are difficult to operationalize. Second, we focus on the issue of attribution. In complex and multi-layered social systems like value chains, not one intervention functions in isolation: many stakeholders, prices and market trends influence value chains that are socially embedded in diverse cultural settings. More so, interventions have various components, implemented with different time frames, in varying combinations that interact with each other. We end with the challenges to generate learning and generalizable conclusions from impact assessment. The following table summarises how the evaluative questions relate to these methodological challenges.

Table 1: Evaluation questions and methodological challenges

	Measuring outcome patterns	Attribution in open systems	Generalization and social embeddedness
Does it work?	++	++	+
How does it work?	+	++	+
Will it work elsewhere?	+	++	++

The information needed to support conclusions on each evaluation question will overlap. To know if some components work in specific conditions, information on outcomes and impact will be very useful; to test if something worked, a statistical model must specify a logical causal model of the characteristics needed to make the intervention work. In social research, a useful distinction is made between Data-Set Observations (DSOs), typically a result of surveys, file records and time-series, and Causal-Process Observations (CPOs), typically based on discrete qualitative case-studies (Brady et al (2006). To make high validity causal inferences, a combination of these two types of information is needed in a process of 'nested inference' or 'triangulation' (Brady and Collier, 2004). The relative emphasis on each type of observation will change according to the exact evaluation questions. The "Does it work?" question tends to demand more efforts in generating DSOs (data sets), while the "How does it work?" question demands more CSOs (case study material).

Measuring outcome patterns

The first evaluation question, does it work, seeks to measure the change caused by the intervention. The DCED (2009) proposes some basic steps for this: define the impact model; define indicators of change (and projections); measure these indicators; and capture the wider change in the value chain. In value chains, support is often directed at actors and institutions in the environment of (poor) producers, like financial and non-financial business support services, rather than at producers themselves. The interventions will have an explicit or implicit 'theory of change' or impact model that translates the support to these chain actors into behavioural outcomes of chain actors, including producers. The impact model to be evaluated is not necessarily comprehensive and may concentrate on subsets of conditions, components of interventions, specific instruments, and types of outcome patterns. Impact assessments need to monitor (sets of) outcome indicators as 'proxy' for performance in each of these target area and test key assumptions behind this model. To facilitate statistical inferences, preferably, measurable, continuous and quantitative indicators (dependant variable) are selected as proxies for the outcomes. However, impact evaluations also need to capture wider, unexpected outcomes. Wider changes are particularly informative as they verify and build our understanding of impact.

In designing the concepts and indicators in impact assessment, construct validity is a key challenge. Performance of a value chain relates to different layers and dimensions of social interaction in the chain network. Similar to the challenges to assess other abstract attributes of social systems, like 'organisational strength', the immaterial aspect of chain performance makes it difficult to capture and measure. More so, concepts and indicators are often influenced by the disciplinary background, ontological theories and personal interests of the evaluator. Value chain performance will be assessed differently according to the angle chosen and aspects focused on. For example, when looking for outcomes of support to multi-stakeholder chain platforms, an economist trained in transaction economics will look for ‘trust’ and ‘coordination’ between chain actors, while someone specialised in the analysis of group dynamics will focus on ‘inclusion/exclusion’ and ‘synergy’. A political economist will see ‘changing power relations’ and a scholar in strategic marketing will look at ‘innovativeness’ and ‘competitiveness’. All will see some of the outcomes of the intervention, but not the whole picture. It is, therefore, important to carefully select an evaluation team that is able to identify and operationalize the relevant performance indicators (Snodgrass, 2006).

Even apparently straightforward indicators need to be well defined, according to a causal model that is comprehensive enough to include the most important outcomes, but lean enough to facilitate attribution. One of the three ‘universal’ indicators proposed by DCED (2008) is “additional net income (additional sales minus additional costs) accrued to targeted enterprises as a result of the programme per year”. Here, for example, the scope for varying interpretations is considerable. E.g. net additional income as a result of a dairy development intervention, can be restricted to net income growth from fresh milk sales. However, it can also be understood as the net income change of the whole agricultural system of the household, as increasing dairy production and increased animal feed production may impact horticultural production and family income. Positive spill-over effects may exist too, since farmers may have learned about milk quality issues, and, as a result of increased communication with other chain actors, may have improved their entrepreneurial skills and technology beyond diary only. This more comprehensive way of calculating net income introduces a wider range of confounding factors, that complicate the attribution of the impact to the specific intervention: e.g.

prices fluctuate between seasons and are prone to natural conditions, and will influence incomes without any causal relation with the dairy support intervention being evaluated.

Commonly, changes in value chain performance are assessed by subtracting or comparing indicator scores: at least a 'before-after' situation and, if possible, a 'with-without' estimate. Measuring d*ifferences* in indicator scores with some accuracy is more important than measuring the absolute value of the indicator. Relatively small measurement errors in both indicators may translate in large errors in the calculated difference between them. Tracing these measurement errors in indicator averages between non equivalent groups is difficult and often limited to outlier checks only.

To indicate the impact that is attributable to interventions, a comparison is needed with a control group: a group with similar characteristics that did not experience the working of interventions. The comparison between these two groups helps to assess if the outcomes can be attributed to any 'exogenous' or 'unknown' causal factor, not related to the intervention's causal mechanisms. Experimental methods, with random assignment to treatment and control groups (Duflo et al., 2006) are especially designed to facilitate this measurement of outcomes between treated and non-treated groups. However, this design is often impossible and may even be unwanted (Shadish et al., 2002, Bamberger et al., 2006); deliberate exclusion of some groups of stakeholders in the value chain from the benefits of a support intervention (like coordination platforms, value chain financing, certification programs, investment subsidies) is often socially and politically unfeasible. Also, in many cases, there are important spill-over effects from pilot-intervention areas to other areas and chain actors that make the definition of who is a participant and who is not is a gliding scale, and the distinction in 'treated' and 'control' groups unworkable (Ravallion, 2009). Random assignment of the intervention to a defined population is rarely possible and, therefore, other, quasi-experimental methods are, hence, more frequently used. However, research designs that deviate from random assignment face the risk of being affected by a selection bias, introducing differences between the treatment and the control group that are unrelated to the intervention, but important in producing the outcomes (e.g. attitude, resource base, etc.). This is a major threat to the validity of the statistical conclusion. A proper evaluation design will have to consider, limit and control for such a bias in data-set observations.

Generally, a survey ends up in a set of qualitatively distinct variables used as proxies for 'improved livelihood strategies of smallholder households'. Statistical analysis, with a set of distinct dependant outcome variables, generates additional threats to validity of the correlations found. Current software makes consecutive iterations of statistical analysis with changing combinations of variables so easy, that 'significant' correlation between variables may result from 'fishing the data' or 'data mining': repeating statistical tests that analyse the significance of differences between groups by selectively re-grouping respondents, variables etc. Even if the intervention has no effect at all, in complex data sets, one or more significant correlations are likely to appear after a sufficient number of iterations (Shadish et al., 2002). Concluding on causal relations from such correlations may wrongly attribute outcomes to interventions. On the one hand, as conclusions tend to concentrate primarily on significant effects, this results in a bias in impact evaluations towards 'significant' though irrelevant conclusions. On the other hand, non-significant effects can be a result of low statistical power (low sample size) or measurement errors that could have been corrected when more deeply analyzed. The recommended solution against 'fishing' is to specify, ex-ante, the hypothesis or theoretical model that is tested and to increase the threshold (significance level) of the correlation detected through iterative analysis[1].

Only data-set observations from surveys with a sufficient sample size (*statistical power*) will make it possible to detect differences between subgroups in the survey population. Commonly, a minimum subgroup size of 30 is used as a rule-of-thumb (Creevey and Ndiaye, 2008). The sample size will have to consider attrition, some respondents will fall out of the sample due to moving, passing away or changing activities. When one wants to compare between different subgroup locations (*g*) disaggregated on different typology criteria (*c*), this minimum total sample size will, thus, be *N= 30*g*c*. For explorative statistical analysis, and considering attrition, sample sizes are ideally larger than the minimal required size. In the 'real world', however, sample sizes are often restricted by resource constraints (financial, not enough people, too difficult to get to, etc)..

The DCED recommends to capture wider changes than just 'predicted' change by the logical model or intervention theory. The most obvious threat to validity of an evaluative conclusion is that it leaves important factors out of the equation, be it as confounding causal factors or as outcome indicators, and, so, weakening the internal validity of the findings. Unintended changes are unlikely to be captured by pre-established indicators in causal impact models. Additionally, more open and qualitative Causal-Process Observations are needed to assess these unintended outcomes. The emphasis on

[1] However, fishing is difficult to detect as often no ex-ante causal hypothesis exists or, more common, the hypothesis is adjusted during analysis and reporting the data. Interestingly, this temptation is even stronger for academics involved in evaluative research, as the chance of research results to be published in scientific journals is far higher with an argument that is supported with 'significant' statistical evidence. This 'publication bias' creates incentives for ex-post modeling of hypothesis and generates a problem for meta-research as there is an overestimation of 'attribution' of change as result of interventions in reviewed literature.

documenting wider impacts is important; too often, evaluations restrict assessment designs to find proof for impact logic only (European Commission, 2008)

Attribution in open systems

Significant correlations do not indicate causality, but at least indicate that there is, most probably, a relation between the intervention and the outcomes. Data-set observations need causal theories to differentiate between collinearity (it happens together) and causality. Analyses of the logic behind the observed changes are necessary to interpret these correlations and to identify causal relations.

The plea for statistical analysis to test the inference about the mechanism's causal power, the scientific testing if they work or don't work, holds especially for simple or at most complicated systems (Snowden and Boone, 2007, Rogers, 2009) where outcomes can be measured with quantitative indicators. However, this is far less realistic for interventions with a wide constellation of causes. It is even impossible to apply in open systems that behave with increasing levels of complexity or chaos (Lawson, 2003, Pawson, 2002, Hospes, 2008, Snowden and Boone, 2007). If value chain support takes place with a high degree of contingency in system behaviour as a result of unobservable, exogenous factors that cannot be incorporated into a statistical model, experimental and quasi-experimental methods that rely on statistics alone, will have problems in demonstrating the internal validity of causal connections (Heckman, 2005).

The difficulty to grasp complexity of change process in econometric models holds especially for evaluation research designs based on comparing groups through 'matching', like Propensity Score Matching (PSM). In PSM, impact is assessed by measuring the outcome difference in pairs of respondents that ‘match’ the same characteristics, except their adoption of the innovations promoted by the intervention. The characteristics on which matching takes place are, ideally, derived from a model that comprises the whole ‘constellation of factors’ that are expected to lead to the measured outcomes (e.g. adoption of technology that leads to higher income levels). The matching is done through calculation of a ‘propensity score’ for all respondents on a construct of different variables that ‘models’ the context of the respondent. The respondents with a comparable score on the model’s dimensions will form ‘matched pairs’ and are supposed to share the likelihood to have the same outcomes, except the ones that result from the adoption of the innovation promoted by the support intervention. The difference in outcomes between the ‘matching pairs’ of adopters and the non-adopters are considered to be attributable to the intervention. As will be clear from the above, these matching models are heavily theory-laden, and they suppose that the matching is done on all relevant variables that will make the pairs similar in reaction to the interventions incentives. This model to ‘capture context’ is ideally elaborated before the PSM survey data is gathered (because on all characteristic there need to be information from the survey), but, in practice, it is often only constructed after the survey, during data-analysis and limited by the variables in the available data-set. But even when the data collection is so comprehensive, in complex systems, the model used to match respondents will always be incomplete and will suffer from ‘essential heterogeneity’ (Heckman, 2005): it may miss a latent, unobserved external that is key in the constellation of causal factors that determine the reactions of stakeholders to the interventions. Even the more sophisticated econometric methods that explicitly try to correct for the variance due to unobservable factors that influence a respondent’s behaviour will end up testing closed models of reality. Therefore, critics may always challenge the validity claims of causal inferences derived from econometric analysis of survey data, indicating that the model is too simplistic and that the context is far more complex to be captured in mathematical models (Lawson, 2003). A (partial) defence against this threat is to limit PSM to social processes that are relatively simple. Social systems with increased levels of complexity or chaos limit the possibility to generate credible and valid conclusions from quasi-experimental data (Snowden and Boone 2007). Further it is important to indicate clearly and consistently why (the most salient) external factors are considered irrelevant for explaining the observed outcomes, and that the conclusions of the PSM are, therefore, essentially fallible.

All value chain interventions, ultimately, are intended to change attitudes and behaviour in persons. The workings of the intervention are often implicitly assumed in the impact logic. Realist evaluation (Pawson and Tilley, 1997) provides a useful framework for analysing specific mechanisms in an intervention that may be ‘fired’ in a specific context and that trigger behavioural change. It emphasises the need to build ex-ante hypotheses related to the (project) mechanisms that (are assumed to) motivate or influence stakeholders ‘to act differently’ and generate changes in outcomes. Realists propose to test key assumptions in these hypotheses with the concepts “Context-Mechanism-Outcome Configurations” (Table 2).

The realist concept of ‘mechanisms’ opens the black-box between intervention/treatment and outcome/impact. The concept ‘configuration’ indicates that mechanisms will only produce certain outcomes in certain contexts, making key discriminations that automatically limit the generalization domain of the causal inference. This realist emphasis on contextual embeddedness helps to specify (and limit) the policy recommendations on eventual future replicability. In contrast with the mainstream econometric approaches, realist evaluators concentrate on the ‘treatment’ and the incentives for the ‘treated’, without bothering too much about a control group. They emphasize that mechanisms work under specific conditions, part of a wider constellation of causal factors, and therefore, generalisations are difficult.

In evaluations of value chain support interventions, the realist framework can be used to describe the workings of interventions in its context. The detailed description and analysis of a pilot intervention may than feed the intervention theories behind new policy and programmes that want to improve or replicate good principles or practices in value chain development.

Table 2: ‘Realist’ Comparative Case Studies in value chain research

Realist Concept	Domain of application	
	Understanding pilot interventions (Causal theories)	Designing policy and program (Normative theories)
Context	Situation of the value chain stakeholders in the pilot experience	Situation of the value chain stakeholders in another setting where the support intervention will take place
Mechanism	Incentives that condition the behaviour of stakeholders in specific institutional arrangements that have emerged in and around the value chain	Intervention that changes the incentive structure for stakeholders and generates an improved institutional arrangements in and around the value chain
Outcome	Actual performance of these institutional arrangements in the value chain.	Intended outcomes of the intervention on institutional arrangements.
CMO-Configurations	Comparative case descriptions of causal connections between interventions and the performance of specific institutional arrangements.	Defined recommendation domain for replicable policies and interventions that enable effective and sustainable institutional arrangements in the value chain

However, critics of realist evaluation point to the weakness in qualitative case-study research and to the tendency to generate speculative hypothesis without strong empirical evidence. “Realists may be strong in identifying rival explanations for the observed impacts and outcomes, but quite poor in convincingly testing and eliminating the erroneous ones”. Farrington (2003) points to this weakness and argues that with limited time and resources for evaluations, it is difficult to deal with multiplicity of contexts, mechanisms and outcome patterns. Realist evaluations often end up with a multitude of possible causal inferences with limited validity. He argues that qualitative research methods alone lack the necessary procedures to answer the most obvious threats on internal validity. He strongly favours the use of statistical analysis of data set observations for supporting causal inferences.

We agree with Farrington that, when designing impact evaluation research, it would be good to have data-set observations to support the validity claim of inferences from qualitative realist case-studies. Hence, to attribute change to value chain support, we re-emphasize the need for triangulation of information collected as (quantitative) data set observations and as (qualitative) causal process observations. Only in combination, these research methods may provide data that will support the conclusion and provide it with sufficient internal validity (Brady et al., 2006). These different research methods need to be directed to the evaluation of the same mechanisms, processes and outcome patterns, because through different perspectives on reality and different conceptualisations of the way impact is generated, this ‘triangulation’ improves the validity of the evaluative conclusion.

Social embeddedness and generalisation

The third evaluation question, will it work elsewhere, is about scaling-up and extrapolating conclusions to other contexts. In relative simple social systems, the common statistical design to maximise external validity of a causal relation found in data sets is randomisation. By gathering data randomly in a certain population or context, the causal inference derived from the survey data is assumed to hold for the whole population or context from where the sample is randomly taken. In data-set observations collected through survey samples, there is, therefore, no better statistical design than random sampling to defend the claim that findings have external validity in generalisations across populations. However, threats to this claim of external validity arise when the conclusions of an evaluation are not bound to the population samples, but are applied to contexts and conditions that are totally different in space and time. And, this is more rule than exception when agricultural value chain support interventions are concerned; evidence of impact in one commodity chain will not necessarily be relevant for another commodity; evidence in one cultural setting or time period will not be generalizable to another.

Policy makers are especially interested in good practices, as it provides them with a menu of options. Evaluative conclusions to be used by policy makers, therefore, have to maximize the generalisation domain while maintaining validity and credibility. 'Good Practices', 'Best Fit Solutions' or 'Principles' are all concepts used to indicate mechanisms or interventions that proved to have worked in a certain setting, and that might work in others. Instead of strong causal inferences about 'Best Practices', only possible in relatively simple social change processes, these concepts apply to more complicated and complex ones.

To explore the generalisation domain of the conclusions on impact in a specific value chain, we need a process to deal with external validity of findings. Shadish et al (2002) present five principles that are especially useful to consider the external validity of policy recommendations about the replication of 'policies that work'. These principles reduce the threat to validity of a causal connection discovered with the evaluation method and may convince a critical or sceptic audience. They propose to: (I) assess the apparent similarities between study operations and the prototypical characteristics of the target of generalisation (Surface Similarity); (II) identify those things that are irrelevant because they do not change a generalisation (Ruling Out Irrelevancies); (III) clarify key discriminations that limit generalisation (Making Discriminations); (IV) explore the possibilities to apply the results within and beyond the (sampled) range of observations (Interpolation and Extrapolation); and (V) to develop and test theories about the pattern of effects, causes and meditational processes that are essential to the transfer of a causal relationship (Causal Explanation).

4. Conclusions

The increased attention of donors to standardised and rigorous impact assessments that can demonstrate impact of value chain support, builds momentum for the development of lean and effective tools and approaches. The existing lack of evidence does not necessarily reflect a low priority on measuring impact, but rather points to the lack of appropriate and credible instruments to do so, and to the complexity of social processes in value chain development processes.

We apply the concept of threats to validity, developed by Shadish et al (2002) to indicate the importance of 'designing' a combination of methods to support the evaluative conclusion. The research methods in itself do not imply higher or lower viability. The appropriate mix of methods depends on the phrasing of the evaluative question and the 'kind of conclusion' that the stakeholders inducing the evaluation have in mind. Starting from one method of collecting information, the review of the threats to validity in four different dimension (statistical conclusion validity, internal validity, construct validity and external validity) will indicate the need for additional research methods and will lead to an appropriate set of mixed methods that can support inferences with credible evidence.

From the above reflexion on impact evaluation methodology, we also conclude that for evaluating replicability of value chain support, the need for 'theory' is paramount. An impact evaluation needs to build on a logic model that indicates how the intervention is expected to influence the incentives for people's behaviour. In the statistical analysis of data-set observation, this theory feeds the variables and matching models used, while in realist evaluation of causal-process observations, the theories are related to the workings of incentives provided and mechanisms triggered by the intervention. In impact evaluation, we need to make causal inferences about "what has worked for whom under what conditions", and, concerning replicability, "what might work for whom under what conditions". For measuring and attributing impact, we need to have causal models that explain dynamics in empirical reality, while for replicability of the intervention, we will have to develop and refine theories on how an intervention will impact future dynamics n other contexts. Chen (1994) calls these two sets of program theories *causal theories* and *normative theories* of program impact. Causal theories are descriptive of change processes in social systems, while normative theories are more prescriptive and action-oriented and represent the impact model behind an intervention. Obviously, the latter benefits from the first and normative theories improve when more causal theory is generated.

A way to focus on the mechanisms that 'trigger' behaviour during or after an intervention is to use the realist concept of "Context-Mechanism-Outcome Configurations" in comparative case studies, analysing cases along the four aspects. To be useful prospectively, as a normative theory, these pilot case studies need to be written in a way that the contextual requirements for the intervention/mechanisms that triggers performance enhancing behavioural changes by chain actors are sufficiently explicit, and with a credible measurement of key outcome indicators. The case-studies will have to describe and unravel context-dependant processes and practices that generated the impact. In the real world complicated and complex value chain dynamics, at most, these case-studies can suggest 'good practices' or promising 'principles' that may work in a comparable configuration. They can be used as 'food for thought' in a learning process with stakeholders from other contexts. Information to conclude on comparability of the two configurations (the match between the case-study reality and the reality in the new intervention context) will always be incomplete, but the realist question 'What works for whom under what conditions, is helpful to generate that information and underpin the internal and external validity of the 'best practice'. Besides methods that make theories explicit, properly designed data collection tools and qualitative research techniques are needed that quantify or describe the outcomes and impacts of value chain interventions and can be used to test the key assumption inherent in the impact models. Multiple methods are needed to support claims that something does work, and provide information that is useful to explore the real causal processes and compare them with the normative impact models.

Statistically significant differences between groups, or correlations between variables, are not sufficient for concluding on attribution of impact. To 'upgrade' a significant correlation into a causal relation with strong validity claims, additional quantitative and qualitative research methods will have to be incorporated in the impact evaluation design. Statistical analyses of differences in average outcomes between groups, regions and intervention packages are helpful. However, they are not the only way to use survey information. Instead of focussing only on data averages and differences in means, the selection of both illustrative and ‘contrasting cases’ (Lawson, 2009) within the sample and the exploration of their ‘logic’ with qualitative case-study methods, may help to understand the unravel the logic behind observed changes in outcomes and help to clarify the conditional and contextual character of an intervention’s impact.

Impact evaluation demands serious efforts from organisations to invest in critical reasoning while designing interventions, presenting an initial ‘intervention theory’ rather than a logical frame or impact logic, that can be tested and improved through monitoring and evaluation activities. Using a realist method to describe and analyze intervention pilots as comparative case-studies facilitates the exchange of experiences between development agencies with evidence-based research. Its restricted and defined generalisation domain may prevent uncritical embracement of good practices. For example, specific types of contract farming, branding, fair trade labelling prove to be viable and effective in a wide range of situations but are not the panacea, the standard solution, for creating market access; they all involve specific institutional arrangements that invoke specific mechanisms and incentives that depend on the institutional environment and social capital of stakeholders involved. A more explicit delineation of the generalisation domain of evaluative conclusions on interventions that change these (interlinked) institutional arrangements may prevent failures, and help to build context specific and evidence-based theories of change.

For evaluations that intend to conclude on the replicability or scalability of agricultural value chain interventions, we propose an evaluative design that is based on a combination of 1) impact models that reflect the intervention theory, 2) a realist focus on the mechanisms that are assumed to be fired by the intervention and the conditions under which these work, and 3) triangulation of data-set observations and casual process observations to support the evaluative conclusions. The resulting mix of appropriate, lean and credible data-collection methods will provide useful information for accountability purposes, for monitoring on-going interventions and for learning on good principles and good-fit practices that have potential to be effective when replicated in future interventions. The logical link between these three design elements facilitates 'nested inference' with increased scientific strength and limits the threats to validity of the evaluative conclusion.

References

BAMBERGER, M. (2007) Simply the Best? Understanding the Market for 'Good Practice' Advice from Government Research and Evaluations: a framework for assessing validity and utilization of evaluations. *American Evaluation Association 2007.* Baltimore.

BAMBERGER, M., RUGH, J. & MABRY, L. (2006) *RealWorld evaluation: working under budget, time, data, and political constraints*, Sage Publications Inc.

BRADY, H. E. & COLLIER, D. (2004) *Rethinking social inquiry: diverse tools, shared standards*, Rowman & Littlefield Publishers.

BRADY, H. E., COLLIER, D. & SEAWRIGHT, J. (2006) Toward a Pluralistic Vision of Methodology. *Political Analysis,* 14**,** 353-368.

CHELIMSKY, E. & SHADISH, W. R. (1997) *Evaluation for the 21st century: A handbook*, Sage.

CHEN, H. T. (1994) *Theory-Driven Evaluations*, Sage Publications Inc.

CREEVEY, L. & NDIAYE, M. (2008) Common Problems in Impact Assessment Research. *Impact Assessment Primer Series #7.* Washington, USAID.

CREEVEY, L. & WOLLER, G. (2006) Methodological Issues in Conducting Impact Assessments of Private Sector Development Programs. *Impact Assessment Primer Series #2.* Washington, USAID.

DAC (2008) *Evaluating Development Co-operation: summary of key norms and standards,* Paris, OECD.

DCED (2008) Quantifying Achievements in Private Sector Development: Control Points and Compliance Criteria. Donor Committee for Enterprise Development.

DUFLO, E., GLENNERSTER, R., KREMER, M., CENTER, L. & FLOOR, T. (2006) Using Ramdomization in Development Economics Research: A Tool Kit. *MIT Department of Economics Working Paper No. 06-36*

EUROPEAN COMMISSION (2008) Second Strategic Review of Better Regulation in the European Union Brussels, European Commission.

FARRINGTON, D. P. (2003) Methodological Quality Standards for Evaluation Research. *The ANNALS of the American Academy of Political and Social Science,* 587**,** 49-68.

HECKMAN, J. J. (2005) The scientific model of causality. *Sociological Methodology,* 35**,** 1-98.

HOSPES, O. (2008) Evaluation Evolution?- three approaches to evaluation. *The Broker,* 2008**,** 24-26.

HUMPHREY, J. & NAVAS-ALEMAN, L. (2009) Multinational Value Chains, Small and Medium Enterprises, and 'Pro-Poor' Policies: A Review of Donor Practice. *IDS Research Report 63.* Brighton, IDS.

LAWSON, T. (2003) *Reorienting Economics,* London & New York, Routledge.

LAWSON, T. (2009) Applied economics, contrast explanation and asymmetric information. *Cambridge Journal of Economics,* 33**,** 405-419.

MACKIE, J. L. (1965) Causes and conditions. *American philosophical quarterly***,** 245-264.

PAWSON, R. (2002) Evidence-based Policy: The Promise of 'Realist Synthesis'. *Evaluation,* 8**,** 340-358.

PAWSON, R. & TILLEY, N. (1997) *Realistic evaluation*, Sage Publications Inc.

RAVALLION, M. (2009) Should the Randomistas Rule? *The Economists' Voice,* 6**,** 6.

ROCHE, C. J. R. & ROCHE, C. (1999) *Impact assessment for development agencies: Learning to value change*, Oxfam.

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 - 226.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Co. Boston, MA.

SNODGRASS, D. (2006) Assessing the Impact of New Generation Private Sector Development Programs. *Impact Assessment Primer Series #1.* Washington, USAID.

SNOWDEN, D. & BOONE, M. (2007) A leader's framework for decision making. *Harvard Business Review,* 85**,** 68.

TANBURN, J. (2008) The 2008 Reader on Private Sector Development: measuring and reporting results. Turin, International Training Centre - ILO.

WORLD BANK (2007) *World Development Report 2008: Agriculture for Development,* Washington, World Bank.

ZANDNIAPUR, L., SEBSTAD, J. & SNODGRASS, D. (2004) Review of Evaluations of Selected Enterprise development Projects. *MicroREPORT #3.* Washington, USAID.