CREDIBLE EVIDENCE: TACKLING THE CHALLENGES IN EVALUATING AGRICULTURAL VALUE CHAIN SUPPORT

Giel Ton & Sietze Vellema & Marieke de Ruyter de Wildt

Draft version 1 september 2009

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 centrepiece to agricultural policy in developing countries. Value chain support focuses on the capacities and capabilities of value chain actors and the enabling policies and institutions that facilitate change processes that benefit the poor, for example, by increasing the scale of operations, improving service provision to producers, developing capacities to comply with (buyer-driven) quality requirements or addressing issue of inclusion/exclusion of smallholder producers in the process of value creation and 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, and, partly a result of stronger public pressures on aid money to show its worth, convincing evidence is asked for (OECD 2008; SDC 2009). Specifically, questions are being raised about the effect of these partnerships on poverty alleviation. 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). 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’ with donor supported intervention ‘inputs’ is not easy. This 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. 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. Measuring impacts of interventions in this dynamic ‘cloud’ of complex and intertwined set 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 and an assessment system that generates sufficient information to choose effective instruments from the available policy menu. One of the promising initiatives to generate credible and comparable information on value chain interventions is led by the Donor Committee for Enterprise Development (DCED) (2008). The initiative proposes minimum standards for reporting on private sector development in a practical and credible way, in which monitoring income changes and calculating attribution to program interventions is a required practice. When put in practice this would be a great leap forward towards developing a body of evidence on value chain support.

To meet this standard reporting, lean research designs are needed that can face the most common threats to validity (Shadish, Cook et al. 2002; Bamberger, Rugh et al. 2006; Creevey and Woller 2006). The conclusions and policy recommendations derived from evaluative research need to be supported by data and information collected and analyzed in a credible way. To meet these challenges, we need a multi method research strategy to collect and analyze information that can stand up to scrutiny (Brady, Collier et al. 2006). In this paper we add to the discussions on the design of impact evaluations tools and methods, and present promising entry points to assess change in value chain configurations.

The paper consists of four sections. First, we briefly discuss the basic evaluation question in impact assessments and point to the different threats to validity when answering these questions. Second, we dig into three areas that are specific for value chain development and that generate additional design challenges: measuring performance indicators; tracking attribution in ever-changing value chain dynamics; and, making generalizations from specific pilot experiences and concluding on policy recommendations with a defined generalisation domain that consider the context specific social embeddedness of value chains. Third, we present design elements that are useful to assess impact and replicability of chain interventions combining data-set observations and causal-process observations (Brady and Collier 2004). In the final section we reflect on the applicability of these methods to assess impact and replicability of value chain support for development practice. We emphasize the need to better link ex-post impact evaluation processes with ex-ante constructions of plausible impact trajectories and credible outcome measurement.

2. Evaluation research

There are many different reasons for doing an evaluation. Three types of evaluation can be 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. Development evaluations seek to strengthen institutions and the strengthening of agencies and institutions in a particular evaluative area. Although there is an overlap in tools and processes, these three types of evaluation are underpinned by different purposes. The design challenges for evaluation methods will differ according to the questions it want to answer. In this paper, we focus on the first two types of evaluations with three basic questions:

- 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?

These three questions are, in varying wordings, asked by the stakeholders commissioning the evaluation and are intimately related. Each evaluation assignment will have a different emphasis. 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), however, argue that the first question continues to be far too dominant in evaluation research whilst the second question is more productive in providing useful guidance to the stakeholders involved and to generate policy recommendations. Likewise, Ravallion, chief evaluator at the World Bank, points to the dominance of methods that focus on showing if policies work or not, without generating much information on how it works and could work in other settings. He opposes specifically the dominance of econometric impact assessment methods that compare average values of outcomes between treated and control, or participant and non-participant or the average impact in the whole population. These might indicate if something ‘worked’ or not and estimate the effect size of the intervention. However, such averages are not very useful to understand why things happened and address issues of attribution and replication. 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 require different sets of methods to generate and analyse information that only partly overlap. They need different kinds of information, at least with different 'depth and detail' of the information collected. 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, is about formulating generalized inferences and extrapolating 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 therefore stress the need to design precise procedures that (partially) control some of the limitations of used research methods 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 lead 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 defines the generalization domain.

We illustrate these validity types with an example, highlighting their relevance to value chain evaluations. Imagine an evaluation team that is evaluating a subsidy for milk cooling facilitates in a developing country. The local government aims to improve the livelihood strategies of smallholder farmers through a grant facility for cooling tanks. Before the cooling tanks, production used to be scattered, irregular and of varying quality. Collecting the milk in a central place, where it is kept in acceptable conditions, is supposed to have made the smallholders' raw milk production attractive for the urban-based processing plants, stimulating diary production and hence enhancing the economic growth of poor households. The evaluation team is expected to come with a clear policy recommendation on replication of the interventions, a variant of " Yes, it works! Scale it up!".

Aware of methodological standards for impact assessment, the team tries to anticipate the most common threats to validity of that (future) conclusion in its research design. The evaluation team is aware that farmers have a diversified farming system, combining horticulture and dairy production. Farmers tend to increase their herd size in response to market incentives for horticulture, dairy and off-farm employment. The traditional use of milk is home consumption and artisanal cheese processing, sold or bartered locally and fresh milk marketing is hence only one of the available marketing channels that households can choose from. Milking in that region is essentially an activity of women and children, while men are dedicated to animal feed production. Knowing this, measuring change, correctly attributing change to the cooling tank facility and drawing wider conclusion for the local and perhaps national government policy is not at all straightforward. The team wants to combine a household survey and a value chain analysis to support its conclusions.

The survey captures different household characteristics and measures the quantity and quality of milk produced, the commercialisation channels and resulting changes in income. It also assesses the distribution of dairy income within the household, among men and women, and the impact of the changes on the agricultural system, especially horticulture income and division of labour within the households. They use several tests to conclude on the probability of a correlation between characteristics and the outcomes. These tests have varying assumptions and pre-conditions related with the data, like the 'normal distribution of the data' or the 'homogeneity of variance of the different groups' (like in ANOVA). Hence, s*tatistical conclusion validity* is key if the research method involves statistical analysis of data-sets. Just producing an output table that indicates 'significant' relations is insufficient. 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.

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: the cause need 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 'sufficient' to produce the outcome (Mackie 1965). Hence, in our example, the team will have to make plausible that the subsidy facility was indeed necessary for producing the outcomes that we observe. Access to finance for cooling technology is not sufficient to generate 'improved dairy production around a cooling tank'. Other causes are factors such as the social action by farmers, access to credit, availability of heifers, and proper road maintenance for the truck to come and collect. Non-observables, like the power struggles involved in determining milk prices between the processing plant and the farmers or gender relations that define the access to and use of monetary milk revenues, are 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 the cooling tanks subsidy policy, the importance of the subsidy policy in this' cloud' of causal factors has to be made plausible, and alternative explanations have to be discarded as much as possible. The data collection 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 that 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. “the support to the start-up investments in cooling tanks is effective in linking dairy producers to markets”, they immediately face several threats to construct validity: is ‘dairy producers’ a good construct, or do they 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 they 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, the team needs to be precise about the concepts and constructs used.

Even more challenging, as the team seeks recommendations about the replicability of the support, are the threats to *external validity*. Even when the team comes to the conclusion that in the specific context the subsidy policy was a key factor with positive results, this will not necessarily hold in other settings. Hence, the team needs to argument why, and to what extent, the findings can be generalized and remains valid in other contexts and conditions. Like all four types of validity, but stronger than the earlier three, there is no definitive answer. All ‘best practices’ and lessons learnt will result from a ‘peculiarity’ related to the context and conclusions will be limited to conditions such as consumer price margins, civil peace or minimal standards of social capital to handle issues of opportunistic behaviour (e.g. adding water to milk). Hence, we need to design tools that can respond to the most obvious or relevant questioning of the validity of our policy recommendations; we have to generate information to define our ‘generalisation domain’ (Chen 1994).

Few evaluations in international development systematically address issues of validity: “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). The field of value chain support is no exception (Zandniapur, Sebstad et al. 2004; Humphrey and Navas-Aleman 2009).

3. Methodological challenges

We will now discuss validity in three core methodological areas that follow the emphasis of 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 operationalise. 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 information needed to support conclusions on each issue overlap. To know if some components work in specific conditions, information on outcomes and impact will be very useful; to test if something worked, the statistical model must have a coherent causal model of the characteristics needed to make the intervention work. 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). Brady et al state that, to make high validity causal inferences, a combination of these two types of information is needed and call this 'nested inference' or 'triangulation' (Brady and Collier 2004). The following discussion will show that 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 support. 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 associations or buyers, rather than at producers themselves. The interventions will have an explicit or implicit 'theory' or impact model that translates the support to these chain actors into behavioural outcomes of chain actors, including producers. The impact model is not necessarily related to the intervention as a whole ("Did the programme work?"), but may concentrate on subsets of conditions, components of interventions, specific instruments, and the type of outcome patterns. Impact assessments need to select (sets of) outcome indicators that can function as 'proxy' for performance in each target areas. Preferably, measurable, continuous and quantitative indicators (dependant variable) are selected as proxies for the outcomes of a (component of an) intervention. 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 defining concepts and indicators of an impact model, the issue of construct validity is paramount. The previous section already mentioned the difficulty to be sufficiently precise about concepts and constructs that enable generalisations at a later stage. Value chain performance 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 these constructs makes it difficult to capture and measure with quantitative indicators. More so, concepts and indicators are often influenced by the disciplinary background and ontological theories of the evaluator (set aside personal interests). 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 measure the relevant 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 can be considerable. In our dairy example, net additional income can be restricted to net income growth from fresh milk sales to processing plants. However, it can also be seen as the net income change of the whole agricultural system of the household, as increasing dairy production and increased animal feed production will impact horticultural production and family income. Positive spill-over effects may exist, since farmers learned about quality issues, have more intense communication with fellow farmers and buyers and therefore improve entrepreneurial skills and production levels beyond diary. If fresh milk sales are only a small part of the farm enterprise, the differences in measured income impacts will be significant. The more comprehensive way of calculating income impact has some important trade-offs. It 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, this will influence incomes without any causal relation with the intervention to be 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. For example, not just the poverty status of the target group will have to be measured, but the number of people that changed category as a result of the intervention. A 2% increase in non-poor can consist in 10% of respondents that move out-of-poverty with a 8% that moved into a lower category. Generalised inferences, indicating that interventions reduced poverty, will have to be supported by tests about the significance of this difference, considering the variation of impact in the data-set observations.

To translate observed changes in outcome indicators to measures of impact that may be attributable to the intervention, ideally, a comparison is needed with outcomes of a control group, a group with similar characteristics that did not experience the working of the interventions. This is necessary to evaluate if the outcomes can be attributed to any 'exogenous' or 'unknown' causal factor or set of causal factors, not related to the intervention's mechanisms and not incorporated as variables in the data-set. Experimental methods that measure and evaluate effectiveness and impact of interventions with random control groups (Duflo, Glennerster et al. 2006) are often impossible, and often even unwanted in value chain support. 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 generally politically unfeasible or ethically unwanted (Shadish, Cook et al. 2002; Bamberger, Rugh et al. 2006). Also, in many cases there are important spill-over effects from pilot-intervention areas to other areas and chain actors. The definition of who is a participants and who is not is often a gliding scale and can make the distinction in 'treated' and 'control' groups unworkable (Ravallion 2009). Random assignment of the intervention to a defined population is rarely possible, and quasi-experimental methods, are therefore more frequently used than random control groups. However, research designs that deviate from random assignment face the risk of 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 statistical conclusion validity. 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 correlation founds. Current software makes consecutive iterations of statistical analysis with changing combinations of variables so easy that 'significant' correlation between some of the variables may result from 'fishing the data' or 'data mining': repeating statistical tests on significance of differences between groups by selective re-grouping respondents and/or variables, etc. Even if the intervention has no effect at all, in complex data sets, one or more significant correlations are likely to appear always after a sufficient number of iterations (Shadish, Cook et al. 2002). Reasoning back, from the detected correlation to a causal hypothesis, induces to conclude that some changes occur as a result of the intervention, while in reality these are unrelated. The recommended solution against 'fishing' is to specify ex-ante the theoretical model of the causality that will be tested with the statistical analysis and to increase the threshold (significance level) of the correlation detected after iterative analysis. 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[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, the reality that respondents will fall out, e.g by moving, passing away or changing their activities in way that their survey result are no longer useful in the analysis. For explorative statistical analysis, and considering attrition, sample sizes are ideally much larger than the minimal required size. However, in the 'real world' sample sizes are often restricted by resource constraints (financial, not enough people, too difficult to get too etc) and subgroup comparisons limited by a low statistical power.

The last step to measure impact proposed by DCED is the intent to capture wider changes. The most obvious threat to validity of an evaluative conclusion is that it left important factors out of the equation, be it as confounding causal factors or as outcome indicators, not capturing the intended and unintended change process as a result of the intervention, and therefore threatening the internal validity of the findings. Unintended changes are unlikely to be captured by pre-established indicators in causal impact models. Additional critical, Causal-Process Observations are needed to assess these unintended outcomes and rule-out irrelevant ones. The emphasis on documenting wider

[1] 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.

impact is important, as many evaluations tend to find proof for their impact logic only (European Commission 2008)

Attribution in open systems

The second evaluation question, how does it work, focuses on the causality between the support and the observed changes. Significant correlations do not indicate causality, but at least indicate that there is, most probably, a relation between the intervention and 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 only for simple and closed social systems 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 (Pawson 2002; Lawson 2003; Hospes 2008). 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 mathematical models holds also 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 construct with different variables that 'models' the context of the respondent. The respondents with a comparable score on the model's questions/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. These matching models are heavily theory-laden: it supposes that the matching is done on variables that indeed 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 can also be constructed after the survey during data-analysis[1]. The model used to match respondents based on background characteristics is always 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, and that are not related to the intervention, will end up testing closed models of reality. Therefore, critics may always threaten the validity claims from statistical and econometric causal inferences of survey data by 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 indicate clearly and consistently why (the most salient) external factors are considered irrelevant for explaining the observed outcomes, and that that the conclusions of the PSM are, therefore, credible and useful, though essentially fallible.

Realist evaluation, specifically concerned with causal process tracing, provides a useful framework (Pawson and Tilley 1997) for analysing which specific mechanisms in an intervention trigger behavioural change. It emphasises the need to build a hypothesis related to the (project) mechanisms that (are assumed to) motivate or influence stakeholders ‘to act differently’ and generate changes in outcomes. All value chain interventions, ultimately, are intended to change attitudes and behaviour in persons. The workings of the support are often implicitly assumed in the impact logic, such as “the availability of cooling tanks will increase the interest of urban-based processing plants in small-scale fresh milk production”. Realist evaluation proposes to test these key assumptions with the concepts “Context-Mechanism-Outcome Configurations”. The concepts are useful to understand real cases and to precise impact models (Table 1). The detailed description and analysis of a pilot intervention feeds the theories behind the design of policy and programmes.

Table 1 - Realist Value Chain Intervention Case Study Format

Realist Concept	Domain of application	
	Understanding pilot interventions (Causal theory)	Designing policy and program (Normative Theory)
Context	Situation of the value chain in the pilot experience	Situation of the value chain 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

The 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. The realist emphasis on contextual embeddedness helps to specify (and limit) the policy recommendations on eventual future replicability. In their analysis, realist evaluators concentrate on the ‘treatment’ and the incentives for the ‘treated’, without bothering too much about a control group. As mechanisms work under specific conditions, causal inferences about them tend to be bound to each ‘case’ or discrete observation.

However, it is difficult in realist evaluations to demonstrate that rival explanations of the occurrences can be ‘eliminated’. Qualitative tools, often used in causal-process observations, are good in identifying and assessing rival explanations, but quite poor in convincingly eliminating them. Farrington (2003) points to this weakness. He argues that with limited time and resources for evaluations, it is difficult to deal with multiplicity of contexts, mechanisms and outcome patterns. Evaluations often end in a multitude of causal inferences with very limited scientific validity, especially if research methods are without 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. He even argues that for assessing causal impact of interventions, instead of the concept ‘mechanism’, one better uses the concepts ‘moderator’ and ‘mediator’ variable, as applied in statistical analysis. The moderator variables show that context matters, and that outcomes are context dependant (e.g. related to a typology of contexts). The mediator variables indicate factors that interact in/with an intervention. Indeed, it would be good to have data-set observations to support the validity claim of inferences from case-studies.

Hence, to attribute change to value chain support, triangulation of data collected with methods for (quantitative) data set observations and (qualitative) causal process observations is necessary to provide data that will support the conclusion and provide it with sufficient internal validity (Brady, Collier et al. 2006). Different methods need to be directed to the evaluation of the same processes and outcome patterns. 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 statistics, the common measure 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. As previously discussed, budget, time, logistics and political constraints are such that the ‘golden standard’ of random surveys, with treatment and control groups and pre- and post measurement of outcomes, is seldom applied (Bamberger, Rugh et al. 2006). Especially in self-selecting populations, e.g. when the location of the treatment and control group is not fixed or the characteristics of the adopting group may change in time, the pre-tests or baseline surveys would have to cover a wide geographical area and a lot of different categories of respondents to be of any use for the ex-post impact assessment of differences in subgroups of the ‘treated’. However, in data-set observations collected through survey samples, there is no better statistical design than random sampling to defend the claim that findings have external validity in generalisations across populations. When correctly executed, it facilitates inferences from a survey sample that are valid for the population the sample is taken from. Threats to this claim of external validity arise especially 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. ‘Good practices’ or ‘emergent practices’ are concepts used to indicate mechanisms or interventions that proved to work in a certain setting, and that might work in others. Policy makers are especially interested in these practices that provide them with a menu of options. Realist evaluation with its focus on “What works for whom under what conditions?” is explicit about the limitations in the generalisation domain of these interventions and offers an approach that strengthens external validity (Pawson and Tilley 1997).

Shadish et al (2002) propose a process to deal with external validity of findings. They present five principles to limit the validity threats in evaluation design, 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 an 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).

Table 2 summarises the validity threats that have been discussed when measuring outcome patterns, when dealing with attribution and when drawing conclusions beyond a specific context. Although the four validity types can not be seen in isolation, they have different levels of relevance to each evaluative question. Statistical conclusion validity seems particularly relevant for the first evaluative question (does it work). Equally so, internal validity seems most weighty for the second question (how does it work) and external validity for the third question (will it work elsewhere). Construct validity, being about concept definitions, is essential throughout the evaluation research.

Table 2 - Relevance of validity threats per methodological area

	Measuring outcome patterns	Dealing with causality in open systems	Social embeddedness and generalisations
Threats to **statistical conclusion validity**, e.g. wrong timing, lack of data, wrong group selection, unconsidered spill-over effects, panel attrition and inadequate sample sizes, data fishing or excluding unforeseen impacts	High	Low	Low
Threats to **internal validity**, e.g. lack of theory supporting causal model, biased control group or absence of experimental methods	Low	High	Low
Threats to **external validity**, e.g. no defined generalisation domain or absence of experimental methods	Low	Low	High
Threats to **construct validity**, e.g. implicitly influencing and narrowing theories, imprecise concepts and constructs, wrong sorts of indicators or too few or too many indicators	High	High	High

4. Discussion

From the above, we conclude that for evaluating replicability of value chain development we need theory. 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. Theories are used to describe causality in past events and to predict causality in current or future realities. 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 impact we need causal models that explain dynamics in empirical reality, while for replicability we anticipate with theories on how an intervention will impact future dynamics. Chen (1994) calls this *causal theories* and *normative theories* of program impact. Causal theories are descriptive of changes 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.

Logframes often fall short of being impact models. Impact models and their supporting theories are often ‘hidden’ in interventions and evaluators regularly find themselves unravelling and reconstructing them. Logframes, in many cases the result of multiple planning sessions and discussions, relate planned intervention to outcomes in a logical sequence of activities/inputs – outputs – outcomes – impact. The logframe is a common management tool when planning interventions and can easily be translated in budget items and monitoring indicators. A disadvantage of this type of planning is that the discussions on how output relates to outcome and outcome to impact is limited. More over, these discussions are hardly systematic and in many occasions poorly documented: the arrows between these four elements remain black boxes. For logframes or impact logics to become more useful for evaluation, underlying theories will need to be made more explicit.

A way to focuses on the mechanisms that ‘trigger’ behaviour during or after an intervention is to use the realist concept of “Context-Mechanism-Outcome Configurations” in case studies, analysing cases along these 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 outcome indicators. The case-studies indicate context-dependant practices, in stead of ‘good practices’ and, at most, can suggest a ‘best fit’ in a comparable configuration. They are 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 extremely 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 are needed that quantify outcomes and impacts of value chain interventions and test the key assumption inherent in the impact models. Measurements are needed to support claims that something does work, and provide information useful to explore the real causal processes and compare them with the normative impact models. As has been discussed in preceding chapters, statistically significant differences between groups, or correlations between variables, are not per sé an indicator of attributable impact. To ‘upgrade’ a significant correlation into a causal relation with strong validity claims, some design features will have to be incorporated in impact assessment surveys that collect information that can be used in appropriate statistical analysis to discard alternative causal explanations of the effect. Statistical analysis 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 analysis of contrasting cases may help to understand the logic and rationale behind observed changes and helps to clarify the conditional and contextual character of an intervention’s impact. “Although means are traditional, the answer to many interesting policy evaluation questions requires knowledge of features of the distribution of program gains other than some mean” (Heckman 2005).

5. Conclusion

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. Checks on validity threats strengthens the understanding of the working of the interventions and delineates the generalisation domain. Threats to validity are especially challenging when the evaluative conclusions are used to decide on replicability and up-scaling.

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. More information of the generalisation domain of interventions that change these (interlinked) institutional arrangements may prevent failures, and help to build context specific and evidence-based theories of change.

We propose a design based on a combination of impact models, triangulation of data-set observations and casual process observations, with a realist focus on the key mechanisms that are assumed to work. This design will provide information that is useful for both accountability purposes, on on-going interventions, and for learning on best-fit practices that can be replicated in future interventions. The logical link between these three design elements facilitate ‘nested inference’ with increased scientific strength and limit the threats to validity of the evaluative conclusion.

7. 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.

Bamberger, M., J. Rugh, et al. (2006). RealWorld evaluation: working under budget, time, data, and political constraints, Sage Publications Inc.

Brady, H. E. and D. Collier (2004). Rethinking social inquiry: diverse tools, shared standards, Rowman & Littlefield Publishers.

Brady, H. E., D. Collier, et al. (2006). "Toward a Pluralistic Vision of Methodology." Political Analysis **14**(3): 353-368.

Chelimsky, E. and W. R. Shadish (1997). Evaluation for the 21st century: A handbook, Sage.

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

Creevey, L. and M. Ndiaye (2008). Common Problems in Impact Assessment Research. Impact Assessment Primer Series #7. Washington, USAID.

Creevey, L. and G. Woller (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., R. Glennerster, et al. (2006). "Using Ramdomization in Development Economics Research: A Tool Kit." NBER Working Paper.

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**(1): 49-68.

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

Hospes, O. (2008). "Evaluation Evolution?- three approaches to evaluation." The Broker **2008**(8): 24-26.

Humphrey, J. and L. Navas-Aleman (2009). Multinational Value Chains, Small and Medium Enterprises, and 'Pro-Poor' Policies: A Review of Donor Practice. Brighton, IDS-Ford Foundation.

Lawson, T. (2003). Reorienting Economics, Routledge.

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**(3): 340-358.

Pawson, R. and N. Tilley (1997). Realistic evaluation, Sage Publications Inc.

Ravallion, M. (2009). "Should the Randomistas Rule?" The Economists' Voice **6**(2): 6.

Roche, C. J. R. and C. Roche (1999). Impact assessment for development agencies: Learning to value change, Oxfam.

Shadish, W. R., T. D. Cook, et al. (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.

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., J. Sebstad, et al. (2004). Review of Evaluations of Selected Enterprise development Projects. MicroREPORT #3. Washington, USAID.