Goals of evaluation and types of evidence

Marielle Berriet-Solliec, Pierre Labarthe, Catherine E. Laurent

Goals of Evaluation and Types of Evidence

Quotation: Berriet-Solliec M., Labarthe P., Laurent C., Goals of evaluation and types of evidence. *Evaluation*. 2014, Vol.20(2) 195-213.

DOI: 10.1177/1356389014529836

The published version of this article is available on SAGE Journals http://online.sagepub.com

Marielle Berriet-Solliec
Professor, Agro-Sup / INRA, Umr 1041, Dijon, France

Pierre labarthe
Scientist, Inra, Sciences for Action and development, Umr 1048, Paris, France

Catherine Laurent
Corresponding author
Senior scientist, Inra, Sciences for Action and development, Umr 1048
16 rue Claude Bernard
75231 Paris Cedex 5
France
Email : catherine.laurent@grignon.inra.fr

Abstract

All stakeholders are urged to pay more attention to the quality of evidence used and produced during the evaluation process in order to select appropriate methods of evaluation. A "theory of evidence for evaluation" is needed to better address this issue. In that aim, this article discusses the relationships between the three main goals of evaluation methods (to learn, measure, and understand) and the various types of evidence (evidence of presence, of difference-making, of mechanism) which were produced and/or used in the evaluation process. It shows the need to clearly distinguish between this approach and that of levels of evidence, which is linked to data collection and processing methods (e.g. single case observations, difference methods, randomised controlled trials…). The analysis is illustrated by examples in the field of agro-environmental policymaking and farm advisory services.

Keywords: evaluation, evidence, evidence-based decision, agricultural extension, agri-environment, agricultural policies, knowledge

Résumé

Tous les acteurs sont appelés à accorder une importance grandissante à la qualité des « preuves » utilisées et produites par les procédures d'évaluation pour choisir les méthodes d'évaluation les plus appropriées. Pour mieux répondre à cet enjeu il est nécessaire de construire une « théorie des preuves pour l'évaluation ». Dans cette perspective, cet article discute les relations entre les objectifs des méthodes d'évaluation (apprendre, mesurer, comprendre) et les différents types de « preuve » (preuves de présence, de mécanisme, d'effet) produits ou utilisés dans les démarches évaluatives. Il montre la nécessité de clairement différencier cette réflexion de celle sur les niveaux de « preuve » qui renvoie aux méthodes de recueil et de traitement de ces données (monographies, méthodes de double différence, essais randomisés contrôlés...). L'analyse s'appuie sur des exemples dans deux champs d'application : les politiques agri-environnementales et le conseil agricole.

Mots clé: évaluation, preuve, evidence-based decision, conseil agricole, agri-environment, politiques agricoles, connaissances

JEL classification:
B49 - Economic Methodology - Other
H83 - Public Administration; Public Sector Accounting and Audits
Q18 - Agricultural Policy; Food Policy
Q58 - Environmental economics - Government Policy

There is a renewal of research into the effects of knowledge characteristics on the dynamics of collective decision-making in public or private organizations. For years, studies have stressed and modelled the diversity of the sources and types of knowledge used in decision-making processes (expertise, theories on causal relations, traditional knowledge, etc.). More recently, some theoretical developments, such as research around "evidence-based decisions", have merged learning from various disciplinary standpoints (philosophy of science, medical studies, economics, ecology...) and opened new debates on "empirical evidence for use" (Cartwright, 2011). Regarding evaluation, a proposal to rank evaluations according to an evidence base (e.g. Lipsey, 2007) has caused heated discussions about what counts as evidence (Donaldson 2008). These debates are calling on decision makers to pay more attention to the quality of evidence for selecting appropriate methods of evaluation and assessing their conclusions. They are also calling for a "theory of evidence for evaluation" (Schwandt, 2008).

This paper aims to contribute to the building of this theory. We shall analyse the relationships between goals of evaluation and types of evidence (i.e. what is object of evidence in different types of evidence). We shall demonstrate how the resulting theoretical advances help to better analyse the trade-offs involved in the use of alternative types of evidence. To do so, we shall focus on the ex-post evaluation of public action programs in agriculture: specifically, for advisory services and agri-environmental policies.

Diversity of the goals of evaluations

A general objective of the public action evaluation process is to organize and analyse information gathered on the program under evaluation. Many methods exist (intervention logic, theory of action, theory of program, theory-driven evaluation, contribution analysis ...), and this diversity may be confusing to users. This difficulty is exacerbated by the similarly wide range of theoretical models upon which the public action programs under evaluation are developed and implemented: Patton (2008) identifies more than one hundred kinds of evaluation and refers to a state of "Babel confusion". Several classifications of these methods were proposed recently (Stufflebeam, 2001; Oliver et al., 2005, Rogers, 2008; Hansen and Rieper, 2009; Fitzpatrick et al., 2011; Stern et al, 2012). In addition Stern (2004) showed the heuristic value of an analysis that links the methods of evaluation with the purpose of the evaluation.

Indeed, regarding evaluation goals, evaluation studies can be classified, in a very stylized way, into three broad groups, keeping in mind that one evaluation procedure may combine several simultaneous or successive goals.

- *Goal 1: to measure.* The evaluation is designed to assess the effects of a program.

A first group of studies focuses on the quantification of program impacts using micro-economic techniques (Rossi et al., 2004), often in line with the work of Heckmann. An emblematic principle of this type of research is the identification of an experimental or quasi-experimental situation in which systematic reference to a counterfactual can be used to identify outcomes which are specific to the program under evaluation (Shadish et al., 2002; Banerjee and Duflo, 2009). This first issue indicates a need to identify if a public intervention works (a measure of effect usually referred to as "impact assessment"). A second group of studies aims at measuring efficiency. It involves measuring the value of goods or services produced through public action programs against the cost of their production. The goal is then to determine whether an organization or initiative has produced as many benefits as possible

given the resources it has at its disposal; this approach takes into account a combination of factors such as costs, quality, use of resources, appropriateness and whether deadlines were met.

- *Goal 2: to understand.* The evaluation identifies and analyses the mechanisms by which the program under evaluation can produce the expected outcomes or may create adverse effects. This second goal is the basis of studies of the theories underlying public programs and analysis of the specific mechanisms by which these programs have made an impact. Chen and Rossi (1983), Chen (1990) and Shadish et al. (1991) introduced the debate in the 1980s and 90s. , and several theoretical works were recently published on these issues (Shadish et al.; 2002, Stame 2004; Donaldson, 2007 ; Donaldson et al.; 2008, Jordan et al., 2008). In practice, this raises the question of what knowledge can be used to provide a reliable empirical basis to implement these approaches (Pawson and Tilley, 1997; Schwandt, 2003, Pawson, 2002 and 2006) but also of what credible claim of the contribution of an intervention to a change can be done in the absence of experimental approaches (Mayne 2012).

- *Goal 3: to learn.* The evaluation is designed as a collective learning process. Many studies emphasize the importance of elements which support the use of evaluation, which is intended to facilitate the implementation of adequate methods and the appropriation of evaluation findings by different types of users (Patton 2008). Evaluation is considered an operational approach intended to improve public action programs and decisions. Emphasis is placed on its instrumental dimension (as a response to an institutional demand) and on the role played by evaluation approaches as an organizational learning process. This goal can lead to the idea of a "learning society" (Schwandt, 2003) and to a new conception of evaluation as a form of inquiry involving pedagogical engagement with real practice. Using diverse participatory methods (stakeholder-based, democratic, collaborative, pluralist, responsive...) (Cousins and Whitemore, 1998, Mertens 1999), this "learning" objective can be paired with the goal of empowerment (Fetterman, 1996; Fetterman and Wandersman, 2005).

This plurality of goals generates an initial question: should we consider the quality of evidence in the same way for all these cases?

Types and levels of evidence

Recent progress in the study of quality of evidence (e.g., Shadish et al., 2002; Laurent et al., 2009; Illari 2011; Cartwright and Hardie, 2012) can help clarify on-going controversy over what counts as good quality evidence for an evaluation. In particular, an explicit distinction needs to be made between *types of evidence*, (depending on the object of evidence), and *data collection methods*, which determine the probative force of the produced evidence.

When an ex-post procedure is used to evaluate a public action program, generally the goal is to produce the best possible knowledge to assess the actual outcome of the program and, to the furthest extent possible, base later action on these outcomes. The 'best' knowledge should be a) *socially relevant* to those concerned and considers adverse effects, b) based on *adequate types of evidence* (in line with what the evaluation entails) and c) *reliable* (produced using rigorous methods, to ensure the highest degree of probative force).

Types of evidence

Broadly speaking, three types of empirical evidence may be necessary to evaluate public policies:

1) *Evidence of presence*: the description and verification of a thing which exists on the ground (e.g. species observed while building a botanical inventory to describe biodiversity). This type of evidence is used to build an agreement among different stakeholders on the state of the world (before and after the program). This can be approached through a proxy (for instance, the number of footprints of individuals belonging to certain species).

2) *Evidence of difference-making*.
- This can be *evidence of effectiveness*: evidence that a given action yields the desired result (e.g. improved biodiversity following the implementation of agro-ecological regulations aimed at biodiversity conservation).
- It can be also *evidence of harm*: obtained when adverse effects of an intervention have been looked for and found (e.g. adverse effects of agro-ecological regulation on the sustainability of small-scale farms (Adams et al., 2004)).

This type of evidence requires the identification of an outcome (O) which can be observed (eg. the return of a species no longer observed in the area) and whose expected value (E) is specific to a certain intervention (T) (here a public program). Thus, the impact (I) of the program for a population can be expressed as

$I = E(O \mid T=1) - E(O \mid T=0)$ (1)

3) Evidence of mechanism for a phenomenon. This is produced when there is evidence that the entities or the activities that make up a mechanism, and the organization of these entities and activities by which they produced the phenomenon, are known (e.g. the bio-chemical reactions needed for an increase in fertilizer (C = cause) to increase crop yield (O = outcome) in a controlled environment).

This type of evidence may confirm a relationship of cause and effect, all other things being equal. It provides information on the causal pathway to intervene upon for the goals of a public program to be achieved.

However, in real life conditions, evaluators are always confronted with complex causal structures in which various mechanisms interfere. In that respect, following Cartwright (2011), a probabilistic theory of causality can be adopted.

"*For each effect-type at a time t, O^t, and for each time t' before t, there is a set of factors $\{C_1^{t'},\ldots,C_n t'\}$ – the causes at t' of O at t – whose values in combination fix the objective chance at t' that O takes value o for any o in its allowed range. A causal structure, $CS^{t'}(O^t)$, for O^t is such a set along with the related objective chances for all values of O^t for all combinations of allowed values, $L_j^{t'}$, of the causes in the set: Prob ($O^t = o/L_j^{t'}$). For simplicity I will usually supress time and other indices and also restrict attention to two valued variables. So a causal structure looks like this: $CS^{t'}(O^t) = <\{C_1^{t'},\ldots,C_n^{t'}\}, \{Prob(O^t/L_1^{t'}),\ldots,Prob(O^t/L_m^{t'})\}>$ "* (2) (p.16) ""

In practice, full knowledge of the causal structure involved in a public program is generally unreachable. It is therefore useful to develop hypotheses on the mechanisms that will play an important role, in order to design an action program and have an effect on "manipulable" factors (Shadish et al., 2002) or to analyse whether an intervention is a

contributory cause to a change (Mayne, 2012). Here, evaluation usually involves the production of both evidence of mechanism and evidence of difference-making, a combination which provides information about causal pathways.

In certain cases, however, an evaluation is based exclusively on evidence of difference-making and therefore says little or nothing about underlying causality if the causal structure is complex.

Observation may, indeed, be focused on one of two elements:

- the production of the expected mechanism, by observing changes which occur at each stage (e.g. whether a financial incentive has led to a shift in practices which has in turn led to the use of a fertiliser that has an impact on crops). In this case, evidence of mechanism will be combined with evidence of difference-making to help clarify causal relationships.
- measurement only of the produced effects (e.g. does income support increase production levels?), without hypothesizing about the causal chain involved (purchasing of consulting services, purchasing of inputs, reduction of risk aversion, etc.). Here, evidence of difference making provides little information on the causal relationships which need to be studied in order to judge how generic obtained results are.

Disentangling various types of evidence highlights the ambiguous relationship between evidence of difference making and causality: in certain cases, these types of evidence reveal nothing about causal pathways. This remains true even when such evidence is produced using methods (such as randomized controlled trials) which may confer high level of proof. Types of evidence and level of evidence are two independent dimensions of the quality of evidence.

Levels of evidence

The assessment of levels of empirical evidence is usually considered a major issue. Whatever the type of evidence, not all findings have the same probative force: they cannot be ranked at the same "level of evidence". In the field of agriculture, for example, levels of evidence of effectiveness can be classified in the following order, from least to greatest, according to the methodology of data collection:

1. the opinions of respected authorities;
2. evidence obtained by single case observations;
3. evidence obtained from historical or geographical comparisons;
4. evidence obtained from cohort studies or controlled case studies;
5. evidence obtained through randomized controlled trials (RCT).

But there is not "one" methodology (e.g. RCT) that could be considered as the gold standard for all situations. Other types of ranking are possible. For instance, if a research aims at understanding a mechanism (e.g. the reasons depending on individual behaviours why children/parents will accept a treatment), then in depth qualitative studies including single cases observations provide higher level of evidence than results of cohort studies based on probabilistic models (Petticrew, Roberts 2003).

In addition the apparent simplicity of the former classification should not conceal that the assessment of the quality of evidence produced at each level could be based on different criteria (study design, quality of study conduct, consistency of results...) (Liberati *et al.* 2001).

It should not conceal either the numerous questions that arise when several types of evidence are involved and need to be combined and/or are in competition (Laurent and Trouvé, 2011).

In other words, the criteria for assessing the level of evidence must be chosen according to the objectives of this assessment.

Invoking the argument that there is no universal rule to rank the level of evidence, some authors reject this very principle and argue in favour of a symmetry of knowledge, putting on the same level opinions from various stakeholders, traditional knowledge gained from experience, empirical evidence resulting from systematic investigation, etc..

Such a renunciation may generate significant adverse effects when it comes to action. In a large number of real evaluation settings, stakeholders want information that is as robust as possible to help them comply with their objectives. This is the case in many areas of public intervention such as agriculture, which involve both private and public organizations and which gather actors who consider that it makes sense to look for the best possible level of evidence for informing their decision (Labarthe and Laurent, 2013).

Therefore, both empirical observations and progresses in the theory of evidence invite to abandon two equally unproductive claims: those pretending that there is a unique methodology for ranking the level of evidence and those rejecting the very principle of assessing the probative force of evidence. Instead, they emphasize the need to define clear principles that will enable various stakeholders to assess the level of available evidence, utilizing the criteria that are relevant for their particular objectives.

The case of agriculture

To analyse the links between goals of evaluation and types of evidence, we shall take examples in agriculture-related policies because in agriculture, like in medicine, evaluating what works and what does not has for long been a source of enquiry, observational tools and analysis. Basing a decision on erroneous conclusions in agriculture or medicine can have serious, irreversible and immediately visible consequences (a person's death, ruined crops and famine, death of a herd, etc.). These decisions are taken at different levels (farm, sector, national policies…) and generate a wide variety of evaluation configurations for the corresponding types of programmes. They concern a wide range of interventional situations, from the 'simplest' of problems (e.g. measuring the impact of a single extra input on yield) to the most complex (e.g. measuring the multivariate impact of agri-environmental policies on communities and ecosystems, discussed in the Millenium Ecosystem Assessment (Carpenter et al. 2006, Mac Neely et al. 2005)).

In agriculture, like in medicine, the context of the decision is very heterogeneous. The phenomena under evaluation involve factors that are physical (e.g. effects of climate, soil differentiations…), biological (e.g. genetic variability generating differences in yields, resistance to diseases…) or of the realm of the social sciences (economic policy, farm advisory services, etc.). Similarly, a single subject of study can be approached using a wide range of complementary or competing interventions based on different theories of action. Two examples are farm advisory services (Davis, 2008) and agro-environmental measures (Kleinj and Sutherland, 2003). Given the eclectic range of subjects and interventions possible, there is a strong incentive to find shared analytical frameworks through which to assess the relative pertinence of alternative evaluation methods.

This tradition thrives all the more because intervention in agriculture (financial public support, regulatory measures, technical support, etc.) is subject to decisions taken jointly at the international level, whether it involves policy frameworks (e.g. the Common Agricultural Policy), health and environmental standards or economic support for production and advisory services. In addition, over the last two decades, evaluation is no more confined to the assessment of the productive performance of farm activity. New stakeholders have joined the discussion with concerns related to the environmental performances of agriculture and to its contributions to rural development and social cohesion issues.

In the case of farm advisory services, for example, a global forum has been created (the Global Forum for Rural Advisory Services, or G-FRAS) to facilitate collective discussion, working groups, reports and evaluation initiatives. In Europe, the European Commission has commissioned an evaluation of the implementation of advisory services in different member countries. These initiatives highlight sensitive issues about the use of evidence according to the goal of the evaluation: i) when measuring the effects of alternative advisory interventions (e.g. debates about the probative force of alternative methods for impact assessment); ii) when assessing the robustness of the causal scheme of these interventions (e.g. does the idea of knowledge diffusion, upon which many of these interventions are based, hold up in the field?); and iii) even when promoting learning through evaluation.

Ideally, an evaluation procedure should be aimed at producing results based on evidence of the best possible quality. However, such a view remains highly theoretical and blind spots subsist in how such evidence is actually produced. As demonstrated below in three kinds of ex-post evaluation, the adequacy of a type of evidence varies depending on the goal of the evaluation.

To measure. Type of evidence for evaluations designed to measure the effects of a public intervention

In this section we consider evaluations aimed at making an impact assessment, to provide empirical evidence of the difference(s) made by a public program, in order to measure as best as possible the actual impact of this program – and *only* that. This impact is defined as the difference between -
the actual situation with the program, and
- the situation that would have occurred without it.

In other words, the evaluation process does not examine in detail the mechanisms by which an action is effective; public programs mobilize a large number of factors and it is often impossible to observe every form of interaction between them. In most cases, evidence of effectiveness is sought in order to prove that the program made a difference, not to describe the mechanisms that made the measure effective, nor to control whether the effects confirm an underlying theory of action. Therefore, the evaluator does not open the 'black box' of the evaluated program. For instance, evidence that an agri-environmental scheme has been effective in maintaining biodiversity can be sought, without analysing the specific ecological, economic and social mechanisms that contributed to that outcome.

Implementation: measuring effectiveness independently of insight into mechanisms

Producing evidence of the effectiveness of a public action program requires identifying that there is an "all else being equal" relationship between two variables: a proxy of the treatment or public program T under evaluation, and a proxy of the desired outcomes of that program T, on a population φ. Here, the main objective is to measure the difference between an observable situation (the level of O, for the population that benefits from the program, T=1) and a counterfactual unobservable one (the level of variable O which would still occur in this same population without the treatment, T=0). In practice, this is done by comparing, through proxies, the levels O in a population which received the treatment and in a control population which did not.

$$I = E(O \mid T=1) - E(O \mid T=0) \quad \textbf{(1)}$$

In other words, ideally, for impact assessment, *"the population φ divides into two groups that are identical with respect to all other features casually relevant to the targeted outcomes, O, except for the policy treatment T, and its downstream consequences*" (Cartwright, 2011, p.18).

The main pitfall in this situation is a selection bias where differences exist between the 'treated' group and the control group (stemming from observable or unobservable factors) which could explain variations in levels O independently of the effects of the program T.

In light of this, evidence-based decision studies in the medical field rank the methods used in terms of their ability to reduce this bias. The smaller the bias, the higher the level of evidence. Traditionally, randomized control trials (RCT) are viewed as the 'gold standard' for measuring the outcomes of a specific program. Selection bias is eliminated by randomly distributing individuals in the treated group and the control group. For this reason, new, experimental evaluation methods (Duflo and Krémer, 2005) are emerging in various sectors (justice, education, the social sciences as well as the environment and agriculture). However, while such methods are widespread in health-related fields, they are less used for other public programs, where the randomization of beneficiaries of a public program can pose technical and ethical problems. In cases where RCT cannot be used, 'semi-experimental' methods such as matching or double differencing are considered the most reliable alternatives (Bro et al. 2004). Matching involves pairing individuals who benefited from the program with individuals who did not and comparing the levels of indicator variables. The goal is to pair individuals based on their most significant similarity, particularly in terms of how likely they are to benefit from the program. The double difference method is a combination of a comparison before and after the implementation of a public program and a comparison with and without the program. Differences in O are measured with proxy variables in both the beneficiary group and the control group. Nevertheless, both matching and double differencing have limitations. Matching makes it possible to pair individuals using only observable variables, with the risk that unobservable ones (skills, attitude, social capital) induce a selection bias. Double differencing relies on the hypothesis that such variables have a constant effect over time.

Such methods have already been used to evaluate farm advisory service policies (Godtland et al.; 2004, Van den Berg and Jiggins, 2007; Davis et al., 2012). But to ensure the empirical reliability of this kind of work, methodological precautions must be taken which may limit the scope of findings. Below are four examples related to farm advisory service programs.

1) The first problem bears on the requirement for a random distribution of farmers who benefited from these programs of advisory services and those who did not (in the case of RCTs). Aside from the ethical issues raised, this requirement is also contrary to the diagrams of causality of certain programs, such as participative and *bottom-up* interventions (e.g. farmer field schools): the effectiveness of such programs theoretically depends on the self-motivated participation of farmers in a collective project.

2) The second problem bears on an essential hypothesis of these methodologies of impact evaluation based on RCT or semi-experimental evaluation: beneficiaries must not be influenced by the fact that non-beneficiaries do not benefit from the program, and vice versa (Stable Unit Treatment Value Assumption – SUTVA). This hypothesis may also be contrary to the diagrams of causality underlying certain advisory service programs, particularly those built on so-called diffusionist models (e.g. the World Bank's Train & Visit program): in theory, their effectiveness resides in the fact that farmers who directly receive advice will share acquired knowledge with those who have not.

3) The third problem is the choice of the indicators. Evaluating the impact of farm advisory services supposes the ability to identify a proxy of the expected results. At which level shall this result be selected (Van den Berg and Jiggins, 2007)? The level of farm performance (yield, income, etc.); the level of the adoption of innovations; or the level most directly affected by farm advisory services: farmers' knowledge and skills? The question then becomes how to express this knowledge and these skills in quantitative variables. In that respect, Godtland et al. (2004) have stressed the difficulties and limitations of their attempt to express farmers' knowledge through knowledge tests. Likewise, the effects of this proxy will have to be observable over relatively short durations (due to costs, RCTs are often used in one- to two-year population tests). However, in the case of farm advisory services, one can wonder whether this short-term measure makes any sense due to certain mid- or long-term dimensions of learning processes.

4) The last aspect is related to the distributive effects of the evaluated policy. In most impact studies, the effect is calculated by looking at the difference between the average obtained by the group of individuals benefiting from the measure in a sample and that of the individuals who do not benefit. However, an average improvement for the target population can hide great inequalities or even aggravate these inequalities. Abadie et al. (2002) have shown for instance that a training program for poor populations could result in an increase in the average income of the target populations, but have no effect on the poorest fraction of this population.

This example of the evaluation of farm advisory services shows that the measurement of the impact of public programs is only rigorous if the methods used are consistent with specific hypotheses associated with the method of data collection (randomization, a lack of diffusion-related effects, etc.).

The evidence issue when assessing effectiveness

Fully understanding the significance and limitations of these approaches is only possible if we accept that they are designed to obtain the highest possible level of evidence of difference-making (effectiveness or harmlessness) – *and only this.*

1) Obtaining high level evidence of difference-making may seem simple or even simplistic. It is in fact quite challenging and involves costly practices which pose significant

methodological and ethical problems. Nevertheless, it is the only way to obtain rigorous evidence of the actual impact of a public action program.

2) The significance of evidence of difference-making should not be overestimated; it does not indicate which mechanisms rendered program action effective; often, several competing explanations emerge concerning the effectiveness of a program.

3) Such results are therefore of limited interest when deciding to extend a public action program to other contexts or periods. This should be done using methods which provide the most reliable hypotheses possible regarding the *mechanisms* that make action effective.

In other words, the experimental settings of the production of evidence of effectiveness are such that they cause many problems of generality and external validity of the knowledge that they contribute to build. This knowledge is only valid for a specific population φ in a particular environment characterized by a specific causal structure $CS^{t'}(O^t)$. And it can only be extended to populations θ that share the same causal structure $CS^{t'}(O^t)$. Some authors propose to solve this "environmental dependence" issue by replicating measures of effectiveness (with RCT) in various contexts, but "*worry that there is little incentive in the system to carry out replication studies (because journals may not be as willing to publish the fifth experiment on a given topic as the first one), and funding agencies may not be willing to fund them either* (Banerjee and Duflo 2009, p. 161). But the problem is not a financial one. In any case, replication alone cannot be a solution; a theory about causal structures is necessary to identify the scale and boundaries of different θ populations that may share a same causal structure. Therefore, it seems necessary to rely on theories generating evidence of mechanism to characterize the causal structure of the target populations of the policies.

To understand. Type of evidence for evaluations aimed at identifying and analyzing the mechanisms by which a program can produce expected results or adverse effects

Many authors have pointed out the importance of basing evaluation on a theory, on a precise understanding of the mechanisms operating in the programs being studied (Chen and Rossi, 1983; Chen, 1990; Shadish et al.; 1991; Pawson and Tilley.; 1997, Pawson, 2002.; Jordan et al., 2008). Their initial acknowledgement underlines the limitations of the two other types of evaluation described in this paper. They insist on the fact that these evaluations, which rely on evidence of effectiveness or on gathering opinions cannot reveal in a reliable way the causal structure - $CS^{t'}(O^t)$ - that explain why a program works or not in a given context, and why it may have different impacts on the various elements of the target population.

Such evaluations focus on understanding (i) the object which is evaluated (ii) the mechanisms of action to be 'revealed' through the analysis and (iii) the context in which the program is implemented. By analyzing, in different contexts, the way in which the impacts are produced, regularities or recurring facts are identified so as to determine the various causes $\{C_1^{t'},\ldots,C_n^{t'}\}$ and the set of causal relations $\{Prob(O^t/L_1^{t'}),\ldots,Prob(O^t/L_m^{t'})\}$ by which the implementation of a public program has expected or unexpected effects. These effects can directly relate to the goal of the program or to its broader context. The evaluation will thus depend on the nature of the problem in question: what is at stake are the specificities of this problem in a particular context and the assessment of the degree of genericity of the proposed solutions for further action.

In certain cases, to improve the quality of the measurement of impacts, the evaluation is constructed using a preliminary analysis of the theory underlying the program (program

theory). A first step is understanding (before the measurement) the causal mechanisms that guided the design of the program. The role of the evaluator consists, more precisely, in putting forth hypotheses on the main features of the causal structure linking a program and its potential subsequent effects. The aim is to build a diagram which traces these patterns of causality and constitutes the theory of the program and is a simplified representation of the comprehensive causal structure. When it is established, such a diagram becomes a reference framework and the basis of the evaluation approach for the evaluator, who proposes indicators that will be useful for measuring impacts.

The analysis of the causal structure of the program allows a better understanding of the distributive effects of a program within the target population and across populations. However, the diagram that is built is *only* a simplified representation of the proposed causal structure. Therefore some of the ways in which evidence on mechanisms is used in the evaluation process raises questions, as we shall see in the following example.

Example: environmental evaluation and coupling between economic models and sustainability indicators

Many public programs aim at encouraging farmers to adopt practices which guarantee better environmental performances (biodiversity conservation, water quality, etc.). This is done by delivering specific financial support or by making changes in farm practices a prerequisite to receiving existing forms of aid (e.g. agri-environmental schemes, cross-compliance for the Common Agricultural Policy in the European Union).

The procedures used to evaluate the environmental impact of these programs almost never rely on the production of evidence of effectiveness, as seen in Kleinj and Sutherland's review on biodiversity (Kleinj and Sutherland, 2003). Measuring the effectiveness of a program for biodiversity conservation would indeed require collecting ecological data according to a specific and elaborate methodological framework (with the possibility of building counterfactuals so as to measure impacts specifically linked to the program). Such methodological frameworks are costly and often regarded as inaccessible. For this reason, many evaluations rely on the diagram of causality at the origin of the public program (an economic incentive A, must cause a change of agricultural practice B, which has an ecological impact C), and make the assumption that if the means were in fact implemented, then the program was effective. Evaluations then focus on measuring B, i.e. the number of farmers who have actually changed their practices. Such approaches have been referred to as the measurement of "policy performances" (Primdahl et al., 2003). In certain cases this information is considered sufficient to draw conclusions on the environmental impact of the program. In other cases, the 'black box' of these changes is opened and additional data are collected (about crop rotation, plant pest management, etc). They are linked to agri-ecological indicators to calculate the potential risks and effects of these changes (for example the use of less chemical inputs is associated with a positive impact on biodiversity) (Mitchell et al.; 1995; Van de Werf and Petit, 2002).

However, it is impossible to identify and take into account the many existing mechanisms that interact in various contexts. Thus, the causal diagram which underlies these actions is only an approximation of a comprehensive causal structure that ideally could allow their effect to be fully predicted. The research articles which examine these types of methods all point out that these measures identify 'potential effects' but fail to measure actual impacts. Nevertheless, these precautions are often absent in the executive summaries of reports which present the results of these evaluations. Variations in the value of an indicator can thus be

presented as evidence of an improvement of environmental performances. This is not only improper from a formal point of view; the few experimental tests carried out on this issue also disprove that it is an acceptable estimate. For instance Kleinj and Sutherland (2003) and Kleinj et al. (2006) show that certain measures which were successful in terms of "policy performance" did not have the expected environmental impact.

Such doubts about the effectiveness of certain agri-environmental schemes can be linked to the weakness of the theoretical models upon which they are based, as well as to a lack of empirical data with which to identify what works and what does not (McNeely et al., 2005). The work done on the eco-millennium assessment demonstrated the importance of these knowledge gaps (Carpenter et al., 2006). This concerns both evidence of difference-making and evidence of mechanism.

The evidence issue when analysing mechanisms

Recourse to evidence of mechanism in evaluation procedures thus takes two principal forms: to produce such evidence to reveal in detail the mechanisms behind the phenomena observed to analyse the way in which it interacts in the theory of the program, in order to structure the evaluation consequently.

1) Identifying the mechanisms by which the actions were effective (or not) is essential to producing generic knowledge that can be used to develop new programs (e.g. a causal relation which can be exploited in various contexts). It can also help assess the genericity of the knowledge used in the program (e.g. to what extent the causal structure of two different populations can be considered similar?) and to raise new issues for the evaluators and stakeholders involved in the evaluation.

2) The issue of level of evidence comes into play for this type of evidence as well. As for evidence of effectiveness, it makes sense to rank results based on the opinions of respected authorities, single case studies, observations on wider samples of situations, etc. in order to assess the robustness of available evidence. However, the use of theoretical models to infer the effective impact of a program, as sophisticated as they may be, is often limited; the causality diagrams formalized in these theoretical models are only ever partial representations of complex causal structures. Their predictive capacities vary according to the object under evaluation and the context; therefore one cannot replace the observation of the real effects (and the production of evidence of effectiveness) by that of expected effects (estimated using an analysis of the means implemented in the program).

3) As mentioned before, under certain conditions, evidence of mechanism can be combined with evidence of difference making to highlight a causal pathway and it would be misleading to associate causality only with one type of evidence.

Learning: Evidence for evaluations primarily designed as a collective learning process

As discussed briefly in section 1, an abundant amount of literature based on different theoretical points of view has shown the importance of associating stakeholders with the evaluation, in order to improve the theories guiding the evaluator's work, improve the quality of the evaluation and allow a more sensible use of results by different stakeholders (Cousin and Whitmore, 1998; Mertens 1999; Fetterman and Wandersman, 2005). Ultimately, these approaches do not call into question the need to review types of evidence (mechanisms, effectiveness, adverse effects) even if they do give rise to new debates, notably concerning the questions for which evidence should be produced and concerning the ways in which data must be collected and interpreted.

However, there are also certain evaluation procedures primarily aimed at promoting consensus through close collaboration between different stakeholders, right from the evaluation design stage, in order to build awareness and encourage new practices, the latter taking precedence over the measurement of a program's outcomes.

Evaluation methods which highlight the educational dimension of evaluation procedures are one such example. These methods "bring to the table" all stakeholders who have a vested interest in the improvement of the program under evaluation. The person in charge of the evaluation begins by drafting as accurately as possible a sociogram of the network of stakeholders which includes information about the nature and intensity of the ties between these actors. The evaluator uses this representation of actor networks to conduct in-depth interviews with stakeholders to gather each person's point of view and suggest ways of improving the program. At each stage of the evaluation, partial conclusions are discussed and analyzed in working groups. In certain cases – service-related programs, for example – evaluators constitute a representative sample of service users. It is the users themselves who then assess the value of the program (after a specific training session).

Here, the heart of the evaluation method is the contributions of program stakeholders to a social construction of representations of an observed reality. While the process may simply mobilize opinions, it also calls upon scientific knowledge (in the field of natural sciences, primarily), often through the tools proposed by researchers (e.g. simulation models). The reliability of evidence used for collective learning is not frequently addressed although it sometimes generates debates (e.g. Van der Sluijs et al., 2008).

In this type of evaluation procedure the evaluator's role is to organize debates, ultimately to obtain the most consensual possible results which can then be used by the largest number of people. These approaches have become highly popular in recent years and take on different forms. They are used for various issues involving collective action (water management, land-use planning) and rely on different methods to promote interaction among actors during the evaluation phase (role playing, multi-agent-based simulation, etc.).

Implementation: the Soft System Methodology (SSM) example

An emblematic example of this type of method can be found in evaluations of public programs offering farm advisory services. A notable example is the relatively widespread use of Soft System Methodology (SSM) to design and evaluate technical advisory programs (Rohs and Navarro, 2008). SSM is designed to help a "human activity systems" (HAS) make the most effective decisions in uncertain and complex contexts (Checkland, 1981) where learning is the priority. Checkland and Scholes (1990) point out that SSM as a model is not intended to establish versions of reality. Instead, it aims to facilitate debate so that collective decisions and action can be taken in problem situations. The seven stages of SSM are (Checkland 1981): i) inquiring into the situation (identifying the problem using different communication techniques: brainstorming, interviews, participant observation, focus groups, etc.); ii) describing the situation (describing the context using a wide variety of sources); iii) defining HAS (identifying program stakeholders, and interviewing them on the transformations they are expecting); iv) building conceptual models of HAS (representing the relationships between stakeholders in the program being designed or evaluated); v) comparing the conceptual models with the real world (preparation of a presentation of the model for a debate with stakeholders); vi) defining desirable and feasible changes; vii) implementation (Rohs and Navarro, 2008).

Corroboration with facts and producing evidence with the best possible level do not appear to be at the heart of this conception/evaluation approach, which instead aims at promoting and structuring debate between program stakeholders to arrive at a consensual solution. In practice, however, significant problems arise (Salner, 2000). In workshops, for example, evidence is provided by different stakeholders verbally, and must be verified. Salner (2000) likens this method to journalism, in that it involves the verification of the opinions of different stakeholders so that "*analysis makes it possible to mount an argument for change which was not simply an intuitive reaction to a conversation held; it was an argument which could be explicitly retraced at any time with links to supporting evidence*" (Checkland and Sholes, 1990: 198-99). Verification is thought to be guaranteed by the open, public and collective nature of the debate. Comparison with 'fact checking' in journalism, however, only holds true if the evidence presented is evidence of presence describing facts known through stakeholder practices. Instead, arguments often go deeper and target the expected or measured impact of programs and even the causality diagram upon which they are based. In other words, these evaluation methods rely not only on evidence of presence but also on evidence of effectiveness and mechanisms but do not formalize this integration. This lack of formalization manifests itself on two levels: (i) in the use of scientific knowledge to formulate hypotheses on the modalities of how public programs function, (ii) in the verification of the level of evidence obtained.

Ultimately, these formalization tasks are implicitly transferred to workshop leaders (often researchers). This situation poses a number of problems as it is assumed that these leaders have extensive skills and means at their disposal (to produce state-of-the-art of available scientific literature, statistical analyses and various types of verifications). For this reason, several authors have pointed out that SSM may be exploited to reinforce a balance of power given the asymmetries of information between stakeholders: "*the kind of open, participative debate that is essential for the success of the soft system approach, and is the only justification for the result obtained, is impossible to obtain in problem situations where there is a fundamental conflict between interest groups that have access to unequal power resources. Soft system thinking either has to walk away from these problem situations, or it has to fly in the face of its own philosophical principles and acquiesce in proposed changes emerging from limited debates characterized by distorted communication*" (Jackson, 1991: 198).

The evidence issue in collective learning

These approaches raise several questions where the issue of evidence is concerned.

1) The issue of level of evidence is often neglected and seen as secondary to collective learning objectives. All contributions are accepted equally and the reliability of evidence is not subject to systematic testing procedures;

2) Very quickly, evidence presented by participants with different interests can be in competition and arbitration is often based on non-transparent criteria;

3) Without a systematic, clear verification procedure for evidence brought to the debate, learning may focus more on the ability to reach consensual positions than on the ability to use the best tools for achieving a given objective and on evaluating outcomes in a rigorous manner.

Conclusion

This article is not intended as a standard-setting tool. Our goal is to contribute build a theory of evidence for evaluation that allows different stakeholders to better judge the quality of evidence they seek depending on their project.

We have shown that while evaluation may have very different objectives (e.g. understanding the mechanisms of public programs, measuring their specific impacts, or supporting collective learning to favour the emergence of an agreement between stakeholders in the programs), each objective leads to a different examination of the question of types of evidence, i.e. what is object of evidence (presence, making a difference, mechanism). This concern must be clearly distinguished from the study of levels of evidence, which deals with data collection and interpretation (single case observations, difference methods, RCT...); each of these methods can be used for producing each type of evidence.

With this in mind, the issue of RCTs must be re-examined, along with the types of evidence for which these methods are used. Experimental economics can be used as a tool to test some hypotheses on mechanisms rather than only be used to assess the impact of a policy in a given environment. Nevertheless, whether RCTs are a relevant tool in that respect is a matter of ongoing discussion both in medical sciences and in economics (Deaton, 2009). A key question in this debate is the importance of heterogeneity and distributive effects across populations, which are not acknowledged by RCTs, but which can be essential for formulating theories in various scientific areas (economics, management science, but also bio-medical sciences and ecology among others).

For each situation, the quality of evidence can be assessed according to three dimensions. Ideally, as mentioned above, one would like to base their decision on evidence that is both socially relevant (addresses phenomena considered by each stakeholder to be important), of a high level (with probative force) and which corresponds to the adequate type for the goals of the evaluation. This ideal is usually inaccessible, for reasons of cost, methodological impossibilities, necessity to select very precise objectives from a large number of possible points of view, etc.

Thus evaluators are permanently confronted with trade-offs. The three examples above show that a better understanding of quality of evidence can help better assess the limits inherent to the conclusions of every kind of evaluation depending on the quality of evidence on which they are based. In the real world, every evaluation process has its own limits and can only produce reliable results for a limited field of interest. Choices should thus be made that will involve institutional issues and possible conflicts of interest. As is the case with any public policy instrument, the final decision depends on a multiplicity of factors which cannot be reduced to a set of known evidence. But a clear specification of the limits of validity of the findings of each evaluation process is thus a prerequisite to avoid misinterpretations. A better shared knowledge of the type and the level of evidence that is used to evaluate the result of interventions can help clarify for various stakeholders what is at stake in making alternative choices.

Acknowledgements

The authors would like to thank the anonymous referees and editors who provided useful and inspiring comments on an earlier version of this article.

Funding and biographical information

This research was conducted in an interdisciplinary research program funded by the French National Agency for Research (program EBP-Biosoc /ADD). It is based on the combination of former research experience on evaluation theories (M. Berriet-Solliec), on international debates on evaluation of farm-advisory services (P. Labarthe) and on quality of evidence (C. Laurent).

References

Abadie, A., J. Angrist, G. Imbens. 2002. Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica* 70 91–117.

Adams W.M., Avelling R., Brockington D., Dickson B., Elliot J., Hutton J., Roe D., Vira B., Wolmer W. 2004. Biodiversity conservation and the eradication of poverty. *Science.* 306, 1147-1149

Banerjee, A.V., E. Duflo. 2009. The Experimental Approach to Development Economics. *The Annual review of economics.* 2009(1) 151-178.

Bro, E., P. Mayot, E. Corda, F. Reitz. 2004. Impact of habitat management on grey partridge populations: assessing wildlife cover using a multisite BACI experiment. *J. Appl. Ecol*. 41 846–857.

Carpenter, S., R. DeFries, T. Dietz, H. Mooney, S. Polasky, W. Reids, R. Scholes. 2006. Millenium Ecosystem assessment : research needs. *Science*. 314 257-258.

Cartwright, N. 2011.Evidence, External Validity and Explanatory Relevance. G.J. Morgan, ed. *Philosophy of Science Matters: The Philosophy of Peter Achinstein*. Oxford University Press, New York, NY.15-28.

Cartwright N, Hardie J, 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*.Oxford University Press.

Checkland, P.B. 1981. *System thinking, system practice*. John Wiley, New-York, NY.

Checkland, P.B., J. Scholes. 1990. *Soft systems methodology in action*. John Wiley & sons, Chister, GB.

Chen, H.T. 1990. *Theory-Driven Evaluation*. Sage, Newbury, CA.

Chen, H.T. and Rossi, P.H. 1983. Evaluating with sense. The theory-driven approach, *Evaluating review*, 7, (3): 283-302.

Cousins, J.B. & Whitmore, E. 1998, Understanding and participatory evaluation**.** *New Directions for Evaluation*, San Francisco, CA :Jossey-Bass.(80), 69-80,

Davis, K.E. 2008. Extension in Sub-Saharan Africa: Overview and Assessment of Past and Current Models, and Future Prospects, 15, (3), 15-28.

Davis, K., E. Nkonya, E. Kato, D.A. Mekonnen, M. Odendo, R. Miiro, J. Nkuba. 2012. Impact of Farmer Field Schools on Agricultural Productivity and Poverty in East Africa, *World Development*, 40(2) 402–413.

Deaton AS 2009. Randomization in the tropics and the search for the elusive keys to economic development. *National Bureau of Economic Search Working Paper:* 14690. Cambridge, M.A. USA. 54 p..

Donaldson, S.I. 2007.*Program theory-driven evaluation science: Strategies and applications*, Routledge.

Donaldson, S.I 2008 *In search of the blueprint for evidence-based global society. in* Donaldson, S.I. Christie C.A. Mark, H.H. 2008. *What counts as credible evidence in evaluation and evidence-based practice?* Thousand oaks, CA: Sage.2-18

Donaldson, S.I. Christie C.A. Mark, H.H. 2008. *What counts as credible evidence in evaluation and evidence-based practice?* Thousand oaks, CA: Sage

Duflo, E., M. Kremer. 2005. Use of Randomization in the Evaluation of Development Effectiveness. G. Pitman, O. Feinstein, G. Ingram, eds. *Evaluating Development Effectiveness*. Transaction Publishers, New Brunswick, NJ, 205-232.

Fetterman, D.M. and Wandersman, A. 2005. *Empowerment evaluation.Principles and Practice*. New York: The Guilford press.

Fitzpatrick, J. L. ; Sanders, J.R., Worthen, B.R. 2011. *Program Evaluation: Alternative Approaches and Practical Guidelines*. Pearson Evaluation.

Godtland, E.M., E. Sadoulet, A. de Janvry, R. Murgai, O. Ortiz. 2004. The Impact of Farmer Field Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes. *Econ. Dev. Cult. Change* 53 (1) 63-92.

Hansen, H.F., Rieper, O. 2009. "The Evidence Movement: The Development and Consequences of Methodologies in Review Practices", Evaluation, vol. 15: pp. 141 - 163.

Illari P. MK (2011) Mechanistic evidence: Disambiguating the Russo-Williamson *Thesis, International Studies in the Philosophy of Science*, 25: 2, 139-157.

Jackson, M. 1991. *Systems methodology for the management sciences*. Plenum Press, New York and London

Jordan, G. B., Hage, J., & Mote, J. 2008. "A Theories-Based Systemic Framework for Evaluating Diverse Portfolios of Scientific Work, Part 1: Micro and Meso Indicators". *New Directions for Evaluation*, (118), 7–24., San Francisco, CA :Jossey-Bass.

Kleinj, D., W. Sutherland 2003. W. How effective are European agri-environment schemes in conserving and promoting biodiversity? *J Appl. Ecol.* 40: 947–969

Kleinj, D., R. A. Baquero, Y. Clough, M. Díaz, J. Esteban, F. Fernández, D. Gabriel, F. Herzog, A. Holzschuh, R. Jöhl, E. Knop, A. Kruess, E. J. P. Marshall, I. Steffan-Dewenter, T. Tscharntke, J. Verhulst, T. M. West, J. L. Yela. 2006. Mixed biodiversity benefits of agri-environment schemes in five European countries. *EcolLett* 9 (3) 243-254.

Labarthe, P., C. Laurent, 2013. Privatization of agricultural extension services in the EU: Towards a lack of adequate knowledge for small-scale farms? *Food Policy*, 38, 240-252.

Laurent, C., A. Trouvé. 2011. Competition of evidences and the emergence of the "evidence-based" or « evidence-aware » policies in agriculture. *122nd EAAE Seminar "evidence-based agricultural and rural policy making: methodological and empirical challenges of policy evaluation"*. Ancona, Italy, February 17-18, 2011.

Laurent, C., M. Berriet-Solliec, M. Kirsch, D. Perraud, B. Tinel, A. Trouvé, N. Allsopp, P. Bonnafous, F. Burel, M.-J. Carneiro, C. Giraud, P. Labarthe, F. Matose, A. Ricroch. 2009. Pourquoi s'intéresser à la notion d'Evidence-basedpolicy ? *Revue Tiers-monde,* 200, 853-873.

Liberati A., Buzzetti R., Grilli R., Magrini N., Monozzi S. 2001. Evidence-based case review. Which guidelines can we trust? Assessing strength of evidence behind recommendations for cilinical practice. *West J Med*. 174, 262-265.

Lipsey, 2007. Method choice for government evaluation. In G. Julnes& D. J. Rog (Eds.) Building the evidence base for methods choice in government sponsored eval. *New Directions for Evaluation*, No. 113, San Francisco, CA :Jossey-Bass.

Mac Neely, J.A. *et al*. 2005. Ecosystems and Human Well-Being, vol 3, Policy responses, Chopra, K., Leemans, R., Kumar, P., Simons, H., Eds (Millenium assessment, island press, Washington DC), 119-172.

Mayne, J. 2012. Contribution analysis: coming of age? *Evaluation*, 18: 270-280

Mertens, D. 1999.Inclusive evaluation: Implications of transformative Theory of Evaluation. *American Journal of Evaluation.* Vol. 20, 1, 1-14

Mitchell, G., A. May, A. McDonald. 1995. PICABUE: a methodological framework for the development of indicators of sustainable development. *Int. J. Sust. Dev. World* 2 104-123.

Oliver, S.; Harden, A.; Rees, R.; Shepherd, J.; Brunton, J.; Garcia, J. and Oakley, A. 2005. An Emerging Framework for Including Different Types of Evidence in Systematic Reviews for Public Policy", *Evaluation*, 11: 428 - 446.

Patton, M. Q. 2008. *Utilization focused evaluation*, 4th Edition. Sage, Thousand Oaks, CA.

Pawson, R., N. Tilley. 1997. *Realistic Evaluation*. Sage, London, UK.

Pawson, R. 2002. "Evidence-based Policy: In Search of a Method", Evaluation, Apr 2002; vol. 8: pp. 157 - 181.

Pawson, R. 2006. *Evidence-based policy: a realistic perspective*. Sage, London, UK.

Petticrews M., Roberts H. 2003. Evidence, hierarchies and typologies: horses for courses. *J Epidemiol Community Health*, 57, 527-529

Primdahl, J., B. Peco, J. Schramek, E. Anderse, J.J. Onate. 2003. Environmental effects of agri-environmental schemes in Western Europe. *J. Environ. Manage*. 67 129–138

Rochs, F., M. Navarro. 2008. Soft System Methodology: an intervention strategy. *Journal of International Agricultural and extension education* 15 (3) 95-99.

Rogers P. 2008 "Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation* 14 (1): 29-48

Rossi, P.H. Lipsey,M.W. & Freeman, H.E. 2004. *Evaluation: a systematic approach*. 7th Edition.Sage, Newbury Park, CA.

Salner, M. 2000. Beyond Checkland& Scholes: Improving SSM. *Occasional Papers on Systemic Development* 11 23-44.

Schwandt, T. 2003. Back to the Rough Ground! Beyond Theory to Practice in Evaluation. *Evaluation, 9* (3), 353–364.

Schwandt, T. 2008. "Toward a practical Theory of evidence for evaluation" in Donaldson, S.I. Christie C.A. Mark, H.H. *What counts as credible evidence in evaluation and evidence-based practice ?*Thousand oaks, CA: Sage, pp 197-212. .

Shadish, W.R., T.D. Cook, L.C. Leviton. 1991. *Foundations of Program Evaluation Theories of Practice.*Sage, Newbury Park, CA.

Shadish WR, Cook TD and Campbell DT, 2002.*Experimental and Quasi Experimental Designs for Generalized Causal Inference.*Boston, New York: Houghton Mifflin Company.

Stame N. 2004. Theory-based evaluation and varieties of complexity. Evaluation 10 (1):58-76

Stern, E. 2004. Philosophies and types of evaluation research. In *The foundations of evaluation and impact research.* Third report on vocational training research in Europe: background report. Luxembourg: Office for Official Publications of the European Communities (Cedefop Reference Serie, 58), 12-42

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R.; Befani, B. 2012. *Broadening the range of designs and methods for impact evaluations*. DFID Working Paper 38, London.

Stufflebeam, D.L. 2001.*"Evaluation models", New directions for evaluation, 89*. San Francisco, CA :Jossey-Bass.

Van den Berg, H., J. Jiggins. 2007. Investing in farmers. The impact of farmer field schools in relation to Integrated Pest Management, *World development*, 35(4), 663-687.

Van der Sluijs, J., J.-M. Douguet, M. O'Connor, A. GuimaraesPeriera, S.C. Quintana, L. Maxim, J. Ravetz. 2008. Qualité de la connaissance dans un processus délibératif. *Nature, Science, Société* 16 265-73.

Van der Werf, H., J. Petit 2002. Evaluation of the environmental impact of agriculture at the farm level: a comparison and analysis of 12 indicators-based method. *Agr. Ecosyst. Environ.* 93 131-145.