Dealing with COMPLEXITY in DEVELOPMENT EVALUATION

A Practical Approach

Michael Bamberger | Jos Vaessen | Estelle Raimondo
Editors

CHAPTER 4

Impact Evaluation Approaches and Complexity

Jos Vaessen, Estelle Raimondo, and Michael Bamberger

In Chapter 1 we introduced the book's conceptual framework for complexity. One of the key dimensions of the framework concerns the nature of causal change. In (development) evaluation, the field of evaluation approaches that specifically deals with causal change is impact evaluation. Impact evaluation looks at the changes in society and the extent to which they are attributable to an intervention, also taking into account other factors. In practice, a number of questions relate to the broader question of impact. In turn, divergent methodological designs are available that are equipped to deal with one or more of these questions. This chapter presents the most prevalent impact evaluation approaches used in development evaluation practice. Subsequently, the strengths and limitations of these approaches are discussed in terms of how they address a number of key complexity issues. The discussion is illustrated with a case study.

1. Key Questions Addressed in Development Impact Evaluations

One of the main purposes of impact evaluation[1] is to assess the extent to which changes in society that the program was designed to influence can be attributed to the program. Addressing this question requires a method of causal inference that seeks to connect causes with effects (outcomes, impacts). In recent years, there has been an extensive debate in international development on this issue (see Cohen & Easterly, 2009). While advocates of certain approaches, particularly randomized controlled trials (RCTs), continue to argue that a certain method or set of methods is the best (the "gold standard"), it is now generally acknowledged that there are a number of different approaches to assess causality. Donaldson, Christie, and Mark (2009), in *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* present a range of approaches that are widely used and considered credible by different disciplines and audiences. Several authors argue that evaluators continue to be narrowly focused on the merits and/or limitations of experimental designs as the appropriate standard for evidence-based evaluation and related debates on quantitative versus qualitative approaches. As a consequence very little serious attention has been given to a wide range of potentially useful research approaches that are used in other branches of the social and physical sciences. Scriven (2009) presents a number of alternative ways to think about causality, while Greene (2009) argues that what some researchers consider as "proof" of causal relations should more modestly be considered as "inklings." Rieper, Leeuw, and Ling (2010) also argue that while there is broad acceptance of the general movement toward evidence-based policy, disciplines differ as to what is considered appropriate evidence. White and Phillips (2012) discuss a range of qualitative methods that are particularly pertinent in the case of evaluating the impact of small "*n*" interventions, that is, those with small target groups for which statistical analysis is not feasible (e.g., the impact of capacity development initiatives on the quality of policy formulation in educational planning units of ministries of education).

Stern et al. (2012, pp. 36–37) identify four impact-related questions of interest to policymakers:

- To what extent can a specific (net) impact be attributed to the intervention?
- Did the intervention make a difference?
- How has the intervention made a difference?
- Will the intervention work elsewhere?

Each of these questions usually requires a different evaluation design, and a design that works well to address one question may not be appropriate for a different question. It is important to ensure that the evaluation design is driven by the questions being asked (issues driven) and not by the researcher's preference for a particular methodology (methods driven). In addition to the evaluation questions guiding the evaluation design, the characteristics of the intervention are another important factor that should inform the design (Stern et al., 2012).

2. Established Evaluation Approaches in the Context of Impact Evaluation

2.1 Overview

There is an extensive literature available on different methods for impact evaluation in the context of international development (see, e.g., Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011; Khandker, Koolwal, & Samad, 2009; Leeuw & Vaessen, 2009; Stern et al., 2012). Table 4.1 provides an overview of the main approaches to impact evaluation based on a recent study commissioned by the

Table 4.1 Main Approaches to Impact Evaluation

Design approach	Specific variants	Basis for causal inference
Experimental	RCTs, quasi-experiments, natural experiments	Counterfactuals, the copresence of cause and effects
Statistical	Statistical modeling, longitudinal studies, econometrics	Correlation between cause and effect or between variables, influence of (usually) isolatable multiple causes on a single effect, control for confounders
Theory-based	*Causal process designs*: Theory of change, process tracing, contribution analysis, impact pathways *Causal mechanism designs*: Realist evaluation, congruence analysis	Identification/confirmation of causal processes or chains Supporting factors and mechanisms at work in context
Case-based	*Interpretative:* Naturalistic, grounded theory, ethnography *Structured*: Configuration, process tracing, congruence analysis, QCA, within-case analysis, simulations and network analysis	Comparison across and within cases of combinations of causal factors Analytic generalization based on theory
Participatory	*Normative designs*: Participatory or democratic evaluation, empowerment evaluation *Agency designs*: Learning by doing, policy dialogue, collaborative action research	Validation by participants that their actions and experienced effects are caused by the program Adoption, customization and commitment to a goal
Review and synthesis	Meta-analysis, narrative synthesis, realist synthesis	Accumulation and aggregation within a number of perspectives (statistical, theory-based, ethnographic, etc.)

SOURCE: Adapted from Stern et al. (2012).

Department for International Development. Needless to say, underlying each of these types of approaches is a multitude of specific data collection and analysis techniques such as surveys, focus groups, participant observation, and so on, which will not be discussed here (see, e.g., De Leeuw, Hox, & Dillman, 2008; Mikkelsen, 2005).

In the remainder of this section we discuss each of these approaches. For each category of approaches, we provide some examples of particular methods. For a more comprehensive discussion of these methods, see, for example, Stern et al. (2012) for an overview, Khandker et al. (2009) and Gertler et al. (2011) on quantitative impact evaluation approaches, Funnell and Rogers (2011) on theory-based evaluation approaches, Byrne and Ragin (2009) on case-based evaluation approaches, Cousins and Whitmore (1998) on participatory evaluation approaches, and Popay (2006) on review and synthesis approaches. Further references on prevalent methods under each of the approaches can be found in the discussion below.

Given the pivotal role of theory-based evaluation in the context of complexity-responsive evaluation, this book includes two chapters on the topic (Chapters 5 and 6). In addition, Chapter 9 is devoted to different approaches to review and synthesis. Consequently, the discussion on this topic in this chapter will be limited, referring the reader to Chapter 9.

2.2 Experimental, Quasi-experimental, and Non-experimental Quantitative Approaches

In this section we discuss quantitative impact evaluation approaches (rows 1 and 2 in Table 4.1). Experimental and quasi-experimental approaches are based on the principle of counterfactual analysis. In various ways they try compare what has happened during the intervention with what would have happened without the intervention. Non-experimental approaches try to capture the effect of an intervention with the help of statistical controls. For example, with the help of multiple regression analysis one can estimate the effect of an intervention variable (which can be dichotomous or continuous) on a dependent variable controlling for all other relevant variables in the regression equation (statistical controls). Because many quasi-experimental techniques also use statistical modeling, we do not discuss non-experimental quantitative approaches separately.

a. Experimental Approaches

For the purpose of defining the evaluation design, the basic causal question can be reformulated as "What would have happened without the intervention?" The conventional way to address this question is to compare the observed world with a theoretical world where the program intervention did not occur. This process is sometimes called a *thought experiment* as it is not possible to observe this theoretical world directly. The established evaluation approach is defining the counterfactual through an *experimental* or *quasi-experimental* design. The experimental approach randomly assigns subjects to the treatment and control groups. If the experiment is well designed, this eliminates (controls for) all factors other than the experimental treatment, and if a statistically significant difference is found between the two groups after the treatment has been administered, this provides initial evidence that the program treatment has contributed to the observed effects. Ideally, the experiment should be repeated several times to determine if the results are robust when replicated in a similar setting or under different conditions. However, in the real world, due to budget and time pressures, decisions about program effectiveness are often based on a single test.

The two most common variations of randomized evaluation designs are (1) the "intention to treat," which compares outcomes for all subjects in the treatment group, some of whom may not participate in the program, with those assigned to the control group; and (2) "treatment on the treated," which compares subjects who actually received the treatment with those who did not (Khandker et al., 2009).

The design is generally considered to be the strongest quantitative evaluation design with respect to attribution in situations to which it can be applied. As Woolcock (2013) points out, development programs with low causal density (few causal pathways) are well suited to RCTs. In this type of intervention, one can expect that the impact can be isolated and studied in conjunction with experimentation on slight variations of the intervention (e.g., different grant sizes for small and medium enterprise investment). For this type of intervention, repeated RCTs can bring us closer to a proof of concept. When the design is properly implemented and the sample is sufficiently large, statistically robust and unbiased estimates of the magnitude of outcomes that can be attributed to the intervention are obtained. The achievement of unbiased estimates is the major benefit of this design as almost every other design is subject to potential selection bias, which affects the validity of the attribution analysis. The rapidly growing body of RCTs means that precise evaluation design protocols now exist for many sectors. It is also possible for most sectors to conduct systematic reviews (see Chapter 9) of significant numbers of studies that have been conducted. The growing interest in RCTs has also challenged evaluators using other designs to assess the potential methodological weaknesses of their approaches and to pay greater attention to evaluation design and threats to validity (see Cook & Campbell, 1979).

RCTs also have a number of general limitations (see, e.g., Cook & Campbell, 1979; Bamberger & White, 2007). First of all, the counterfactual answers only *setting-specific questions* (e.g., Did it work here, for this particular group?) and cannot generalize to other settings (low external validity). Second, the design analyzes only linkages between intervention outputs (causes) and outcomes (effects) and does not examine processes (what happens between intervention outputs and outcomes). It does not explain how the outcomes are achieved or how and why the assumed causes contributed to the outcomes. Third, there are serious constraints to applicability. RCTs work better for certain kinds of interventions and in some kinds of project settings than for others. There are also many constraints on when randomization can be applied. Fourth, the interpretation of findings is complicated by *early preemption* (things that happened before the effects) and *late preemption* (things that happened after the effects). Finally, it is important to note that experimental designs conducted under field conditions are much less methodologically rigorous than laboratory experiments.[2]

b. Quasi-experimental Approaches

Quasi-experimental designs (QEDs) are used when randomization is not possible but when a comparison group can be identified. Sample selection takes place either after subjects have made the decision whether to participate in the program or when an administrative agency has made the decision to provide services to certain subjects or communities and not to others. In either case there is the possibility of systematic differences between the two groups (selection bias), which may significantly affect program outcomes. QEDs match the two groups as closely as possible, using either statistical matching techniques such as propensity score matching or judgmental matching with the comparison group selected using the advice of experts, community leaders, or similar groups and using whatever kinds of secondary data are available.

In strong QEDs the treatment and comparison groups are statistically matched (e.g., through propensity score matching). These designs are statistically weaker than RCTs as there is likely to be a selection bias due either to self-selection or to the selection procedures adopted by the implementing agency. There are a wide range of QEDs that vary in terms of their statistical strength and consequently in the adequacy of the counterfactual for causal attribution (analysis). Examples are regression discontinuity, propensity score matching, difference-in-difference regression, pipeline design, and judgmental matching. In general, quasi-experimental approaches can be characterized by two features: the modality of defining the group comparisons and the number of data points over time. Regarding the latter, the higher the number of data points in time (e.g., annual measurements of household savings and expenditures), the higher the likelihood that one can capture effects over time.

A QED has the advantage that it is more flexible to adapt to the program design as the project and comparison groups are normally chosen after project participants have been selected. This means the evaluation design does not impose constraints on how participants are selected in the way that an RCT does, making the design more acceptable to program managers. The design also has more flexibility to adjust to changes in program design. This is an important practical advantage because the strict program design requirements for using RCTs means that they can probably be applied in perhaps only 1%–2% of projects; the greater flexibility of QEDs means that they can be applied in many more program evaluations (see Bamberger & White, 2007). Quasi-experimental and non-experimental statistical approaches are also useful for looking at larger and more complex interventions (e.g., budget support, sector budget support).[3]

Designs such as the pipeline design are particularly useful for evaluating interventions that are designed to cover the whole target population, meaning there is no comparison group. These designs can be applied in creative ways to take advantage, for example, of planned phased implementation of programs with national coverage or of programs that encounter unanticipated delays in some areas. In both cases the regions or areas where there are planned or unanticipated delays can be used as the comparison groups.

Most of the limitations of RCTs also apply to QEDs. In addition, the issue of selection bias is a major challenge as changes that are assumed to be due to the program intervention may in fact be partially or mainly due to special attributes of the project group.

2.3 Theory-Based Approaches

Theory-based evaluation is discussed extensively in Chapters 5 and 6. The core of theory-based evaluation is the so-called intervention theory, or theory of change, a set of causal assumptions that explain how an intervention works (or is intended to work) and contributes to processes of change in society. These causal assumptions have to be made explicit, refined, and tested using a variety of methods and sources of information. The main approaches and principles of theory-based evaluation explained in Chapters 5 and 6 also apply to impact evaluation. Here, we focus explicitly on theory-based evaluation in relation to the evaluation of impact.

Broadly, one can discern two different approaches to theory-based impact evaluation:

- *The intervention theory (or theory of change) as the overarching framework of the evaluation.* Typically, evaluators reconstruct the intervention theory (or even multiple rival theories). Subsequently, the theory is empirically tested, matching the appropriate methods to particular assumptions of the theory. Theory-based evaluation is not method-specific; any appropriate method (and ideally multiple methods) may be applied to test a particular assumption. For example, assumptions regarding the outreach and accessibility of mobile sexual and reproductive health clinic programs can be studied using information on routes, communities visited, registry data of patients, and visits to (a purposive/random sample of) communities to interview patient and non-patient households. The effect of health services on health indicators could, for example, be studied in a more tightly controlled experimental (RCT) setting.
- *The intervention theory (or theory of change) as a tool for refining and testing the causal logic underlying an intervention eventually resulting in a causal impact narrative.* Realist evaluation fits into this category (e.g., Pawson & Tilley, 1997). Contribution analysis (Mayne, 2001) is another example. The main difference with the previous approach is that the refinement of the theory is the focus of the evaluation and also often the output of the evaluation (a refined theory). The emphasis is on explanation through an iterative process of revising the theory and collecting new empirical evidence. Another variant is process tracing (see below). Here, the emphasis is on systematically assessing each causal step in the theory using four tests. Finally, qualitative comparative analysis (QCA; which is discussed later as a separate approach to impact evaluation) is about identifying causal packages, sets of independent and dependent variables, which recur across settings. Understanding that processes of change are about the confluence of a number of factors (including the intervention) influencing a number of causal pathways is at the heart of this method. QCA can help to identify theories around these patterns of association.

To elaborate on the first bullet point above, as argued by Cook (2000), the choice between quantitative counterfactual analysis and theory-based evaluation is a clearly false one as the two complement each other in many ways:

- The intervention theory will help indicate which of the intervention components are amenable to quantitative counterfactual analysis through, for example, (quasi)experimental evaluation and how this part of the analysis relates to other elements of the theory.
- The intervention theory approach will help identify key determinants of impact variables to be taken into account in a quantitative impact evaluation.
- The intervention theory approach can provide a basis for analyzing how an intervention affects particular individuals or groups in different ways. Although quantitative impact evaluation methods typically result in quantitative measures of average net effects of an intervention, an intervention theory can help to support the analysis of distribution of costs and benefits.
- The intervention theory can help strengthen the interpretation of findings generated by quantitative impact evaluation techniques.

2.4 Case-Based Approaches

Our discussion of case-based approaches is purposely limited to two relatively novel approaches. The reason for this is threefold. First, the number and diversity of methods under this approach is high. It would take up a lot of space to adequately capture this diversity. Second, most approaches are well described in the literature (see, e.g., Byrne & Ragin, 2009). Finally, we have noted that there has been increased interest in the evaluation community to apply approaches that we discuss below: process tracing and qualitative comparative analysis.

a. Process Tracing

Process tracing (PT) is a method of inquiry premised on the idea that a theory can be tested based on the evidence in a case against new factors or new evidence in the same case. This method shares some similarities with detective work and relies heavily on Bayesian logic, particularly with regard to the requirement of constantly updating prior knowledge based on new evidence. It is also closely related to theory-based evaluation and in fact can be considered an application of it (see Chapters 5 and 6). In PT causal inference about the relations between an intervention and an outcome is thought to be mediated by a causal mechanism with several components, each of which is a necessary part of a complete causal mechanism that in itself may be sufficient for the effect to occur but might not be necessary (since the effect might be reached through other causal mechanisms).

PT is essentially about analyzing trajectories of change and causation. It relies on careful description as well as examination of diagnostic evidence of a causal mechanism. In practical terms, an evaluator using PT describes and articulates with great detail the causal chain mediating the relationships between the intervention and the outcome. In PT, the quality of the causal inference depends on how fine-grained the descriptions of the micro-mechanisms joined in the causal chain are (Befani, 2012).

The basic logic underlying process tracing is that tracing the processes that may have led to an outcome helps narrow the list of potential causes. By doing so, it seeks to eliminate a large number of alternative explanations for an effect, more than most other methods that eliminate single causes one by one. Rather than operating with single variables, process tracing methods eliminate rival causal chains (George & Bennett, 2005). PT enables evaluators to assess transparently and in a systematic manner the confidence that can be placed in the causal mechanism underlying the theory of change of an intervention. In particular, its application enables evaluators to confirm or disconfirm a hypothesis about why an intervention does or does not work in a particular context (based on Bennett & Elman, 2006; Collier, 2011).

To test causal relationships, PT relies on four empirical tests to determine whether a particular condition is necessary and/or sufficient to affirm causal inference. Table 4.2 reproduces Collier's (2011) presentation of the four tests. These tests share the objective of progressively eliminating rival hypotheses, but they differ in their capacity to do so. Further discussion of different variants of process tracing can be found in Beach and Pedersen (2013).

Table 4.2 Four Tests Used in Process Tracing to Assess Causal Inference

		Sufficient for Affirming Causal Inference	
		No	Yes
Necessary for Affirming Causal Inference		1. Straw in the Wind	3. Smoking Gun
	No	a. Passing: Affirms relevance of the hypothesis, but does not confirm it	a. Passing: Confirms hypothesis
		b. Failing: Hypothesis is not eliminated, but is slightly weakened	b. Failing: Hypothesis is not eliminated but is somewhat weakened
		c. Implications for rival hypotheses: Passing: *slightly* weakens them Failing: *slightly* strengthens them	c. Implications for rival hypotheses: Passing: *substantially* weakens them Failing: *somewhat* strengthens them
	Yes	2. Hoop	3. Doubly Decisive
		a. Passing: Affirms relevance of hypothesis but does not confirm it	a. Passing: Confirms hypothesis and eliminates others
		b. Failing: Eliminates hypothesis	b. Failing: Eliminates hypothesis
		c. Implications for rival hypotheses: Passing: *somewhat* weakens them Failing: *somewhat* strengthens them	c. Implications for rival hypotheses: Passing: eliminates them Failing: *substantially* strengthens them

SOURCE: Adapted from Collier (2011, p. 825).

b. Qualitative Comparative Analysis

Qualitative comparative analysis (QCA) refers to a family of methods that seeks to identify causal packages. It focuses on a limited number of empirical cases, for which configurations of effects (outcomes, impacts) and conditions for effects are captured in a *truth table*. In this table, each configuration of conditions or factors is represented by a series of zeros and ones that translates into the absence or presence of a given condition.

QCA sees cases as complex systems and does not attempt to decompose the causal configurations into variables with equal causal power (Byrne & Ragin, 2009). To start with, QCA is an approach that considers cases in their entirety rather than simply harvesting variables across a large number of cases, as is done in variable-based approaches. It also takes for granted that it is a combination of causal conditions that eventually generates an outcome, not simply one particular cause. In development processes, there are a number of *ground-preparing causes* that are necessary elements of development success, but are not sufficient by themselves (Befani, 2013). For example, three conditions without a fourth may not lead to any meaningful change, but the presence of the four factors together might allow a program to go from poor performance to excellent results, in a nonlinear causal pattern (Befani, 2012). QCA is also grounded in an embedded contextual view of reality: Depending on the context, a given set of conditions may very well lead to different outcomes. Finally, QCA relies on the idea that multiple causal chains coexist and lead to the same effects (equifinality) and considers as relevant all the potential causal paths that can lead to a given outcome. The result of QCA is the identification of a number of causal paths that are sufficient to produce a given outcome.

As a family, QCA encompasses three main types of techniques, each relying on a different set-theory. Crisp-set QCA (csQCA) applies Boolean logic to the various conditions by dichotomizing each condition into absence or presence (0 or 1). Multi-value QCA (mvQCA) allows for multiple category conditions. Finally, fuzzy-set QCA (fsQCA) enables the researcher to assign a degree of membership to each condition rather than a dichotomized membership (Ragin, 2000). Box 4.1 illustrates a simple application of QCA to the assessment of the effectiveness of an irrigation assistance project in Nepal (Lam & Ostrom, 2010).

BOX 4.1 APPLICATION OF QCA

Lam and Ostrom (2010) evaluated the impact of an innovative irrigation assistance project that was undertaken in 19 irrigation systems in Nepal starting around 1985. This project had various innovative components, including provision of technical and financial assistance, partial funding for physical infrastructure, extensive involvement of farmers in the decision-making process, and farmer-to-farmer training. For the evaluation, data were collected in three time periods (at the start of the program in 1985, in 1991, and in 1999). The availability of structured information over time allowed the evaluators to look at how the irrigation effectiveness

(Continued)

(Continued)

had changed over the years using statistical analysis. This analysis revealed fluctuating patterns across systems and time periods. The authors identified three main sources of complexity that needed to be addressed: (1) the effect of a particular factor is contingent and combinatorial (it is the articulation of several factors that produce the outcome), (2) the effects are not linear, and (3) the complex dynamics of institutional change needed to be captured.

The authors therefore used QCA to identify a set of causal conditions, amid the diversity of experiences, conducive to sustained intervention effects. Through an in-depth literature review and interviews with the farmers, they identified five key conditions (continual assistance on infrastructure improvement, existence of formal rules for irrigation operation, provisions of fines, consistent leadership, and collective action among farmers for system maintenance) that could explain why high performance was sustained in some systems but not in others. All conditions were dichotomized as being either present or absent from the system. One of the outcome variables was the availability of water during the winter. The truth table below summarizes the 11 unique configurations of factors for the 15 systems for which water supply measurements were available.

Five causal conditions					Number of systems	
Assistance (A)	**Rules (R)**	**Fines (F)**	**Leadership (L)**	**Collective action (C)**	**Sustained performance**	**Not sustained performance**
Absent	*Present*	*Absent*	*Present*	*Present*	1	1
Absent	*Present*	*Present*	*Present*	*Present*	2	0
Present	*Present*	*Absent*	*Present*	*Present*	2	0
Present	*Present*	*Present*	*Present*	*Present*	2	0
Absent	*Absent*	*Absent*	*Absent*	*Present*	1	0
Absent	*Present*	*Absent*	*Absent*	*Present*	1	0
Absent	*Absent*	*Present*	*Present*	*Present*	0	1
Present	*Absent*	*Absent*	*Absent*	*Absent*	0	1
Present	*Present*	*Present*	*Absent*	*Absent*	1	0
Present	*Present*	*Absent*	*Absent*	*Present*	1	0
Present	*Present*	*Present*	*Absent*	*Present*	1	0

The fsQCA software was used to operate the Boolean minimization and come up with a parsimonious solution that related the sustained performance of irrigation as measured by the availability of water in winter and the various causal conditions. By going back and forth between the cases and the truth table, the authors identified the following equation: W = AR (+lF) + CLRF + Calf.[4]

Three groups of explanatory configurations emerged. Here we present only one of them. The first configuration showed that ongoing infrastructure investment can enable sustained performance only if farmers have developed rules (AR go together). These combined factors are a necessary but not sufficient part of success. They should be present in a context where either collective action takes place or fines are imposed in a context of weak leadership.

SOURCE: Adapted from Lam & Ostrom (2010).

2.5 Participatory Approaches

Participatory evaluation designs involve a wide range of stakeholders in the design, implementation, interpretation, and use of the evaluation. Participatory approaches may be used for methodological reasons, to strengthen data quality and validity, or for ideological reasons (Cousins & Whitmore, 1998). Participatory approaches are often used in many mixed methods designs to triangulate among different sources of data to increase reliability and validity of the data. In contrast, the ideological dimension of participatory approaches is central to empowerment, feminist, or equity-oriented evaluation as part of a process of political and social empowerment.

A potential downside is the risk that participatory processes may be monopolized by politically or socially more powerful groups. With the increasing use of mobile phones and other new information technology, there is also the risk of selection bias as people who have access to mobiles and other devices are likely to be the wealthier and better educated groups. There may also be a gender bias as, in some contexts, women may have less access to mobile phones or more generally are not in a position to speak freely.

There are many participatory approaches to evaluation (see Kumar, 2002). Examples of participatory techniques in the context of impact evaluation include the following:

- *Outcome mapping:* This focuses on outcomes as behavioral change (Earl, Carden, & Smutylo, 2001). It recognizes that external partners do not directly produce outcomes, but rather they work with boundary partners (local agencies) that directly produce the changes. As most programs involve multiple boundary partners, each with its own interests and priorities, programs are likely to produce a wide range of outcomes, not all of which were planned or even necessarily desired by the external agencies. Outcome mapping involves three stages: intentional design (designing the program in a participatory way in collaboration with boundary partners), outcome and performance monitoring, and evaluation planning.
- *Outcome harvesting:* This approach, which builds on outcome mapping, "enables evaluators, grant makers and managers to identify, formulate, verify and make sense of outcomes" (Wilson-Grau & Britt, 2012, p. 1). Information is gleaned from reports, personal interviews, and other sources to document how a given program has contributed to outcomes. Outcomes can be positive or negative, intended or unintended, but the connection to the intervention must be verifiable. Wilson-Grau and Britt (2012, Box 1) draw the analogy with forensic science as a wide range of techniques is used to "sleuth the answers" by generating evidence-based answers to the following questions:
  - ○ What happened?
  - ○ Who did it (or contributed to it)?
  - ○ How do we know this? Is there corroborating evidence?
  - ○ Why is this important? What do we do with what we found out?

- *Most significant change:* "The process involves the collection of significant change (SC) stories emanating from the field level, and the systematic selection of the most significant of these stories by panels of designated stakeholders or staff. The designated staff and stakeholders are initially involved by 'searching' for project impact. Once changes have been captured, various people sit down together, read the stories aloud and have regular and often in-depth discussions about the value of these reported changes. When the technique is implemented successfully, whole teams of people begin to focus their attention on program impact" (Davies & Dart, 2005, p. 8).

2.6 Review and Synthesis Approaches

Review and synthesis approaches involve the practice of identifying and selecting existing evaluation studies, reviewing and extracting information, and aggregating and synthesizing information into an overall perspective on what works (for whom and under what circumstances). Over the last decade, with the increasing availability of impact evaluation studies, there has been a marked increase in the application of review and synthesis studies in the context of international development cooperation. Chapter 9 discusses some of the prevalent approaches in review and synthesis, while Chapters 16 and 20 present examples on microcredit interventions and community accountability and empowerment initiatives, respectively.

3. Strengths and Limitations of Established Impact Evaluation Approaches in the Context of Complexity

In this section we discuss some of the comparative advantages and limitations of different methodological approaches in terms of addressing complexity. A few qualifying remarks are in order:

- In Chapter 1 we distinguished between a general complexity and restricted complexity perspective. Our discussion of complexity in relation to established impact evaluation approaches is mainly framed within the latter.
- Below we distinguish between several aspects of complexity in the light of causal change. While we try to discuss these aspects separately for each of the six impact evaluation approaches, it should be noted that in reality they are closely related. For example, the occurrence of multiple causal pathways, multiple (un)intended effects, emergence, and other aspects of the nature of causal change (e.g., uncertainty) are all closely linked to each other.

The core of the impact evaluation debate in the context of complexity revolves around the discussion of the nature of causal change and how it relates to development interventions, one of the dimensions of the book's conceptual framework.

In order to assess the strengths and limitations of the impact evaluation approaches discussed in this chapter with respect to this dimension, we focus on the following issues:

- *Attribution:* This refers to the extent to which a particular change can be attributed to an intervention, taking into account other variables. Here are a couple of important impact evaluation questions: Has the intervention led to change? To what extent has it made a difference?
- *Explanation:* Development interventions aim to change the behavior of individuals and organizations. At the same time, the likelihood and nature of change is dependent on the behavior of a multitude of actors affected by underlying contextual conditions. Here are a couple of important impact evaluation questions: How do interventions work? How are they affecting the behavior of different actors?
- *Multiple causal pathways:* An intervention (especially if the intervention encompasses multiple activities at different levels) can trigger multiple causal pathways of change. This idea is more in line with the concept of contribution, that is, a confluence of factors affecting a particular change or multiple changes. In the latter case one can speak of causal packages.
- *Nature of causal change:* Causal change is often path dependent yet at the same time can be highly uncertain, nonlinear (abrupt, gradual, or both, over time), and emergent (see below).
- *Emergence:* The principle of emergence is also an element of the nature of causal change yet deserves particular attention. A development intervention has only imperfect control over the possible achievement of its objectives, especially because the program changes the conditions that made the program work in the first place. Consequently, the most successful programs and organizations are those that adapt to this emergent change. Emergence is also closely linked to the concepts of uncertainty and dynamics in both implementation (interventions change over time, they are not stable) and changes in society.
- *Scope of effects:* Interventions may affect multiple processes of change at different levels, resulting in a number of intended and unintended outcomes. The extent to which the evaluation is able to capture all effects is important here.

Our succinct discussion of how the main impact evaluation approaches deal with complex causal change is presented in Table 4.3.

It is very clear that the different methodological approaches have comparative advantages in dealing with particular aspects of causal change. In practice it is therefore important to adopt a mixed methods approach. Chapter 8 explains the different principles and variations of mixed methods evaluation within the framework of complexity. In Box 4.2 we illustrate various aspects of complexity in causal change in relation to method choice using an example of an evaluation on the topic of payments for environmental services in Latin America.

Table 4.3 Strengths and Weaknesses of Established Impact Evaluation Methods With Regard to Complexity in Causal Change

Methodological approach	**How is causal change addressed?**
Experimental	*Attribution:* Very strong on determining the effect of an intervention on a limited number of effect (outcome and impact) variables. In Chapter 7 we discuss the principle of unpacking complex interventions into evaluable parts. Within this framework experimental and quasi-experimental designs can be compatible with complexity-responsive evaluation.[5] *Explanation:* RCTs are not designed to explain why certain changes occur as they control for (but do not measure or theorize on) all other observable and non-observable characteristics that may influence the causal change process. In quasi-experimental designs, to the extent that confounding factors are measured and included in the (regression) model, some aspects of causal exchange may be explained. Due to the variable-based approach there are also serious challenges in terms of construct validity. *Multiple causal pathways:* Limited options for dealing with multiple causal pathways, for example, through randomized experiments incorporating multiple treatments. *Nature of causal change:* The important factor here is the number of data points, the availability of data over time. Posttest-only or pretest-posttest designs are inherently limited in terms of dealing with nonlinearity. Multiple data points (longitudinal designs) can show the patterns of change (a limited number of variables) over time. *Emergence:* Effect (outcome and impact) variables remain constant over time. Moreover, posttest-only or pretest-posttest designs are inherently limited in terms of dealing with dynamics over time. In case there are changes in the intervention over time and/or how the intervention influences change processes, this cannot be captured by quantitative methods for two reasons. First, quantitative methods are not designed to detect such changes. Second, the definition and selection of variables for which data are collected over time are determined before the first data point (e.g., ex ante baseline survey) and remains constant over time. *Scope of effects (unintended effects):* Focus on a limited number of effect variables. Very reductionist in focus. No attention to unintended effects.
Statistical	*Attribution:* Strong on determining the effect of an intervention on a limited number of effect (outcome and impact) variables. *Explanation:* In multivariate designs, to the extent that confounding factors are measured and included in the (regression) model, some aspects of causal change may be explained. Due to the variable-based approach there are also serious challenges in terms of construct validity. *Multiple causal pathways:* Limited options for dealing with multiple causal pathways.

Methodological approach	**How is causal change addressed?**
	Nature of causal change: See *experimental*. Longitudinal designs can deal with nonlinearity to some extent. *Emergence:* See *experimental*. *Scope of effects (unintended effects):* Focus on a limited number of effect variables and the relationships with a number of independent and confounding variables. Reductionist in focus. No attention to unintended effects.
Theory-based	*Attribution*: Strong on making explicit the causal assumptions that could explain how interventions are expected to lead to change. The quality of the theory is highly dependent on the resources and specific methods for the reconstruction and refinement of the assumptions. The quality of attribution analysis is highly dependent on the underlying methods for testing particular causal assumptions and the corresponding data (see Chapter 7 on unpacking). The potential for a good macro-perspective on causal change (i.e., explaining the different steps in causal change processes, theorizing on causal change) is high, but not necessarily with respect to specific causal linkages between intervention outputs and outcomes. *Explanation:* See *attribution*. The strength lies in explaining causal change processes, but this is highly dependent on the quality of underlying data and methods for looking at particular causal assumptions. *Multiple causal pathways:* Can clarify multiple causal pathways between different intervention components and different processes of change. The same disclaimer as above applies. *Nature of causal change:* Theorizes on the nature of causal change (taking into account existing knowledge from multiple sources), which then may be used to guide data collection and analysis to empirically capture this. Data collection over time or access to long-term data is important. *Emergence:* Theories of change should be periodically updated and revised to reflect changes in the dynamic and complex reality of a development intervention. To the extent that this is done, theory-based approaches can address emergence. If they are not periodically revised, it is likely that a discrepancy between the dynamic reality and the theory will arise, making the theory less and less useful as an abstraction of reality and a framework for complexity-responsive evaluation. *Scope of effects (unintended effects):* Once a theory of change has been reconstructed, it can cause bias in terms of how evaluators view the intervention and its context. Such bias may draw attention away from the complexity of causal change in practice. For the same reasons, unintended effects may go undetected. Multiple theories of change and/or multiple iterations to adapt the theory on the basis of new data and insights are important.

(Continued)

Table 4.3 (Continued)

Methodological approach	How is causal change addressed?
Case-based	*Attribution:* Different principles for dealing with attribution are available in this set of approaches. Process tracing is potentially very strong on attribution. Rich description of complex causal change processes, part of many qualitative case-based methods, can also strengthen attribution claims in specific situations. Generalization of findings to the overall target population may be challenging. *Explanation:* Case-based methods should ideally be theory-driven and in this sense are closely aligned to theory-based evaluation (some case-based methods are often classified as part of the theory-based evaluation tradition). A general constraint in most case-based methods is the quality of the (initial) theory of causal change. A theory-based framework may be helpful in reconstructing realistic initial theories of change that guide further data collection and analysis. *Multiple causal pathways:* This is the core of QCA, the identification of multiple causal packages. Measurement issues and model specification are potentially important constraints. Other qualitative case-based methods are strong on identifying multiple causal pathways through in-depth analysis of the case and its embeddedness in the wider context. *Nature of causal change:* Qualitative case-based methods are strong on the in-depth analysis of the case and its embeddedness in the wider context. Data collection over time or access to long-term data is important. *Emergence:* Same as above. Process tracing and QCA are not particularly strong on this point. Using particular underlying methods and data (i.e., that provide evidence on respectively the causal assumptions and the variables for process tracing and QCA) that are sensitive to emergence to some extent can be helpful. *Scope of effects (unintended effects):* Some methods that rely heavily on rich description and in-depth context-specific data collection are more likely to capture the full range of potential effects of an intervention at a particular level (e.g., household, community). Often, this information cannot be generalized beyond the case. In multilevel, multisite interventions, depending on the unit of analysis (i.e., the case), this may be an important limitation.
Participatory	*Attribution:* Taking on board different stakeholder perspectives can significantly increase the evaluator's understanding of the nature, diversity, and extent of changes brought about by an intervention. Rich descriptive information may be helpful in understanding complex processes of change and attribution. Generalizability of the findings may be an important constraint. Biases such as groupthink, knowledge limitations, and cognitive bias need to be taken into consideration.

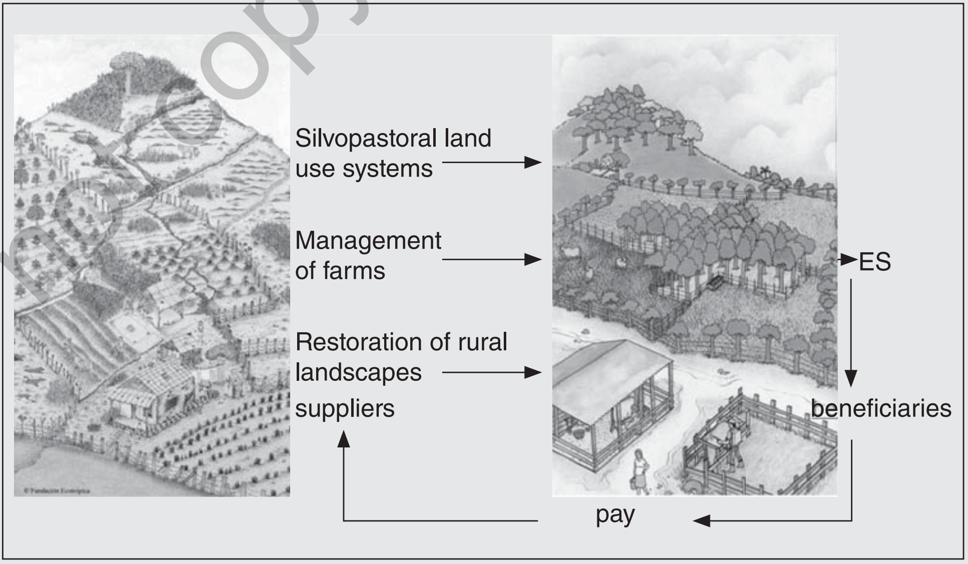
Methodological approach	**How is causal change addressed?**
	Explanation: Perspectives from different stakeholders can generate a unique multi-angle perspective of an intervention. *Multiple causal pathways:* Same as above. *Nature of causal change:* Same as above. Data collection over time or access to long-term data is important. *Emergence:* Participatory methods can be particularly strong on detecting emergence in terms of evolving patterns of implementation and change. This potential becomes stronger with higher degrees of participation and involvement of stakeholders in data collection and analysis over time. This is especially true for implementation but not necessarily for processes of change. These may occur at levels of analysis (e.g., regional employment effects, climate change, biodiversity, inequality) that may not be directly perceived by stakeholders. *Scope of effects (unintended effects):* Some changes induced by an intervention may affect (or be of importance to) only one particular stakeholder group. Involving a broad range of stakeholders enhances the likelihood of generating a comprehensive perspective on the effects of interventions. Effects at higher levels (the detection of which requires other data and methods) may go undetected by stakeholders.
Review and synthesis	*Attribution:* The extent to which effects can be attributed to an intervention, and the extent to which intervention change processes can be explained, is largely reliant on the type of review and synthesis approach (i.e., systematic review may be strong on the first; realist synthesis may be strong on the second) and the underlying evidence base. Systematic review (using meta-analysis) is very strong on attribution for a very limited set of effect (outcome and impact) variables. *Explanation:* Systematic review using meta-analysis is usually very weak on explanation. By contrast, realist synthesis focuses on how interventions work and affect the realities of different stakeholders. *Multiple causal pathways, nature of causal change, emergence, scope of effects:* By and large, for systematic review many of the same strengths and limitations of experimental and statistical approaches apply. By contrast, narrative reviews and realist syntheses are more similar to case-based approaches and their strengths and limitations. In general, the option of triangulating evidence from multiple studies enables the evaluator to strengthen the validity of claims on these criteria as well as attribution and explanation.

NOTE: Methodological approaches are often implemented in combination. For example, theory-based evaluation constitutes the framework for many of the other approaches. In any case, all the assessments refer to the specific methodological approach, not taking into account that some of the shortcomings in practice are compensated for through complementarity of methods in mixed methods designs.

BOX 4.2 ILLUSTRATING THE COMPLEXITY OF EVALUATING CAUSAL CHANGE: THE REGIONAL INTEGRATED SILVOPASTORAL APPROACHES TO ECOSYSTEM MANAGEMENT PROJECT

The Regional Integrated Silvopastoral Approaches to Ecosystem Management Project (RISEMP) was implemented in the period 2002–2008. It was a GEF-World Bank project, designed as an innovative pilot initiative, which would promote silvopastoral practices through technical assistance and payments for environmental services (generated by these practices). The project was implemented in three countries: Nicaragua, Costa Rica, and Colombia.

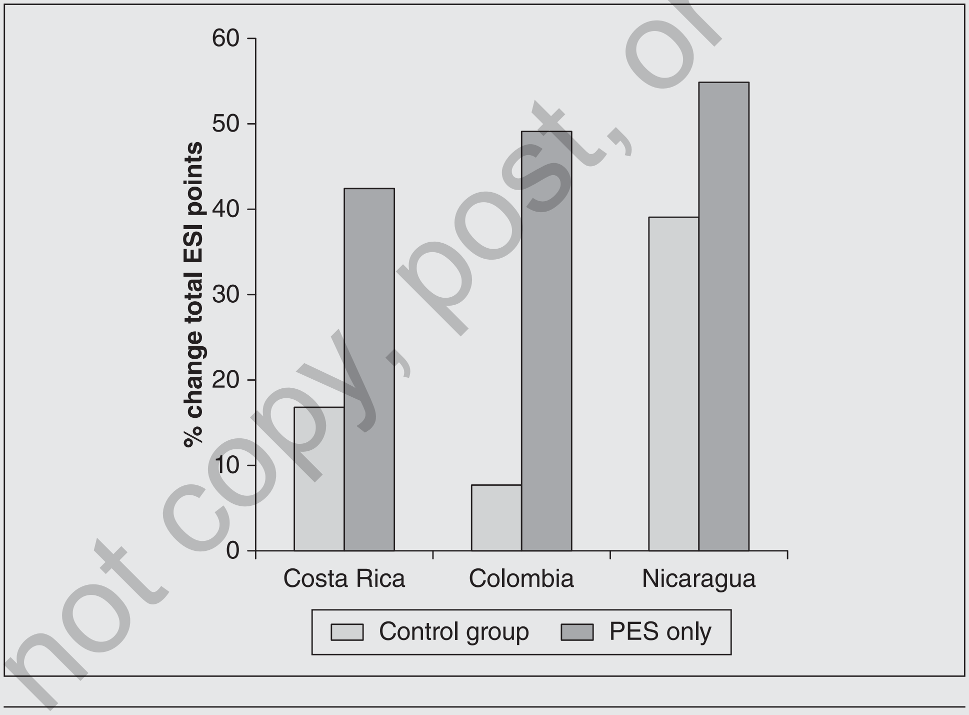
At the level of the three pilot sites, the project focused on three main areas of work: improvements of silvopastoral land use systems, improved management of farms, and restoration of rural landscapes. The funding model was innovative. Improvements in land use (LU) were expected to generate environmental services, particularly improved biodiversity (e.g., agrobiodiversity and regional biodiversity through improved connections of habitats and improved ecosystems between protected areas and the private farms in between protected areas [i.e., the corridor function]) and carbon sequestration (e.g., in the soil and the vegetation). Farmers were paid for the environmental services (biodiversity, carbon sequestration) generated by improvements in LU. Detailed studies analyzing the relationships between different LUs and environmental services resulted in indices that provided the basis for payments for environmental services to land users. Findings were recorded in academic and policy-oriented publications. The principle of payments for environmental services also closely related to the idea that eventually the beneficiaries of environmental services (e.g., tourists, the general public) would compensate the land users for generating them. This idea of market creation was not tested in this project (i.e., the project represented the beneficiaries of environmental services).


Silvopastoral land use systems
Management of farms
Restoration of rural landscapes
suppliers
ES
beneficiaries
pay

In 2009 the GEF Evaluation Office commissioned an assessment of the project's monitoring and evaluation framework and its potential for assessing the project's effects. One of the main reasons for this was that the project was based on a randomized experiment that was expected to generate rigorous evidence on the project's outcomes and impacts. More specifically, both payments for environmental services and technical assistance were randomly allocated to farmers. Through the principle of randomization and group comparison, it was expected that changes in LU (and subsequent changes in environmental services as well as economic effects) could be attributed to different project incentives, controlling for all other (observable and non-observable) factors (Vaessen & Van Hecken, 2009).

In light of the substantial external pressure on the GEF to build in randomized experiments in the design and monitoring and evaluation (M&E) frameworks of GEF projects, the assessment was intended to provide an objective view on the strengths and weaknesses of randomized experiments in the context of GEF projects. Below we discuss some elements of complexity with respect to the intervention and its context and how they could be addressed.

Attribution



SOURCE: Vaessen & Van Hecken (2009).

NOTE: ESI refers to environmental services index, which captures the relationship between land use systems and environmental services generated.

The assessment of the project's randomization model and corresponding measurements of LU, environmental services, and other factors concluded that, to some extent, the project was able to generate rigorous evidence on the effects of different incentives (see Vaessen & Van Hecken, 2009, for a discussion on the threats to validity of findings

(Continued)

(Continued)

on attribution). The figure above shows the differences in changes in the environmental service index over time between (randomly determined) groups of farmers receiving payments for environmental services and control groups.

Explanation

A base theory of change was developed (see next page) that makes explicit the relationships between project incentives and the conditions under which particular types of farmers were expected to change their LUs and eventually generate environmental and economic benefits. This theory was developed by the evaluators. Several surveys and semistructured interviews were undertaken which in principle would allow the evaluators to analyze the question: What types of farmers, and under what circumstances, will undertake particular LU changes?

Multiple Causal Pathways

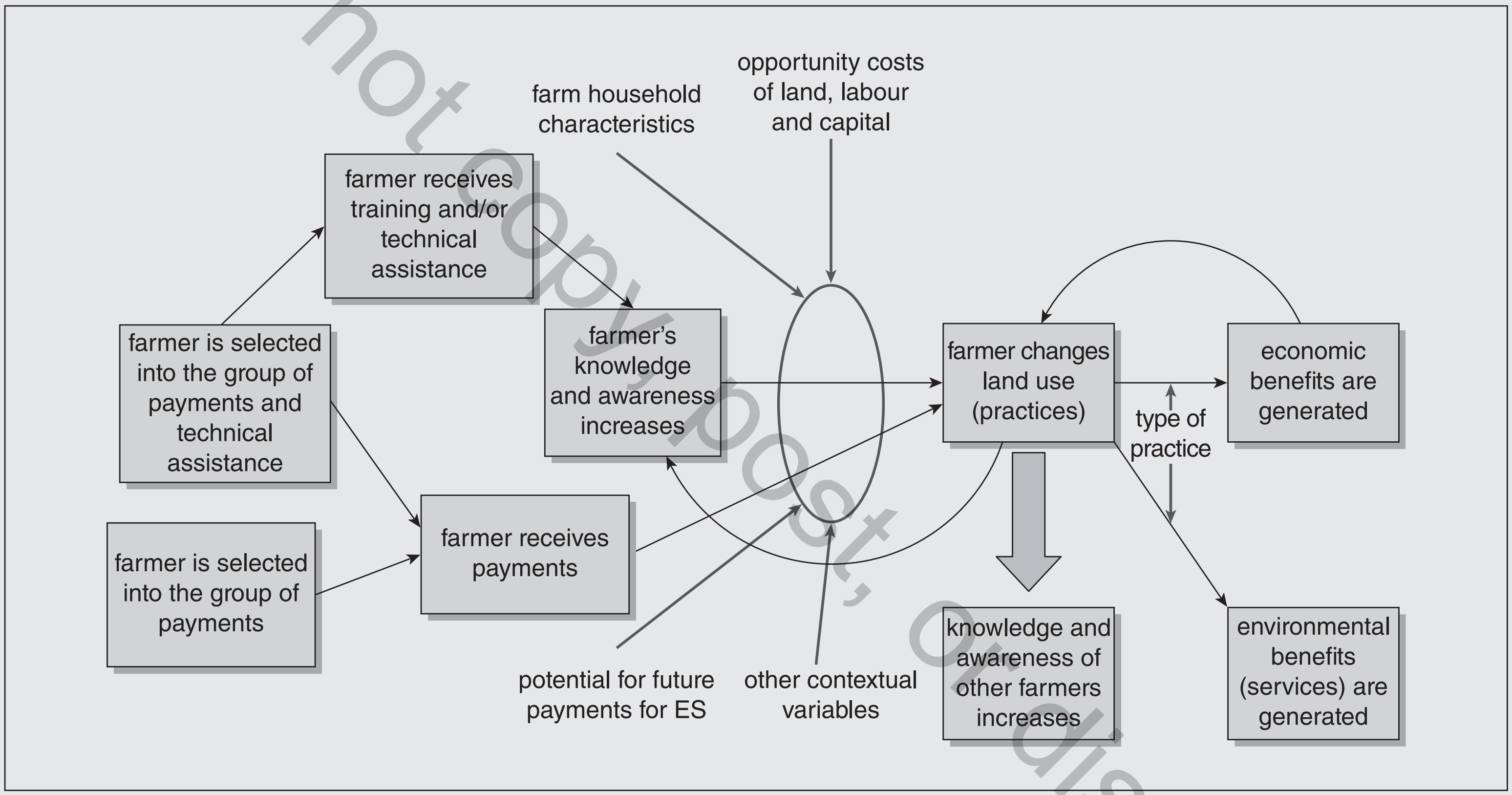
Through field observation and interviews with stakeholders at different levels, the evaluators identified three main levels at which the project was expected to contribute to change: farm, regional, international. The implicit theory of change of the project, which was made more explicit by the evaluators, was limited to effects at the farm/household level. There were no implicit (or explicit) theories of change for changes at the regional level or the international level. However, there were intended effects at the regional and international levels. In addition, the evaluators identified a number of unintended effects. These could not be captured through the project's M&E framework (see discussion below).

Nature of Causal Change

The project's framework for data collection and analysis as well as its randomization model were well equipped to address direct effects (e.g., LU) and indirect effects (e.g., environmental services). However, in general environmental change is difficult to capture. The project's multiyear monitoring of LUs, the studies on ecosystems, and biodiversity was very detailed and able to unravel a lot of the complexity regarding the nonlinear relationships between, for example, LU and species abundance. However, the corridor biodiversity function of LU in agricultural landscapes in between protected (biodiversity-rich) areas was very difficult to capture. At what point and in what ways will biodiversity at the regional level benefit from improved ecosystems in farms? Given the nonlinear and emergent nature of these processes, the analysis would have required very long monitoring over time and detailed studies beyond the project area level.

Emergence

Apart from the emergent nature of environmental effects (especially regarding biodiversity), there were a number of other aspects that can be characterized as emergent. For example, the relationships between changes in LU, increases in production levels of certain crops and livestock products, evolutions in prices (inputs and products), and availability of labor affect household incomes and potentially local and regional economic growth. These interactions are complex and difficult to capture (e.g., through system mapping and modeling). Another interesting aspect was that despite the rigid restrictions on implementation (for the randomized experiment to work, implementation should be homogeneous across farmers and over time), there were changes in implementation over time. For example,


farm household characteristics
opportunity costs of land, labour and capital
farmer receives training and/or technical assistance
farmer is selected into the group of payments and technical assistance
farmer's knowledge and awareness increases
farmer changes land use (practices)
economic benefits are generated
type of practice
farmer receives payments
farmer is selected into the group of payments
potential for future payments for ES
other contextual variables
knowledge and awareness of other farmers increases
environmental benefits (services) are generated

SOURCE: Vaessen & Van Hecken (2009).

(Continued)

(Continued)

there were differences (between countries but also within project areas) and changes over time in payment levels and in the modalities of technical assistance delivery (e.g., working through farmer groups vs. working with individual farmers). Apart from affecting the validity of the experimental findings, these changes also affected change processes in ways that were more difficult to trace and understand in comparison to homogenous implementation across farmers over time.

Scope of Effects

The assessment exercise identified the (likely) existence of a number of unintended effects, yet there were no data or data collection exercises planned to evaluate these in more detail. The effects not covered by the project's M&E framework can be summarized by the following causal assumptions. Please note that for each assumption there were empirical indications that these effects were in fact occurring.

Farm/household-level effects:

1. IF farmers are selected for the control group or the group without technical assistance THEN they may still learn from other farmers and implement LU practices. (unintended)
2. IF farmers are selected for the control group THEN they may change LU practices based on expectations about future payments and/or motivated by competition. (unintended)
3. IF farmers from different treatment groups (e.g., technical assistance vs. control) change their LU practices THEN it is likely that there are differences in quality in application. (intended)

Regional effects and effects outside the project area:

4. IF LU practices are implemented THEN land prices may rise; IF land prices increase THEN farmers may sell their land. (unintended)
5. IF LU practices are profitable THEN employment opportunities for external labor may increase. (intended)
6. IF farmers own land outside the project region THEN environmentally destructive LU may be displaced. (unintended)

National and international effects

7. IF innovative knowledge about the relationship between LU and environmental services or other topics is generated and published THEN this may contribute to replication of (parts of) the project elsewhere. (intended)
8. IF project staff and GEF or World Bank staff disseminate knowledge about the project THEN this may contribute to replication of (parts of) the project elsewhere. (intended)

It should be noted that a much wider range of methods would have to be applied to look into these different assumptions about possible effects. For example:

- Assumption 3: field observation and interviews with farmers
- Assumption 5: surveys and system modeling
- Assumption 7: bibliometric analysis and interviews based on a purposive sample of (inter)national stakeholders

4. Practical Applications

- There are considerable limitations in using established quantitative impact evaluation approaches in the context of complexity-responsive evaluations. However, unpacking complexity is often possible and the effects of particular intervention activities may be addressed by these approaches (see Chapter 7). Other methods are needed to shed light on, for example, context-specific implementation and change processes embedded in different systems of norms, beliefs, and values as well as interactions between different intervention processes and different stakeholder groups.
- Impact evaluations usually rely on combinations of methodological approaches. Very often, a theory of change (or multiple theories of change) constitutes the basis for framing the evaluation design and the choice of methods to look at particular causal assumptions (see Chapter 8).
- Different impact evaluation approaches each have their comparative advantages in terms of helping to address particular aspects of complexity in causal change.
- Evaluators should be open to the possible occurrence of unintended effects. The nature of complexity (e.g., emergence, unintended effects) generally makes it more difficult to plan for data collection processes over time. Next to the measurement of key variables over time, there should be space for exploratory qualitative research at different points in time during an intervention.
- A lot of the methodological debates on complexity are about how and with which tools to look at the empirical reality surrounding development interventions. Insufficient attention is given to data collection over time. Increasing the number of data points in time improves the likelihood of detecting patterns of change in key variables that are influenced by an intervention.

References

Bamberger, M., & White, H. (2007). Using strong evaluation designs in developing countries: Experience and challenges. *Journal of Multidisciplinary Evaluation, 4*(8), 58–73.

Beach, D., & Pedersen, R. (2013). *Process-tracing methods: Foundations and guidelines*. Ann Arbor: University of Michigan Press.

Befani, B. (2012). Models of causality and causal inference. In E. Stern, N. Stame, J. Mayne, K. Forss, R. Davies, & B. Befani (Eds.), *Broadening the range of designs and methods for impact evaluation* (Working Paper No. 38, pp. 103–126). London, UK: Department of International Development.

Befani, B. (2013). Between complexity and generalization: Addressing evaluation challenges with QCA. *Evaluation, 19,* 269–283.

Bennett, A., & Elman, C. (2006). Qualitative research: Recent developments in case study methods. *Annual Review of Political Science, 9,* 455–476.

Byrne, D., & Ragin, C. (Eds.). (2009). *Sage handbook of case-based methods*. Thousand Oaks, CA: Sage.

Cohen, J., & Easterly, W. (Eds.). (2009). *What works in development? Thinking big and thinking small*. Washington, DC: Brookings Institution Press.

Collier, D. (2011). Understanding process-tracing. *Political Science and Politics, 44,* 823–830.

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (pp. 27–34). San Francisco, CA: Jossey-Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5–23). San Francisco, CA: Jossey-Bass.

Davies, R., & Dart, J. (2005). *The "most significant change" technique: A guide to its use.* Retrieved from http://www.mande.co.uk/docs/MSCGuide.htm

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.). (2008). *International handbook of survey methodology*. New York, NY: Lawrence Erlbaum.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage.

Earl, S., Carden, F., & Smutylo, T. (2001). *Outcome mapping: Building learning and reflection into development programs*. Ottawa, Ontario, Canada: International Development Research Center.

Elbers, C., Gunning, J. W., & De Hoop, K. (2008). Assessing sector-wide programs with statistical impact evaluation: A methodological proposal. *World Development, 37,* 513–520.

Funnell, S., & Rogers, P. (2011). *Purposeful program theory*. San Francisco, CA: Jossey-Bass.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact evaluation in practice*. Washington, DC: World Bank.

Greene, J. C. (2009). Evidence as "proof" and evidence as "inkling." In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice*? (pp. 153–167). Thousand Oaks, CA: Sage.

Khandker, S. R., Koolwal, G. B., & Samad, H. (2009). *Handbook on quantitative methods of program evaluation*. Washington, DC: World Bank.

Kumar, S. (2002). *Methods for community participation: A complete guide for practitioners.* London, UK: ITDG.

Lam, W. F., & Ostrom, E. (2010). Analyzing the dynamic complexity of development interventions: Lessons from an irrigation experiment in Nepal. *Policy Science, 43,* 1–25.

Leeuw, F. L., & Vaessen, J. (2009). *Impact evaluations and Development: NONIE guidance on impact evaluation*. Washington, DC: Network of Networks on Impact Evaluation.

Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *Canadian Journal of Program Evaluation, 16*(1), 1–24.

Mikkelsen, B. (2005). *Methods for development work and research*. Thousand Oaks, CA: Sage.

OECD-DAC. (2002). *Glossary of key terms in evaluation and results based management*. Paris, France: Author.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Popay, J. (2006). *Moving beyond effectiveness: Methodological issues in the synthesis of diverse source of evidence*. London, UK: National Institute for Health and Clinical Excellence.

Ragin, C. (2000). *Fuzzy-set social science.* Chicago, IL: University of Chicago Press.

Rieper, O., Leeuw, F. L., & Ling, T. (2010). *The evidence book: Concepts, generation and use.* New Brunswick, NJ: Transaction.

Scriven, M. (2009). Demythologizing causation and evidence. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 134–152). Thousand Oaks, CA: Sage.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluation* (Working Paper No. 38). London, UK: Department of International Development.

Vaessen, J., & Van Hecken, G. (2009). *Assessing the potential for experimental evaluation of intervention effects: The case of the Regional Integrated Silvopastoral Approaches to Ecosystem Management Project (RISEMP)* (Impact Evaluation Information Document No. 15). Washington, DC: GEF Evaluation Office.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation, 16*, 153–164.

White, H., & Phillips, D. (2012). *Addressing attribution of cause and effect in small impact evaluations: Towards an integrated framework* (Working Paper 15). New Delhi, India: International Initiative for Impact Evaluation.

Wilson-Grau, R., & Britt, H. (2012). *Outcome harvesting*. Cairo, Egypt: Ford Foundation. Retrieved from http://usaidlearninglab.org/sites/default/files/resource/files/Outome%20Harvesting%20Brief%20FINAL%202012-05-2-1.pdf

Woolcock, M. (2013). Using case studies to explore the external validity of "complex" development interventions. *Evaluation, 19,* 229–248.

Notes

1. The OECD-DAC (2002) defines impacts as "positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (p. 24). However, when we look at the body of research under the banner of impact evaluation, a substantial part of it is not on long-term results nor on indirect and unintended results. In fact, a lot of impact evaluation is about analyzing the attribution of short-term outcomes to a particular intervention. For a wider discussion on the different interpretations of impact evaluation and the term impact, see White (2010).

2. Different kinds of unintended behavioral effects may affect the experiment that have nothing to do with the intervention (see Vaessen & Van Hecken, 2009, for an example).

3. See, for example, Elbers, Gunning, and De Hoop (2008).

4. Note that in Boolean algebra capital letters signal the PRESENCE of a condition and noncapital letters signal the absence of a condition. Addition is equivalent to OR, and multiplication means conjunction of causal factors.

5. Quantitative counterfactual designs are compatible with a restricted complexity perspective (see Chapter 1 for a succinct discussion) but have been criticized by scholars whose ontology and epistemology is situated within the general complexity perspective.