Methodological and conceptual challenges of evaluating the impact of development interventions

Methodological and Conceptual Challenges of Evaluating the Impact of Development Interventions

Jos Vaessen
Doctoral dissertation

ISBN 978 905278 968 2

Published by Datawyse / Universitaire Pers Maastricht

Methodological and Conceptual Challenges of Evaluating the Impact of Development Interventions

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen op donderdag 16 september 2010 om 14.00 uur

door

Jozef Leonardus Vaessen

UNIVERSITAIRE PERS MAASTRICHT

Promotor:
Prof. dr. Frans Leeuw

Copromotor:
Prof. dr. Robrecht Renard (Universiteit Antwerpen)

Beoordelingscommissie:
Prof. dr. Chris de Neubourg (voorzitter)
Prof. dr. Lex Borghans
Prof. dr. Valentina Mazzucato
Prof. dr. Arie de Ruijter (Universiteit van Tilburg)

Anything that happens, happens.
Anything that, in happening, causes something else to happen,
causes something else to happen.
Anything that, in happening, causes itself to happen again,
happens again.
It doesn't necessarily do it in chronological order, though.

Douglas Adams
(*Mostly Harmless*, "the fifth book in the increasingly inaccurately named Hitchhiker's trilogy"
The Hitchhiker's Guide to The Galaxy)

TABLE OF CONTENTS

PREFACE 9

ACKNOWLEDGMENTS 11

CHAPTER 1
Introduction – Methodological and conceptual challenges in impact evaluation 13

CHAPTER 2
Methodological challenges of evaluating the impact of the Global Environment Facility's biodiversity program. 43

CHAPTER 3
Interventions as theories: closing the gap between evaluation and the disciplines? 55

CHAPTER 4
Assessing the potential for experimental evaluation of intervention effects: The case of the Regional Integrated Silvopastoral Approaches to Ecosystem Management Project (RISEMP) 79

CHAPTER 5
Evaluating training projects on low external input agriculture: lessons from Guatemala. 123

CHAPTER 6
Programme theory evaluation, multicriteria decision aid and stakeholder values: a methodological framework. 137

CHAPTER 7
Reflections on the practice and methodology of impact evaluation in development 159

SAMENVATTING 169

CURRICULUM VITAE 175

PREFACE

This dissertation has come to a conclusion at a time when impact evaluation features prominently at the center of the development debate. At the time the 'oldest' study of this collection of essays was initiated (1998, Guatemala), impact evaluation was still mostly confined to the periphery of development research. By and large, donors and implementing agencies were not that interested in analyses of the effects resulting from their interventions, and most of them were content to focus on funding, implementation and outputs only. With hindsight I am therefore quite grateful for having been part of a project which was supervised by an enlightened administrator who decided to allocate funds to impact evaluation at a time when this was far from obvious. My tasks as a researcher/evaluator in this project were to conduct a baseline (1998) and ex post survey (2001) with the purpose of analyzing the causal links between intervention activities and particular intended effects.

Until the very end of the process leading to this dissertation I was in doubt about whether or not to include this specific study in the present collection of essays. My arguments against inclusion were quite straightforward. The impact design of the Guatemalan study was quite rudimentary and not really innovative, the analysis was not particularly sophisticated, with no advanced statistical analysis whatsoever (I had previously done more complicated work on larger data sets), and finally, the short time span and limited budget made it difficult to expand the scope of the analysis beyond that which was considered as strictly necessary. Yet in the end, arguments in favor prevailed. I asked myself the following questions. Is it not the purpose of policy evaluation to produce the best possible and reliable results within a context of multiple practical constraints (which distinguish an evaluation from research for purely academic purposes)? And if so, does this question not merit the attention of an academic audience?

When I started the Guatemalan assignment in 1998, I had already implemented three different survey studies in Central America. In these previous experiences I had struggled in order to master all aspects of survey research (e.g. including such aspects as designing the sample framework and questionnaire, training and managing interviewers, setting up a system of quality control of data, handling logistics, etc.). The sheer complexity and number of challenges to be addressed in a survey research process up to the production of the disk with the coveted treasure –the data set- has often struck me with awe. Thus far, I have not encountered anything that is more challenging in research than conducting a successful survey (and generating reliable data) in a rural context within a developing country such as Guatemala or Nicaragua. In Guatemala I was armed with knowledge and experience on how to address key challenges in survey research: how to obtain reliable information on sensitive topics from potentially distrustful respondents; how to maintain a representative coverage of the target population when denied access to a particular part of the territory; how to manage a group of interviewers with divergent talents and shortcomings, how to use different methods of inquiry to improve the quality of

inference, etc. In the end, I was quite proud of the whole endeavor. I carried out the field work in 1998 and 2001 and subsequent analyses with great enthusiasm and with rewarding results. My first experience in impact evaluation brought me a great sense of satisfaction.

In subsequent years, through different policy-oriented research assignments I gradually learned more about the specificities of policy evaluation in general and impact evaluation in particular. When over the last four years or so debates on impact evaluation in development started to flourish, my reaction was somewhat ambivalent. On the one hand, as a young scholar in development evaluation I enjoyed studying the expanding literature, which stimulated my learning process and the development of my own ideas about the subject. In addition, the growing attention for impact in the development community brought new opportunities for research and evaluation. I was fortunate to benefit from some of the new and exciting developments in the field of impact evaluation; in various capacities I participated in the work of the Network of Networks on Impact Evaluation (NONIE), The International Initiative for Impact Evaluation (3ie), and the impact evaluation work carried out by the Evaluation Office of the Global Environment Facility. On the other hand, I felt a sense of disappointment. The good thing was that debates on impact evaluation highlighted the need for rigorous assessment of impact. On the other hand, the meaning of rigor was reduced to a few things: the need for experimental and quasi-experimental designs in evaluation and new advances in econometric techniques. In fact, the latter aspect was regarded as second-best within the growing movement advocating 'rigorous' impact evaluation; rigor being equated with experimental (and quasi-experimental) designs. As a young scholar riding the learning curve of impact evaluation methodologies and practices I welcome the expansion of randomized experiments (and its close derivatives) in the practice of development evaluation. Yet at the same time I abhor the narrow focus of the debate. The Guatemalan case and other subsequent assignments have taught me a lot about the divergent challenges that need to be addressed in the process of generating a reliable understanding of impact. This dissertation aims to make explicit some of these challenges.

ACKNOWLEDGMENTS

This dissertation represents a collection of multiple strands of research completed in various episodes over the last twelve years. Whereas most of my research efforts in this period (especially in the first five years or so) are not included in this dissertation, in a way I find it gratifying that the dissertation spans (almost) my entire career as a 'young' scholar of development policy. Throughout this journey I have been fortunate to meet many inspiring people in a variety of institutional contexts in different countries. The list of people who have provided guidance, assistance or inspiration for my research efforts is probably longer than I am able to recall at present. Thanks to all I may have missed below.

I remember when I came back from my first extended trip to the tropics. It was 1995, and I just spent five months or so doing fieldwork in a FAO project in Costa Rica within the framework of my studies at Wageningen University. One of my first phrases to my parents was: "As soon as I graduate [I still had two years ahead of me] I will be off". I would like to thank my parents for all their support, patience and quiet encouragement throughout the years. 'In ways not measurable nor attributable' I am who I am because of them. Erna, my partner since 2002, has been a major catalytic force in my professional life. Her love, her character and steady presence have been the counterweight for my impulses. Life is yin and yang. Thank you.

I would like to extend my thanks to the following people who have been my main mentors in the 'early days'. Ruerd Ruben, supervisor of my Master thesis at Wageningen University taught me the value of methodological rigor. In addition, he opened my first doors into the fascinating and challenging world of smallholder agriculture in Central America. Jan de Groot invited me to work with him in Guatemala, where not only I was able to develop my research and evaluation skills, but also witness his inspiring leadership in a context of fragile recovery after the civil war. Johan Bastiaensen shared his enthusiasm and knowledge of Nicaragua with me. His critical views challenging vested paradigms of thought are a continuous source of inspiration.

The Institute of Development and Management at the University of Antwerp has been my main hub of operations. I would like to extend my gratitude to my co-supervisor, Prof. Robrecht Renard, current Chair of the Institute. His balanced and nuanced knowledge regarding the world of development cooperation, as well as his continued support and guidance throughout the years have been very important to me. I also would like to thank my colleagues at the Institute for their support, encouragement and for creating a pleasant enabling environment for research and other professional activities. 'Contemporaries' including An Ansoms, Björn van Campenhout, Ben D'Exelle, Filip Meheus, Sara Dewachter and more recently Geovanna Benedictis, Gert van Hecken, Martin Prowse, Ana Rivas, Lodewijk Smets, Karel Verbeke and Inge Wagemakers in their own divergent ways have contributed to my endeavors. Greet Annaert, Hugo Decraen, Joelle Dhondt, Nicole Dierckx, Patrica Franck, An Vermeesch and other support staff have offered valuable and

often last minute assistance in countless occasions. Finally, I would like to thank members of the Aid Policy group, Danny Cassimon, Nathalie Holvoet, Nadia Molenaers, and other colleagues with whom I have collaborated over the years, German Calfat, Tom De Herdt and Stefaan Marysse, for their constructive and pleasant collaboration.

At Maastricht University I would like to thank colleagues at the capacity group of Foundations and Methods of Law, especially Jaap Hage, for graciously accepting me within their ranks. Despite the fact that the content of my work is quite distinct from that of my colleagues, they have made me feel at ease and welcome. In addition, I would like to thank Chang Chiung, Han Aarts and Rogier Creemers for their enjoyable and fruitful collaboration.

I owe an enormous debt of gratitude to my supervisor, Frans Leeuw who has played a pivotal role in the process leading up to this dissertation. Available at all hours, seven days a week, he has spared no effort to provide quiet (and not so quiet) encouragement, offer intellectual feedback, help develop new pathways of thought, and help set up research and teaching partnerships. He has been an invaluable guide into the world of policy evaluation, a terrific colleague, mentor and boss. Thank you.

Over the past ten years or so, many colleagues beyond the institutions just mentioned have been of great help and inspiration. I would like to express my sincere thanks to the following persons. At the Global Environment Facility Evaluation Office: Rob van den Berg, Lee Risby, Claudio Volonte. A special word of thanks to David Todd who has collaborated with me on a number of projects. At the World Bank: Victoria Gunnarsson, Gunars Platais and Andrew Warner. At Nitlapán (Nicaragua): Miguel Alemán, Ligia Gómez, René Mendoza, Elias Ramirez and Tomás Rodriguez. At CATIE (Costa Rica): Claudia Sepúlveda and Muhammad Ibrahim.

I also would like to thank other colleagues and individuals spread across the globe who have helped me in various ways: Freddy Amador, Mateo Ambrosio, Marleen Boelaert, Anneliese Bruner, Orlando Cortez, Osvaldo Feinstein, Amanda Fitoria, Melvin Guevara, Antonie De Kemp, Hans Keune, Eva Llopis, Lissette Maliaño, Hans Nusselder, Gustavo Siles, Sandra Speer and Johan Springael.

Finally, my thanks go to Chris de Neubourg, Lex Borghans, Valentina Mazuccato and Arie de Ruijter, members of the assessment committee of this dissertation. I would also like to thank the additional members of the public defense panel of this dissertation.

To all brothers and sisters in arms gathered under the banner of the Holy Triangle (data, method and theory): may the quest continue…

CHAPTER 1

Introduction – Methodological and conceptual challenges in impact evaluation

1.1. The rising star of impact evaluation in development

The question of 'what works and why' in development assistance has received considerable attention over the past few years. The major reason is that many observers outside of development agencies believe that achievement of results has been poor, or at best not convincingly established. In the last decades of the previous century part of the development assistance paradigm was about 'thinking big', e.g. structural adjustment policies as key factors in generating stability and growth in developing countries. Correspondingly, a lot of intellectual effort went into analyses of development at the macro level, with scores of economists working on growth regressions, trying to identify the key factors that were determining country growth of GDP and the role of development assistance therein. Growing pessimism within the international community about the effectiveness of macro interventions, in part fuelled by the lack of decisive evidence from the academic community, gradually led to a shift in development paradigm towards more 'thinking small' (Easterly, 2001; Cohen and Easterly, 2009). Yet, the state of evidence on the effectiveness of concrete policy interventions (programs, policies) was far from promising either. Towards the end of the previous century the realization grew that, given the evidence base, it was hard to determine the extent to which interventions were making a difference (Baker, 2000). In 2006, an influential paper published by the Center for Global Development -"When will we ever learn?" (CGD, 2006)- pointed at an evaluation gap in development; despite enormous investments in development policy, the evidence base on what works was diagnosed as weak. According to the paper, too much of the evaluative work in development focused on process instead of results and credible evaluations of results were scarce. Fortunately, at the time of publication of this paper, the tide had already been gradually turning. A number of key events such as the endorsement of the Millennium Development Goals by the global community in 2001, the 2002 Monterrey Conference on Financing for Development, the 2005 Paris Declaration on Development Effectiveness and the 2008 High-Level Forum on Aid Effectiveness in Accra were signs of a growing results-focus in the development community, gradually paving the road for more attention to the assessment of effects of development interventions.

As a result of this evolution, in recent years debates on impact evaluation as well as its funding have flourished. Impact evaluation can be roughly defined as the (growing) field of evaluative practices aimed at assessing the intended and unintended effects of policy interventions. One of the particularly productive areas in impact evaluation is (quasi-)experimental impact evaluation, in particular randomized experiments. A number of initiatives, most notably the Poverty Action Lab (J-PAL), Innovations for Poverty Action and the World Bank's Spanish Impact Evaluation Fund, are creating a growing body of evaluative evidence based on randomized experiments.[1] The comparative advantage of the latter methodology,[2] as

1 Another fairly recent initiative, the International Initiative for Impact Evaluation, is funding proposals for research based on a broader palette of methodological designs, usually a combination of quantitative methods embedded in a theory-based approach.

has been argued widely, is its inherent strength to address the attribution problem in evaluation through counterfactual analysis. The basic idea of a randomized experiment (RE)[3] is that the situation of a participant group (receiving benefits from/affected by an intervention) is compared over time with the situation of an equivalent control group that is not affected by the intervention. Allocation to either of these groups[4] is random. Consequently, in sufficiently large samples the probability that both groups are equivalent on all observable and non-observable characteristics except for intervention participation is very high (see for example Shadish et al., 2002; Morgan and Winship, 2007). This inherent strength of REs can resolve the selection bias problem in evaluation. People that participate in or are affected by an intervention usually differ from the population at large due to self-selection or targeting. As a result, simple comparisons between participants and people not covered by the intervention will be biased. Randomization of intervention benefits or participation addresses this issue. This particular feature of REs has been lauded by growing groups of researchers and decision makers in development. Indeed, for some the rise of REs signifies the beginning of a new era in development evaluation: "[c]reating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th" (The Lancet Editorial quoting Esther Duflo, 2004: 731). Recent examples of randomized experiments include Miguel and Kremer (2004) on deworming treatment in Kenya, Banerjee et al. (2007) on education in India, Olken (2007) on monitoring corruption in Indonesia, McKenzie and Woodruff (2008) on returns to capital and access to finance in Mexico, or Karlan and Zinman (2009) on microcredit in the Philippines.

While the growing body of evidence based on REs certainly seems promising, the fact that some of the protagonists of the RE tradition, dubbed 'randomistas', perceive REs as the only way to produce rigorous evaluative evidence has led to a storm of critique.[5] Ravallion (2009a: 1) warns about the consequences of the growing influence of the 'randomistas': "Researchers are turning down opportunities to evaluate public programs when randomization is not feasible. Doctoral students are searching for something to randomize. Philanthropic agencies are sometimes unwilling to fund non-experimental evaluations." A key argument against the alleged supremacy of REs[6] expressed by critics (see below), has been that if one randomly controls for all observable and non-observable confounders, one cannot generalize

2 To keep things simple, I distinguish between methodologies and methodological designs on the one hand and methods on the other. The former can comprise multiple methods, while the latter refer to specific tools of data collection and analysis.

3 Given the scope of the interventions that are currently assessed with this methodology, I prefer the term randomized experiments (RE) instead of the "treatment-oriented" term of randomized controlled trials.

4 In fact, there may be more than two groups, as multiple "treatment" groups may be defined and monitored over time.

5 See Cohen and Easterly (2009) for a collection of essays by eminent development scholars in favor and against REs as principal tools for assessing development effectiveness.

6 Including the corresponding hierarchy of methodological designs that places randomized controlled trials at the top of the "food chain" followed by several quasi-experimental designs. A well-known example of such a hierarchy of designs is the Maryland Scale of Scientific Methods (Sherman et al., 1997).

conclusions about effectiveness beyond the specific sample, as one does not know exactly in what aspects the experimental sample differs from the population at large (see for example Deaton, 2009).

The phenomenon of REs being perceived as superior to other methods of impact evaluation by a part of the development community is not new and has also occurred in other disciplines and policy fields. REs have been applied since the early twentieth century in the fields of education, crime, health and social welfare mostly by psychologists, sociologists, economists and health scientists (see Oakley, 2000; Leeuw, 2009). Especially in evidence-based medicine REs have gained a prominent role, yet have also been criticized (see for example Worrall, 2007). In the field of international development, REs have been embraced mostly by economists. Rodrik (2008) and Deaton (2009) argue that this is due to mainly two reasons. First, they note a certain sense of disappointment among economists regarding the failure of economic theory in providing accurate and generalizable guidance on effectiveness. Second, there is growing consensus on the shortcomings in econometric methods for explaining effectiveness, mainly due to the identification problem.[7] As REs have been enthusiastically taken up by development economists, sharp critique has come from the same discipline, most notably from Rodrik (2008), Ravallion (2008, 2009a) and Deaton (2009).[8] Banerjee and Duflo (2008), two prominent 'randomistas', acknowledge much of the critique on REs. However, at the same time they counter the critique by arguing that most of it is not specific to randomized experiments but also applicable to non-experimental observational studies. However, this is exactly a point that critics refer to. The fact that randomized experiments can be justifiably criticized on a number of key issues (see below) undermines the alleged claim to epistemological supremacy, a claim that many 'randomistas' explicitly or implicitly endorse.

1.2. Scope and outline

1.2.1. Scope

The controversy surrounding REs unfortunately has kindled a sense of polarization within the development research and evaluation community, and indeed also within practitioner and policymaker circles (see for example Cohen and Easterly, 2009), which is counterproductive to achieving the important goal that so many endorse: to promote a growth of knowledge on what works and why in development. In this Chapter and elsewhere, when discussing the multiple conceptual and methodological challenges in impact evaluation, I will recurrently refer to REs and their potential to address a particular challenge. The discussion will demonstrate that REs are not equipped to address *all of these challenges*. However, this should not lead to erroneous conclusions about the utility of REs. Whereas the potential

7 Generally, this refers to the problem that multiple values of parameters or multiple models might explain a certain pattern in the data.
8 See also Cohen and Easterly (2009).

role of REs (and quasi-experimental designs)[9] in impact evaluation is probably overestimated by 'randomistas' and in some cases also policymakers, they do possess a comparative advantage in addressing the issue of attribution (internal validity), as also argued in this Chapter and elsewhere (Chapter 4).

The purpose of this dissertation is threefold:

- to illustrate that the key question of what works and why in development cannot be fully addressed by exclusively relying on *one* methodological design only;[10]
- to illustrate some limitations in the applicability of randomized experiments;
- to illustrate that impact evaluation can benefit from using an intervention theory as a guiding framework, both for randomized experiments and for other designs.

As a structure to the discussion I discern three key challenges in impact evaluation, each of which is further classified into sub issues. The classification is based on three pillars. First of all, it relies on an elaborate discussion in 2008-2009 with a community of practitioners and scholars working on impact evaluation in development.[11] The second pillar concerns a review of the current literature on impact evaluation and development. Finally, my own empirical and conceptual work on impact evaluation also served as inspiration for the structure employed below.

Challenge 1: Delimitation. The scope of an impact evaluation can widely differ depending on the nature of the evaluand, the types of effects that might occur, as well as the choices that are made about the aspects to be assessed in detail. These choices can be determined by decision makers and/or researchers and may include the priorities of other stakeholder groups such as target groups.

Challenge 2: Attribution versus explanation. REs have a comparative advantage in determining the net effects[12] of an intervention with a high degree of internal validity. Theory can help to strengthen the internal validity of findings by elucidating how and why certain changes occur. In addition, theory can strengthen the external validity of findings.[13]

---

9 Quasi-experiments do not rely on randomization but on other principles for establishing participant and control groups (which are generally considered as inferior to randomization in terms of their potential to generate unbiased estimates of impact).

10 In principle, there is nobody who disagrees with the message that there is no method which fits all needs. Moreover, all research, even research that is narrowly focused on establishing causality (internal validity), to some degree is multi-method in nature (see Cook (2006) for a compelling argument). In this dissertation I will elucidate the different dimensions and reasons why it is important to think in terms of multi-method approaches.

11 In 2008-2009, the Network of Networks for Impact Evaluation (http://www.worldbank.org/ieg/nonie , last consulted January 11, 2009) commissioned an assignment with the purpose of revising and adding new content to existing guidelines on impact evaluation. The result was a new Guidance on impact evaluation (Leeuw and Vaessen, 2009).

12 This refers to the effects of an intervention adjusted for what would have happened if the intervention had not taken place, a concept often used in counterfactual analysis.

13 Internal validity is about establishing a causal relationship between intervention outputs and processes of change leading to outcomes and impacts. External validity concerns the generalizability of findings to other settings. In addition to these two dimensions it can also be argued that theory can strengthen the construct validity (i.e. the extent to which variables adequately represent the phenomena they refer to) and statistical conclusion validity (i.e. the degree of confidence about the existence of a relationship between intervention and effect variable) of findings (see for example Cook and Campbell, 1979; Shadish et al., 2002).

Challenge 3: Impact evaluation in practice. Good impact evaluation is good research. Whereas many of the current debates on impact evaluation tend to center on arguments pro and against REs, in practice the validity of findings of any type of impact evaluation, including REs, heavily depends on the extent to which a number of key design and implementation challenges have been appropriately addressed.

In this dissertation the focus will be on methodological design and implementation aspects of impact evaluation. Less attention will be paid to the properties of statistical analysis within the context of REs or other methodological designs (see for example Heckman, 1992; Shadish et al. 2002; Cook, 2006; for development interventions see for example Deaton, 2009). The individual Chapters in this dissertation each address a number of conceptual and/or methodological challenges of impact evaluation. The scope of the dissertation in terms of the types of interventions, policy domains and contexts covered by the individual Chapters is by no means meant to be comprehensive in the sense of covering the entire landscape of development intervention and evaluation. Instead, to ensure sufficient depth of analysis as well as a certain level of logical coherence between Chapters, three common focal points or elements, besides impact evaluation, characterize this dissertation. First of all, Chapters 4, 5 and 6 are based on empirical analyses of interventions directed at smallholder (sustainable) agriculture in the tropics. References to environmental and agricultural interventions (specific or more in general) can also be found in the other Chapters 2 and 3. Second, the geographical focus of the empirical analyses is Central America, more particularly, field work in Nicaragua (and Costa Rica; Chapter 4) and Guatemala (Chapters 5 and 6). A third element concerns the institutional context. Chapters 2 and 4 discuss interventions within the framework of the Global Environment Facility, a global funding mechanism that finances projects in developing countries with respect to global environmental issues such as biodiversity, climate change and others.

Together the Chapters provide illustrations of a range of important challenges concerning the design and implementation of impact evaluations. The links between the individual Chapters and current debates in the literature are elucidated in sections 1.3. to 1.5. The next section presents short summaries of the individual Chapters.

1.2.2. Outline

This volume comprises seven Chapters including this introduction. Chapter 2 explores some of the methodological challenges that evaluators face in assessing the impacts of complex intervention strategies. These challenges are illustrated, from the context of impact evaluation of one of the six focal areas of the Global Environment Facility; its biodiversity program. The discussion is structured around the concepts of attribution and aggregation, offering the reader a framework for reflection. Subsequently, the Chapter discusses how theory-based evaluation can provide a basis for addressing the attribution and aggregation challenges presented.

Subsequently, Chapter 3 takes up the theme of theory-based evaluation, presenting a detailed discussion of this approach: what is it about, what are the roles of theory in evaluation and how can we meaningfully theorize about impact. It elabo-

rates on the core idea that interventions can be perceived as theories and evaluations are about testing these theories. The second part of the Chapter focuses on how a theory on processes of change and impact can be usefully reconstructed on the basis of Coleman's work on social mechanisms (1986, 1990).

Chapter 4 discusses the merits and limitations of an experimental design as a basis for impact evaluation for the specific case of a recently completed project in Latin America funded by the Global Environment Facility. The purpose of the design underlying the project was to facilitate the analysis of the impact of project incentives provided to farmers (i.e. payments for environmental services and technical assistance) on land use changes and ultimately on environmental and socio-economic benefits. The Chapter illustrates how the analytical potential of experimental designs can be compromised if one does not pay sufficient attention to design and implementation and the corresponding potential threats to the validity of findings resulting of an experiment. It is also argued that the combination of theory-based evaluation with an experimental design approach can maximize the attribution and explanatory potential of an impact evaluation.

Chapter 5 discusses an impact evaluation study of a training project on external input agriculture (LEIA) in Guatemala. The Chapter illustrates how a mixed methods perspective, combining a simple quasi-experiment with qualitative methods such as semi-structured interviews can provide a useful basis for data collection and analysis. Moreover, the combination of triangulation between insights from different methods and existing research informing the design and analysis of the evaluation, constitutes a strong framework for assessing the impact of small-scale interventions given the constraints of budget, data and time.

Chapter 6 presents a methodological framework for systematically incorporating stakeholder values in the ex ante evaluation of the potential effects of alternative policy options. Nowadays there are a number of evaluation approaches that specifically focus on the elicitation of stakeholder values. Multicriteria decision aid enables evaluators to go one step further by systematically showing how different stakeholder values would affect evaluative outcomes and subsequent policy decisions about future policy interventions. An impact theory of an existing intervention can constitute a valuable framework for assessing the potential effects of future policy interventions with similar characteristics.

Finally, Chapter 7 is the epilogue to this volume. In this short Chapter some of the insights that can be derived from the previous Chapters on the methodology and practice of impact evaluation are synthesized into succinct topics for further reflection and discussion.

The contributions of the individual Chapters to the various debates on methodological and conceptual challenges in impact evaluation are described below, for each of the three main challenges that are discussed.

1.3. Challenge 1: delimitation

1.3.1. The importance of stakeholder values

A first important criterion for delimitation in impact evaluation concerns the question of 'impact according to whom'. There is a strong movement in development research and practice which endorses the idea that impact evaluation is not only about assessing the effects of an intervention but also about underlying questions of what types of processes of change and effects are valued as important (either positive or negative) and by whom?[14]

This line of thought is most manifest in the evaluation tradition of participatory impact evaluation. Nowadays, participatory methods have become 'mainstream' tools in development in almost every area of policy intervention. The roots of participation in development lie in the rural sector, where Chambers (1995) and others developed the now widely used principles of Participatory Rural Appraisal.[15] Participatory evaluation approaches (see for example, Cousins and Whitmore, 1998) are built on the principle that stakeholders should be involved in some or all stages of the evaluation. In the case of impact evaluation participation includes aspects such as the determination of objectives, indicators to be taken into account, as well as stakeholder participation in data collection and analysis. In practice it can be useful to differentiate between stakeholder participation as a process and stakeholder perceptions and views as sources of evidence (Cousins and Whitmore, 1998).

Randomized designs are not about stakeholder participation or elicitation of stakeholder values. However, there is no reason to assume that REs may not be combined with participatory processes and methods of data collection (Karlan, 2009). In practice a wide variety of methods is available (see for example IFAD, 2002; Mikkelsen, 2005; Pretty et al., 1995; Salmen and Kane, 2006). Stakeholder participation in impact evaluation can be beneficial in many ways, i.e. by enhancing the ownership and (possibly) utilization of an evaluation, improving the quality of the data collected from target populations, or strengthening local processes of governance. At the same time, participatory methods have been criticized on many grounds. Often mentioned critical aspects concern the limited applicability of impact evaluations with a high degree of participation especially in large-scale, comprehensive, multi-site interventions. In such contexts, organizing processes of stakeholder participation may not be feasible due to high costs and logistical barriers. In addition, there is some doubt about the reliability of information based on stakeholder perceptions (e.g. due to risks of strategic responses, manipulation of information or advocacy by stakeholders).[16]

---

14 The importance of stakeholder values may differ according to the type of intervention. For example, whereas stakeholder views on what is important in the evaluation of the effects of reforestation programs may widely differ, this may be less the case for health indicators in the case of nutrition programs.

15 Participatory Impact Assessment is an extension of Participatory Rural Appraisal and involves the adaptation of participatory tools combined with more conventional statistical approaches specifically to measure the impact of humanitarian assistance and development projects on people's lives (Catley et al., 2008).

16 In turn, proponents of participatory evaluation approaches are often very critical of the reliability of survey-based research and corresponding data analyses within the framework of experimental or non-experimental methodological designs. Especially well-known is Robert Chambers' (1983) discussion of what

An alternative approach for eliciting stakeholder values which does not rely on stakeholder participation is values inquiry and "refers to a variety of methods that can be applied to the systematic assessment of the value positions surrounding the existence, activities, and outcomes of a social policy and program" (Mark et al., 1999: 183). Values inquiry exercises may be more useful than participatory evaluation approaches in situations where policy makers are interested in a representative picture of the value positions of large groups of beneficiaries dispersed over large territories.

1.3.2. The impact of what?

When talking about the scope and delimitation of impact evaluation it is useful to address the following two questions: the impact *of* what and the impact *on* what?[17] Regarding the impact *of* what, today more than ever one can speak of a 'continuum' of interventions. At one end of the continuum are relatively simple projects characterized by 'single strand' initiatives with explicit objectives, carried out within a relatively short timeframe, where interventions can be isolated, manipulated and measured. An impact evaluation in the agricultural sector for example, will seek to attribute changes in crop yield to an intervention such as a new technology or agricultural practice. In a similar guise, in the health sector, a reduction in malaria will be analyzed in relation to the introduction of bed nets. For these types of interventions, experimental and quasi-experimental designs may be appropriate for assessing causal relationships. At the other end of the continuum are comprehensive programs with an extensive range and scope (increasingly at country, regional or global level), with a variety of activities that cut across sectors, themes and geographic areas, and emergent specific activities. Many of these interventions address aspects that are assumed to be critical for effective development yet difficult to define and measure, such as human security, good governance, political will and capacity, sustainability, and effective institutional systems.

One of the trends in development is that donors are moving up the 'aid chain'. Whereas in the past donors were very much involved in 'micro-managing' their own projects and (sometimes) bypassing government systems, nowadays a sizeable chunk of aid is allocated to national support for recipient governments. Attention to some extent has shifted from micro-earmarking (e.g. donor money destined for an irrigation project in district *x*) to meso-earmarking (e.g. support for the agricultural sector) or macro-earmarking (e.g. support for the government budget to be allocated according to country priorities). Besides a continued interest in the impact of individual projects, donors, governments and nongovernmental institutions are increasingly interested in the impact of comprehensive programs, sector strategies or country strategies, often comprising multiple instruments, stakeholders, sites of intervention and target groups (see Jones et al. (2008) for a recent inventory of impact evaluations in different sectors of development intervention).

---

he labeled as 'survey slavery', criticizing among other things the costs (also for respondents), inefficiencies, rigidities and data problems of survey research in rural development contexts.

17 In Chapter 2 they are called respectively the independent and the dependent variable problem.

In most countries donor organizations are (still) the main promoters of impact evaluation. The partial shift of the unit of analysis to the macro and (government) institutional level requires impact evaluators to pay more attention to complicated and more complex interventions at national, sector or program level. Multi-site, multi-governance and multiple (simultaneous) causal strands are important elements of this (see Rogers, 2008). At the same time, the need for more rigorous impact evaluation at the level of 'single strand' projects or activities remains as important as ever since they are the building blocks of higher-level programs and policies. Furthermore, the ongoing efforts in capacity-building on national M&E systems (see Kusek and Rist, 2004; Morra and Rist, 2009) and the promotion of country-led evaluation efforts stress the need for further guidance on impact evaluation at 'single intervention' level.

Within the light of the heterogeneous landscape of interventions, critics and (most) proponents of REs alike acknowledge the limitations in applicability of REs. What is problematic is that the special status attributed to REs by some researchers and policymakers is likely to generate a bias in terms of too much evaluative focus on interventions that are amenable to this approach. Already development evaluation is biased in terms of what Ravallion calls a " 'myopia bias' in our knowledge, [with evaluation] favoring development projects that yield quick results" (Ravallion, 2008: 6). Similarly, Blattman (2008) refers to the 'overevaluation' of certain economic, educational and health interventions and the 'underevaluation' of interventions on peace-building, crime reduction, and governance issues (e.g. public management, decentralization; see also Jones et al., 2008). REs are most readily applicable in case of discrete, homogenous interventions with clearly delineated target groups[18] rather than more complicated interventions, interventions that evolve during implementation or full-coverage interventions such as laws or macroeconomic policies (Bamberger and White, 2007; Rossi et al., 2004). Even in the case of rather simple interventions, quantitative researchers such as those handling REs find themselves opposed by anthropologists and sociologists who criticize the rather simplistic view of deconstructing interventions into neatly delineated packages of activities, benefits or treatments (see for example Hulme (2000) for a discussion). In contrast, they emphasize the complexity in social development interventions. Staff in such interventions "are not functionaries dutifully providing a standardised service, such as immunising babies or distributing food rations; they are instead engaged in extensive face-to-face interaction with villagers over many months, making innumerable discretionary decisions. In many respects 'the project' is itself a dynamic decision-making process rather than a static 'product', and as such attempts to make causal claims regarding overall impact must address endemic unobserved heterogeneity bias. In short, on both the 'demand side' (local context) and the 'supply side' (front-line project staff) there is, by design, enormous variation" (Woolcock, 2009: 5). Indeed, identifying what the intervention exactly is, where it begins and where it ends can be rather challenging (Pawson and Tilley, 1997).

18 This in line with what Ravallion (2009b) calls assigned programs.

In case of complicated interventions (comprising multiple (interacting) intervention components) at best sometimes only some of the components may be amenable to a RE. In case of interventions that focus on the institutional level (e.g. capacity-building, technical assistance, administrative reform), corresponding evaluations look at one or a few units of analysis (i.e. the institution) -Bamberger and White (2007) call this the small *n* problem- a situation that is not amenable to a RE. Homogeneity of the intervention is another frequently referred to condition for REs. A proper RE requires a high degree of homogeneity in intervention, target groups and context over time, conditions which are unlikely to hold in many cases. Often the nature of the intervention changes over time through adaptive learning, political pressures or to obtain more funding (e.g. adding training to subsidy programs). In addition, there might be changes in contextual factors that affect participants differently than control group members. For example, rising output prices might speed up technology adoption processes among participants of a training program more than in the case of control group members who are facing a knowledge constraint. Given the above, if anything, a narrow focus on REs would reinforce an already existing evaluation bias towards particular types of interventions.

1.3.3. The impact on what?

1.3.3.1. Institutional versus beneficiary level effects

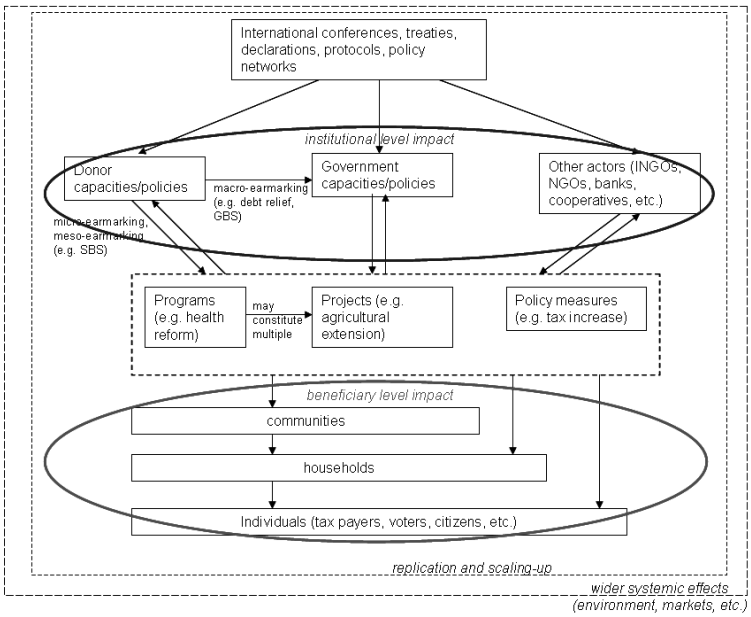
The second issue, the impact *on* what, concerns the type of effects that we are interested in. The causality chain linking policy interventions to ultimate policy goals (e.g. poverty alleviation) can be relatively direct and straightforward (e.g. the impact of vaccination programs on mortality levels) but also complex and diffuse. Impact evaluations of for example sector strategies or general budget support potentially encompass multiple causal pathways resulting in long-term direct and indirect impacts. Some of the causal pathways linking interventions to impacts may be fairly straightforward (e.g. from training programs in alternative income generating activities to employment and to income levels), whereas other pathways are more complex and diffuse in terms of going through more intermediate changes, and being contingent upon more external variables (e.g. from stakeholder dialogue to changes in policy priorities to changes in policy implementation to changes in human welfare).

Given this diversity it is useful for purposes of ´scoping´ to distinguish between two principal levels of impact: impact at the institutional level and impact at the beneficiary level (see Figure 1).[19] It broadens impact evaluation beyond either measuring whether objectives have been achieved or assessing direct effects on intended beneficiaries. It includes the full range of impacts at all levels of the results chain, including ripple effects on families, households and communities, on institutional, technical or social systems, and on the environment. Interventions that can be labeled as institutional primarily aim at changing second-order conditions (i.e. the capacities, willingness, and organizational structures enabling institutions to

19 In addition, one can discern other "levels" such as replicatory effects and systemic effects.

design, manage and implement better policies for communities, households and individuals). Examples are policy dialogues, policy networks, training programs, institutional reforms, and strategic support to institutional actors (i.e. governmental, civil society institutions, private corporations, hybrids) and public private partnerships. Other types of interventions directly aim at/affect communities, households, individuals, including voters and taxpayers. Examples are fiscal reforms, trade liberalization measures, technical assistance programs, cash transfer programs, construction of schools, etc.

Figure 1. Levels of intervention, programs and policies and types of impact



Source: Leeuw and Vaessen (2009)

1.3.3.2. Intended versus unintended effects

A widely endorsed reference to impact evaluation concerns the OECD-DAC definition, which defines impacts as (OECD-DAC, 2002: 24) "[p]ositive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended". However, when we look at the body of research under the banner of impact evaluation, a large part of it is not on long-term results, nor on indirect and unintended results. In fact, the body of evaluation research based on REs and quasi-experiments is usually about analyzing the attribution of short-term outcomes to a particular intervention (White, 2009). Moreover, in a difference in difference estimation of the net effects of an intervention, only indicators of effects that are expected to be important are taken into ac-

count. As a result of the rather limited and rigid set of indicators employed within the framework of REs and quasi-experiments they are not very useful for identifying and analyzing unanticipated effects (see for example Davidson, 2006; Bamberger and White, 2007).

Policymakers are very much interested in the indirect diffusion, replication or scaling-up effects of interventions. Whether intended or unintended they usually are 'under' the radar of REs as the analysis of these effects requires broadening the scope of indicators as well as the sampling framework of an impact evaluation. Replicatory effects in terms of behavioral changes in actors beyond the original target group can stem from market responses (given that participants and non-participants trade in the same markets), the (non-market) behavior of participants/non-participants or the behavior of intervening agents (governmental/NGO). For example, "aid projects often target local areas, assuming that the local government will not respond; yet if one village gets the project, the local government may well cut its spending on that village, and move to the control village" (Ravallion, 2008: 10). Another example concerns displacement effects of environmentally damaging land use towards other areas beyond the grasp of an intervention; deforestation may increase elsewhere as land use becomes more restricted in certain areas.

1.3.3.3. Short-term versus long-term effects

In some types of interventions, effects emerge quickly. In others effects may take much longer to become manifest, and change over time. The timing of the evaluation is therefore important. Development interventions are usually assumed to contribute to long-term development (with the exception of humanitarian disaster and emergency situations). However, focusing on short-term or intermediate outcomes often provides for more useful and immediate information for policy- and decision-making. However, intermediate outcomes may be misleading, often differing markedly from those achieved in the longer term. Many of the impacts of interest from development interventions will only be evident in the longer-term, such as environmental changes, or effects on subsequent generations. Searching for evidence of such effects too early might mistakenly conclude that interventions have failed.

In this context, the exposure time of an intervention to be able to make an impact is an important point. A typical agricultural innovation project that tries to change farmers' behavior with incentives (training, technical assistance, credit) is faced with time lags in both the adoption effect (farmers typically are risk averse and face resource constraints and start adopting innovations on an experimental scale) as well as the diffusion effect (other farmers want to see evidence of results before they copy). In such gradual non-linear processes of change with cascading effects, the timing of the ex post measurement (of land use) is crucial. Ex post measurements just after project closure could either underestimate (full adoption/diffusion of interesting practices has not taken place yet) or overestimate impact (as farmers will stop investing in those land use practices that are not attractive enough to be maintained without project incentives).

Woolcock (2009) has recently highlighted a related problem. He argues that REs, and especially those that are based on a limited number of data points in time (e.g. before and after only), do not take into account the nature and dynamics of processes of change induced by an intervention. Processes of change are often not linear. In practice, processes of change often resemble j-curves or step functions. Examples of such processes are the effects of microcredit on empowerment (e.g. initial resistance by men until persistent and collective pressures lead to a shift in norms) or the adoption of new agricultural technologies (e.g. Rogers, 2003). The implication for REs is that for example if ex post measurement happens to take place when the change curve has hit a (temporal) low, then estimates of net effects will be entirely unrealistic. REs that are not supported by theory or data from additional methods of inquiry are not equipped to address the abovementioned issues.[20]

1.3.4. The contribution of this dissertation

Chapter 2 discusses the methodological challenges in evaluating the impact of a portfolio of biodiversity interventions. The objective of the evaluation is to assess the effects of biodiversity interventions on environmental variables. Given the heterogeneity within the portfolio and the limited budgets for evaluation, striking a balance between scope and depth was an important driver in the evaluation. As a result, delimitation issues such as the 'impact of what' and 'impact on what' merit a lot of attention. A range of delimitation issues are discussed. Of particular interest is the level of analysis. This aspect is discussed from two angles. First of all, in the choice of the level of analysis, administrative levels such as projects and operational programs are compared with alternative levels of analysis such as strategic priorities and policy instruments that recur across interventions. Second, the alternative foci of effects at the levels of institutional and individual behavior versus effects at the level of environmental benefits and costs are discussed. Causality between interventions and the latter level of effects is much more indirect than in the case of institutional and individual behavioral change.[21] Moreover, environmental change is often long term, non-linear, uncertain and therefore difficult to capture. The Chapter discusses the methodological approach of theory-based evaluation as a framework for delimitation and analysis at different levels (e.g. recurring policy instruments, recurring intervention strategies).

Chapter 3 provides an in-depth account of theory-based evaluation. Regarding the challenge of delimitation, it is illustrated that an intervention theory can provide a useful framework for deconstructing processes of change into different types of mechanisms. More particularly, the deconstruction of processes of change into macro-micro, micro-micro and micro-macro linkages can be helpful in structuring the collection and analysis of evidence and facilitate a better understanding of how interventions induce processes of change.

---

20 It is important to mention that REs using a simple difference in difference estimate of net impact (ex ante versus ex post) are more prone to error than designs that rely on multiple waves of measurement within participant and control groups.

21 Behavioral changes are mediator variables of environmental changes.

Chapter 4 discusses a RE underlying a concrete project with respect to its potential to assess the impact of monetary incentives and technical assistance on land user behavior of farmers. Apart from the RE's potential to address the challenge of attribution, the Chapter also explores its potential to address the issue of impact in a broader sense, i.e. as defined in the OECD-DAC definition of impact. Several limitations of REs in addressing impact in a broad sense are illustrated from the perspective of the project. It is argued that a number of aspects of interest to policy-makers would require complementary or additional data collection and analysis embedded in a theory-based approach. Examples of important issues beyond the scope of the RE are displacement effects (e.g. degenerative land use by farmers increasing elsewhere, beyond the scope of the experiment), the factors that explain how and why project incentives affect farmers in different ways, and the (expected) sustainability of land use changes and corresponding environmental effects.

Chapter 6 presents an alternative methodological perspective on impact evaluation as it focuses on ex ante impact evaluation of policy alternatives. A key argument is that evidence on (potential) effectiveness is not enough as a basis for decision-making. In practice, stakeholder values play an important role. The Chapter discusses an original methodological framework which enables decision makers to systematically discuss policy alternatives based on assessments of alternatives on different criteria corrected for the relative importance of these criteria according to groups of stakeholders.

1.4. Challenge 2: Attribution 'versus' explanation

1.4.1. Addressing the attribution challenge with randomized experiments

The OECD-DAC definition of impacts mentioned above refers to the 'effects produced by' an intervention, stressing the attribution aspect. This implies an approach to impact evaluation which is about attributing impacts rather than 'merely' assessing what happened.[22] Multiple factors can affect the livelihoods of individuals or the capacities of institutions. For policy makers as well as for stakeholders it is important to know what the added value is of the policy intervention apart from these other factors. The attribution problem is often referred to as the central problem in impact evaluation. The central question is to what extent can changes in outcomes of interest be *attributed* to a particular intervention?[23] Attribution refers both to isolating and estimating accurately the particular contribution of an intervention and ensuring that causality runs from the intervention to the outcome. In most contexts, adequate empirical knowledge about the effects produced by an intervention requires at least an accurate estimate of what would have occurred in the absence of the intervention (the counterfactual) and a comparison with what has occurred with the intervention implemented.

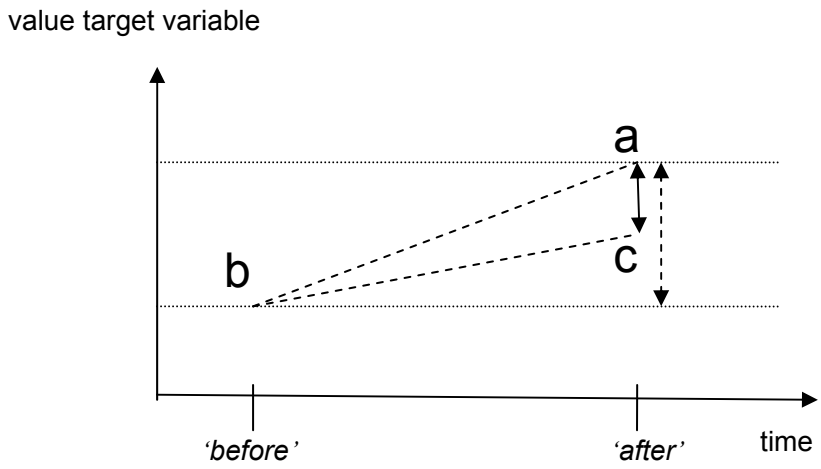
22 Goal achievement can be assessed without the need for attribution analysis as no differentiation is made between whether changes are to due to the intervention or other factors.

23 Attribution is about causal relationships between intervention and effects. How and to what extent the attribution challenge has been addressed in an impact evaluation determines the internal validity of findings.

Box 1. The attribution problem

Analyzing attribution requires comparing the situation 'with' an intervention to what would have happened in the absence of an intervention, the 'without' situation (the *counterfactual*). Such comparison of the situation 'with and without' the intervention is challenging since it is not possible to observe how the situation would have been without the intervention, and has to be constructed by the evaluator. The counterfactual is illustrated in the Figure below. The value of a target variable (point a) after an intervention should not be regarded as the intervention's impact, nor is it simply the difference between the before and after situation (a-b, measured on the vertical axis; the dotted arrow). The net impact (at a given point in time) is the difference between the target variable's value after the intervention (a) and the value the variable would have had in case the intervention would not have taken place (c).

Figure - Graphical display of the net impact of an intervention



Source: adapted from Leeuw and Vaessen (2009)

Usually interventions target particular groups, and in addition self selection effects may occur as more motivated or socially well-positioned individuals or groups gain access to a particular program. Consequently, one cannot simply compare the situation of participants over time with the population at large. Estimates would be distorted due to this selection bias problem. The safest way to avoid selection effects is a randomized selection of intervention group and control before the experiment starts. When the experimental group and the control group are selected randomly from the same eligible population, in sufficiently large samples both groups

will have similar average characteristics (except that one group has been subjected to the intervention and the other has not). Consequently, in a well-designed and correctly implemented RE a simple comparison of average outcomes in the two groups can adequately resolve the attribution problem and yield accurate estimates of the net effect of the intervention on a variable of interest: by design, the only difference between the two groups was the intervention.

This powerful feature of REs explains the increasing popularity of this methodology. With a long tradition in medicine and public health and a much younger tradition in policy fields such as education and crime and justice (Leeuw, 2009), REs are now also increasingly applied in the context of (social) development interventions. Randomization is not always feasible, but a wide variety of quasi-experimental designs is available to ensure a high internal validity of findings. Basically, designs differ in terms of the technique used for creating comparable groups (e.g. regression discontinuity, propensity score matching, pipeline approaches) as well as in terms of the structure of periodic measurement within participant and control groups (e.g. simple ex ante – ex post participant group design, interrupted time series design; see for example Campbell, 1969; Cook and Campbell, 1979; Shadish et al., 2002, Morgan and Winship, 2007; for development interventions see for example Bamberger et al., 2006).[24]

1.4.2. Internal versus external validity

REs are typically equipped to address the question of what works within the particular confines of the experiment; they are strong on internal validity. However, policymakers are often interested in other questions (Heckman, 1992; Heckman et al., 1997; Ravallion, 2009b). Does this intervention also work in other regions or contexts? What happens when we go to scale with a particular intervention? What are the determinants of effectiveness? Another typical question that might not be easily answered with REs or quasi-experiments, is whether and how people are differently affected by an intervention.[25] This question can be answered with additional quantitative data analysis, if (large) data sets are available which allow for extensive modeling of confounders and interaction effects. Alternatively (or in addition), many qualitative methods such as case study methods can help evaluators to study in detail how interventions work differently in different situations.

Without further information results of a RE cannot be generalized beyond the experimental setting, as important confounders that are controlled for in the experiment are not revealed by the experiment itself. Moreover, "[t]he people who are normally attracted to a program, taking account of the expected benefits and costs to them personally, may differ systematically from the random sample of people

24 Most quasi-experimental techniques are useful when selection characteristics are known and can be measured. Even in the case of unobservable characteristics which might differ between groups and affect intervention effects, this may not be a problem. If these characteristics are time invariant, in principle they can be controlled for by double differencing or multiple data points in time.

25 This is an issue that is closely related to the idea of external validity. If one knows how an intervention affects groups of people in different ways, then one can more easily generalize findings to other similar settings.

who were included in the trial" (Ravallion, 2008: 17). Critics of the 'alleged superiority' of REs argue that internal validity is not typically the question that policy makers are interested in and argue for more attention to external validity of findings, an aspect on which REs enjoy no comparative advantage (e.g. Rodrik, 2008; see Shadish et al. (2002) for a discussion on experiments and external validity). They argue that in order to be able to generalize findings from a RE to other settings, one needs to know how an intervention works, what are the determinants of the processes of change (possibly induced by an intervention), how an intervention might affect people in particular circumstances in different ways and what the time path and nature of the changes might be. In order to answer these types of questions (and others, see Ravallion, 2009b) one needs an informative explanatory theory (e.g. based on research within the social sciences) and other (additional) methods of data collection and analysis (e.g. Van der Knaap et al., 2008; for the case of development interventions see Deaton, 2009).

'Randomistas' have asserted that external validity of findings can be enhanced by doing a series of experiments in different contextual settings (Banerjee and Duflo, 2008). Yet as Ravallion argues, "the feasibility of doing a sufficient number of trials—sufficient to span the relevant domain of variation found in reality for a given program, as well as across the range of policy options—is far from clear. The scale of the randomized trials needed to test even one large national program could well be prohibitive" (Ravallion 2008: 19). Moreover, as Rodrik argues, "[f]ew randomized evaluations—if any—offer a structural model that describes how the proposed policy will work, if it does, and under what circumstances it will not, if it doesn't. Absent a full theory that is being put to a test, it is somewhat arbitrary to determine *under what different conditions* the experiment ought to be repeated." (Rodrik, 2008: 21-22; italics added). A similar point is made by Deaton who asserts that "repeated successful replications of a 'what works' experiment is both unlikely and unlikely to be persuasive. Learning about theory, or mechanisms, requires that the investigation be targeted towards that theory, towards why something works, not whether it works" (Deaton, 2009: 31).

A special type of generalizability concerns the inference from small-scale pilots to up-scaled programs with broad coverage. In development contexts REs are often implemented on small scale pilots. While the outcomes of REs are considered as useful inputs in the decision to go to scale, REs alone are insufficient to support such a decision.[26] In case of scaling-up, conditions at both the institutional level (of implementing agencies) as well as the beneficiary level invariably change, making the new intervention altogether different from the small-scale pilot it was derived from (Deaton, 2009). At the institutional level for example, the organizational setup might change completely, with new people with divergent capacities and interests managing and implementing the new intervention, or corrupt public officials suddenly being drawn to a scaled-up intervention which has appeared on their radar. In addition, new contextual conditions and characteristics of target groups may con-

26 In fact, there are good examples of REs convincing policymakers to scale up and replicate interventions. For example, the spread of conditional cash transfer programs over Latin America is in part fuelled by the evidence produced by REs.

found intervention effects, this new heterogeneity not being covered by the original small-scale experiment (Ravallion, 2009a). On the other hand, it can be argued that a RE which has covered a sizeable sample can be quite informative on the likelihood that the same policy instrument will produce similar results when scaled up within the same population the RE sample was taken from.[27]

1.4.3. Interventions as theories

Most critics of REs subscribe to the point of view that there is nothing that precludes REs from being more theory-based or -driven. Of course, in such cases it would be the combination of theory (or theories) and RE that would increase the external validity (and also internal validity) potential of REs rather than the RE format alone. Recently, more examples of theory-based REs can be found in the literature (see for example Banerjee (2005) or Banerjee and Duflo (2008) for illustrations).[28] Moreover, as argued by Cook (2006) REs are never completely theory-empty.[29] Among other things, they require substantive theory as guidance for selecting or constructing suitable measures of effects.

While REs in principle are not or need not be theory-empty, the abovementioned critique of Deaton and others on REs remains valid. REs are geared towards the question of whether an intervention works and do not shed much light on how and why interventions work. In order to look at the latter question, evaluation researchers need to look into the black box between output, outcome and impact indicators and reconstruct the underlying causality (Pawson and Tilley, 1997). Two types of theory are of importance here. A reconstructed intervention theory should adequately represent the main assumptions of decision makers, target groups and other stakeholders about causal pathways running from intervention to effects (see for example, Chen, 1990, Rogers et al., 2000; Leeuw, 2003). The intervention theory can be further enriched or tested by taking into account existing explanatory theories (from the social and behavioral sciences) on intervention contexts, processes of change and potential effects.

The idea that existing explanatory theories can improve the quality of evaluations is not a new one and has been discussed quite extensively in the literature (e.g. Riggin, 1990; Lipsey, 1993; Donaldson and Lipsey, 2006). For example, Riggin (1990) discusses how Etzioni's theory of compliance can improve the quality of an evaluation of an employment assistance program both at the stage of conceptualization as well as at the interpretation stage of an evaluation. A first important role for 'theory' lies in the conceptualization of an evaluation question. Substantive theories help to point the evaluator toward the relevant constructs and relationships between

27 Thereby assuming that first a random sample was taken from the population, with subsequent randomized allocation of the intervention to treatment and control groups within the sample.

28 One example is Duflo and Hanna (2008) on using an experiment to develop a model of teacher behavior.

29 Theory-empty refers here to a situation in which the causal relationships between intervention outputs, outcomes and impact are not made explicit in a theoretical framework. Such a framework may be based on reconstructions of stakeholders' assumptions and/or existing research relevant to the causal pathways in question. For example, an analysis in which a claim to a relationship or even a causal relationship between two variables is based on statistical association only (whether in an experimental setting or not), can be called theory-empty (for a discussion see for example Coleman, 1986).

these constructs in order to make useful abstractions of the reality of a policy intervention, its intended effects and the wider context in which it is embedded, aspects which subsequently can be tested through empirical research. Substantive theory from past (evaluation) research to some extent can help to anticipate these effects, which subsequently can be taken into account by means of additional data collection. A second important role for theory lies in reinforcing the causal analysis, the analysis of how and to what extent changes in target variables can be attributed to a policy intervention. Relevant substantive theories can shed light on the nature of the causal relationships between interventions and (un)intended processes of change and help to rule out rival explanations for changes in target variables.[30] Finally, there is an important role for theory in the interpretation of evaluation findings. Theory can provide a useful framework for helping us to understand why certain changes have come about or provide insights into the relevant (contextual) variables which are likely to influence patterns of results across settings.

1.4.4. The law of comparative advantages: theory-based and multi-method evaluation

My account so far has gradually led us to the realization that REs are potentially strong on internal validity yet miss out on providing (strong) evidence on other aspects such as for example unanticipated and long-term effects and the external and construct validity of findings. This brings us to the important realization that impact evaluations by default cannot and should not exclusively rely on *one methodology* only.[31] The remainder of this section will add some more illustrative power to this argument.

As discussed elsewhere (Leeuw and Vaessen, 2009), the intervention theory constitutes the guiding framework of impact evaluations. Typically, multiple causal assumptions emerge that require further testing in order to be able to claim whether the intervention has induced certain changes, and in what circumstances. However, not all causal assumptions can be tested with the same methodological design or specific method. For example, consider the effects of subsidies for sustainable land use on biodiversity. One can envisage a useful RE which tests the effects of subsidies on the land use behavior of farmers. The subsequent causal step from land use behavior to biodiversity cannot be tested so easily by means of a RE. One of the main reasons is that biodiversity depends on plot-specific variables (e.g. the type of vegetation on a certain plot of land) as well as on determinants at higher levels such as the connectivity between systems of land use, the proximity of certain bio-

30 An example of such an approach is Scriven's (2008) General Elimination Methodology. See also Pawson's (2006) discussion on adjudication between rival theories. In an evaluation of the impact of social funds, Carvalho and White (2004) reconstruct a theory and an 'anti-theory', which are both put to the test empirically, in order to arrive at a better understanding of the nature of change processes.

31 "[I]t is trivial to argue about whether evidence-based research should be multi-method or not. Even causal research is deepened by learning about non-causal issues, such as what the substantive theory behind a program design is, who gets to participate in it or not, how well the program is implemented, who gets greater exposure, and what the program costs. So nearly all causal studies require multiple methods that complement each other. Multi-method, complementary research is desirable even when a causal claim is centrally at issue" (Cook, 2006: 1).

spheres, the geographical location with respect to migration routes of birds and other species, and so on. Moreover, as argued above, there are other challenges such as the non-linearity, uncertainty and sustainability of changes.

Ideally, the intervention theory should provide guidance on the choice of causal assumptions to be analyzed (see Weiss (2000) for a discussion on this) and correspondingly, different designs and methods can be used to assess specific assumptions. In addition, other complementary perspectives on the use of multiple methods can be discerned in the literature (for a general discussion see for example Tashakkori and Teddlie, 2003).

Let us briefly illustrate three 'logics' for multi-method approaches in impact evaluation. The first starts out from Campbell's framework of different types of validity. As suggested earlier in this Chapter, specific methods have comparative advantages in ensuring a high degree of internal/external/construct validity. Consider a similar example as introduced above, i.e. an intervention that provides monetary incentives and training to farmers in order to promote land use changes leading to improved livelihoods conditions as well as other effects. Simplified, a comprehensive methodological design could be the following:

- E.g. a randomized experiment with the use of survey data on participant and control groups could be used to assess the effectiveness of different incentives on land use change and/or socio-economic effects of these changes *(potentially strengthens internal validity of findings);*
- E.g. further survey data (multivariate) analysis and case studies could tell us how incentives have different effects on particular types of farm households *(potentially strengthens internal validity and increases external validity of findings);*
- E.g. targeted syntheses of existing research, semi-structured interviews and focus group conversations could tell us more about the nature of effects in terms of production, consumption, poverty, environment etc. *(potentially enhances construct and internal validity of findings).*

A second framework of multi-method evaluation has been labeled the 'shoestring' approach (see Bamberger et al., 2004). It refers to a number of scenarios of multi-method approaches which are adapted to particular conditions of budget, time and data constraints. These scenarios all rely on an intervention theory model as a basis for different methodological strategies to simplify evaluation design (e.g. in relation to text book REs), reduce costs of data collection and analysis, and integrate qualitative and quantitative methods.[32]

A third illustration of a multi-method perspective is Woolcock's account of different options to gain a better understanding of the nature of causal pathways. "There are at least three entry points, each of increasing degrees of sophistication. The first is simply raw experience: seasoned project managers should have a good sense of how long and in what form the impacts associated with a particular project in a particular context should take to materialise. [...] Astute intuition and seasoned field experience combined with solid theory should provide a second avenue: the

32 See also Bamberger et al. (2009) for further discussion on mixed methods in the context of impact evaluation.

very essence of a good theory should be that it provides a sense and a justification of the conditions under which, and mechanisms by which, certain project outcomes should be expected. [...] Thirdly, the regular collection of empirical evidence can itself be a basis for determining the shape of the project's impact trajectory, and is ultimately (for researchers at least) the most defensible basis on which to do so" (Woolcock, 2009: 7-8).

1.4.5. The contribution of this dissertation

Chapter 4 discusses the analytical potential of a RE for a concrete project in Latin America. The Chapter shows how different threats to validity may compromise the analytical potential of the design. This is the case for several of the group comparisons the RE was designed for (see below). However, for one of the group comparisons which is (relatively) unaffected by threats to (internal) validity, the Chapter shows how group-based comparisons over time can be very useful for providing unbiased estimates of the project's effect on land use changes, a powerful comparative advantage of REs. In addition, the Chapter shows how the analytical potential of the RE can be further strengthened by additional methods of inquiry and using an intervention theory as framework for argumentation and analysis. In fact, both Chapters 3 and 4 illustrate how impact evaluations and in particular REs (Chapter 4) can benefit from a theory-based approach to evaluation, potentially enhancing the internal, external and construct validity of findings.

Chapter 3 provides an in-depth account of theory-based evaluation. Intervention theories or theories of change refer to the causal assumptions regarding how a particular intervention works and relates to processes of change. They can be reconstructed and further refined or tested on the basis of stakeholder interviews, reviews of policy documents, existing substantive knowledge relevant to an intervention and its context, or on the basis of further empirical research. The Chapter illustrates how an intervention theory embedded in a theory of social action (Coleman, 1986) can constitute a useful framework that facilitates a better understanding of the potential processes of change of an intervention (i.e. embedded in a theory of individual and social behavior); it provides a structure for argumentation and interpretation[33] and empirical data collection and analysis.

Chapter 5 is an example of a theory-informed mixed method impact evaluation in the spirit of Bamberger et al.'s (2004) 'shoestring' approach, tailored to the local, budgetary, time and data constraints of the evaluation context.

The idea that particular methods have comparative advantages in addressing specific aspects or questions of an impact evaluation is also illustrated by the methodological perspective adopted in Chapter 6. The usual counterfactual in a RE is the case when there is no intervention. Often this is not the counterfactual policymakers are interested in; for them the counterfactual is not no intervention, but a different intervention (Ravallion, 2009b).[34] Chapter 6 presents a methodological framework

33 The initial theory also provides a useful structure for consulting other useful substantive theories from academic research, as the Chapter also illustrates.

34 REs can also address this issue. Increasingly, REs comprise multiple 'treatment packages' in a multi-group comparative design.

based on three pillars – multicriteria analysis, elicitation of stakeholder values and theory-based evaluation – which allows for a systematic assessment of alternative policy scenarios. In the framework, policy alternatives are ranked with respect to how much better or worse they do than alternative policy alternatives (rather than the traditional without scenario). An important difference with other Chapters is that in this case we are dealing with an ex ante evaluation of (expected) effects of policy alternatives. The tested assumptions representing the intervention theory of an existing intervention (one of the policy alternatives) provide the basis for determining the expected effects of alternative policy options.

1.5. Challenge 3: Impact evaluation in practice

1.5.1. Threats to attribution analysis in experimental settings

The main threats to the internal validity of findings from REs (and certain quasi-experiments) are well-known and widely discussed (Campbell and Stanley, 1963; Campbell, 1969; Cook and Campbell, 1979; Shadish et al., 2002):

- Selection bias: refers to the problem of under- or overestimating intervention effects due to uncontrolled differences between different (treatment) groups that would lead to differences in effect variables if none of the groups would have received benefits.
- Contagion (or treatment diffusion): refers to the problem of groups that are not supposed to be exposed to (or receiving) certain benefits are in fact benefiting from an intervention in one or more ways: by directly receiving the benefits from the intervention, by indirectly receiving benefits through other participants, or by receiving similar benefits from other organizations.
- Aging/Maturation: an effect that arises when participants grow older, wiser, more tired, more self-confident due to factors other than the intervention.
- History: the effect is caused by some event other than the intervention occurring during the same time period as the intervention.
- Testing: the pre-test measurement causes a change in the post-test measure.
- Instrumentation: the effect is caused by a change in the method of measuring the outcome.
- Regression to the Mean: where an intervention is implemented on units with unusually high scores (e.g. unusually high student performance scores), natural fluctuation will cause a decrease in these scores on the post-test which may be mistakenly interpreted as an effect of the intervention.
- Attrition: changes in effect measurements due to drop-outs.
- Causal Order: it is unclear whether the intervention preceded the outcome.
- Behavioral responses: several unintended behavioral responses not caused by the intervention or 'normal' conditions might inhibit the reliability of comparisons between groups and hence the ability to attribute changes to the intervention. An example is expected behavior or compliance behavior: beneficiaries' behavior is not caused by intervention outputs (e.g. a subsidy) but due to rea-

sons of compliance with a (formal/informal) contract between beneficiary and implementing organization, due to the (longstanding) relationship with a particular organization (delivering intervention outputs), or due to certain expectations about future benefits from the organization (not necessarily the intervention).

While each of these issues might be potential problems that render claims on attribution less valid, the fact that they have been systematically identified and addressed in the literature (see for example Cook and Campbell, 1979) enhances the scientific rigor of experimental and quasi-experimental designs.

A key underlying determinant of the internal validity of findings from REs is the extent to which those managing the experiment are capable and willing to safeguard the design from the threats to validity described above. This is certainly the case for selection bias and treatment diffusion issues; well-designed REs may be safeguarded from these problems. More complicated is the potential threat from unintended behavioral responses among participant or control groups. As argued by several authors (e.g. Deaton, 2009; Cohen and Easterly, 2009) randomization can lead to hard feelings between the different groups involved in the experiment. "[S]ubjects may fail to accept assignment, so that people who are assigned to the experimental group may refuse, and controls may find a way of getting the treatment, and either may drop out of the experiment altogether. The classical remedy of double blinding, so that neither the subject nor the experimenter know which subject is in which group, is rarely feasible in social experiments" (Deaton, 2009: 36). In some cases researchers cannot (and indeed should not) withhold the information from stakeholders that they are part of an experiment. Banerjee and Duflo (2008) provide several examples of mechanisms (e.g. lotteries) which facilitate the collaboration of local populations in an experiment. Yet, even in situations where institutions and target groups agree to randomized allocation of benefits, it is well-known in the literature (e.g. Campbell, 1969; Shadish et al., 2002) that experimentation may lead to a range of unintended behavioral responses within the participant and treatment groups which can affect the validity of findings resulting from an experiment. An example of unintended behavioral responses comes from the famous PROGRESA study (see for example Skoufias and McClafferty, 2001). "One issue with the explicit acknowledgement of randomization as a fair way to allocate the program is that implementers will find that the easiest way to present it to the community is to say that an expansion of the program is planned for the control areas in the future (especially when it is indeed the case, as in phased-in designs). This may cause problems if the anticipation of treatment leads individuals to change their behavior. This criticism was made in the case of the PROGRESA programs, where control villages knew that they were going to eventually be covered by the program" (Banerjee and Duflo, 2008: 22).

1.5.2. Other design and implementation challenges

A key issue regarding the success of REs in practice revolves around the ethics and feasibility of randomization in practice and the corresponding reactions by stakeholders. Experiments can cause resentment as people do not understand or

support the differences in benefits allocated to different groups. Random allocation in many cases is also unacceptable to policy makers (e.g. Bamberger and White, 2007; Ravallion, 2008). Interventions are often intended to be targeted to specific groups and outreach is a direct concern to policy makers. Randomization would limit outreach and at the same time is often considered as unethical in view of the pressing needs of target populations.[35]

Banerjee and Duflo (2008: 22) argue that collaboration with institutions in developing countries is becoming less of an issue in developing countries. "Randomization that takes place at the level of location can piggy-back on the expansion of the organization's involvement in these areas limited by budget and administrative capacity, which is precisely why they agree to randomize. Limited government budgets and diverse actions by many small NGOs mean that villages or schools in most developing countries are used to the fact that some areas get some programs and others do not and when a NGO only serves some villages, they see it as a part of the organization's overall strategy. When the control areas [are] given the explanation that the program had only enough budget for a certain number of schools, they typically agree that a lottery was a fair way to allocate it---they are often used to such arbitrariness and so randomization appears transparent and legitimate".

Another way to enhance the goodwill among implementing agencies is to forge a long-standing relationship between the latter and RE researchers in which a series of experiments will constitute the basis of a cumulative process of learning. It still remains to be seen, however, what the costs and benefits of such a relationship would be in divergent institutional contexts, especially in view of the previously discussed issues. Institutional willingness to undertake a RE is not a black and white issue, as interventions typically comprise multiple institutional partners and multiple layers of management, from headquarters to field level. The idea of institutional willingness is also closely related to institutional capacities and incentives. REs are intrinsically linked to intervention implementation processes and the question to what extent well-trained researchers are de facto present and able to ensure the quality control of experimental conditions, is an empirical one that merits further inquiry. There is a marked difference between an experienced research team undertaking REs (as is the case for most of the published work on REs in development) and the idea of mainstreaming REs in the design of selected projects of donor and developing country agencies' portfolios. In the latter case invariably not all the tasks pertaining to the design and implementation of a RE are managed by experienced and motivated researchers.

Apart from institutional capacities, willingness to collaborate and ethics, other challenges remain. A first obvious condition for REs is the active involvement of researchers or evaluators in the intervention design and implementation phase. This involvement is essential for baseline data collection as well as quality control of randomization. In practice, however, many impact evaluations are commissioned after an intervention has been implemented and baseline data continue to be a problem (Bamberger and White, 2007). Although preferably double difference (partici-

35 This often provides a compelling argument for using quasi-experimental designs. For example, regression discontinuity is very useful when targeting is based on clear and easy to measure criteria of selection.

pant-control group comparisons over time) designs should be used, it is more usual that impact assessments are based on less rigorous – and reliable – designs, where baseline data are reconstructed or collected later during the implementation phase, or baseline data are collected only for the treatment group, or there are no baseline data for neither treatment nor control group (for options on how to address these constraints see Bamberger et al., 2004; Bamberger et al. 2006; or Bamberger and White, 2007).

A second aspect is the costs of doing REs. Early experiences of REs (and quasi-experimental studies) in development by the World Bank turned out to be rather costly (OED, 2005), but these studies were often quite ambitious in scope. More recent experiences seem to suggest that REs do not necessarily have to be more expensive than other non-experimental observational studies that are based on original fieldwork with multiple data points in time. However, given the narrow focus of REs, when large programs with multiple intervention activities need to be evaluated, REs need to compete with less expensive non-experimental methodologies which can cover a broader scope of activities with less budget. Proponents of REs will need to justify whether the potential gains in terms of the high internal validity of findings delivered by an RE will be worthwhile the investment given the loss in scope. Alternatively, adversaries of experimentation or others that choose scope over depth will have to justify that alternative methodological designs adequately address crucial issues such as attribution.

A third issue concerns the quality of the data. Bamberger and White (2007) argue that problems of data quality, although relevant for any type of methodological design, might be particularly problematic in case of REs as they rely on a limited number of indicators. Banerjee and Duflo (2008) contest this idea. In their view, if data is being collected especially for the purposes of a RE, then given the limited number of observations that are usually necessary for reliable estimates, researchers are able to dedicate special attention to data collection and measurement issues. Ensuring high data quality is particularly challenging in rural contexts in developing countries.[36] Surveys continue to be the main instruments generating impact evaluation data. Potential measurement errors due to problems of recall, sensitivity of certain topics, intercultural communication, translation errors are just a few of the problems that affect data quality (see for example De Leeuw et al., 2008; Mikkelsen, 2005; Bamberger et al., 2004; Bamberger, 2000). Moreover, surveys may not be appropriate for addressing sensitive topics (see for example Bamberger et al., 2009) such as for domestic violence, household income or local norms and in these cases are more likely to generate unreliable data. Unfortunately, data quality is not as sexy a topic as methodological design, especially if the latter is the acclaimed basis for delivering 'rigorous scientific evidence'. Indeed, in this dissertation data issues do not receive prominent attention either. Nevertheless, it is the personal opinion of this author that rather than being banished to the fringes of the debate, dismissed by some as routine or not worthy their intellectual attention, the capacity to generate high-quality data probably remains the most crucial and at the same

36 For a rather critical perspective on data quality in (rural) developing country contexts, see Gill (1993).

time best-hidden skill that determines the quality of findings of an impact evaluation.

1.5.3. The contribution of this dissertation

Chapter 4 assesses the different threats to validity that affect the analytical potential of attribution analysis based on a RE in a concrete project in Latin America. More specifically, the threats of selection bias, treatment diffusion and several unintended behavioral responses caused by the experiment are analyzed in a systematic manner. Examples of the latter are compliance behavior, behavior motivated by expectations about future benefits, or rivalry behavior. The idea of differentiation between groups, each group receiving different benefits from the project caused quite some resentment among farmers in the project; part of which was due to communication problems at the start of the project. At the root of many of the threats to validity, affecting the analytical potential of the RE, lies the institutional setup and quality control of the RE in practice. While the project had excellent field staff in terms of professional experience in ecology, agronomy and other related disciplines, field staff members were not researchers and did not have the incentives nor capacities to manage a randomized experiment in practice. The Chapter provides some practical recommendations for the design and implementation of REs in similar projects.

Chapter 5 presents an impact evaluation study of a training project in organic agriculture in the highlands of Guatemala, at that time (1998-2001) a post-civil war region. In a context of high levels of distrust and language barriers (Spanish versus local Mayan languages) between researchers (including local staff) and farmers, the analysis and findings are based on triangulation between survey research within a simple quasi-experimental design, semi-structured interviews and field visits. In addition, existing research on smallholder adoption behavior and peasant economics (see Ellis, 1988) is used for framing the analysis and interpreting the findings. The rudimentary quasi-experimental design comprises three samples: participant group ex ante, participant group ex post and control group ex post. The reason why no control group was covered in the ex ante study is the following. The design was based on paired sample comparisons with low sample sizes (due to time and budget restrictions). Given the high levels of temporary and permanent migration in the region, it was considered too costly to include a control group of which a sufficiently large proportion would also be present and willing to cooperate in the ex post study. By contrast, participant farmers were committed to the project and therefore also to the ex ante and ex post survey. As mentioned previously Chapter 5 presents an example of a mixed method theory-informed impact evaluation analogous to Bamberger et al.'s (2004) 'shoestring' approach.

References

Baker, J.L. (2000) Evaluating the impact of development projects on poverty, World Bank, Washington D.C.

Bamberger, M. (2000) “Opportunities and challenges for integrating quantitative and qualitative research”, in: M. Bamberger (ed.) *Integrating quantitative and qualitative research in development projects*, World Bank, Washington D.C.
Bamberger, M. and H. White (2007) “Using strong evaluation designs in developing countries: Experience and challenges”, *Journal of Multidisciplinary Evaluation* 4(8), 58-73.
Bamberger, M., J. Rugh, M. Church and L. Fort (2004) “Shoestring Evaluation: Designing impact evaluations under budget, time and data constraints”, *American Journal of Evaluation* 25(1), 5-37.
Bamberger, M. J. Rugh and L. Mabry (2006) *Realworld evaluation*: Working under budget, time, data, and political constraints, Sage Publications, Thousand Oaks.
Bamberger, M., V. Rao and M. Woolcock (2009) “Using mixed methods in monitoring and evaluation: Experiences from international development”, MIMEO, World Bank, Washington D.C.
Banerjee, A. (2005) “‘New development economics and the challenge to theory”, *Economic and Political Weekly* 40, October 1-7, 4340-4344.
Banerjee, A.V. and E. Duflo (2008) “The experimental approach to development economics”, *NBER Working Paper* 14467, Cambridge.
Banerjee, A., S. Cole, E. Duflo and L. Linden (2007) “Remedying education: Evidence from two randomized experiments in India”, *Quarterly Journal of Economics* 122(3), 1235-1264.
Blattman, C. (2008) “Impact Evaluation 2.0”, Presentation to the Department of International Development, February 14, 2008, London.
Campbell, D.T. (1969) “Reforms as experiments”, *American Psychologist* 24, 409-429.
Campbell, D.T. and J.C. Stanley (1963) “Experimental and quasi-experimental designs for research on teaching”, in: N. L. Gage (ed.) *Handbook of research on teaching*, Rand McNally, Chicago.
Carvalho, S., and H. White (2004) “Theory-based evaluation: The case of social funds”, *American Journal of Evaluation* 25(2), 141–160.
Catley, A., J. Burns, D. Abebe and O. Suji (2008) *Participatory Impact Assessment: a guide for practitioners*, The Feinstein International Center, Tufts University, Medford.
CGD (2006) *When will we ever learn? Improving lives through impact evaluation*, Report of the Evaluation Gap Working Group, Center for Global Development, Washington, DC.
Chambers, R. (1983) Rural development: putting the last first, Wiley, New York.
Chambers, R., (1995) “Paradigm shifts and the practice of participatory research and development”, in: S. Wright and N. Nelson (eds.) *Power and participatory development: Theory and practice*, Intermediate Technology Publications, London.
Chen, H.T. (1990) *Theory-driven evaluation*, Beverly Hills, Sage Publications.
Cohen, J. and W. Easterly (eds.) (2009) *What works in development? Thinking big and thinking small*, Brookings Institution press, Washington D.C.
Coleman, J.S. (1986) "Theory, social research and a theory of action", *American Journal of Sociology* 91(6), 1309-1335.
Coleman, J.S. (1990) *Foundations of social theory*, Belknap Press, Cambridge.
Cook, T.D. (2006) “Describing what is special about the role of experiments in contemporary educational research: Putting the ‘Gold Standard’ rhetoric into perspective”, *Journal of Multidisciplinary Evaluation* 3(6), 1-7.
Cook, T.D. and D.T. Campbell (1979) Quasi-experimentation: Design and analysis for field settings, Rand McNally, Chicago.
Cousins, J.B. and E. Whitmore (1998) “Framing participatory evaluation”, in: E. Whitmore (ed.) *Understanding and practicing participatory evaluation,* New Directions for Evaluation 80, Jossey-Bass, San Francisco.
Davidson, E.J. (2006) “The RCTs-only doctrine: Brakes on the acquisition of knowledge?” *Journal of Multidisciplinary Evaluation* 3(6), ii-v.
Deaton, A. (2009) “Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development”, *NBER Working Paper* 14690, Cambridge.
De Leeuw, E.D., J.J. Hox and D.A. Dillman (eds.) (2008) *International handbook of survey methodology*, Lawrence Erlbaum Associates, London.
Donaldson, S.I. and Lipsey, M.W. (2006) “Roles for theory in contemporary evaluation practice”, in: I. Shaw, J.C. Greene and M.M. Mark (eds.) *The SAGE handbook of evaluation*, Sage Publications, Thousand Oaks.

Duflo, E., and R. Hanna (2008) "Monitoring works: Getting teachers to come to school", *NBER Working Paper* 11880, Cambridge.
Easterly, W. (2001) The elusive quest for growth: Economists' adventures and misadventures in the tropics, MIT Press, Cambridge.
Ellis, F. (1988) Peasant economics: Farm households and agrarian development, Cambridge University Press, Cambridge.
Gill, G. (1993) Ok the data is lousy, but it's all we've got (Being a critique of conventional methods), *Gatekeeper Series* 38, Sustainable Agriculture Program, International Institute for Environment and Development, London.
Heckman, J. (1992) "Randomization and social program evaluation," *NBER Technical Working Paper* 107, Cambridge.
Heckman, J., J. Smith and N. Clements (1997) "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts", *Review of Economic Studies* 64, 487-535.
Hulme, D. (2000) "Impact assessment methodologies for microfinance: Theory, experience and better practice", *World Development* 28(1), 79-98.
IFAD (2002) Managing for impact in rural development: A practical guide for M&E, IFAD, Rome.
Jones, N., C. Walsh, H. Jones and C. Tincati (2008) Improving impact evaluation coordination and uptake - A scoping study commissioned by the DFID Evaluation Department on behalf of NONIE, Overseas Development Institute, London.
Karlan, D. (2009) "Thoughts on randomised trials for evaluation of development: presentation to the Cairo evaluation clinic, *Journal of Development Effectiveness* 1(3), 237-242.
Karlan, D. and J. Zinman (2009) "Expanding microenterprise credit access: Using randomized supply decisions to estimate impacts in Manila", *CEPR paper* 7396, London.
Kusek, J and R.C. Rist (2004) Ten steps to a results-based monitoring and evaluation system: A handbook for development practitioners, World Bank, Washington D.C.
Leeuw, F.L. (2003) "Reconstructing program theories: Methods available and problems to be solved", *American Journal of Evaluation* 24(1), 5-20.
Leeuw, F.L. (2009) "On the contemporary history of experimental evaluations and its relevance for policy making", in: O. Rieper, F.L. Leeuw and T. Ling (eds.) *The evidence book: concepts, generation, and use of evidence*, Transaction Publishers, New Brunswick.
Leeuw, F.L. and J. Vaessen (2009) *Impact evaluations and development – NONIE guidance on impact evaluation*, Network of Networks on Impact Evaluation, Washington D.C.
Lipsey, M.W. (1993) "Theory as method: Small theories of treatments," in: L.B. Sechrest and A.G. Scott (eds.), *Understanding causes and generalizing about them*, New Directions for Program Evaluation 57, Jossey-Bass, San Franscisco.
Mark, M.M., G.T. Henry and G. Julnes (1999) "Toward an integrative framework for evaluation practice", *American Journal of Evaluation* 20(2), 177-198.
McKenzie, D. and C. Woodruff (2008) "Experimental evidence on returns to capital and access to finance in Mexico", *World Bank Economic Review* 22(3), 457-482.
Miguel, E. and M. Kremer (2004) "Worms: Identifying impacts on education and health in the presence of treatment externalities," *Econometrica* 72(1), 159-217.
Mikkelsen, B. (2005) *Methods for development work and research*, Sage Publications, Thousand Oaks.
Morgan, S.L. and C. Winship (2007) Counterfactuals and causal inference – methods and principles for social research, Cambridge University Press, Cambridge.
Morra, L.G. and R.C. Rist (2009) The road to results: designing and conducting effective development evaluations, World Bank, Washington D.C.
Oakley, A. (2000) Experiments in knowing: Gender and method in the social sciences, Polity Press, Cambridge.
OECD-DAC (2002) Glossary of key terms in evaluation and results based management, OECD-DAC, Paris.
OED (2005) *OED and impact evaluation: A discussion note*, Operations Evaluation Department, World Bank, Washington D.C.
Olken, B. (2007) "Monitoring corruption: Evidence from a field experiment in Indonesia", Journal of Political Economy 115(2), 200-249.
Pawson, R. (2006) *Evidence-based policy: A realist perspective*, Sage Publications, London.

Pawson, R. and N. Tilley (1997) *Realistic Evaluation,* Sage Publications, Thousand Oaks.
Pretty, J.N., I. Guijt, J. Thompson and I. Scoones (1995) *A trainers' guide to participatory learning and action*, IIED Participatory Methodology Series, IIED, London.
Ravallion, M. (2008) "Evaluation in the practice of development", *Policy Research Working Paper* 4547, World Bank, Washington D.C.
Ravallion (2009a) "Should the Randomistas rule?" *Economists' Voice*, February 2009.
Ravallion, M. (2009b) "Evaluating three sylised interventions", *Journal of Development Effectiveness* 1(3), 227-236.
Riggin, L.J.C. (1990) "Linking program theory and social science theory", in: L. Bickman (ed.) *Using program theory in evaluation*, New Directions for Program Evaluation 33, Jossey-Bass, San Francisco.
Rodrik, D. (2008) "The new development economics: We shall experiment, but how shall we learn?" MIMEO, Harvard University, Cambridge.
Rogers, E.M. (2003) *Diffusion of innovations*, New York, Free Press.
Rogers, P. J. (2008) "Using programme theory for complex and complicated programs", *Evaluation* 14(1), 29-48.
Rogers, P.J., Hacsi, T.A., Petrosino, A., and Huebner, T.A., (eds.) (2000) *Program theory in evaluation: Challenges and opportunities*, New Directions for Evaluation 87, Jossey-Bass, San Francisco.
Rossi, P.H., Lipsey, M.W., and Freeman, H. E. (2004) *Evaluation: A systematic approach*, Sage Publications, Thousand Oaks.
Salmen, L. and E. Kane (2006) Bridging diversity: Participatory learning for responsive development, World Bank, Washington D.C.
Scriven, M. (2008) "Summative evaluation of RCT methodology: & an alternative approach to causal research", *Journal of Multidisciplinary Evaluation* 5(9), 11-24.
Shadish, W. R., T.D. Cook and D.T. Campbell (2002) *Experimental and quasiexperimental designs for generalized causal inference*, Houghton Mifflin Company, Boston.
Sherman, L.W., D.C. Gottfredson, D.L. MacKenzie, J. Eck, P. Reuter and S.D. Bushway (1997) *Preventing crime: what works, what doesn't, what's promising,* US Office of Justice Programs, Washington D.C.
Skoufias, E. and B. McClafferty (2001) "Is PROGRESA working? Summary of the results of an evaluation by IFPRI", *FCND Discussion Paper* 118, IFPRI, Washington D.C.
Tashakkori, A., and C. Teddlie (eds.) (2003) *Handbook of mixed methods in social and behavioral research*, Sage Publications, Thousand Oaks.
The Lancet Editorial (2004) "The World Bank is finally embracing science", *The Lancet* 364(9436), 731-732.
Van der Knaap, L.M., F.L. Leeuw, S. Bogaerts and L.T.J. Nijssen (2008) "Combining Campbell standards and the realist evaluation approach – the best of two worlds?", *American Journal of Evaluation* 29(1), 48-57.
Weiss, C.H. (2000) "Which links in which theories shall we evaluate?" in: P.J. Rogers, T.A. Hacsi, A. Petrosino and T.A. Huebner (eds.) *Program theory in evaluation: Challenges and opportunities*, New Directions for Evaluation 87, Jossey-Bass, San Francisco.
White, H. (2009) "Some reflection on current debates in impact evaluation", *Working Paper* 1, International Initiative for Impact Evaluation, New Delhi.
Woolcock, M. (2009) "Toward a plurality of methods in project evaluation: a contextualized approach to understanding impact trajectories and efficacy", *Journal of Development Effectiveness* 1(1), 1-14.
Worrall, J. (2007) "Why there's no cause to randomize*", The British Journal for the Philosophy of Science* 58(3), 451-488.

CHAPTER 2

Vaessen J. and D. Todd “Methodological challenges of evaluating the impact of the Global Environment Facility's biodiversity program”, *Evaluation and program planning*, 31(3), 231-240.

Evaluation and Program Planning 31 (2008) 231–240

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Methodological challenges of evaluating the impact of the Global Environment Facility's biodiversity program ☆

Jos Vaessen [a,*], David Todd [b,1]

[a] Institute of Development Policy and Management, University of Antwerp, Lange Sint-Annastraat 7, 2000 Antwerp, Belgium
[b] GEF Evaluation Office, Global Environment Facility, 1818 H Street, NW MSN G6-604, Washington, DC 20433, USA

ARTICLE INFO

Article history:
Received 10 January 2007
Received in revised form
11 January 2008
Accepted 14 March 2008

Keywords:
Global Environment Facility
Biodiversity
Impact evaluation
Theory-based evaluation

ABSTRACT

In this paper, we explore some of the methodological challenges that evaluators face in assessing the impacts of complex intervention strategies. We illustrate these challenges, using the specific example of an impact evaluation of one of the six focal areas of the Global Environment Facility; its biodiversity program. The discussion is structured around the concepts of attribution and aggregation, offering the reader a framework for reflection. Subsequently, the paper discusses how theory-based evaluation can provide a basis for addressing the attribution and aggregation challenges presented.

1. Introduction

In recent years, several developments such as the universal endorsement of the millennium development goals and growing demands for accountability by citizens have contributed to an increasing interest among multilateral and bilateral donor organizations to generate evidence on the results of the policies, programs and projects they support. A recent report by the Center for Global Development (CGD, 2006) has been influential in further promoting a growing practice of impact evaluation within the development community. In line with this trend, the Global Environment Facility (GEF) recently prepared its approach to an impact evaluation of its biodiversity program.

The GEF is an international funding mechanism which supports developing countries and countries with economies in transition to implement policies, programs and projects that protect the global environment. The GEF is funded through replenishments by member states on initially a 3-year basis and more recently on a 4-year basis. The latest fourth replenishment amounted to US$3.15 billion. The GEF was established in 1991 and since then has provided grants for more than 1300 projects in 140 countries. GEF grants support projects in six focal areas: biodiversity, climate change, international waters, land degradation, the ozone layer, and persistent organic pollutants.

The biodiversity program constitutes the largest portfolio of interventions financed by the GEF. In the period 1991–August 2006 approximately $2.22 billion of GEF funding with some $5.16 billion of co-financing was allocated to biodiversity projects (GEF PMIS data base, August 30, 2006). The impact evaluation aims "to evaluate the long-term results of GEF interventions, a few years after GEF support is concluded and to assess the sustainability and replication[2] of the support as well as to extract lessons learned" (GEF, 2006a, p. 6). While a comprehensive treatment of all intervention activities lies beyond what is practically feasible, at the same time it is expected that the exercise should transcend the limitations of isolated perspectives on the impacts of particular projects or activities within projects.

A number of particular characteristics of the GEF and its biodiversity program set the stage for the impact evaluation. First of all, responsibilities for managing and implementing GEF interventions are shared between different institutions. The sometimes complex institutional set-up of interventions is in part due to the GEF's principle of maximizing co-financing by other institutions

☆ This article discusses an impact evaluation managed by the GEF Evaluation Office. The content of the article is the sole responsibility of the authors and does not commit the GEF Evaluation Office or any other actors involved in the evaluation to the authors' views.

* Corresponding author. Tel.: +32 3 275 5929; fax: +32 3 275 5771.
E-mail addresses: jos.vaessen@ua.ac.be (J. Vaessen), dtodd@thegef.org (D. Todd).
[1] Tel.: +1 202 473 6028; fax: +1 202 522 1691/3240.

[2] The GEF evaluation office has initiated a separate evaluation on the catalytic role of the GEF. As a result, the impact evaluation will not provide an in-depth coverage of this dimension.

0149-7189/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.evalprogplan.2008.03.002

as well its specific focus on global environmental benefits. The GEF Council (as the supreme decision-making organ of the GEF) and the GEF Secretariat are the main GEF bodies responsible for strategic and budgetary decisions. Individual GEF interventions are mainly managed by the GEF's Implementing Agencies, the World Bank, The United Nations Development Program (UNDP), and the United Nations Environment Program (UNEP); and (to a lesser extent) by its Executing Agencies (e.g. the Asian Development Bank, the Inter-American Development Bank). In practice, additional institutional layers can be distinguished, as GEF operations in the countries and regions of intervention are to a great extent handled by governmental, non-governmental or private sector organizations in collaboration with the above-mentioned agencies. As a result, GEF interventions are often part of or 'blend in' with broader intervention strategies of other institutions.

The GEF biodiversity program serves the overarching objective of protecting globally important biodiversity. In practice however, there is still uncertainty regarding to what extent and how this objective is achieved by the projects to which the GEF contributes. More specifically, this uncertainty is fuelled by the fact that:

- projects often encompass a wide range of discrete activities while it is often not clear how these different activities contribute to project objectives and higher-level program goals (GEF, 2004a);
- data on the outcomes and impacts of biodiversity projects (including appropriate indicators for measuring these effects) are scarce (GEF, 2004a);
- environmental change processes are complex and changes may only become apparent years after a project has been completed (MEA, 2005).

These constraints as well as the inherent complexity of linking specific interventions to global biodiversity gains pose particular methodological challenges to the impact evaluation.

The purpose of this article is twofold. First, we will discuss the principal methodological challenges that evaluators are facing in this study. Subsequently, we will argue that theory-based evaluation can provide a basis for meeting some of the challenges presented. Our discussion is inspired by White's (2003) triple-A assessment of development agency performance. The three A's are: attribution, aggregation and alignment. Although White applies the concepts in the assessment of the quality of agency performance reports, they also represent key concerns in evaluation and are particularly relevant to impact evaluation. Applied to the context of impact evaluation they can be defined as follows. Attribution refers to the problem of establishing a causal link between intervention outputs and observed changes in impact variables. In order to be able to isolate the effect of an intervention on a particular target from the influence of other variables (e.g. the policy environment, socio-economic trends), evaluators often rely on the principle of a counterfactual scenario (what would have happened without the intervention). Aggregation concerns the question of how micro-level impact data can be meaningfully aggregated across interventions. This is crucial in impact assessment studies of clusters of interventions (as opposed to a single (site-specific) intervention), like the biodiversity portfolio. Related to this, alignment touches upon the issue of whether data collected at micro-level are relevant with respect to an agency's overall objectives. Our discussion is mainly constructed around two of the three concepts, the challenges surrounding the issues of attribution and aggregation.

2. Methodological challenges

2.1. The problem of the independent variable

A first important question evaluators need to raise is the question of impact *of* what? Two key issues come to mind: the delimitation issue and the choice of level(s) of analysis. Regarding the former, Pawson and Tilley (1997), among others, illustrate that an exact delimitation of an intervention can be problematic. Rather than constituting a clearly delineated mechanism, an intervention resembles more an open system, a social system embedded in a larger social system in which it is often not easy to determine where an intervention ends and the 'external' world begins. In the case of the GEF evaluators are furthermore confronted with the issue of 'blending' of interventions. For example, in the case of the World Bank as implementing agency, GEF grants are often blended with World Bank loans where the GEF project de facto is part of a bigger intervention package. In principle the GEF grant is designed to account for the incremental costs associated with generating global benefits as opposed to the local benefits generated by the loan package. In practice however, it is often very difficult to clearly distinguish between the two. Similar problems sometimes occur when GEF projects are part of broader intervention strategies of other implementing or executing agencies.

The second problem, the choice of level(s) of analysis, is one of the key issues to be resolved in portfolio (impact) evaluations. What level(s) of analysis is/are appropriate for making (in the simplest and most straightforward manner possible) plausible and coherent statements about attribution? For example, should we analyze the impact of projects, activities within projects, operational programs, etc.? Correspondingly, at what level(s) of analysis should evidence about impact be aggregated? For the GEF Council it would be desirable that impact evidence could be aggregated to the portfolio level and put into perspective with global trends, so that changes put into motion by GEF funding can be identified. Apart from that however, one can raise the question whether other types of aggregation of evidence on impact would be useful, especially from the point of view of knowledge management.

This is not easily resolved as different levels of analysis each present their own advantages and disadvantages. The project is the basic administrative unit of intervention and as such presents a natural choice as a focus for impact assessment exercises. In addition, data on performance, outputs and (to a lesser extent) biodiversity-related data are collected and reported at this level. A disadvantage is the fact that projects are not always clearly articulated to higher-level program objectives (GEF, 2004a). This makes it difficult for many projects in the portfolio to aggregate evidence at this level to higher levels of analysis. A second choice would be the level of the operational program (see Appendix A), since projects within the portfolio are classified according to the operational program (e.g. Forest Ecosystems) they adhere to. While the ecosystem-related program categories are meaningful at one level, many projects adhere to multiple operational programs. Most importantly, the variety in terms of objectives, activities and institutional structures between projects within one operational program makes this unit of analysis difficult to use for evaluation purposes. The more recent strategic objectives (see Appendix A) could be useful for evaluative purposes, as the categories (e.g. catalyzing the sustainability of protected area systems) represent different groups of projects with (as a group) more coherent objectives and strategies. Nevertheless, the categories are quite general, still harboring a substantial variety of intervention activities. In addition, projects can serve multiple strategic objectives. Finally, the strategic objectives have only

recently been introduced and as such are not particularly useful as units of analysis in the impact evaluation, which focuses mainly on completed projects.[3]

Given the difficulties associated with these (what can be called) traditional levels of analysis, the evaluators have considered alternatives. An interesting level of analysis is the thematic area of intervention. To a large extent projects (and intervention activities within projects) can be categorized in a coherent manner on the basis of the main theme addressed. Examples of thematic areas of intervention are: protected area management, alternative livelihoods, research on innovative practices, and particular mainstreaming[4] models (e.g. sector-specific legislation). A second interesting level of analysis is that of policy instruments. Policy instruments are the basis of public intervention everywhere. Examples of generic policy instruments are: economic incentives (e.g. tax reductions, subsidies), regulations (e.g. laws, restrictions), and information (e.g. education, technical assistance). As argued by several authors (e.g. Pawson, 2006; Salamon, 1981; Vedung, 1998), a classification of different policy instruments recurring throughout the portfolio in relation to specific purposes and contexts can constitute a useful tool for the assessment of effectiveness of interventions as well as institutional learning. "Rather than focusing on individual programs, as is now done, or even collections of programs grouped according to major 'purpose' as is frequently proposed, the suggestion here is that we should concentrate on the generic tools of government that come to be used, in varying combinations in particular public programs" (Salamon, 1981, p. 256). Acknowledging this central role of policy instruments enables evaluators to take into account lessons from the application of particular (combinations of) policy interventions elsewhere, in the first place relating to the field of environmental protection and development, but also beyond (see Bemelmans-Videc & Rist, 1998).

2.2. The problem of the dependent variable

A second key issue is the question of impact *on* what? An important consideration concerns the question on which point in the causal chain between intervention output and final (desired) impact one should focus? The primary objective of the GEF is to generate global environmental benefits. Ideally, the effects of all GEF interventions should therefore be traceable up to changes in global environmental benefits, e.g. in the case of biodiversity (positive) changes at the levels of ecosystems, species and gene pools. However, if the impact evaluation were to concentrate on impact at these levels, its utility would be severely hampered by the substantial challenges of attribution (establishing to what extent environmental changes can be shown to result from GEF interventions) and aggregation (the extent to which localized biodiversity changes resulting from interventions can be seen to contribute to higher-level (ideally) global changes). More specifically the following complicating factors play a role.

2.2.1. The nature of environmental change

The complexity of processes of environmental change continues to be a challenging and elusive area of scientific inquiry. Large-scale scientific efforts such as the recent Millennium Ecosystem Assessment (MEA) have contributed to strengthening the scientific consensus on a number of issues regarding environmental processes, more particularly the key role of ecosystems in sustaining life on earth. A promising framework linking ecosystem services,[5] their underlying drivers of change and different aspects of human well-being has been developed by this project. On the other hand, once again many of the limitations in our understanding of these processes have been pointed out. In particular the non-linear nature of environmental change, the time scales over which changes occur and the interaction effects between different drivers of change (e.g. the interplay of climate change and economic activity and the effects on various ecosystem services) are often not well understood. A specific complicating factor is the irreversibility of many environmental change processes (e.g. habitat loss, species extinction); once a certain point of change (a threshold) has been passed a process of restoration towards the former state of the environment is no longer possible (MEA, 2005; Rao, 2000).

2.2.2. The concept and measurement of biodiversity

The GEF has adopted the CBD's definition of biodiversity, which encompasses the diversity in species, gene pools and ecosystems. Biodiversity as a whole as well as the three subcomponents cannot be easily captured by simple indicators and requires multiple indicators representing the different aspects of (genetic, ecosystems, species) biodiversity (Duelli & Obrist, 2003). Comprehensive indicators are often contested as they are clearly value-laden (i.e. including specific dimensions of biodiversity with certain relative weights) and involve adding up different types of biodiversity which (in some cases) might be negatively correlated (Duelli & Obrist, 2003). Regarding the latter, for example there is sometimes a negative correlation between species diversity (e.g. the number of different fish in a lake) and species abundance (e.g. the amount of fish of one species). This trade-off can become problematic when minimal thresholds for species survival are threatened. Despite all this, comprehensive indicators can be very useful in determining priorities for resource allocation. Under the new country-based resource allocation framework the GEF uses a comprehensive indicator of biodiversity encompassing species and ecosystem biodiversity. The GEF biodiversity indicator can be broken down into two components: representation of ecosystems and species diversity, and threats to ecosystem quality and species. The impact evaluation will primarily focus on the second component as GEF interventions are mainly focused on reducing biodiversity threats (GEF, 2006b). Regarding the effect of GEF interventions on levels of biodiversity (representation), data on biodiversity aspects are often not readily available (see below). Measurement of the different aspects is often not straightforward nor easy and therefore can be very costly. Consequently, it can be worthwhile to choose proxy indicators, which are highly correlated with multiple aspects of biodiversity (Duelli & Obrist, 2003). Further analysis is needed to reveal what proxy indicators might be useful in such a role. In the case of the GEF, the intensity of land use or the number of hectares of protected area could be useful proxies for biodiversity (GEF, 2004a).

2.2.3. Current data availability

A major input to impact evaluation studies is the existing information base. As a result, evaluators first of all inquire whether there is useful existing evaluative evidence (at project level) to inform impact evaluation studies. Second, the question

[3] Since many of the processes of change induced by GEF interventions are likely to produce observable effects on biodiversity only after a certain period of time, a focus on relatively older projects seems justified.

[4] Mainstreaming biodiversity "involves integrating the values and goals of biodiversity conservation and sustainable use into economic sectors and development policies and programmes" (GEF, 2004b, p. 2).

[5] The MEA distinguishes four main groups of ecosystem services: provisioning (e.g. food, water, fiber, fuel), regulating (climate regulation, water, disease), cultural (spiritual, aesthetic, recreation, education), and supporting (primary production, soil formation) (MEA, 2005).

arises how existing evaluative evidence (at project level) can be usefully aggregated to inform impact assessment at portfolio level. Recent studies have pointed out the lack of information on impact in existing end-of-project evaluations (GEF, 2004a, 2005). In addition, the same studies have reported that in general projects do not have adequate (standardized) reporting systems on biodiversity impact data. Without the existence of reliable empirical data on results, projects often tend to overestimate the positive impact on biodiversity (or amplify the threats to biodiversity). The recently introduced strategic objectives and corresponding tracking tools to monitor performance and impact-related indicators represent a positive development towards such a reporting system and increasingly, projects are systematically collecting data on biodiversity conservation and sustainable use.

2.2.4. The type of intervention supported by the GEF

In recent years, in the GEF biodiversity portfolio there has been a relative shift from site-specific interventions to interventions that support a broader agenda of advancing biodiversity concerns not directly related to a specific site. Many of the latter group encompass intervention activities at national or regional (group of countries) level while also including localized intervention activities as pilot and demonstration sites of some of the principles promoted at higher levels of administration. The causal chain connecting this type of intervention activities to biodiversity variables is often more indirect and diffuse than in site-specific intervention activities, making it more difficult to resolve the attribution problem. For example, it is hard to establish clear causal links between enhanced political will of a national government to put biodiversity on the political agenda (e.g. as a result of a GEF-funded national policy dialogue process) and actual changes in biodiversity indicators, given the large number of intermediate steps (from political will to resource allocation to policy design to policy implementation, etc.), the influence of other variables (e.g. other policy priorities, resource constraints, institutional alliances, institutional capacities, etc.), and the uncertain time path of these processes. At the same time, there is an explicit interest from within the GEF to measure impact of its interventions at intermediate levels of the causal chain towards biodiversity conservation and sustainable use. Given the increasing importance of interventions focusing on issues like awareness and political will, policy design and implementation capacity, institutional collaboration and coordination, there is a growing demand for knowledge about in what ways and to what extent the GEF has achieved positive results in these fields.

2.2.5. Additional considerations

The problems of attribution and aggregation discussed above have led some previous studies to conclude that the impact of GEF activities is best measured at the level of behavioral changes of actors (e.g. GEF, 2003). First, as discussed in the previous section, this refers to the behavior of individuals and institutions that influence policies and markets, which in turn (in)directly affect biodiversity variables. Second, it refers to the behavioral changes among end users of natural resources (e.g. farmers, fishermen, the public, etc.), more specifically, behavioral changes in the conservation, sustainable use and benefit sharing of biodiversity. Analogous to the institutional level, impact at the level of behavior of end users of natural resources represents an important intermediate level of impact relevant to actors within (and outside) the GEF network. From that point onward one can venture further down the causal chain towards changes in biodiversity. In some cases the links between certain patterns of behavior and biodiversity are straightforward and attribution issues can be resolved relatively easily; for example, the link between an increase in intercropping systems and on-farm biodiversity (e.g. in terms of plants, insects and birds). In other cases, one can only assume that there is a positive causal link between behavior and biodiversity on the basis of existing (scientific) evidence. In the worst case, the causality is highly contested as the current state of the art of knowledge about the interplay of different variables and their effect on biodiversity is insufficient to draw conclusions about causality and attribution (from human behavior to environmental change).

A final note regarding the 'dependent variable' concerns the dimension of sustainability. Previous studies, in line with their conclusions on impact assessment, have signaled the potential difficulties in sustainability assessment (GEF, 2004a, 2005). Questions about the sustainability of impacts are often even more shrouded in fog than questions of attribution of changes to an intervention. Sustainability is a highly controversial concept that is difficult to pin down in terms of indicators or fixed goals (Mog, 2004). Nevertheless, evaluators can make headway by looking into the factors that make it more or less likely for particular changes to be sustainable (Mog, 2004). In doing so, they should distinguish between different relevant units of assessment (e.g. institutions, ecosystems) and different dimensions of sustainability (e.g. financial sustainability, ecological sustainability).[6] Examples of questions evaluators could ask are the following. Are particular institutional structures (e.g. management structures of protected areas) likely to be financially sustainable? Are technological innovations (e.g. intercropping systems) likely to be appropriated and integrated into existing practices? Are particular enabling environments (e.g. political dialogue, institutional collaboration, legislation on biodiversity) likely to be politically sustainable? Are particular practices (e.g. selective harvesting of non-timber forest products) likely to have a lasting positive influence on biodiversity variables (e.g. ecosystem quality)? What are the main contextual variables obstructing/enabling these processes? To some extent, these questions can be translated into measurable indicators. However, the scope of such questions is almost endless and, as a result, a challenge for evaluators lies in the 'economical' incorporation of sustainability concerns in the overall impact exercise. This is a particular challenge for the GEF, where the concept of a global environmental benefit implicitly incorporates the concept of sustainability, since it attempts to counter current unsustainable patterns of natural resource use.

2.3. Methodological responses to the attribution and aggregation challenge

Before we introduce the basics of the methodological approach applied in the GEF impact evaluation it is worthwhile to reflect briefly on the current methodological debate on impact evaluation.

In several policy fields such as health, education and criminal justice, and to a lesser extent development interventions, 'rigorous impact evaluation' is mostly equated with randomized controlled trials (RCT) or close derivatives (quasi-experiments). The core idea is that observed changes can only be interpreted if they are objectively compared to a counterfactual situation (i.e. that which would have happened without the intervention). In the case of RCTs this works by randomly separating an 'intervention' group from a control group for the duration of an intervention. As a result, differences in target variables between the two groups can be attributed to the intervention as for all other variables conditions are the same (due to the random participation in the

[6] The biodiversity program study (GEF, 2004a) briefly discusses different dimensions of sustainability relevant to biodiversity interventions.

intervention). If random assignment is not possible, control groups are constructed to reflect intervention groups as closely as possible in order to be able to attribute differences in target variables to the intervention (quasi-experimentation). Several variations of this principle are applied,[7] depending, among other things, on data availability (before and after) and budgetary constraints (see IEG, 2006; Rossi, Lipsey, & Freeman, 2004; Shadish, Cook, & Campbell, 2002).

Nowadays, in the fields of evaluation and applied policy analysis one can notice a strong current in favor of more applications of (quasi-)experimental impact evaluation (CGD, 2006), proponents perceiving this type of methodology (either as a stand-alone procedure or embedded in a mixed method design) as the most rigorous and trustworthy way to resolve the attribution issue. Moreover, by capturing impact in terms of effect sizes they generate indications of the magnitude of impact. Very importantly, results of single impact evaluation procedures can be relatively easily aggregated by using quantitative meta-analysis.

Why then are there so few applications of this type of rigorous impact evaluation in development intervention (including the particular field of environment and development in which the GEF operates)? A few reasons can be stated that apply to impact evaluation in general. First of all, the results stemming from rigorous impact evaluation studies are usually freely available and to a large extent very useful to different organizations working on similar intervention activities. Consequently, individual organizations are facing a disincentive to engage in rigorous in-depth impact evaluation,[8] as useful results might be produced by others and therefore available at little or no cost. In addition, there might be other disincentives such as the fear of finding negative impacts or insufficient positive evidence which might put in jeopardy future support for funding (Pritchett, 2002).

Other reasons why there are relatively few (quasi-)experimental evaluations are the following. First of all, they can be very costly and time-consuming. For example, the World Bank, between 1980 and 2005 has conducted only 23 of this type of evaluations with costs ranging between US$200,000 and US$900,000 while taking sometimes more than 2 years to complete (OED, 2005). Second, there are a number of technical and practical considerations which raise the threshold of doing this type of evaluation. These include the high demands in terms of statistical analysis skills, and planning and organization of experimental designs. Regarding the latter, studies are mostly of a quasi-experimental nature as randomization in social policy is often simply not possible (people cannot be excluded at random) or unethical to implement (withholding benefits from particular people while providing them to others). Another constraint concerns the fact that impact evaluations are often not part and parcel of the regular policy cycle and are often commissioned ad hoc. Rigorous (quasi-)experimental evaluation on the other hand (ideally) requires careful planning from the start of an intervention, enabling an adequate set-up of the design as a basis for reliable baseline and ex post data (Rossi et al., 2004). Finally, there are also critical signals stemming from academic debate which raise doubts about the 'superiority' of quasi-experimental evaluation from a conceptual-methodological point of view, and as a result dampen enthusiasm for application. An important critique comes from the field of 'realist evaluation'. This critique is mainly centered around the reductionist nature of quasi-experiments and meta-analysis. It highlights elements such as the incorrect equation of apparently similar intervention mechanisms (e.g. several projects on health education) which in reality might work in different ways, the oversimplification of outcomes, and the concealment of intervention contexts (Pawson, 2002; see also Pawson, 2006).

Ferraro and Pattanayak (2006) put forward a convincing case for more rigorous (quasi-) experimental evaluation of biodiversity interventions. Many of the instruments and approaches to biodiversity conservation are untested and the potential gains from rigorous assessment could be substantial. Yet, there are several reasons why in this particular evaluation (quasi-) experimental evaluation was not chosen as a *starting point* for the evaluation.

First of all, the limited budget for the GEF impact evaluation would not permit conducting an extensive rigorous quasi-experimental evaluation unless one would be willing to accept a substantial loss in scope, reducing the range of lessons which can be generated to help improve future performance. As a result, one would prefer to start out from a more comprehensive methodological framework which would allow for addressing a broader range of evaluative questions to be answered on the basis of different sources of ('weak' and 'strong') evidence for several of the predominant intervention strategies within the biodiversity portfolio.

In addition, as discussed earlier, the growing importance of GEF interventions directed at awareness building, natural resource management systems, legislation design, capacity building, political support at national or regional level, as well as other catalytic effects central to the GEF's role in supporting the global environment cannot be adequately assessed on the basis of quasi-experimental designs. These interventions and their intended effects are completely different from the relatively well-delineated site-specific interventions with clearly identifiable target groups which usually are the subject of quasi-experimental impact evaluation. In global environment interventions it is much more difficult to isolate the intervention from the wider institutional and policy environment, while effects are complex, diffuse and uncertain, making it very difficult if not impossible to determine counterfactuals. Given this complexity and the uncertainty surrounding GEF impacts on biodiversity, the evaluation should be at least as much about generating insights about the nature of processes of change instigated and/or influenced by GEF interventions, as about the actual demonstration of change attributable to the GEF. In practice, the latter cannot be established in a reliable manner without the first.[9]

Consequently, the impact evaluation should begin by mapping different processes of change related to different intervention activities within the portfolio. Then, at different levels of analysis, more precise data can be gathered to establish more precise claims of attribution. Theory-based evaluation constitutes a suitable framework for this type of approach.

Currently, the first phase of the impact evaluation is underway. This pilot phase focuses on intervention strategies that fall under the first strategic objective of the biodiversity program (i.e. catalyzing the sustainability of protected area systems).

[7] Basically, one can discern two groups of approaches. The first group of approaches employs experimental design as a basis for isolating intervention effects (preferably before an intervention has started, enabling ex ante–ex post comparisons). The second group relies primarily on advanced statistical analysis to isolate intervention effects from other influences. The latter group of approaches is mostly applied in cases where there are insufficient or no experimental design controls.

[8] Instead opting for cheaper less in-depth studies.

[9] In general, one can question, at least for the type of intervention activities sketched above, whether the determination of attribution of changes to GEF interventions is at all realistically possible. Accordingly, some authors talk about contribution instead of attribution, which basically entails a more comprehensive perspective on causality without a claim on determining the precise (magnitude of) causal effect from the intervention to change the dependent variable (Van den Berg, 2005).

In line with the foregoing, the methodological approach comprises two parallel evaluation exercises. The main exercise constitutes a theory-based evaluation of GEF support to protected areas in East Africa. This exercise is expected to generate useful lessons for further theory-based impact evaluation of other major intervention strategies within the portfolio such as particular mainstreaming approaches. Second, due to the availability of existing data it has been possible to implement in a more ad hoc manner at relatively low cost a quasi-experimental evaluation of the impact of GEF support to protected areas on avoided deforestation in Costa Rica.

In the next section we will concentrate on the main approach underlying the current and future phases of the impact evaluation, i.e. theory-based evaluation, and discuss how this approach can help to resolve some of the methodological challenges sketched so far in this article.

3. A theory-based impact evaluation approach

Over the past two decades theory-based evaluation has developed into an important methodological current in evaluation theory and practice (see for example Donaldson, 2003; Rogers, Hacsi, Petrosino, & Huebner, 2000; Weiss, 1997). Although particular theory-based approaches[10] differ in terms of the way theory is perceived and handled in evaluation, all approaches share the basic idea of theory as a set of assumptions underlying the way an intervention is supposed to work (i.e. the intervention theory). Consequently, the task of evaluators lies in reconstructing the main assumptions that underlie an intervention and subsequently, testing whether these assumptions are valid.

3.1. Addressing the attribution challenge

A common interpretation of an intervention theory is that it starts out from a systematic representation of the expectations and assumptions held by intervention staff and decision makers. These intentions and assumptions are only in part made explicit in formal documents (such as formal logical frameworks) and thus require further reconstruction. In other words, it is not a priori altogether evident what an intervention actually 'is', what it is meant to achieve and how it is meant to achieve it. There are several methodologies available for reconstructing intervention theories (for an overview see Leeuw, 2003; for interesting applications in development interventions see Carvalho & White, 2004 and Klein Haarhuis & Leeuw, 2004).

Subsequently, the question arises to what extent the assumptions that make up the theory are valid. In practice, evaluators have at their disposal a wide range of methods and techniques to test the intervention theory. In this sense, theory-based evaluation is not method-specific. For example, an intervention theory can constitute the basis for a quasi-experimental evaluation. Yet in many theory-based evaluation exercises the principle of analyzing causality and attribution works quite differently from quasi-experimental evaluations. Rather than trying to control for all the possible exogenous influences on impact variables in order to determine an intervention's effect on these variables, the evaluator relies primarily on logical argumentation, by carefully tracing all the assumptions underlying the theory (from inputs to outputs to impacts). Depending on the type of assumption, different sources of evidence come into play. For many types of assumptions (e.g. the influence of local norms and beliefs on institutional performance) 'hard' evidence is difficult to come by, nor can evidence always be guaranteed to be collected in a 'scientifically rigorous' manner. As a result, theory-based evaluations rely on the principle of triangulation of methods and sources of information, bringing as much evidence as possible into play (from different perspectives) in the assessment of hypotheses and assumptions. In many cases, there is no clear distinction between reconstruction and assessment as evaluators start out from a simple intervention theory and gradually work towards a more refined, empirically tested intervention theory that will help draw conclusions on attribution as well as serving as a basis for institutional learning.

In the GEF impact evaluation intervention theories can be usefully reconstructed at two levels: project level and portfolio level intervention theories. Here we largely focus on the latter. From the perspective of the portfolio, the most coherent evaluand at the highest level of aggregation is the 'thematic area' of intervention (e.g. protected areas, ecotourism, biosafety), representing major intervention strategies. A first step in the impact evaluation is to identify the main thematic areas and their respective weights in the biodiversity portfolio. For example, in a desk study in preparation of the impact evaluation, covering a sample of 30 projects from the biodiversity portfolio, it was found that, 20 out of 30 projects comprise intervention activities on land, water or species management. Furthermore 15 out of these 20 projects include activities on protected area management and 11 out of 20 develop activities on compatible resource use.

Fig. 1 shows the generic results chain underlying the logic of many conservation projects (Salafsky, Margoluis, Redford, & Robinson, 2002). This model provides a useful starting point to elaborate more specific intervention theories of strategies at portfolio level. The basic idea is one of a network of actors collaborating in the implementation of a number of actions designed to induce certain (behavioral) changes in organizations and/or individuals which ultimately affect biodiversity.

For the purpose of impact evaluation it is useful to distinguish process theory and impact theory (Rossi et al., 2004). The former refers to the assumptions and expectations underlying the processes of inputs leading to outputs, while the latter concerns the assumptions regarding particular outputs inducing processes of change resulting in final impacts. This differentiation is important in order to determine whether an observed lack of change is (mainly) due to problems of implementation, often referred to as implementation failure, or whether the concept of intervention (the idea that particular outputs lead to desired impacts) is fallacious, which is called theory failure (Suchman, 1967). As a result, in impact evaluation evaluators should have at their disposal substantial data about intervention outputs and the implementation process producing these outputs as a basis for further analysis. Only then can the more complex question of attribution in impact theory, i.e. the interaction between intervention outputs and external variables, the (potentially) complex and diffuse causal chain linking outputs to impacts, be addressed.

In this article we focus on impact theory, particularly the question of how common combinations of policy instruments under certain circumstances contribute to processes of change in institutions and the behavior of end users of natural resources and ultimately affect biodiversity variables. In practice, there is often a strong association between the thematic area of intervention and certain combinations of policy instruments that occur throughout projects. For example, protected area management relies on policy instruments like: capacity building intended to strengthen the institutional and financial sustainability of the protected area framework; imposing restrictions on land use and natural resource exploitation with the intention of generating changes

[10] Theory-based evaluation (introduced by Weiss) is probably the most commonly used term. Other terms are used in the literature to refer to broadly similar approaches.

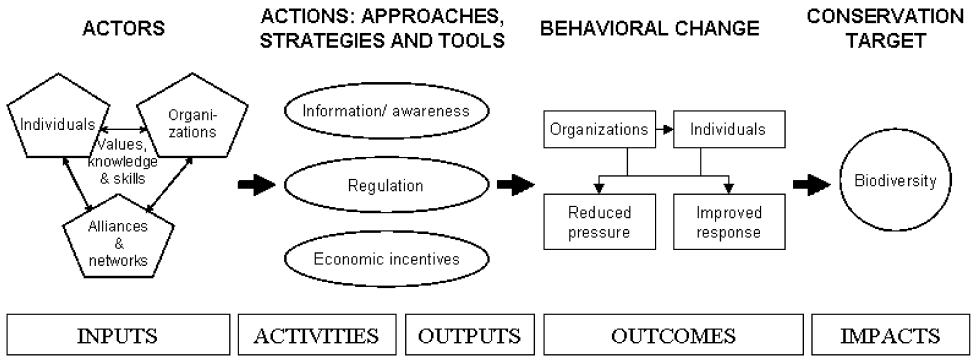


Fig. 1. Basic intervention theory conservation projects. *Source*: Adapted from Salafsky et al. (2002) (see also GEF, 2006c).

in land use and natural resource exploitation; and awareness raising on natural resources in order to support changes in land use and natural resource exploitation. This connection between the thematic area of intervention and policy instruments is crucial as theories on the effectiveness of (combinations of) policy instruments constitute essential building blocks of intervention theories at thematic area level.

Some work on establishing this type of connection has already been done. For example, in another desk study covering medium-sized and full-sized projects from three different portfolios[11] an inventory of site-specific interventions related to agriculture was made (GEF, 2006d). First, major thematic areas were identified. It was found for example that approximately 59% of all the projects included alternative livelihoods activities, 40% ecotourism, 49% sustainable land use techniques, and 14% reforestation. In addition, an inventory was made of different policy instruments. For example, 14% of the projects included microgrants (economic incentives), 24% microcredit (economic incentives), 42% education and awareness building (information/awareness), 26% technical assistance (information/awareness), 32% community-based natural resource management (regulation), and 8% land user agreements (regulation). Further analysis would make it possible to identify the major patterns of thematic areas linked to particular combinations of policy instruments.

The causal chain linking GEF intervention outputs to changes in the behavior of end users of natural resources can be relatively short and straightforward, for example in site-specific GEF interventions that directly target end users of natural resources (e.g. technical assistance resulting in crop diversification). In other cases, such as in GEF intervention activities directed (at least in the first instance) at changes at institutional level (e.g. capacity building resulting in improved legislation), the subsequent causal effects on end users of natural resources are more diffuse and difficult to capture. In the latter case evaluators require a workable model in order to better conceptualize these potentially complex processes.

The same goes for the human behavior–environment interface which is the most complex part of the impact theory. In a few cases the linkages between human behavior and environmental change are straightforward and the assumptions connecting changes in institutional and individual human behavior to changes in environmental benefits can be reconstructed and tested in a relatively simple manner. In other cases, the evaluator is facing the frontier of the state of the art of research in the natural sciences (e.g. biology, ecology) and can only make very rough assumptions about these complex causal relationships. Several models available in the literature can assist evaluators in the reconstruction of useful intervention theories. An example is the so-called pressure-state-response model (PSR) developed by the OECD in the 1990s (OECD, 2003). The basic logic of the model is represented in Fig. 1 and more elaborately illustrated in Fig. 2. The model helps to classify the effects of GEF interventions in terms of reducing pressures on the environment, improving the state of the environment or improving responses by institutional actors. In addition, for each category of effects the GEF's contribution can be assessed in relation to other influencing variables. In short, evaluators can conduct PSR analyses for selected projects and on this basis will be able to articulate an intervention's influence on wider processes of environmental change.

3.2. Addressing the aggregation challenge

We can make a distinction between the aggregation of quantitative data from project level to higher levels of intervention (thematic area, portfolio, global trends) and the process of generalization or theory-building from project level intervention to impact theories at the level of thematic areas of intervention. Regarding the latter, intervention theories can potentially constitute a powerful basis for institutional learning and knowledge management on impacts of different types of GEF interventions in different settings. To develop a feeling of how this might work, let us briefly highlight a few points of attention on intervention theory-building relevant to the impact assessment of the biodiversity program.

Starting out from a generic results chain as shown in Fig. 1 and using information from different project documents of projects pertaining to a particular thematic area as well as other sources of information (e.g. existing literature, interviews with project staff), crude intervention theories can be reconstructed representing the basic causal linkages between inputs, activities, outputs, outcomes and impacts. Subsequently, evaluators further refine and test these intervention theories on the basis of multiple sources of information, the main sources being: information available at project level (progress reports, end-of-project evaluations, field visits, staff interviews), existing evaluative evidence within the GEF (e.g. program studies, thematic cross-cutting studies, overall

[11] Biodiversity, Land Degradation and Multi-Focal Areas. All medium-sized and full-sized projects approved between January 2000 and June 2005 ($n = 332$).

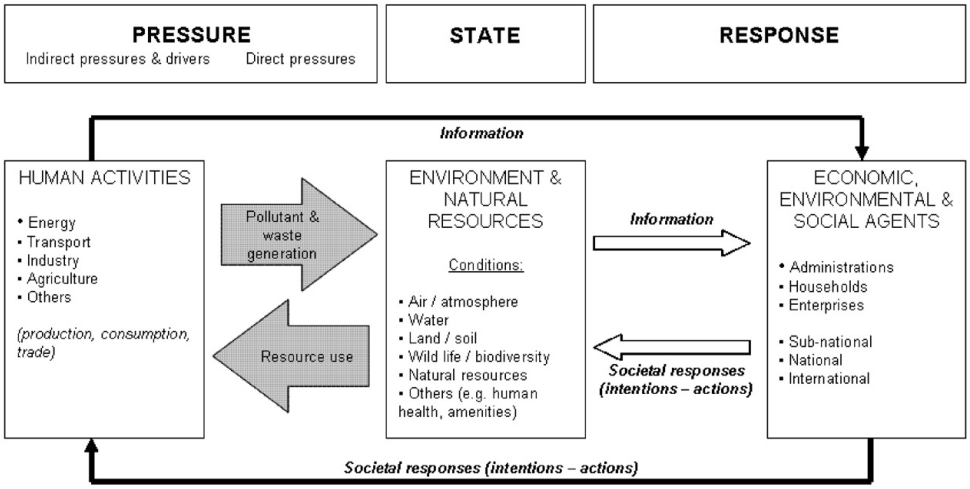


Fig. 2. The pressure-state-response model. *Source*: OECD (2003).

performance studies), studies on similar interventions elsewhere, expert interviews, and academic literature.

The process of intervention theory refinement basically works as follows. After the initial reconstructions, evaluators identify key assumptions to be tested and correspondingly define more focused study questions and indicators. These assumptions, questions and indicators will provide a structure for more systematic empirical data collection at project level. In this process the thematic area intervention theory is more and more refined taking into account contextual factors of different projects pertaining to a thematic area. Pawson and Tilley (1997) introduced the principle of context-mechanism-outcome (CMO) as the basic ingredients for theory-building about what works (for whom) under what circumstances. This works roughly as follows. Evaluators first look for outcome patterns at thematic area level, the major mechanisms contributing to these patterns (institutional structures, combinations of policy instruments) and the particular contextual settings (at project level) that condition these mechanisms. This leads to a first reconstruction of CMO theories. Project level data (and other empirical data) are subsequently fed into the theory in order to further refine the CMO assumptions. As a result, new theories on CMO emerge, which subsequently can be subjected to further refinement when new empirical data become available. This process can be repeated in an iterative manner until the best possible explanatory model is achieved.

Intervention theories can be developed in relation to different levels of impact. As discussed earlier, depending on the type of intervention, and state of knowledge and available data on processes of environmental change, impacts are captured at three levels: institutional changes (e.g., capacities, awareness, political will), behavioral changes of end users of natural resources (e.g. land use changes, reductions in harvesting of natural resources), and changes in biodiversity variables (e.g. species diversity, ecosystem quality). The evaluators will concentrate on developing causal theories that link particular thematic areas of intervention to the levels of impact which (in first instance) are deemed most relevant for these areas. For example: under what circumstances have the main thematic strategies of the GEF subscribing to the (broad) purpose of institutional change (e.g. enabling activities, biosafety projects) induced positive and sustainable changes at institutional level? Under what circumstances have the main thematic strategies of the GEF subscribing to the (broad) purpose of directly influencing the behavior of end users of natural resources (e.g. projects on mainstreaming) induced positive and sustainable changes at this level? Subsequently, the causal analysis is extended to biodiversity targets by using among other things the PSR model introduced earlier.

In practice, the majority of interventions comprise objectives and activities, which explicitly aim at inducing institutional change, behavioral changes of (groups of) individual users of natural resources as well as indirect (catalytic) processes bearing on biodiversity variables elsewhere. The above-mentioned evaluation strategy of linking thematic areas to levels of impact therefore should be triangulated through case study analyses (and other sources), of the type discussed earlier, in order to analyze the interplay between different intervention mechanisms (policy instruments, institutional structures), specific contextual variables and effect patterns at all levels of change.

The second dimension of aggregation, the definition and measurement of impact indicators is related to the foregoing. The intervention theory is the principal basis for indicator development. In addition, it constitutes the basis for other types of empirical assessment regarding the contribution of GEF interventions to processes affecting biodiversity variables and intermediate levels of change.

Existing problems of alignment and aggregation of project level information cannot be fully compensated by the evaluation's data collection activities. The number and diversity of GEF interventions makes it too costly to be able to define relevant indicators and collect data for all types of interventions. As mentioned earlier the knowledge deficit on particular environmental processes poses an additional barrier to indicator development. This is the main variable which determines the type and precision of indicator to be measured.

Roughly, we can distinguish between three levels of precision. The scope for collecting and aggregating precise quantitative data on different types of biodiversity is limited. This may be feasible with regard to specific endangered species, such as the giant panda, which has been the target of a number of GEF projects, including the Qinling Forest Reserve project in China (GEF, 1995). For projects with broader aims, such as the protection of

mangrove forests, it is likely to be far more difficult to define a specific set of impacts with corresponding measurable indicators. In projects in which data on biodiversity are not available but where it is relatively easy to identify the nature of the causal linkages (i.e. the intervention theory) between intervention outputs and processes of (environmental) change, the PSR model illustrated in Fig. 2 can constitute the basis for defining questionnaires and indicators in order to collect ordinal data concerning intervention effects on environmental variables. This type of data can also be relatively easily aggregated across interventions and compared to national or international trends. Finally, in projects that comprise intervention activities about which little is known regarding possible causal patterns towards environmental change, more attention should be paid to intervention theory reconstruction (preferably complemented by field assessments, expert interviews, and consultation of academic literature). This will provide the basis for collecting data on relevant proxy indicators of biodiversity (see Duelli & Obrist, 2003) such as the number of hectares of protected area, the intensity of land use or agricultural diversification. In addition, it is advisable in these cases to focus more on intermediate levels of impact.

3.3. Additional considerations

In the complex impact evaluation under discussion it is important to arrive at usable abstractions of the GEF's strategies and their effects on the global environment. However, one has to keep in mind the reductionist nature of such abstractions. The complexity surrounding each individual GEF project context can only be captured in a limited way by the higher-level intervention theories. So, the evaluator team may also need to develop some detailed case study analyses to complement the process of reconstructing and refining intervention theories. A narrative historical approach can be very useful to generate additional understanding about the complex linkages between GEF interventions, the context in which they operate and possible outcomes and impacts. Such an approach would focus on the evolution of a particular GEF intervention or a series of GEF interventions in a particular region or country (e.g. in a biodiversity 'hot spot'). The in-depth illustration of the embeddedness of current GEF interventions in past interventions and strategies of other institutional actors will be particularly useful. In addition, of particular interest will be qualitative analyses of the sustainability of effects and patterns of replication, since these concepts are multi-dimensional and relatively difficult to analyze empirically. The evaluators will thereby gain a more detailed understanding of the complex interactions between interventions and social-institutional and environmental dynamics and will be able to develop a more detailed picture of the induced processes of change and new impacts on biodiversity. Another option would be a focused quasi-experimental study on the effectiveness of a particular instrument or type of support. The quasi-experimental exercise referred to earlier in which the effectiveness of GEF support to protected areas on avoided deforestation in Costa Rica is analyzed, constitutes a useful example of such an exercise with a more narrow focus. Such an exercise can rigorously test specific assumptions of the overall impact theory on GEF support to protected areas.

4. Lessons learned

In our discussion about the methodological challenges in the GEF impact evaluation of the biodiversity program, three important lessons emerged. First, theory-based evaluation provides a useful overall methodological framework, which subsequently can be elaborated into specific methods and include divergent qualitative and quantitative sources of evidence that focus on particular aspects of the intervention theory. In this article, we did not enter into that level of detail, as we focused on the key issues at portfolio level. Second, intervention theories should be reconstructed at the level of thematic area of intervention, the highest aggregate level of intervention representing more or less coherent intervention strategies. Initial preliminary reconstructions based on multiple sources of information (primarily staff interviews, project documents and field visits) will provide a basis for further empirical data collection in order to develop claims on attribution, while at the same time offering a structure for aggregating evidence and lessons on impact across interventions. Third, due to practical limitations and knowledge constraints as well as the nature of GEF interventions, evaluators should analyze causality at various points in the causal chain between outputs and impacts. In practice, this implies that evaluators address causal links between intervention outputs and changes at institutional level, behavioral changes at the level of end users of natural resources, or changes in biodiversity variables.

The first phase of the impact evaluation that is currently underway is expected to generate useful methodological lessons. Though limited in terms of the number of project-specific studies to be undertaken and the number of intervention theories to be reconstructed and tested in a detailed manner, the approach could form an important model for future phases of the impact evaluation as well as other GEF evaluation studies, and a foundation for institutional learning and knowledge management on GEF intervention strategies. In addition, theory-based evaluation can be usefully applied in other evaluative exercises. Without being exhaustive we mention a few options. First, conventional meta-evaluations of end-of-project evaluations focus on extracting lessons on both the content (e.g. performance of projects) as well as the quality of project level evaluations. In the future, as more and more GEF projects start reporting systematically on outcome and impact-related data, such exercises could be expanded by synthesizing impact-related evidence into portfolio-level statistics as well as more qualitative intervention theories covering major thematic areas of intervention. Improved guidelines on the type of information that should be reported at project level would benefit this type of analysis. Second, there are a number of learning projects on biodiversity issues currently funded by the GEF and managed by the Implementing Agencies. The insights produced by these projects would constitute important ingredients of knowledge management activities on GEF impact at portfolio level, complementary to the insights generated by existing thematic and meta-evaluations.

Acknowledgements

An earlier version of this article was presented at the UKES-EES Conference in London in October 2006. The authors would like to thank participants at the conference, Frans Leeuw, Robrecht Renard, Osvaldo Feinstein and the anonymous reviewers for their comments.

Appendix A

The structure of the biodiversity program is the following.

Operational programs

1. Arid and Semi-arid Zone Ecosystems
2. Coastal, Marine, and Freshwater Ecosystems

3. Forest Ecosystems
4. Mountain Ecosystems
5. Conservation and Sustainable use of Biological Diversity important to Agriculture

Strategic objectives

1. Catalyzing the sustainability of protected area systems.
2. Mainstreaming biodiversity conservation and sustainable use into production landscapes/seascapes and sectors.
3. Safeguarding biodiversity through: (i) building country capacity to implement the Cartagena Protocol on Biosafety (CPB) and (ii) prevention, control, and management of invasive alien species.
4. Capacity building to support the implementation of the Bonn Guidelines on access to genetic resources and benefit-sharing.

Further information on the GEF biodiversity program as well as GEF policies and strategies can be found on http://thegef.org (last consulted, January 9, 2008).

References

Bemelmans-Videc, M. L., & Rist, R. C. (Eds.). (1998). *Carrots, sticks and sermons: Policy instruments and their evaluation*. New Brunswick: Transaction Publishers.

Carvalho, S., & White, H. (2004). Theory-based evaluation: The case of social funds. *American Journal of Evaluation*, *25*(2), 141–160.

CGD (2006). *When will we ever learn? Improving lives through impact evaluation*. Report of the Evaluation Gap Working Group. Center for Global Development, Washington, DC.

Donaldson, S. I. (2003). Theory-driven program evaluation in the new millennium. In S. I. Donaldson, & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium*. London: Lawrence Erlbaum Associates.

Duelli, P., & Obrist, M. K. (2003). Biodiversity indicators: The choice of values and measures. *Agriculture, Ecosystems and Environment*, *98*, 87–98.

Ferraro, P. J., & Pattanayak, S. K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *Plosbiology*, *4*(4), 482–488.

GEF (1995). *People's Republic of China: Nature reserves management project*. Project appraisal document. Global Environment Facility, Washington, DC.

GEF (2003). *Measuring results of the GEF biodiversity program*. Monitoring and evaluation working paper 12. Global Environment Facility, Washington, DC.

GEF (2004a). *Biodiversity program study*. Global Environment Facility, Washington, DC.

GEF (2004b). *Mainstreaming biodiversity in production landscapes and sectors*. Discussion paper. Global Environment Facility, Washington, DC.

GEF (2005). *OPS3: Progressing toward environmental results*. Third overall performance study. Global Environment Facility, Washington, DC.

GEF (2006a). *Four year work program and budget of the Office of Monitoring and Evaluation: FY06-09 and results in FY05*. Global Environment Facility, Washington, DC.

GEF (2006b). *GEF impact evaluations: Initiation and pilot phase-FY06*. Approach paper. Global Environment Facility, Washington, DC.

GEF (2006c).*GEF impact evaluation: Final report on a proposed approach*. Working document prepared for the GEF Evaluation Office by Foundations of Success. Global Environment Facility, Washington, DC.

GEF (2006d). *Sustaining global environmental benefits through changes in farmers' behavior: A review of GEF-funded activities*. Progress report, unpublished. Global Environment Facility, Washington, DC.

IEG. (2006). *Conducting quality impact evaluations under budget, time and data constraints*. Washington, DC: Independent Evaluation Group, World Bank.

Klein Haarhuis, C. M., & Leeuw, F. L. (2004). Fighting governmental corruption: The new World Bank programme evaluated. *Journal of International Development*, *16*, 547–561.

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation*, *24*(1), 5–20.

MEA. (2005). *Ecosystems and human well-being: Synthesis—Millennium Ecosystem Assessment*. Washington, DC: Island Press.

Mog, J. M. (2004). Struggling with sustainability: A comparative framework for evaluating sustainable development programs. *World Development*, *32*(12), 2139–2160.

OECD. (2003). *OECD environmental indicators: Development, measurement and use*. Paris: OECD.

OED. (2005). *OED and impact evaluation: A discussion note*. Washington, DC: Operations Evaluation Department, World Bank.

Pawson, R. (2002). Evidence-based policy: In search of a method. *Evaluation*, *8*(2), 157–181.

Pawson, R. (2006). *Evidence-based policy: A realist perspective*. London: Sage Publications.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage Publications.

Pritchett, L. (2002). It pays to be ignorant: A simple political economy of rigorous program evaluation. *The Journal of Policy Reform*, *5*(4), 251–269.

Rao, P. K. (2000). *Sustainable development: Economics and policy*. Oxford: Blackwell Publishers.

Rogers, P.J., Hacsi, T.A., Petrosino, A., & Huebner, T.A., (Eds.) (2000). *Program theory in evaluation: Challenges and opportunities. New directions for evaluation*. Vol. 87. San Francisco: Jossey-Bass.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage Publications.

Salafsky, N., Margoluis, R., Redford, K. H., & Robinson, J. G. (2002). Improving the practice of conservation: A conceptual framework and research agenda for conservation science. *Conservation Biology*, *16*(6), 1469–1479.

Salamon, L. (1981). Rethinking public management: Third party government and the changing forms of government action. *Public Policy*, *29*(3), 255–275.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage Foundation.

Van den Berg, R. D. (2005). Results evaluation and impact assessment in development co-operation. *Evaluation*, *11*(1), 27–36.

Vedung, E. (1998). Policy instruments: Typologies and theories. In M. L. Bemelmans-Videc, & R. C. Rist (Eds.), *Carrots, sticks and sermons: Policy instruments and their evaluation*. New Brunswick: Transaction Publishers.

Weiss, C. H. (1997). Theory-based evaluation: Past, present and future. In D. J. Rog, D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice and methods. New directions for evaluation*, Vol. 76. San Francisco: Jossey-Bass.

White, H. (2003). Using the MDGs to measuring donor agency performance. In R. Black, & H. White (Eds.), *Targeting development: Critical perspectives on the Millennium Development Goals*. London: Routledge.

CHAPTER 3

Vaessen, J. and F.L. Leeuw (2009) "Interventions as theories: closing the gap between evaluation and the disciplines?", in: J. Vaessen and F.L. Leeuw (eds.) *Mind the gap: perspectives on policy evaluation and the social sciences*, New Brunswick, Transaction Publishers.[1]

[1] With minor alterations. References to the original book have been deleted and in some cases references to the current volume of papers have been added.

Introduction

This chapter discusses the idea of perceiving interventions as theories. In particular it focuses on the interplay between two types of theories; on the one hand theory as a systematic reconstruction of the underlying assumptions of an intervention, and on the other, theory as an abstraction of a particular aspect of social reality, product of a body of research within the social and behavioral sciences (see Chapter 1). Within the framework of this perspective, it addresses the issue of the uneasy relationship between evaluation and the social and behavioral sciences [1]. One of the questions which come to mind is how to improve this uneasy relationship. How can evaluators make better use of relevant substantive knowledge generated within the social sciences? How can evaluators contribute to knowledge development on interventions?

Undoubtedly several different recommendations and scenarios can be developed depending on the purpose and context of a particular evaluation exercise. Consider the different traditions of thought and practice in evaluation. One that immediately comes to mind in relation to the questions posed above is the theory-oriented tradition, most notably the work done under the banners of theory-driven, theory-based, theory of change or program theory evaluation. Reading through major contributions within the theory-oriented evaluation tradition, one unmistakably comes to the conclusion that the two questions posed above, if not at the heart of the tradition, are certainly of special interest to theory-oriented evaluators. Indeed, a quest to resolve the questions above would almost automatically become an exercise which falls within in the theory-oriented evaluation tradition.

The main purpose of theory-oriented evaluation is to help the evaluator to better understand processes of implementation in interventions as well subsequent induced processes of change. In other words, theory-oriented evaluators are very much concerned with the question of why particular interventions work in certain circumstances by looking inside the 'black box' of interventions. In doing so, they focus on reconstructing chains of assumptions which help explain how interventions work. These chains of assumptions have received names such as policy theories (e.g. Hoogerwerf, 1990), program theories (e.g. Bickman, 1987), intervention theories (e.g. Vedung, 1997) or theories of change (e.g. Weiss, 1997). In this chapter we prefer the use of the word intervention as a generic indication for policies, programs and projects, including so-called 'tools of government' (such as subsidies, levies, information campaigns, inspections and loans) and accordingly will apply the term "intervention theory" to refer to the assumptions underlying interventions.

The theory-oriented focus on underlying assumptions behind interventions has recently come under renewed attention within the light of knowledge accumulation about interventions (e.g. Pawson, 2003). Yet, perhaps more importantly, different authors have argued and illustrated how a theory-oriented focus can provide the basis for, or enhance the quality of, different types of evaluation exercises (Bickman, 1987; Weiss, 1997; Donaldson, 2003). Insights from social science research

play an important role in the realization of both types of potential advantages of theory-oriented evaluation exercises.

In this chapter we will first explore the main characteristics of the theory-oriented evaluation tradition, thereby paying special attention to social science theory. In what ways does social science theory come into play the practice of (theory-oriented) evaluation? What are the functions of social science theory in theory-oriented evaluation? This overview, comprising two sections following this introduction, provides the basis for a concrete illustration, which is divided into two parts. The first part sketches a generic framework for reconstructing intervention theories, inspired by the work of Coleman (1986, 1990) and others on social mechanisms. Subsequently, the presented framework will be further illustrated by means of an example. Finally, the chapter concludes with a reflection on the foregoing and implications for evaluation practice.

The theory-oriented evaluation tradition

In terms of the 'history of ideas', the very concept of assumptions underlying policies and programs goes back to the work of the German-English sociologist Karl Mannheim, who, in 1934, coined the concept of the 'principia media', i.e. the time and space restricted assumptions on what makes a policy effective (Mannheim, 1951). Nowadays, these principia media would be characterized as underlying intervention theories. A second and rather distinct *'ideengeschichtliches'* element of this work can be found in the work of (cognitive) psychologists who during the 1940s and 1950s started to unpack 'lay theories of behavior' (hence the people applying them were called 'lay psychologists'). Here the focus was on unpacking the beliefs and causes which lay persons attribute to their own behavior and that of other (corporate) actors. Finally, the sociological tradition of *ethnomethodology* focused on how to uncover the implicit rules, 'regulations', arrangements that people use in day-to-day life to understand why they use them (Garfinkel, 1967). These are the more important 'ideengeschichtliche' backgrounds of earlier work closely resembling current practices of uncovering the underlying assumptions behind interventions by theory-oriented evaluators. It is important to know that, with Mannheim as an exception, there was no focus at all on interventions as such.

Over the last forty years or so a number of influential publications (e.g. Suchman, 1967; Chen and Rossi, 1980; Chen, 1990; Leeuw, 1991; Weiss, 1995; Pawson and Tilley, 1997; Rogers et al., 2000) have contributed to the creation of what can be called a theory-oriented evaluation tradition, mostly known under names such as theory-based, theory-driven or program theory evaluation. In a way this tradition represents a 'systemization' of a lot of evaluative work carried out by policy researchers and evaluators, who for decades have been focusing on the underlying assumptions of interventions (policies, programs, projects). One of the distinguishing features of theory-oriented evaluation in comparison with earlier work is the particular focus on how underlying assumptions behind interventions can be reconstructed in a valid and reliable manner (Leeuw, 1991; Leeuw, 2003). The increased

attention to looking inside the 'black box' of interventions can in part be attributed to a sense of dissatisfaction with the large quantity of evaluations which focused on measuring and demonstrating outcome and impact without paying due attention to the complex interplay of factors which might be causing the (lack of a) manifestation of particular effects. This can be illustrated by the experience of criminal justice program evaluations based on randomized designs at the end of the 1950s in the UK. Since the establishment of the Research Unit within the British Home Office in 1957, slowly but steadily experimental evaluations were conducted. Sometimes these studies were carried out by Home Office researchers, sometimes by others. Examples include the effectiveness studies of social work in prisons, the effectiveness of probation programs (the 'Impact Experiment') and the effectiveness of a therapeutic community in a residential school for delinquent children between 10 and 17 (Nuttal, 2003: 273). These studies "mark both the beginning and the end of random allocation experiments in Home Office research" (Nuttal, 2003: 274). Problems the evaluators had in doing these studies, the mixed bag of results and the problem that "the experiment[s] might have been able to say what had happened but [it] could not answer how or why" (Nuttal, 2003: 277) are the reasons.

In contrast to the latter type of evaluations (which are sometimes roughly characterized as method-oriented), theory-oriented evaluators are less concerned with optimizing the internal (and external) validity of established causal links between intervention and effects and, as a result, are less concerned with developing methods and tools for this purpose. Theory-oriented evaluation is not method-specific; different types of assumptions call for particular types of evidence in order to test them, in turn implying divergent methods of data collection and analysis. This has two important consequences. First, and we will touch upon this point later on, a theory-oriented focus can be complementary to (and reinforce the quality of) evaluation exercises which employ certain methodological designs for the purpose of say determining the impact of a particular intervention [2]. Second, the flexibility towards methods and tools in theory-oriented evaluation studies has led to quite some variety in application, making it more difficult to arrive at clear unequivocal practical guidelines. This problem is further compounded by the disparities between the different currents of practice that make up the theory-oriented tradition.

At the risk of oversimplification we will briefly recapitulate the main sources of heterogeneity. In doing so, at the same time we are in fact covering the main defining characteristics of theory-oriented work. Broadly these fall into two categories: the nature of intervention theories and the methods used to construct and test them.
A first major issue in theory-oriented evaluation concerns the *sources* of the assumptions. Broadly, one can discern three major sources of assumptions. First, politicians, decision makers, intervention staff, all harbor their own beliefs and expectations as to how an intervention should and will work [3]. Second, the assumptions held by other groups of stakeholders. Finally, assumptions are based upon social science research regarding intervention-specific topics as well as more generic behavioral and social theories (Chen, 1990; see also Donaldson, 2003). The balance between the three as inputs for intervention theory is contested, resulting in divergent practices.

Some authors (e.g. Wholey, 1987) have argued for a focus on intervention staff as principal determinants of the intervention theory, as they are the major actors who continuously shape the implementation of the program. This is especially relevant for two purposes. First of all, from a policy point of view, a systematization of the intentions and expectations of decision makers and staff provides a relevant reference point for the rest of the evaluation study (e.g. comparing the theory with practical implementation and data on output and outcome). Second, the exercise of uncovering stakeholder intentions can be an essential part of a formative evaluation strategy aimed at advancing a shared and solid (tested) vision on what the intervention should achieve and how it should be achieved. Especially within the light of the latter purpose, the reconstruction of stakeholder theory is narrowly related to the work on program logic (e.g. Coffman, 1999). The difference between the two is not always clear-cut. Program logic is first and foremost a representation of articulated intervention strategy whereas stakeholder theory is reconstructed for the purposes of evaluation and based on articulated as well as tacit assumptions about the intervention (requiring a method for uncovering these hidden assumptions). In addition, theory-oriented evaluators are inclined to give more attention to contextual variables, and accordingly are prone to develop more systemic representations of the intervention [4].

Some twenty-five years ago, an alternative current in theory-oriented evaluation emerged, emphasizing the role of substantive social science theory in the definition and development of intervention theories (e.g. Chen and Rossi, 1980). Rossi, Chen and others see evaluation primarily as applied social research. According to this view, evaluators should be able to offer insights to decision makers and staff whereby "the evaluator should actively search for and construct a theoretically justified model of the social problem in order to understand and capture what a program really can do for a social problem" (Chen and Rossi, 1980: 111). More than a decade and a half later, Weiss voiced a similar concern. "Evaluators are currently making do with the assumptions that they are able to elicit from program planners and practitioners or with the logical reasoning that they bring to the table. Many of these theories are elementary, simplistic, partial, or even outright wrong. Evaluators need to look to the social sciences, including social psychology, economics and organization studies, for clues to more valid formulation, and they have to become better versed in theory development themselves" (Weiss, 1997: 51). Most theory-oriented work nowadays adheres to an integrative view of theory reconstruction. Depending on the purposes of the evaluation, both stakeholder theory and insights from the social sciences are used to make sense of interventions. Nevertheless, the words of Weiss echoing Chen and Rossi remain as relevant as they were ten years and twenty five years ago. An evaluator is often well versed in terms of methodology (e.g. interview skills, statistical design), knowledge of a field of intervention (e.g. health policy), and knowledge of modes of intervention (e.g. vaccination programs). In contrast, evaluators often fall short in their understanding about, among other things, the motivations that drive people in different settings, a basic element of many of the assumptions behind interventions. It is especially this type of knowl-

edge that the state of the art substantive theories from economics, sociology and psychology can provide.

A second issue which characterizes theory-oriented evaluation refers to the *focus* of the assumptions. In 1967, Suchman was the first to distinguish between implementation failure and theory failure (Suchman, 1967; see also Weiss, 1972). An intervention can fail either because of weak implementation, or, in the case of smooth implementation, a weak theory regarding the types of activities and outputs needed in certain circumstances to bring about a desired effect. This problem has been analyzed in research on what has come to be known as the performance paradox. The main idea is that high positive scores on process criteria, on 'auditability', on 'inspectorability' of interventions do not guarantee high positive scores on (final) impact and effectiveness (Van Thiel and Leeuw, 2002).

A related and very useful distinction is that between process theory and impact theory (Rossi et al., 2004). Process theory concerns the assumptions behind the way in which the implementation of an intervention takes place, with the focus on delivery of outputs. It stresses processes that have to be 'steered'. It can be further decomposed into program organizational plan and service utilization plan, referring respectively to the assumptions regarding activities, resources and persons, and the nature of interactions with target populations, which are expected to ensure a successful delivery of intervention outputs. Impact theory is concerned with the set of assumptions on how outputs induce processes of change, eventually resulting in particular intended (and unintended) effects.

Though the distinction is useful in practice, the two are evidently related. For example, the influence of the characteristics of the service delivery process (part of process theory) goes beyond the delivery of output as such and in part determines the potential outcome of an intervention. To briefly illustrate this, consider the following. It makes quite a difference whether a community training program on non-agricultural income-generation activities (e.g. in rural areas in a developing country) is designed and implemented in a participatory manner or not. Two training programs with the same course content might lead to completely divergent outcomes depending on whether participants are consulted or invited to decide upon, for example, program organizational issues (schedules, venues, entry rules) or not.

Chen's work on theory-driven evaluation (Chen, 1990, 2005) has generated a substantial part of the analytical underpinnings of theory-oriented evaluation. He has probably produced the most elaborate and in-depth discussion of the *nature* of assumptions, a third important issue and source of heterogeneity in theory-oriented evaluation. As a first classification, Chen distinguishes between descriptive and prescriptive theory. Descriptive theory concerns the set of assumptions regarding the underlying causal relationships between an intervention's activities and outputs and the alleviation of a social problem. These assumptions can be 'politicians/decision makers/staff/stakeholder-based' or inspired by other sources of (e.g. social science research) evidence. The same applies for prescriptive theory, which captures the assumptions regarding what should be done in terms of objectives, activities and outputs in order to achieve a desired result. In Chen's (1990) first comprehensive account of theory-driven evaluation he carefully distinguishes between

six types of theory-oriented evaluations, three of a prescriptive and three of a descriptive nature. The practical use of these different types of theories has, however, been limited. Indeed, as remarked by Weiss (1997), most of the intervention theories are descriptive in nature. Thorough elaboration and testing of the theories is expected to generate prescriptive recommendations as to what type of intervention is needed to successfully alleviate social problems in particular circumstances.

A second dimension in the discussion on the nature of assumptions concerns the nature of causality. The concept of causality is crucial to the reconstruction of intervention theories and detailed accounts on types of causality in theory-oriented evaluation abound in the literature (e.g. Rogers, 2000; Chen, 2005). A relatively recent current in theory-oriented evaluation, called realist evaluation, distinguishes itself on the basis of its particular notion of causality. Realist evaluators adhere to the principle of generative causality. Accordingly, interventions should not be treated as independent variables or treatments. They are embedded in complex processes of human interaction. It is essentially these processes of human agency that can make seemingly identical programs work in one place and fail in the other (Pawson and Tilley, 1997). In other words, interventions are perceived as opportunities which may or not be acted upon in certain circumstances. Evidently, economic, sociological and psychological theories on human behavior can play an important role in the process of making sense of the complex embeddedness of interventions in processes of human interaction (both in the implementation sphere as well as in the impact sphere). Consequently, one of the merits of realist evaluation is that is has brought human behavior to the centre stage of the evaluation and can be regarded as an important move towards reconciliation between professional evaluation and disciplinary research.

A final element in our succinct overview of theory-oriented evaluation is the discussion of *methodologies* for reconstructing and assessing the assumptions. The range of systematic methodologies (including different methods) on theory reconstruction is rather limited, although it is expected to grow along with the growing number of applications of theory-oriented evaluation (for examples see Leeuw, 2003; Trochim, 1989; Rosas, 2005; Wholey, 1987). In practice, theory reconstruction is often described as highly non-linear, with no clear consecutive strategies of data collection and analysis (Donaldson and Gooler, 2003). One of the crucial elements in theory reconstruction about which no consensus exists (and indeed cannot exist due to distinct evaluation purposes and contexts) concerns the issue of addressing multiple divergent stakeholder perspectives. Leeuw (2003) describes three rather different methods each suited to particular purposes and contexts (e.g. summative retrospective evaluation versus formative prospective evaluation). Once reconstructed, an intervention theory can provide the basis for a broad range of evaluation activities to further validate and test the theory. Besides conventional empirical process, output and impact evaluations, intervention theories can be tested by looking at (e.g.): the internal logic of the theory, the needs of target groups, relevant existing empirical evidence and substantive social science theories relevant to intervention effectiveness (see Smith, 1989; Rossi et al., 2004). In some approaches no distinction is made between reconstruction and assessment, the evaluation study

being constituted of a highly interactive process between evaluator and stakeholders towards the definition of a refined intervention theory (e.g. Pawson and Tilley, 1997).

It is beyond the scope of this chapter to review the whole landscape of methodologies available for building and testing intervention theories. An important point we would like to highlight is the fact that there is quite some diversity among theory-oriented approaches in terms of how social science theory is brought into the analysis. For example, in the so-called policy-scientific approach a clear distinction is made between the phases of theory reconstruction and theory assessment (Leeuw, 2003). First, the evaluator carefully reconstructs stakeholder theory on the basis of project documents and interviews. Subsequently, the reconstructed theory is assessed on the basis of multiple sources of evidence including substantive knowledge generated by social science research. An alternative procedure has been described by Pawson and Tilley (1997), who talk about an iterative procedure of intervention theory reconstruction. The evaluator starts with an initial theory which is defined on the basis of multiple sources (e.g. impressions, observations, existing empirical evidence, relevant substantive theories from the social sciences) and subsequently confronts different stakeholders with this theory. A process of iterative discussion and adaptation of the theory starts, with the objective of confirming and falsifying parts of it, eventually producing a satisfactory refined intervention theory. Rossi et al. (2004) also talk about an iterative process of intervention theory-building in which the theory is gradually reconstructed by moving back and forth between stakeholders, documentation and other sources of evidence.

The role of social science theories in theory-oriented evaluation

Although the heterogeneity in thought and practice within the theory-oriented evaluation tradition is rather striking, it is not surprising. Different purposes and contexts call for distinct approaches. Indeed, the very disciplinary and technical backgrounds of the evaluators determine to a large extent the way abstractions are made about intervention reality and in what way they are validated and tested. From the foregoing however, it becomes clear that social science research plays an important role either as an inspiration or as a benchmark in theory-oriented evaluation.

Let us briefly discuss what roles social science theory can play in theory-oriented evaluation. We will approach this issue from the perspective of the potential advantages of adopting a theory-oriented evaluation approach. Several authors have illustrated the potential advantages of theory-oriented design. Broadly, they fall into three categories: advantages for knowledge accumulation about human behavior and behavioral change, advantages for intervention planning (see Chapter 6), and advantages for evaluation design and implementation (Birckmayer and Weiss, 2000; see also Bickman (1987) and Donaldson (2003) for illustrations of the potential advantages of theory-oriented evaluation). A fourth more holistic perspective concerns the advantages of theory in terms of strengthening claims on construct, internal and external validity of evaluation findings. Let us briefly discuss this.

A first role is that of social science theory serving as an important source of hypotheses and tested substantive knowledge on relationships between different variables regarding intervention implementation and outcome. In this way, among other things, the construct validity of (theory-oriented) evaluation studies can be strengthened. Riggin (1990) uses the example of Etzioni's theory of compliance to show how this theory could have improved the specification of an intervention theory on employment assistance. He argues that pattern matching of an 'Etzioni-inspired' intervention theory with empirical data on the intervention would prove to show a better match with intervention practice than an intervention theory that would not have been inspired by this theory. As a result, the construct validity claims of the intervention theory would be stronger with the insights of Etzioni's theory than without them.

Internal validity claims of causal relations between intervention output and outcome variables can also be strengthened for example by using specific behavioral theories to help explain the causal mechanism behind a statistically significant association which has been found between output and outcome. The explanation of causal mechanisms in function of strengthening internal validity is one of the key elements to be illustrated in the next sections of this chapter.

Finally, external validity claims of evaluation findings can be strengthened if the latter show similarities with social science research evidence on (particular dimensions of) similar social problems and interventions. Strong similarities are indications of a high degree of robustness to the findings which are likely to be valid beyond the specific reality of the intervention under evaluation.

Other specific advantages of an intervention theory approach are the following. First of all, and especially relevant in early stages of the intervention cycle, a well-specified intervention theory inspired by relevant substantive theories can help anticipate potential weaknesses in implementation and predict whether it is likely that desirable effects are going to be achieved (Chen and Rossi, 1980). Second, as mentioned earlier, social science theory can constitute an important source of evidence for testing and refining particular assumptions, in combination with empirical data or in case empirical data are too costly or difficult to obtain. Third, as envisioned by Chen and Rossi and others, evaluation can contribute to knowledge development in the social sciences as substantive theories, as part of the intervention theory, are tested and refined in empirical evaluation exercises. Finally, the question of what works (for whom) in what circumstances has guided theory-oriented evaluators in exploring ways to accumulate knowledge *across* interventions. A very interesting perspective is being developed by realist evaluators who use the structure of linking mechanisms to outcomes and contexts (also referred to as the CMO principle) as a basis for theorizing about what works across interventions and contexts (see Pawson and Tilley, 1997; Pawson, 2002; Pawson, 2006). Relevant behavioral theories and topic-specific and context-specific social science research can strengthen external validity claims of CMO constructions. The structure is both useful in single intervention evaluations (i.e. as a framework for organizing and interpreting data and comparison with other interventions) as well as in synthesis studies.

A framework for reconstructing impact theories: Coleman's model of social mechanisms

Our high-velocity excursion through the different aspects of theory-oriented evaluation and the linkages with social science theory has now brought us to an illustration of how within the framework of theory-oriented evaluation social science theory and evaluation can be brought together in a mutually beneficial way. Focusing on impact theory we will illustrate a concrete framework for intervention theory reconstruction. The framework as such is inspired by social science theory, more specifically Coleman's model of social mechanisms. In addition, the framework will offer a useful structure for intervention theory specification and further incorporation of useful strands of social science theory to deepen the understanding of change processes induced by an intervention.

Impact theory is concerned with the set of assumptions on how outputs induce processes of change eventually leading to certain impacts. It can constitute the basis for an empirical retrospective impact evaluation. Alternatively, it can be used in the appraisal or interim evaluation of interventions, using existing evidence to better understand and appraise the potential for intended or unintended consequences.

The following questions are typically posed in impact theory-oriented evaluations. Given the (successful) delivery of outputs (e.g. a training program, a subsidy scheme) under what conditions is it likely that the desired effects envisaged by decision makers are in fact generated? What are the principal (contextual) variables influencing the realization of desired outcomes? What other unintended effects are likely to be generated in certain circumstances? How can we explain the (lack of/partial/full) manifestation of expected outcomes? To what extent can changes in outcome variables plausibly be attributed to an intervention?

One of the classical analytical problems in impact evaluation is to grasp how policies and programs lead to collective outcomes. Interventions, be they policies, programs or projects, all aim at some particular change at an aggregate level such as a reduction in poverty levels, reduction in income inequality, conservation of natural resources, etc. Looking for mere statistical association between a policy measure and a collective outcome (e.g. a change in average income or income equality) is insufficient in order to be able to make judgments about the effectiveness of the intervention. The danger of type I and type II errors [5] (in part caused by underlying design errors) is always present, which makes interpretation without a strong explanatory model dangerous or, in some cases, even meaningless.

The potential complexity of causal processes (embedded in dynamic social, institutional and physical realities) between an intervention and collective outcomes further underlines the importance of constructing such an explanatory model. In our case, the first question that should be posed is how can we meaningfully break down the rather complex causality between intervention delivery and the (un)intended effects on (collective) human behavior? The beginning of an answer can be found in the work of sociologist James Coleman. Coleman (1986) highlights the deficiencies in analyses of what can be called 'macro-macro' relationships, for example the effect of introducing (and 'enforcing') freedom of speech on the evolu-

tion of religious movements. He argues that analyses which fail to introduce an element of purposive human action in their explanations are often rendered meaningless. As an alternative, Coleman presents a model of social mechanisms which rests on the principle of methodological individualism. The main analytical implication of this model is the decomposition of macro-macro relationships into a macro-micro, a micro-micro and a micro-macro explanation. Put simply, each change in the 'macro' environment can be logically related to individual reactions, which in turn are connected to collective outcomes.

The purpose of this model, introduced by Coleman (1986) and further developed by Coleman (1990) and Hedström and Swedberg (1998), is to provide a structure to help explain how changes in social phenomena (e.g. ideologies) can lead to changes in collective social behavior. In this chapter we employ a somewhat broader interpretation of the model. The explanatory model, meaningfully connecting macro-sociological events to collective social outcomes, can easily be expanded to further the understanding of causal relationships between any type of (policy) change in the (international, national, regional, local) environment and collective outcomes. In the remainder of this section we will briefly explain the basics of Coleman's model from this somewhat broader perspective.

The core of the social mechanisms model is constituted by three types of mechanisms, summarized in figure 1 (Hedström and Swedberg, 1998):

A) Situational mechanism ('macro-micro')

The situational mechanism captures the macro-micro transition of how a certain change in the environment affects the opportunity structure, values and beliefs of individual actors. Environment can be interpreted broadly as all the social, political, institutional, economic and physical conditions that surround individual actors. To make the discussion more operational one usually focuses on one type of 'environmental' variable, in this case the introduction of a policy intervention. An example of a situational mechanism is the effect of land tenure security (e.g. a legalization program) on the opportunity structures and attitudes of landowners. In this case an important moderating variable would be whether access to credit is conditional on legal land ownership or not (i.e. as collateral to the loan). Another example is a development organization which supplies food to community members in an emergency situation. The 'sudden' availability of this new resource might affect the legitimacy of existing norms of reciprocity between community members (e.g. Coleman, 1990).

B) Action-formation mechanism ('micro-micro')

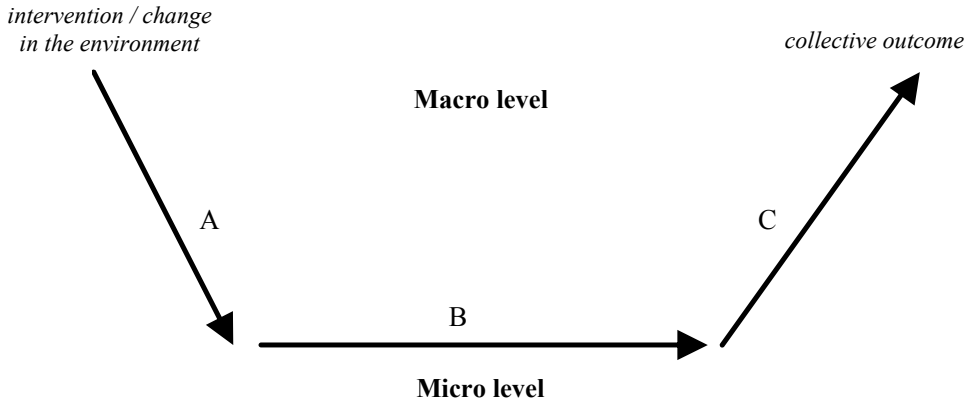
The action-formation mechanism aims to explain how a specific combination of beliefs and opportunities (including a change in the policy environment) triggers individual action. An example is cognitive dissonance theory. The theory has shown that individual actors are inclined to ignore new pieces of information if there are substantial contradictions with existing beliefs and attitudes. Research work on the topic of social capital has substantiated several action-formation mechanisms (e.g.

Coleman, 1990, Putnam, 1993). An example is the existence of shared norms of reciprocity in a community which encourage individual actors to share information and resources [6]. Another example is the existence of a minimum generalized level of trust, which lowers the threshold for individual actors to engage in transactions with other community members.

C) Transformational mechanism ('micro-macro')

The transformational mechanism offers an explanation of how changes in individual behavior translate into collective outcomes. Transformation is not equal to aggregation. Even in small-scale interventions, a change in collective human behavior cannot be understood just in terms of the total of individual changes in human behavior as the former is heavily influenced by social dynamics among individual actors. Micro-macro transformation potentially can be very complex and qualifies among the least understood dimensions in processes of social changes. Examples include: individual dissatisfaction leading to revolution, the exchange of goods between people leading to market prices; individual actors refraining from logging leading to the preservation of forests. As shown by Coleman (1986) an explanation of these micro-macro processes starts with an understanding of how people are linked to each other, e.g. through markets, through (informal) social networks, or through hierarchies. In the social science literature different strands of social theory can be found to guide the evaluator's process of making sense of the micro-macro mechanism. Collective action theory (e.g. Olson, 1965), for example, can help to explain the relation between group characteristics and collective cooperation; (evolutionary) game theory can help to understand how consecutive interactions between individuals can lead to the emergence of social norms (e.g. Axelrod, 1984); general equilibrium economic theory explains how individual preferences can lead to market prices (e.g. Arrow and Debreu, 1954). Obviously, one cannot demand of an evaluator to be well-versed in all the different strands of theory relating to micro-macro mechanisms (the same being valid for the other two mechanisms). This is a question we will return to later on when we discuss application issues of the framework.

Figure 1. Typology of mechanisms



A = situational mechanism
B = action-formation mechanism
C = transformational mechanism

Source: Adapted from Coleman (1986), Hedström and Swedberg (1998)

The three mechanisms constitute the basic structure of a framework for impact theory reconstruction. The framework can be further improved by looking at two additional elements. First, an evaluator should ask the basic question of what exactly it is in an intervention that is expected to change human behavior. Interventions can have multiple 'active ingredients'. For example, an intervention can include training courses (e.g. on administrative and managerial skills) as well as subsidies (e.g. to facilitate the start-up of new businesses). Both 'active ingredients' are assumed to trigger their 'own' mechanisms. In addition, the combination of the two also affects change processes. Vedung (1998) has developed a useful classification of policy instruments [7] which helps the evaluator to rapidly identify different types of 'active ingredients' in an intervention. An additional advantage is that each type of policy instrument can subsequently be linked to the body of literature on the effectiveness of the instrument in different contexts.

Second, the evaluator would benefit from additional guiding principles on how to capture relevant contextual variables at the level of each of the three mechanisms. As suggested by the previous discussion, different strands of social science theory that offer partial explanations of change processes constitute the principal sources for identifying relevant contextual variables. In addition, abstractions of interventions and their contexts are also influenced by the evaluators' experience and, not unimportantly, (preliminary) communication with stakeholders. In addition, one can find literature on a priori classifications of different contextual variables which, depending on the focus and purpose of the evaluation exercise, might turn out to be insightful. For example, an interesting a priori non-exhaustive classi-

fication of (mainly social) contextual variables is offered by realist evaluators. They distinguish between individual capacities of stakeholders, interpersonal relationships 'in and around' the intervention, institutional setting, and the wider infrastructural and welfare system (Pawson and Tilley, 1997).

Evaluators are (usually) not content specialists, at least not for all the potentially relevant dimensions of an impact chain induced by an intervention. Initially, the basic model of impact mechanisms helps evaluators to ask the right questions to decision makers, stakeholders and social scientists (either by interviewing them or by consulting relevant literature). Subsequent feedback from these sources helps refining these questions and raising additional ones. In the process, the evaluator refines the causal mechanisms linking output to impact. Finally, different strands of social science literature as well as other sources of (empirical) evidence are consulted to test and/or validate the model. To briefly illustrate, let us give a few examples of relevant questions that arise in the reconstruction of each of the three mechanisms underlying the impact theory.

Situational mechanisms: How and to what extent does an intervention output reach particular people? What are the important contextual variables influencing these processes? Questions will cover issues like intervention outreach (who is affected, who benefits), communication and interaction with target population (e.g. consultation, participation, joint decision-making), (in economic policy) value chains and economic networks (e.g. marketing channels), etc.

Action-formation mechanisms: How do changes in values, beliefs, opportunities and incentives affect individual behavior? What other relevant variables (e.g. socio-economic characteristics, existing values and beliefs, capacities, individual positions in social structures, geographical location, etc.) determine the nature and outcome of these processes? In what ways?

Transformational mechanisms: How are beneficiaries/clients/affected people linked to each other and to other groups of people likely to be (indirectly) affected by the intervention? Are there many differences within the group of affected people? What are the proportions of different groups of affected people defined on the basis of variables that are deemed relevant in determining action-formation processes? In what ways do affected people differ from the population at large? Under what circumstances are people likely to simulate or reject the behavior of others? Questions cover issues of population composition (e.g. socio-economic characteristics), social structures (social relationships, informal networks, organizations, markets), (shared) norms and values in social interaction, etc.

The depth and breadth of the range of questions to be posed is largely dependent on operational parameters of the evaluation exercise (e.g. purpose, budget, time, topic) and in many cases is rather limited. In the next section we will illustrate how the three mechanisms can be applied. Per mechanism, strands of social science theory can be instrumental in raising and refining questions, providing tentative answers to questions, and providing guidance on (empirical) data to be collected and analyzed in order to answer the questions.

Illustrating the framework

In this section we will illustrate the framework introduced in the previous section for the concrete case of an impact evaluation of the effects of a policy measure. More specifically, we will look at the example of a government introducing an important restriction on maize. Rather than presenting a comprehensive analysis, we will focus on some of the key issues that arise when addressing the evaluative challenges of this case.

Consider the situation of a low- or middle-income country with a substantial agricultural sector, both in types of share in gross domestic product as well as in employing a substantial share of the economically active population. One of the main crops is maize, both produced and consumed domestically by large sections of the population. As a result, policy interventions in maize crop production can have major repercussions on the economic and social conditions of the country. Accordingly, interventions in maize production and marketing are perceived by the government as important tools in achieving social and economic policy objectives.

Suppose the government introduced an import restriction on maize in order to achieve two important objectives: an aggregate increase in the domestic production of maize and a reduction in income inequality. The basic theory harbored by government officials has been the following. First of all, a restriction on maize imports is expected lead to an increase in the market price of maize. Consequently, farmers will be motivated to increase their production levels. Given the fact that the major part of national maize production is produced by poor farmers, they are expected to benefit more in relation to the rest of the population, causing a reduction in income inequality (as well as in poverty levels).

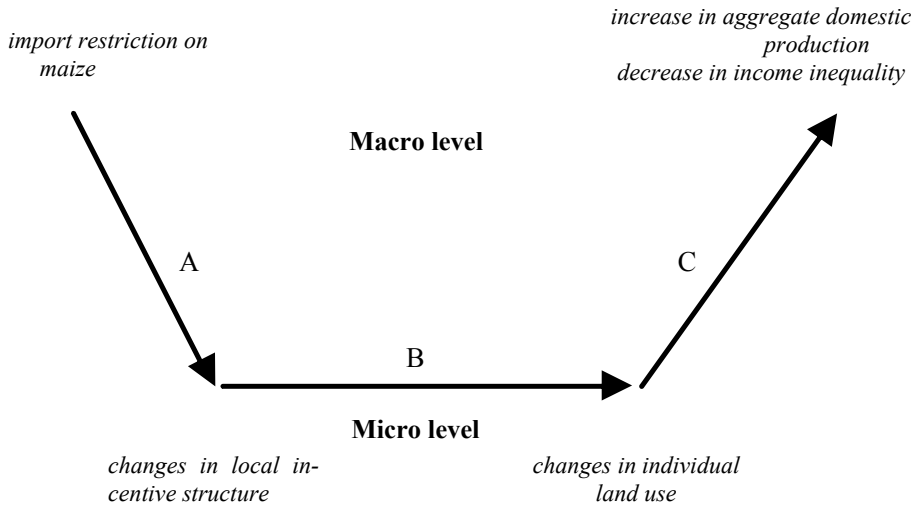
Suppose that after some time tentative macro-economic data show that domestic maize production has increased but income inequality has worsened. As a result, the Minister of Agriculture is worried and has hired a team of evaluators to assess why (at first view) only one of the two objectives has been achieved. More specifically, the team is to establish to what extent the increase in domestic production can be attributed to the policy measure and whether the policy measure, contrary to expectations, might have contributed to an increase in income inequality, or, as hoped by government officials, has counteracted (albeit insufficiently) other developments that have caused income inequality to worsen.

The evaluation team decides to apply a theory-oriented evaluation approach. First, they reconstruct the impact theory connecting the policy measure to possible outcomes (including the ones envisaged by the government). In this particular case, existing research evidence (theory and empirical research) and diagnostic data are identified as the major sources for the reconstruction of the (impact) intervention theory. Subsequently, the team plans to test and validate the theory by means of additional data collection and analysis activities [8].

The basic model of the impact theory for this example is shown in figure 2. In the remainder of this section we will illustrate the further reconstruction of the theory, thereby highlighting some of the more relevant analytical issues (as inspired by strands of social science theory), instead of providing a comprehensive and in-depth

account of the evaluation exercise (which would be somewhat beside the point and take up too much space).

Figure 2. Basic overview of intervention theory on import restriction on maize



Source: Adapted from Coleman (1986), Hedström and Swedberg (1998)

At first sight we are looking at a standard economic policy analysis problem. One of the analytical options suggested by conventional economic policy analysis is the construction of a quantitative supply response model as a basis for analyzing the impact of the policy measure. Such an assessment indeed captures complex linkages between different economic variables and would constitute a valuable input in any kind of impact evaluation on this type of policy intervention. However, as we will show below, a real world assessment goes beyond the economic perspective, taking into account other important dimensions which condition potential processes of change. An evaluator's task is to reconstruct a broader impact intervention theory taking into account different types of potential change processes. In sum, while an evaluator would undoubtedly benefit from consultations with an economist on how the policy measure affects micro-level supply response, the economist in turn could learn from the impact theory reconstruction exercise (based on wider social and institutional realities). Not only might this help to enrich the economic model, in addition, some of the underlying assumptions of the model might be reconsidered or refined.

A) Situational mechanism: from import restriction on maize to changes in local incentives for farmers

The basic mediating variable from policy measure to local incentives is an increase in the market price of maize. Standard economic theory dictates that the reduced supply due to import restrictions will raise the market price of maize. This is indeed an important assumption held by the evaluators. From that point onwards, however, the problem becomes more complex. The basic question is perhaps how and to what extent increased market prices will influence farm gate prices [9]. Several questions can be formulated.

- What are the main characteristics of different markets (e.g. integration, market power, transaction costs) that determine the translation of market price signals to farm gate prices in different regions? What are the characteristics of maize commodity chains? Who are the main actors? How is added value distributed? How well are markets integrated? Are there large regions which are (semi-) autarkic when it comes to maize?
- Beyond the structure and characteristics of regional and local markets, what other economic (e.g. prices of other products) and non-economic (e.g. civil unrest) developments that might impede changes in local incentive structures to occur (e.g. in particular regions) can be identified?
- To what extent do market price signals reach farmers in different regions? Are poor farmers more/less likely to face more beneficial farm gate prices than non-poor farmers?

The above questions already raise some doubts on whether the price signals will reach the end of the chain and consequently will affect poor farmers' opportunity structures. Literature on market integration and characteristics of different maize marketing channels (distribution of value added, power relations, diffusion of price information) can give insights into what variables are relevant to take into account in order to assess the transfer of price signals to local levels in different parts of the country. Several hypotheses can be found regarding why the market price increase might not reach the farm gate. First of all, information on prices in many regions is scarce or monopolized by traders. The latter are in the position to exploit this advantage and capture a lot of the gains in the price increase. The presence of local organizations (governmental or non-governmental) which disseminate price information to some extent might offset this problem. Second, poor farmers often sell small quantities of their produce. Transaction costs for marketing are often very high, so instead of directly selling in regional or district markets they sell at low prices to traders. The presence of cooperative organizations for poor farmers might lead to economies of scale in marketing and a reduction in transaction costs. Third, poor farmers are often risk-averse and prefer to sell their harvest well before the season to traders at lower prices, thereby reducing price risks as well as getting the necessary capital to be able to cultivate their crop. The presence of micro insurance schemes as well as microfinance schemes might offset the need for pre-harvest sales. Given the aforementioned considerations, it might well be possible that non-poor farmers are more likely to experience a substantial positive incentive to produce more maize than poor farmers.

B) Action-formation mechanism: from changes in local incentives to changes in individual land use

An important question here is whether changes in farm gate prices are sufficient for poor farmers to (substantially) raise production levels. Again several questions can be raised.

- Given a higher farm gate price, is it likely that poor farmers will substantially raise maize production levels? What are the most important constraints that restrict production levels (capital, labor, land, risk aversion)?
- Will the price change make it more attractive for poor farmers to adopt modern high-yielding varieties? What variables constrain the adoption of these varieties? What other individual characteristics determine the predisposition to adopt new varieties?
- Are non-poor farmers more/less likely to raise production levels of current varieties of maize? Are non-poor farmers more/less likely to adopt high-yielding varieties?

At least two strands of social science theory are well worth consulting for the team of evaluators: the literature on farm household economics (e.g. Ellis, 1988) and the literature on adoption and diffusion of innovations (Rogers, 2003). Farm household economics literature shows that poor farmers are more risk-averse than well-off farmers and therefore less likely to significantly increase the area for maize cultivation. They might do so gradually if capital, labor and land availability constraints allow for such a gradual increase. With a higher market price for maize, many non-poor farmers (whether maize producers or not) might find it sufficiently profitable to reserve more land for maize production. They are less constrained by resource issues and are likely to have better access to input markets. In the case of switching from traditional maize varieties to high-yielding varieties we are facing a diffusion of innovations problem. A whole range of system characteristics come into play, which determine the extent and depth of moving from traditional to new varieties. For poor farmers the lack of access to input markets, the fact that high-yielding varieties are potentially more vulnerable to diseases and weather, and the fact that traditional (lower-yielding) varieties may taste better (poor farmers consume a substantial part of their produce) are all examples of factors that will limit the adoption of high-yielding varieties. Non-poor farmers may be more likely to be prepared to allocate more land to high-yielding varieties. They have definite input market access advantages and are more active (and stronger) market players.

C) Transformational mechanism: from changes in individual land use to increased domestic maize production and reduced income inequality

Several questions come to mind with respect to the last mechanism. The following is a non-exhaustive list.

- What are the main variables influencing the social dynamics of adoption of high-yielding maize varieties? On average at country level, do input markets have sufficient outreach among the poor? Is farmer to farmer extension a wide-

spread practice within the country? Does information on maize production and varieties circulate widely among different regions and groups of farmers?

- What is the proportionate size of different groups of farmers in different regions of the country? Which types of farmers (in which regions) are likely to be responsible for the increase in aggregate domestic production?
- How many poor farmers and other poor are net consumers of maize? Are the losses in welfare of poor net consumers offset by the (possible) gains in welfare of poor net producers of maize? Have non-poor farmers benefited more from the price increase (in terms of per capita and aggregate welfare) than poor farmers?
- What is the evolution of growth per capita in other crop and livestock production sub sectors as well other non-agricultural sectors of the economy? Which groups in society (poor versus non-poor) have benefited most from positive growth in different sectors? What other factors have influenced income inequality? What is the relative contribution of the 'maize price effect' versus other factors on income inequality?

The questions posed above are organized according to different dimensions of the micro-macro problem: the social dynamics of diffusion of high-yielding varieties; the explanation of increased domestic maize production, taking into account the analysis of different variables influencing action-formation mechanisms and the proportionate sizes of farmer groups defined on the basis of these variables [10]; the income inequality effect of the policy measure assessed at sub sector level; the income inequality level assessed at country level taking into account developments in other sectors and effects on the poor and non-poor.

Let us briefly discuss two of the issues. First of all, the income inequality effect of the policy measure is quite complex. Economists use potentially elaborate models to understand these processes. The evaluation team should definitely talk with country and sector economists about the availability of these models as inputs to the intervention theory. The argumentation provided below is a simple example of how such inputs are translated into a basic theory which can be connected to the other strings of intervention theory (and can also easily be portrayed graphically). From the literature on farm household economics (e.g. Ellis, 1988) and agricultural price policy analysis (e.g. Mellor, 1975) we can construct useful hypotheses which constitute the basis for further analysis. First of all, many poor farmers in different developing countries are net consumers of certain staple foods such as maize. An increase in the market price of maize might have a net negative effect on their economic welfare and food security situation and might push them deeper into poverty. In addition, all other things being equal, a higher market price of maize has a negative effect on the welfare situation of other groups of poor people (e.g. the rural landless, the urban poor, who are also net consumers of maize), enhancing poverty. All of this needs to be offset by the welfare gain in the group of poor farmers who are net producers of maize in order for the policy measure to have a (potentially) positive effect on income inequality. However, this is not the end of it. Subsequently, the net welfare gains among the poor need to be higher than welfare gains among non-poor farmers. Finally, other factors that affect income inequality, be-

yond the sub sector of maize production, need to be taken into account to assess the 'maize price effect' described here vis-à-vis other factors.

Second, the social dynamics of diffusion of high-yielding varieties constitutes a part of the impact theory which feeds into the 'income inequality' assumptions as well as the 'production increase' assumptions. As stated earlier, diffusion and innovation theory, as treated in the comprehensive seminal work of Rogers (2003), constitutes an important input to deepen the understanding of the individual action-formation mechanism of adoption of new varieties. In addition, the theory is very insightful on the social dynamics of innovation and diffusion. Rogers discusses several crucial dimensions: the nature of communication channels, the nature of social networks, the role of opinion leaders (people influencing the choices of others), the role of change agents (e.g. organizations actively promoting new maize varieties), etc. The potential influence of these dimensions needs to be assessed by the evaluation team, preferably at the lowest level of analysis (e.g. regional, district) that is feasible given the operational constraints of the evaluation exercise. Subsequently, an assessment can be made of how all the factors have influenced adoption behavior in different regions, which, in combination with an assessment of individual determinants of adoption behavior and the proportionate sizes of groups of farmers in different regions can lead to an overall assessment of adoption of high-yielding varieties, by whom, and, as a result, its contribution to the overall increase in domestic production. The combination of individual determinants and social determinants of diffusion processes constitutes the basis for reconstructing the partial impact theory on adoption of new varieties which in turn will be the basis for the kind of assessment sketched above.

Three important insights can be deduced from the above illustration. First, we have shown how strands of social science theory provide useful hypotheses for understanding causal processes underlying the impact of the policy measure. For example, in this case it is not altogether unlikely that the increase in domestic production of maize is largely due to increases in cultivated area of high-yielding varieties produced by non-poor farmers. In addition, (leaving aside other developments) the increase in income inequality can in part be explained by the fact that many of the welfare gains of the price increase are occurring among non-poor farmers. Moreover, if many of the poor are net consumers of maize, they may have experienced a decrease in their real income. Accordingly, the effect on poverty levels may have been negative, too.

Second, the intervention theory helps evaluators to explain why and where a causal chain breaks down. In other words, weaknesses in the theory can be more easily identified and explained using an intervention theory framework based on Coleman's theory of social action. In this case, at the end of the exercise, the evaluators, guided by the framework, will be able to state whether the failure to achieve a reduction in income inequality is largely due to (for example):

- increased market prices not leading to increased farm gate prices for poor farmers (situational mechanism);
- increased farm gate prices not being sufficient (or sufficiently high) to substantially influence individual decisions on land use (action-formation mechanism);

- increases in maize production and sales by poor farmers not sufficiently offsetting negative effects on the real income of poor net consumers of maize, positive effects among non-poor farmers, or other factors affecting income inequality (transformational mechanism).

Finally, the third message is an implication of the first two. We have seen that the deconstruction of a policy intervention problem into different mechanisms and associated questions generates a useful structure for making sense of the causal processes between intervention and potential effects. In addition, spelling out the different assumptions and questions helps to select the aspects that should be analyzed in more detail and the corresponding appropriate quantitative and qualitative methods to do so.

Conclusion

Theory-oriented evaluation comprises different currents of thinking and practice, all of which revolve around the broad idea of uncovering, understanding and testing underlying assumptions behind interventions. Social science substantive theory plays a key role in making sense of interventions, either as a source of assumptions, as an input for refining assumptions (e.g. as sources of relevant contextual variables), or as a benchmark for testing assumptions. Sometimes these roles are clearly delineated within the processes of reconstructing and testing intervention theories, sometimes they overlap, and in other cases no clear distinction at all can be discerned between these roles, as reconstruction and assessment of the intervention theory are completely locked together in one and the same (iterative) process of theory refinement.

Coleman's typology for social mechanisms facilitates the meaningful decomposition of change process induced by an intervention into three different analytical phases. Consequently, the basis for a social science theory-informed explanation of change processes rooted in methodological individualism can be developed. Another important advantage of this approach is the fact that the three types of mechanisms are explicitly linked to distinct types of substantive theories (i.e. macro-micro, micro-micro, micro-macro). This is definitely a step forward as it potentially brings evaluators and social scientists closer together. Yet, for meaningful two-directional traffic to occur in terms of better social science theory-informed evaluations (i.e. beneficial to the evaluation community) and more social science theory-testing in evaluations (i.e. beneficial to the social science community) the gap still looms large.

Notes

1 For purposes of readability, we will use the term social sciences or social science theory throughout the text when referring to the social and behavioral sciences.

2 For example, the advantages of 'making' (quasi-)experimental evaluation more theory-oriented have been widely discussed (e.g. Lipsey, 1993; Cook, 2000; Bamberger et al. 2004).

3 Sometimes called ‘pet theories’ (Bogue, 1974), or ‘theories-in-use’ (Argyris and Schön, 1978).
4 It has to be added that already in 1990 the US General Accounting Office in a prospective evaluation of the (possible) impact of two competing proposals of teenage pregnancy laws, a distinction was made between the reconstruction of the operational logic of the laws and the conceptual logic (US GAO, 1990).
5 Respectively, the detection of significant association while none exists (i.e. rejecting the null hypothesis while in reality it holds) and failing to detect a significant association while in reality it exists (i.e. failing to reject the null hypothesis).
6 As stated in the previous paragraph, such norms can be weakened (or reinforced) by external actors.
7 Vedung (1998) describes three major categories of policy instruments: sticks (regulation), carrots (economic means), and sermons (information).
8 In theory-oriented evaluation, one can find discussions on the issue of which links in a theory to evaluate. See Weiss (2000) for a discussion on the different criteria that can be applied in setting priorities for theory testing.
9 The price received by farmers, as opposed to for example the price that consumers pay for their maize.
10 To keep things simple and directly related to the objectives of the policy measure (reduction in income inequality, poverty reduction) we have highlighted the dimension poor/non-poor as the key moderating variables in all three ‘stages’ of explanation. In reality other characteristics are likely to play an important role, such as geographical region (and agro-physical conditions), ethnic and cultural background, etc.

References

Argyris, C. and Schön, D. (1978) Organizational Learning: A Theory of Action Perspective, Reading, Addison-Wesley.

Arrow, K.J. and Debreu, G. (1954) “The Existence of an Equilibrium for a Competitive Economy”, Econometrica, XXII: 265-290.

Axelrod, R. (1984) The Evolution of Cooperation, New York, Basic Books.

Bamberger, M., Rugh, J., Church, M. and Fort, L. (2004) “Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints”, American Journal of Evaluation, 25(1): 5-37.

Bickman, L. (ed.) (1987) Using Program Theory in Evaluation, New Directions for Program Evaluation, 33, San Francisco, Jossey-Bass.

Birckmayer, J.D. and Weiss, C.H. (2000) “Theory-Based Evaluation in Practice”, Evaluation Review 24(4): 407-431.

Bogue, D. (1974) “Policy implications of theory and research on motivation and induced behavior change for fertility and family planning”, Paper presented at the AID Asia Population officers Conference, Honolulu, 1974.

Chen, H.T. (1990) Theory-Driven Evaluation, Beverly Hills, Sage Publications.

Chen, H.T. (2005) Practical program evaluation: Assessing and improving planning, implementation, and effectiveness, Thousand Oaks, Sage Publications.

Chen, H.T. and Rossi, P.H. (1980) “The Multi-Goal, Theory-Driven Approach to Evaluation: A Model Linking Basic and Applied Social Science”, Social Forces 59: 106-122.

Coffman, J. (1999) Learning from logic models: an example of a family/school partnership program, Harvard Family Research Project, Cambridge.

Coleman, J.S. (1986) "Theory, Social Research and a Theory of Action", American Journal of Sociology 91(6): 1309-1335.

Coleman, J.S. (1990) Foundations of Social Theory, Cambridge, Belknap Press.

Cook, T.D. (2000) “The false choice between theory-based evaluation and experimentation”, in: P.J. Rogers, T.A. Hacsi, A. Petrosino and T.A. Huebner (eds.) Program theory in evaluation: challenges and opportunities, New Directions for Evaluation, 87, San Francisco, Jossey-Bass.

Donaldson, S.I. (2003) "Theory-driven program evaluation in the new millennium", in: S. I. Donaldson and M. Scriven (Eds.) Evaluating social programs and problems: Visions for the new millennium, Mahwah, Lawrence Erlbaum.
Donaldson, S.I., and Gooler, L.E. (2003) "Theory-driven evaluation in action: Lessons from a $20 million statewide work and health initiative", Evaluation and Program Planning 26: 355-366.
Ellis, F. (1988) Peasant Economics: Farm Households and Agrarian Development, Cambridge, Cambridge University Press.
Garfinkel, H. (1967) Studies in Ethnomethodology, Prentice Hall, Englewood Cliffs.
Hedström, P. and Swedberg, R. (1998) Social Mechanisms: An Analytical Approach to Social Theory, Cambridge, Cambridge University Press.
Hoogerwerf, A. (1990) "Reconstructing policy theory", Evaluation and program planning 13(3): 285-291.
Leeuw, F.L. (1991) "*Policy Theories, Knowledge Utilization*, and Evaluation", Knowledge and Policy 4(3): 73-92.
Leeuw, F.L. (2003) "Reconstructing Program Theories: Methods Available and Problems to be Solved", American Journal of Evaluation 24(1): 5-20.
Lipsey, M.W. (1993) "Theory as Method: Small Theories of Treatments," in: L.B. Sechrest and A.G. Scott (eds.), Understanding Causes and Generalizing about Them, New Directions for Program Evaluation 57, San Francisco. Jossey-Bass.
Mannheim, K. (1951) Man and Society In an Age of Reconstruction, London, Routledge.
Mellor, J.W. (1975) "Agricultural Price Policy and Income Distribution in Low Income Nations", World Bank Staff Working Paper 214, Washington, D.C.
Nuttal, C. (2003) "The Home Office and random allocation experiments", Evaluation Review 27(3): 267–290.
Olson, M. (1965) The Logic of Collective Action, Cambridge, Harvard University Press.
Pawson, R. (2002) "Evidence-based policy: The promise of 'realist synthesis'", Evaluation 8(3): 340-358.
Pawson, R. (2003). "Nothing as Practical as a Good Theory." Evaluation 9(4): 471-490.
Pawson, R. (2006) Evidence-based policy: A realist perspective, London, Sage.
Pawson, R. and Tilley, N. (1997) Realistic evaluation, London, Sage.
Putnam, R. (1993) Making Democracy Work: Civic Traditions in Modern Italy, Princeton, Princeton University Press.
Riggin, L.J.C. (1990) "Linking program theory and social science theory", in: L. Bickman (ed.) Using Program Theory in Evaluation, New Directions for Program Evaluation 33, San Francisco, Jossey-Bass.
Rogers, P.J. (2000) "Causal Models in Program Theory Evaluation", in: P.J. Rogers, T.A. Hacsi, A. Petrosino and T.A. Huebner (eds.) Program Theory in Evaluation: Challenges and Opportunities, New Directions for Evaluation 87, San Francisco, Jossey-Bass.
Rogers P.J., Hacsi, T.A., Petrosino, A. and Huebner, T.A. (eds.) (2000) Program Theory in Evaluation: Challenges and Opportunities, New Directions for Evaluation 87, San Francisco, Jossey-Bass.
Rogers, E.M. (1962, 2003) Diffusion of Innovations, New York, Free Press.
Rosas, S.C. (2005) "Concept Mapping as a Technique for Program Theory Development", American Journal of Evaluation 26(3): 389-401.
Rossi, P.H., Lipsey, M.W. and Freeman, H.E. (2004) Evaluation: A systematic approach, Thousand Oaks, Sage Publications.
Smith, M.F. (1989) Evaluability Assessment: A Practical Approach, Boston, Kluwer Academic Publishers.
Suchman, E.A. (1967) Evaluative research: Principles and practice in public service and social action programs, New York, Russell Sage Foundation.
Trochim, W.M.K. (1989) "An introduction to concept mapping for planning and evaluation", Evaluation and Program Planning 12: 1-16.
US GAO (1990) Prospective Evaluation Methods: The prospective evaluation synthesis, US General Accounting Office, Washington, D.C.
Van Thiel, S. and Leeuw, F.L. (2002) "The performance paradox in the public sector", Public Productivity and Management Review 25(3): 267-281.
Vedung, E. (1997) Public Policy and Program Evaluation, New Brunswick, Transaction Publishers.

Vedung, E. (1998) “Policy instruments: Typologies and theories”, in: M.L. Bemelmans-Videc and R.C. Rist (eds.), Carrots, sticks and sermons: Policy instruments and their evaluation, New Brunswick, Transaction Publishers.
Weiss, C.H. (1972) Evaluation Research: Methods of Assessing Program Effectiveness, Englewood Cliffs, Prentice Hall.
Weiss, C.H. (1995) “Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families”, in: J.P. Connell, A.C. Kubisch, L.B. Schorr and C.H. Weiss (eds.), New approaches to evaluating community initiatives, 1, Washington, DC, The Aspen Institute.
Weiss, C.H. (1997) “Theory-Based Evaluation: Past, Present and Future”, in: D.J. Rog and D. Fournier (eds.) Progress and Future Directions in Evaluation: Perspectives on Theory, Practice and Methods, New Directions for Evaluation 76, San Francisco, Jossey-Bass.
Weiss, C.H. (2000) “Which Links in Which Theories Shall We Evaluate?” in: P.J. Rogers, T.A. Hacsi, A. Petrosino and T.A. Huebner (eds.) Program Theory in Evaluation: Challenges and Opportunities, New Directions for Evaluation 87, San Francisco, Jossey-Bass.
Wholey, J.S. (1987) “Evaluability Assessment: Developing Program Theory”, in: L. Bickman (ed.) Using Program Theory in Evaluation, New Directions for Program Evaluation 33, San Francisco, Jossey-Bass.

CHAPTER 4

Vaessen, J. and G. Van Hecken (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, GEF Evaluation Office, Washington D.C.

List of acronyms

CATIE	Centro Agronómico Tropical de Investigación y Enseñanza
CG	control group
CIPAV	Centro para la Investigación en Sistemas Sostenibles de Producción Agropecuaria
ESI	environmental services index
FAO-LEAD	Food and Agriculture Organization – Livestock, Environment and Development
FDL	Fondo de Desarrollo Local
FONAFIFO	Fondo Nacional de Financiamiento Forestal
GEF	Global Environment Facility
GIS	geographic information system
NGO	non-governmental organization
PAD	project appraisal document
PES	payments for environmental services
RISEMP	Regional Integrated Silvopastoral Approaches to Ecosystem Management Project
TA	technical assistance
WB	World Bank

1. Introduction

1.1. The Regional Integrated Silvopastoral Approaches to Ecosystem Management Project

This report presents the findings of an evaluation commissioned by the GEF Evaluation Office within the framework of its Annual Report on Impact (see for example GEF, 2007a). The Regional Integrated Silvopastoral Approaches to Ecosystem Management Project (RISEMP) was selected as a case study because it is one of the few recently completed conservation projects based on an experimental impact design, allowing (in theory) for an assessment of the net effects of an intervention.[1] This evaluation analyzes the strengths and weaknesses of the project's underlying experimental design.

The Regional Integrated Silvopastoral Approaches to Ecosystem Management Project (RISEMP) was initiated in 2002. It was a full-sized 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. It was managed by the World Bank and coordinated by CATIE, an international research institute in Costa Rica. Country pilot sites were managed by national non-governmental organizations (Nitlapán, CATIE, and CIPAV). The intended total cost of the project was US$8.72 million; of which US$4.77 million was financed by a GEF grant and US$3.95 million through co-financing (from FAO-LEAD, Nitlapán, CATIE and CIPAV and other local donors). The project closed in January, 2008.

The main development objectives of RISEMP were to demonstrate and measure; a) the effects of the introduction of payment incentives for environmental services (PES) to farmers, based on their adoption of integrated silvopastoral farming systems in degraded pasture lands; and b) the resulting improvements in ecosystems functioning, global environmental benefits, and local socio-economic gains resulting from the provision of these services (see also the summary logical framework in Annex 1). There were four project components.[2] The first component aimed at strengthening local development organizations (especially the managing NGOs: CATIE, CIPAV and Nitlapán) to assist farmers in establishing and maintaining improved silvopastoral systems, and in the technical and institutional aspects of silvopastoral systems. The second component concerned developing and implementing an improved monitoring system to provide accurate information and understanding on the potential of intensified silvopastoral systems in providing global environmental services and local socio-economic benefits. The third component was about creating and implementing a payment mechanism to provide incentives

1 With respect to PES it might be the only completed PES project based on an experimental design (Wunder et al., 2008).

2 A fifth component is project management activities, see also Annex 1.

for establishing and maintaining improved silvopastoral systems on farms. The fourth component aimed to support policy formulation and dissemination, specifically developing a replication strategy, including exploration of potential sustainable financing mechanisms, to ensure the long-term sustainability of the project.

Important differences with other prominent PES programs, such as the Costa Rican national PES program,[3] should be noted. First of all, RISEMP focused on landscape restoration in agricultural landscapes whereas most other programs focus on land use conservation (e.g. forest conservation). This has implications in terms of costs of implementation (e.g. land use monitoring, market development) as well as the sustainability of the generated environmental services (threats to degradation). Second, relatively little attention was devoted to financial sustainability of payment mechanisms. In contrast, the project focused on testing the effectiveness of payment mechanisms on land use changes in agricultural landscapes and analyzing the relationships between different land uses and the generation of environmental services. Both issues are to a large extent unexplored territories of inquiry and thus illustrate the innovative nature of the project.

Box 1. Key concepts defined

Outcomes are defined as short-term, immediate effects attributable (in part) to intervention outputs.

Impacts refer to the "[p]ositive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (OECD-DAC, 2002: 24).

We use the term effects in the more generic sense, to refer to the direct and indirect changes that are (in part) the result of an intervention. Effects comprise both outcomes and impacts.

The RISEMP project was in essence a research and innovation project. Apart from providing incentives to farmers to adopt silvopastoral practices in function of generating multiple environmental services, the project was designed to investigate:

- on the one hand, the effects of different types of incentives on land use changes and the sustainability of these changes;
- and on the other hand the effects of land use changes in terms of (global and local) environmental services and (local) socio-economic benefits

Thus, to some extent the project in itself was about outcome and impact assessment. As part of the project's objectives, the project teams (in the three coun-

3 Also supported by the GEF.

tries) in collaboration with World Bank staff developed their own system of research and monitoring, the results of which have been published in project documents and books and journals (see for example Pagiola et al. 2004; Ibrahim et al., 2007; Pagiola et al., 2007).

The project was based on the experimental mechanism of targeting groups of farmers with different incentives. In principle, this would offer a solution to the attribution problem in impact assessment, as differences between otherwise similar groups could then be attributed to the differences in incentives received from the project.

1.2. Logic and comparative strengths of (quasi-)experimental designs for evaluating intervention effects

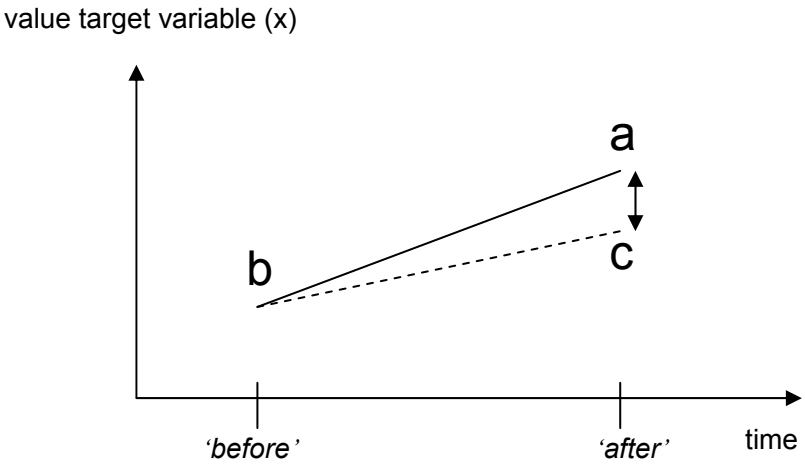
A fundamental problem in outcome and impact evaluation is attribution. Can changes in certain variables be attributed to an intervention or are they the result of other factors? The project's underlying experimental design targeted exactly this question.

The most widely known and advocated types of methodological approaches that address the attribution problem are experimental and quasi-experimental approaches. The idea of (quasi-) experimental counterfactual analysis is that the situation of a participant group (receiving benefits from/affected by an intervention) is compared over time with the situation of an equivalent control group that is not affected by the intervention. Several designs of combinations of ex ante and ex post measurements of participant and control group have been used in this type of analysis (see for example Cook and Campbell, 1979; Shadish et al., 2002; for development interventions see for example Baker, 2000; Bamberger, 2006; White, 2006; Bamberger and White, 2007). Randomization of intervention participation is considered to be the best way to create equivalent groups. In case of random assignment to either the participant and control group, in sufficiently large samples the probability that both groups are equivalent on all observable and non-observable characteristics except for intervention participation is very high.[4]

The attribution issue can be briefly illustrated as follows. Consider a target variable x. In Figure 1 line b-a is the evolution of variable x in the participant group while line b-c represents the evolution within the control group. Randomization of membership of either of the two groups assures for the fact that the differential evolution between the two lines can be attributed to the intervention (since other factors can be considered equal). Consequently, the net effect of the intervention is the difference between a and c.

4 As a second-best alternative, several matching techniques (e.g. propensity score matching) can be used to create control groups that are as similar to participant groups as possible (see below). Finally, regression-based approaches following the same logic of counterfactual analysis can be used in case design-based data are not available or designs are not feasible in practice, see for example GEF (2007b).

Figure 1. Graphical display of the net effect of an intervention.



Source: own elaboration

In sum, the approach offers the following advantages:

- It provides a robust estimate of the net effect of an intervention 'controlling' for other factors
- It provides an indication of the magnitude of an effect

Whether or not these advantages are realized in practice depends on the extent to which major threats to validity are effectively addressed by the project (see section 2.3.). We discuss this for the case of the Nicaraguan pilot site below.

1.3. The RISEMP project's experimental design for outcome and impact assessment

The project was designed as a pilot project to test the effects of incentives (payments for environmental services and technical assistance) on land use changes and ultimately on environmental and socio-economic benefits. The experimental design was to be the basis for being able to attribute changes to different types of incentives. More specifically the following hypotheses were to be tested on the basis of the design (operational manual RISEMP):

- Adoption of silvopastoral practices can be attributed to payments for environmental services (PES)
- Adoption of silvopastoral practices can be attributed to technical assistance (TA)
- Adoption of silvopastoral practices can be attributed to both payments and technical assistance

In addition, a fourth hypothesis was that different payment schemes (2 years and 4 years) would affect the speed and intensity of adoption behavior. Farmers receiving only two years of payments were expected to invest more heavily in their farms in order to benefit as much as possible from the PES payments (and get as close as possible to the maximum payment of 6,000 US$ per farm). Moreover, the comparison between the two groups would have to shed light on the question of sustainability of land use changes. It was hypothesized that the 2 years group would initially invest more and subsequently (after payments had ceased) less in land use changes (or even reverse some of the changes).

In order to be able to test these hypotheses, it was envisaged that the following 'treatment' groups should be established: a group of farmers receiving only PES ('PES only'), a group receiving PES and technical assistance ('PES + TA') and a control group. In addition, the two PES groups were again subdivided into a group receiving PES for 2 years ('PES 2 years') and another group receiving PES for 4 years ('PES 4 years'). The selection of farmers for the different 'treatment' groups ('PES only', 'PES + TA', 'PES 4 years', 'PES 2 years', control group) was to be done at random to assure equivalent groups. Consequently, the effects of different types of incentives can be directly deduced from a simple comparison of means (of changes over time)[5] between 'treatment' groups.

The experimental design can be used to compare groups in terms of the changes in land use over time. In addition, land use changes can be directly linked to environmental impact on the basis of an environmental services index (ESI). Consequently, the experimental design can be used in a fairly straightforward manner for estimating environmental impact (i.e. by multiplying the environmental value of a particular land use with the area of application and comparing this outcome between groups). The ESI was based on past research, and as part of the project's research activities was successfully validated and adjusted to better reflect the relationships between different types of land uses and environmental benefits. Annex 2 presents the ESI of the different land uses for biodiversity and carbon sequestration.

2. Objectives and methodology of the evaluation

2.1. Delimitation and objectives

On the basis of the project's strategy and logical framework one can discern three principal dimensions of project effects:

- Effects at field level: this refers to the processes of change induced by the project at pilot site level (among and beyond the participant farmers), from project outputs to outcomes and impacts;
- Institutional effects: in a narrow sense referring to the learning processes induced by the project at the level of the three implementing organizations (CATIE, Nitlapán, CIPAV) and the World Bank, reinforcing capacities and knowl-

5 Also called double difference.

edge to further innovate as well as implement similar interventions; this is closely related to the third dimension;

- Replicatory effects: which refers to the indirect processes of change in terms of diffusion and uptake of lessons learned; the nature and extent to which these lessons are taken up in research, policy design and implementation by research communities, governments and other institutional actors.[6]

As an innovative pilot project, all three dimensions are of importance and there is a link between the three. This report exclusively focuses on the first dimension, which in a sense provides the ingredients for the other two dimensions (the substantive content for institutional learning and replication). It focuses on the fundamental question of how to evaluate the effectiveness of payments for environmental services (and other incentives) in generating and sustaining such outcomes as land use changes and such impacts as environmental and socio-economic benefits.

The current evaluation focuses on the case of the Nicaraguan pilot site, one of the three pilot sites of the project. The Nicaraguan case was selected for its learning potential, as the implementation of the design experienced more problems than in the other two countries.

The objectives of the evaluation are the following:

- To assess the strengths and weaknesses of the experimental design underlying the RISEMP project; its design and implementation;
- To assess the potential of the experimental design as a basis for analyzing the effectiveness of project incentives on land use changes;
- To suggest alternative and/or complementary methods for outcome and impact assessment;
- To draw lessons on the viability and utility of (quasi-) experimental designs as a future evaluation component in similar projects as the RISEMP project, in which the GEF could play a role.

2.2. Methodology

2.2.1. Data collection

The evaluation relied on a variety of sources of data and methods of data collection including stakeholder interviews, interviews with farmers, document review and secondary data. Stakeholder interviews were conducted with staff from Nitlapán, CATIE and the World Bank (see Annex 3 for a list of interviewees). In addition, in Nicaragua a total of 29 farmers were interviewed (see next paragraph and Annex 3). Document review included documents produced by stakeholders (within the framework of the project) but also 'external' literature on impact evaluation. Project data on land use changes and adoption of silvopastoral practices were used in the inter-group comparisons that constitute the basis of experimental evaluation of intervention effects.

6 Evidently, replicatory effects can also occur at the farmer level. In this case, we classify these under field level effects.

Some methodological observations on the interviews with farmers are in order. In the Nicaraguan pilot site of Matiguas-Rio Blanco 29 farmers were selected using a maximum variability sampling procedure. The interviews were conducted in name of the University of Antwerp in order to avoid being associated with the project and therefore eliciting socially desired responses. The latter phenomenon is quite common among farmers given the expectations they have vis-à-vis the large number of organizations offering support in the region, as well as more specifically, the long history of cooperation between several farmers and Nitlapán. Basically, the first part of every interview was devoted to getting to know the life history and livelihood strategies of every farmer (on average 1 hour), after which the interview was gradually directed to the topic of projects and institutions with which the respondent had been collaborating, eventually talking about the RISEMP project (on average 45 minutes).

Interviews with stakeholders and farmers were semi-structured and covered a list of topics described in the next section on the conceptual framework that was used for evaluating the project's experimental design. Triangulation between opinions and findings from different interviews, documents and data analysis was used to validate findings.

2.2.2. Methodological framework

The basic idea of the experimental design is that one compares the intervention situation with the counterfactual, the situation that would have occurred without the intervention, in order to determine whether and to what extent changes in variables of interest can be determined to the intervention. More specifically, one compares a participant group (affected by/receiving benefits from the intervention) with a control group, a group that exactly resembles the participant group in all aspects but for participation in the intervention.

The validity of the tests of the main hypotheses underlying the experimental design (see section 1.3.) depends on the extent to which the group comparisons actually represent unbiased estimates of the net effects of particular incentives. In other words, in order to be able to analyze in a credible and valid way (and subsequently accept or reject) these hypotheses, the following three inter-group comparisons would need to be bias-free:[7]

- comparing the average change over time of the 'PES only' group with the CG;
- comparing the average change over time of the 'PES only' group with the 'PES + TA' group;
- comparing the average change over time of the PES 2 years group with the PES 4 years group.

Several aspects of design and implementation of an experimental design in development interventions can potentially threaten its validity. In analyzing the strengths and weaknesses of the experimental design the following aspects were

7 Not affected by any of the problems described

taken into account, which are deemed most pertinent as threats to the validity and utility of an experimental design in projects such as the RISEMP project:[8]

- Selection bias: refers to the problem of under- or overestimating project results due to uncontrolled differences between different (treatment) groups of farmers that would lead to differences in result variables if none of the groups would have received project benefits (Rossi et al., 2004; Shadish et al., 2002). One can differentiate between selection bias on the basis of observable variables (e.g. farm size, education level) and unobservable variables (e.g. motivation, risk aversion).
- Contagion (or treatment diffusion): refers to the problem of groups of farmers that are not supposed to be exposed to (or receiving) certain project benefits are in fact benefiting from a project in one or more ways: by directly receiving the benefits from the project, by indirectly receiving benefits through other participating farmers (e.g. knowledge transfer), or by receiving similar benefits from other organizations (see Shadish et al., 2002).
- Behavioral responses: several unintended behavioral responses not caused by project incentives or 'normal' conditions might disrupt the validity of comparisons between groups and hence the ability to attribute changes to project incentives. The most important are the following (see Shadish et al., 2002):
    - Expected behavior or compliance behavior: participants react in accordance with project staff expectations for reasons of compliance with the established contract, due to the (longstanding) relationship with staff, or due to certain expectations about future benefits from the organization (not necessarily the project).
    - Compensatory equalization: discontent among staff or recipients with the inequality between groups might result in compensation of groups that receive less than others.
    - Compensatory rivalry: differentiation of incentives between groups of farmers might result in social competition between those receiving (many) project benefits and those that receive less or no benefits.
- Other aspects that might weaken attribution analysis:
    - Characteristics of the intervention.
    - Quality of the data collected.
    - Timing of the data collection activities.
    - Characteristics of the design.

8 For detailed discussions see for example Shadish et al. (2002) or Morgan and Winship (2007).

3. Assessing the design and implementation of the experimental framework in the Nicaraguan pilot site

3.1. Introduction: project design and preparation[9]

In Nicaragua, in 2002, in order to start project activities farmers were selected in a systematic manner and assigned to groups receiving different types of incentives. After the selection of communities (seven communities in two watersheds in the Matiguas-Rio Blanco region) in which the project would intervene, a census was held among all farmers. On the basis of the census the project staff invited farmers to meetings to explain the rules of the game and promote the project. The project objectives were explained to farmers and farmers were selected for participation on the basis of the following criteria (operational manual RISEMP):

- small and medium farmers;
- secure land tenure;
- livestock as principal income activity;
- willingness to sign a contract with the project;[10]
- willingness to collaborate with project monitoring activities regarding the following information: socio-economic, carbon, water, biodiversity data;
- willingness to participate in training and receive technical assistance;
- willingness to develop a farm development plan in order to generate environmental services and improve productivity;
- willingness to continue to manage silvopastoral systems after project closure.

In addition, in practice the following criteria for selection were applied:

- proximity to the road;
- farmer should live in the farm;
- farmer should have between 8 and 100 hectares of land.

At the time of project initiation meetings with farmers, the message was that all farmers participating in the project would receive payments for environmental services generated by their changes in land use. Some farmers left as they did not believe that benefits would come forth or they lost interest in the project.[11] Consequently, apart from the formal selection criteria a kind of natural selection process took place in which the most motivated farmers, i.e. those that continued to attend the meetings, would be the first to qualify for project benefits.

The interested and selected participants were then assigned to two groups, those that would receive payments and technical assistance (PES + TA) and those that would receive payments only (PES). For the two groups preliminary quota were established per community. Subsequently, independent of the previous subdivision,

9 Findings in this section are primarily based on interviews with staff from Nitlapán, and to a lesser extent interviews with staff from CATIE and farmers.

10 Among other things the contract stipulated some land use restrictions such as the prohibition to burn fields or forested areas, or the prohibition to deforest.

11 Other reasons for leaving/not participating were the following: general feeling of distrust towards institutions, reluctance to sign a contract with the project, reluctance to take risks when investing in the farm, resistance to experimentation (in terms of trying out new practices).

the total group of people receiving PES (with or without TA) was again divided into two groups, one receiving only payments during the first two years of the project, and one receiving payments for four years (until the end of the project). The control group was established later on (see below).

The number of farmers per category was more or less fixed beforehand (PES: 30; PES + TA: 70; CG: 30; see operational manual RISEMP). Actual numbers of farmers per group in 2003 are depicted in Table 1 (see also Annex 4 for the other countries).

Table 1. Subdivision of farmers according to type of incentive, Nicaraguan pilot site

Group	2003	2007
PES	30	28
2 years	7	6
4 years	23	22
PES + TA	77	70
2 years	24	20
4 years	53	50
CG	29	25
Total	136	123

Source: RISEMP data

3.2. Validity of comparison of means: 'PES only group' versus 'control group'

3.2.1. Selection bias

The groups receiving PES only, PES + TA, and the crosscutting groups receiving PES for a period of 2 or 4 years were established more or less at random from the population of farmers attending project meetings and falling within the pre-established criteria described in the previous section.

The control group was selected after the other groups and its subdivisions (as described above) had already been established. The urgency to find a sufficiently large group of willing farmers and the timing of the selection made it impossible for project staff to select farmers randomly (from the same population as farmers selected for the other groups) or even on the basis of certain selection criteria (see section 3.2.3.). In the end, the control group comprised farmers who had continued to attend the project meetings (but did not comply with selection criteria for PES), farmers who had ceased to attend the meetings, and others. As a result, comparisons between the 'PES only' group and the CG would be biased due to severe problems of selection bias on the basis of observables as well as unobservables.

Selection bias on the basis of observables. Landowners not complying with selection criteria for receiving PES or TA and therefore rejected for receiving PES, in some cases were asked to become part of the control group. This was part of a

pragmatic solution to rapidly define groups of sufficient size. The downside of this type of measure was that it introduced a clear selection bias on the basis of observable characteristics (see Table 2). CG farmers had on average more land, livestock and a relatively smaller proportion of the CG (in comparison with the 'PES only' group) had a history of receiving TA prior to the project. In addition, in the case of land and livestock the standard deviation is much higher in the control group because to a large extent the CG contained farmers with properties that were either too small (less than 8 hectares) or too large (more than 100 hectares) to be considered eligible for receiving PES. Especially the latter type of farmer was quite different from the average participant. Apart from having more land and assets and thus more capacity to invest, several of these farmers did not spend much time on their farms, instead having a manager to run their farm for them. In other words, decision-making in these cases was divided between owners and managers. Apart from variables such as farm size, assets, living and working on the farm, there were other differences. In the PES groups there were subgroups of people that were very well organized. This was mostly due to the fact that the initial group of potential project beneficiaries included networks of members of some farmer associations with a previous experience of working with and benefiting from projects implemented by Nitlapán. There was no such social structure linking control group farmers to each other.

Selection bias on the basis of unobservables. Some of the farmers that had lost interest in the project at the time of preliminary meetings (before the experimental design was established), were later asked to become part of the CG. While some of the CG farmers were likely to be more reluctant to adopt innovations than the average participant, in practice a subgroup of the CG was triggered by the project to invest in silvopastoral practices (see discussion below).

Table 2. Evidence of observable selection bias between the different 'treatment' groups of the RISEMP project, Nicaragua pilot site

	Farm size (ha) mean (std.dev.)	Units of livestock mean (std.dev.)	Received technical assistance in the 3 years prior to project (% yes)	Received credit in the 5 years prior to project (% yes)
Control group (n = 25)	46,7 (37,1)	51,0 (41,1)	12,0 %	60,0 %
PES + TA (n = 69)	31,9 (25,8)	30,2 (29,0)	42,0 %	47,8 %
PES only (n = 28)	29,5 (25,1)	32,6 (35,6)	21,4 %	64,3 %
PES 2 years (n = 26)	27,2 (22,7)	22,9 (23,9)	26,9 %	46,2 %
PES 4 years (n =71)	32,7 26,4)	33,8 (32,7)	39,4 %	54,9 %

Source: RISEMP baseline data 2002

3.2.2 Treatment diffusion

Several types of contagion affected the experiment, reducing the differences between 'treatment' groups and rendering part of the comparison between groups invalid.

There was no contagion effect with respect to PES. Payments were restricted to the groups selected for payments. However, payments were expected to alleviate the capital constraint and consequently boost adoption of silvopastoral practices. Any evidence of this effect could be distorted by selection bias problems; i.e. a substantial number of CG farmers (see above) did not face a capital constraint and would be able to invest without receiving PES.

The objective of technical assistance was to relieve the knowledge constraint for being able to invest in silvopastoral practices. The potential of using the experiment as a framework for isolating technical assistance effects from the influence of other effects was in practice completely compromised by the following types of contagion.

The first type of contagion refers to the issue of farmers that were not supposed to receive TA from the project, in fact benefiting directly from project TA activities. Project staff admirably tried to separate the PES + TA group from the other groups ('PES only' and CG). Nevertheless, given the proximity between the farmers, the social relationships among farmers and also the social relationships between farmers and staff, this was very difficult. Information about project activities was widely available. Although to a large degree, workshops, exchanges and personal visits were restricted to the technical assistance group, other farmers from other groups,

although not formally invited, would also sometimes attend the sessions. It was difficult for project staff to prevent these farmers from attending.[12] In other occasions, information and advice on land use techniques was given to farmers from whatever group when they asked for it.

The second type of contagion refers to the diffusion of knowledge from farmer to farmer. Farmers from the different groups were often neighbors, friends, members of the same networks or even family members. Farmers from the technical assistance group would often share their newly acquired knowledge on silvopastoral techniques with others. Evidently, when some of the silvopastoral practices began to manifest their benefits, interest from other farmers increased. What in development projects is usually considered as an important benefit, i.e. the diffusion of project knowledge beyond the participant group, turned out to be a substantial impediment on the validity of the experimental design.

The third type of contagion refers to knowledge acquired by other institutions. Before the project initiated its activities in the region, several institutions, including Nitlapán, had already been working with farmers on livestock production and land use systems. In fact, most of the practices promoted by the RISEMP project were not new. During the project implementation period, Nitlapán successfully negotiated a division of labor with the principal institution delivering TA in the region, Technoserve. The implication was that Technoserve would focus on other topics and other farmers than Nitlapán. Despite this agreement, basic knowledge of silvopastoral practices continued to be available to farmers through different channels. While both the 'PES only' group and the CG were meant not to have access to TA, both groups in fact had access to knowledge about silvopastoral land use practices through the three mechanisms described above. It is unclear whether the 'PES only' group on average benefited more from project TA than the CG. Probably, the problem of treatment diffusion of TA has had no substantial effect on the validity of the comparison between the groups as an indicator of the net effect of payments on land use changes. In contrast, treatment diffusion rendered group comparisons between the 'PES + TA' group and 'PES only' group invalid, as this comparison was precisely designed to test the effectiveness of TA on land use changes (see section 3.3.2.).

3.2.3. Unintended behavioral responses

Previously, we identified several unintended behavioral responses that might negatively affect the validity of group comparisons. The first type of unintended response concerns expected behavior or compliance behavior: participants react in accordance with project staff expectations for reasons of compliance with the established contract or the relationship with staff or due to certain expectations about future benefits. Although unintended[13] from an experimental design point of view, project staff in fact intentionally tried to influence participant behavior through other mechanisms than payments and technical assistance. For example, the practices of burning plots and deforestation were by formal agreement prohibited and

12 Nor did they feel very motivated to exclude farmers from courses.

13 We use the term unintended to refer to other responses than those generated by project incentives.

farmers not complying with this directive would be expelled from the project. The implication of the latter is that any change or lack of change in forest cover cannot be uniquely attributed to project incentives given this contractual obligation.

Another way in which project staff tried to ensure contract compliance as well as stimulate adoption of silvopastoral practices was the promise of a second phase. The impression was created that if farmers did well they would be eligible for a second phase[14] of payments and project benefits. The promise of a second phase was also used to motivate farmers to join the CG. This had a clear behavioral effect in terms of motivating several CG farmers to start investing in silvopastoral practices with the expectation of becoming a future participant eligible for payments.

The longstanding relationship between several farmers receiving PES (and TA) and Nitlapán was another element that triggered unintended behavioral responses. For example, participant farmers in the community of San Ignacio had been working with Nitlapán for years and by the end of the project were likely to continue the collaboration in the future. Compliance with contractual obligations as well as land use changes in this community were substantially influenced by a variety of mechanisms such as trust, reputation and friendship that were part of the ongoing collaboration between the farmers and Nitlapán.[15]

The second type of behavioral response is called compensatory equalization: the idea that discontent among staff or recipients with the inequality between groups might result in compensation of groups that receive less than others. The selection of the control group in many aspects had been problematic. The late introduction of the idea of a control group made it extra hard to find farmers willing to be part of this group.[16] Apart from the promise of eligibility for a second phase of the project, in order to arrive at a sufficiently large number of farmers willing to be in the control group, Nitlapán relied on the following mechanisms:

- (in some communities) the longstanding relationship of collaboration between certain farmers and Nitlapán was invoked to persuade farmers to be part of this group;
- farmers were paid a small fee (10 US$ per year) for collaborating with the data collection activities in the farm;
- farmers were offered a map of their farm, based on a satellite photograph and GIS information; this was considered to be useful for further on-farm planning and production.

A final unintended behavioral effect is compensatory rivalry: differentiation of incentives to groups of farmers might result in social competition between those receiving (many) project benefits and those that receive less or no benefits. In the Nicaraguan case, this has probably been the strongest unintended behavioral effect. Among a number of CG farmers there was substantial resentment for being rejected for participating in the project.[17] As a result, these farmers wanted to show the pro-

14 In fact, project staff and WB staff themselves were convinced that they would be able to successfully define a follow-up project.

15 In fact, adoption of silvopastoral practices in this community has been more successful than in other communities.

16 Farmers would be deprived from previously promised benefits.

17 Interviews with farmers and Nitlapán field staff.

ject staff and the other participating farmers that they could innovate without project support. The fact that they lived near farmers who were receiving TA (and PES), which facilitated learning about the new land use practices, made it easier for CG farmers to compete.

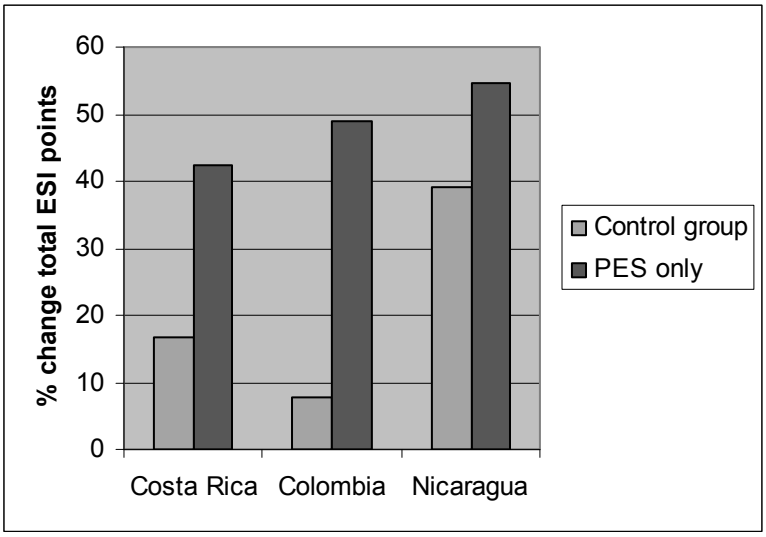
3.2.4. Conclusion on validity of comparison of means

Brief summary on threats to validity:

- selection bias: problematic
- treatment diffusion: not problematic
- unintended behavioral responses: problematic
- validity of comparison of means: low

Figure 2 shows a comparison of increases in ESI points (see Annex 2) per hectare for the three countries.

Figure 2. Incremental ESI points per hectare (2003-2007), three countries – PES only versus CG



Source: own calculations based on RISEMP project data, January 2008

In theory, the net effect of the project would be the difference between the average incremental ESI values of the 'PES only' group and the CG. On the basis of the previous discussion we can conclude that given the problems in the design and implementation of the experiment, both the CG and 'PES only' group values are distorted in such a way that an unbiased comparison is impossible. With respect to the CG, we do not know to what extent problems of selection bias, contagion and unintended responses have distorted the value of the CG from the value it would have had in the absence of these problems. We can only conclude that several elements

have contributed to a relatively high increase in ESI points in the CG. For example in the case of Nicaragua:

- Selection bias: several CG farmers on average were wealthier than 'PES only' farmers, with access to capital (and labor) to implement changes (positive effect on adoption of silvopastoral practices).
- Contagion: several CG farmers had access to technical assistance from other organizations as well as the project itself (positive effect on adoption). The fact that the 'PES only' group also had access to TA makes contagion less problematic for this particular comparison.
- Unintended responses: several CG farmers harbored the expectation of becoming a future beneficiary of the project (positive effect on adoption).

With respect to the 'PES only' group, unintended responses such as compliance behavior or the expectation of maintaining access to future benefits of Nitlapán positively affected adoption behavior.

To conclude, the true net effect of the project for this variable (difference in incremental ESI points) is different from what we can deduce from the inter-group comparisons of 'PES only' and CG farmers. Given the systematically higher incremental change in ESI points (or environmental value of the land use) among 'PES only' farmers in comparison to CG farmers in the three countries we can safely conclude that PES has had an effect on land use changes. Nevertheless, for the Nicaraguan case, given the problems affecting the design, and on the basis of these comparisons only, we cannot say whether payments have been important or even decisive in bringing about land use changes and corresponding indirect effects.

It is worthwhile to mention that the lower values in the CG in Costa Rica and Colombia suggest that the experimental designs in these sites were probably less problematic than in Nicaragua. However, also in these cases the differences between 'PES only' and CG values do not represent unbiased estimates of the net effect of the project. Given the scale and diversity in problems affecting the experimental design in Nicaragua it is unlikely that the other two sites were unaffected by any of these problems (see Annex 5).

3.3. Validity of comparison of means: 'PES only group' versus 'PES plus technical assistance group'

3.3.1. Selection bias

The allocation of farmers to the 'PES + TA' group, the group that was considered to be the most attractive to farmers, was not entirely random. Interviews with staff and farmers confirmed that there was some favoritism towards the most motivated farmers (those that had attended all the meetings) and farmers with a history of collaborating with Nitlapán on previous projects. The higher percentage of previous access to TA among 'PES +TA' farmers appears to confirm this. Nevertheless, at group level, the two groups ('PES + TA' and 'PES only') are largely similar with little observable selection bias (see Table 2). In addition there is likely to be some unobservable selection bias as assignment of farmers to either of two groups was not entirely done in a random manner.

In sum, the principle randomization of groups was respected to some extent and there were no substantial differences between the 'PES only' and PES + TA' group that might seriously invalidate the comparison of means as a measure for the net effect of the project.

3.3.2. Treatment diffusion

On the basis of the discussion in section 3.2.2 we can conclude that group comparisons between the 'PES + TA' group and the 'PES only' group are rendered useless due to the different treatment diffusion problems. Knowledge on silvopastoral practices was widely available through the three mechanisms described earlier. 'Treatment' differences between the two groups are considered to be too small to allow for any meaningful interpretation of changes attributable to the project TA on the basis of the experimental design.

3.3.3. Unintended behavioral responses

Although the experiment was severely affected by unintended behavioral responses, the effect was primarily on the 'PES only' group versus CG comparison. There are no reasons to assume that there were any systematic differences in unintended behavioral responses between the 'PES only' and 'PES + TA' groups.

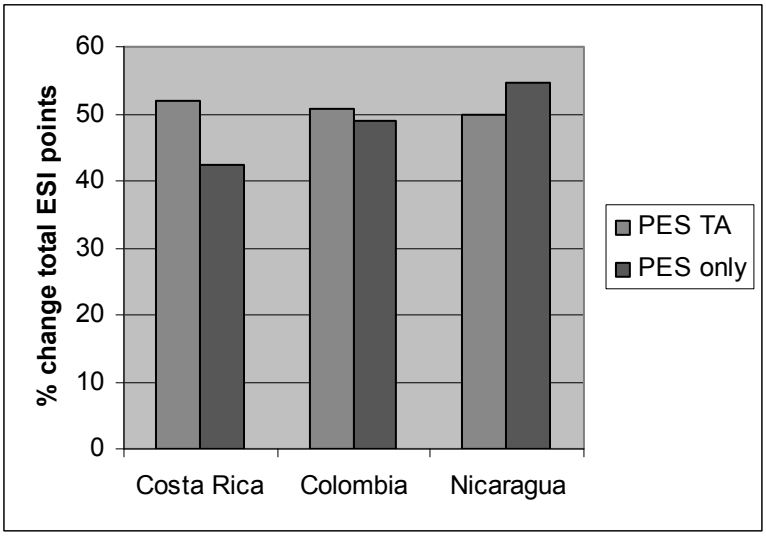
3.3.4. Conclusion on validity of comparison of means

Brief summary on threats to validity:

- selection bias: not problematic
- treatment diffusion: problematic
- unintended behavioral responses: not problematic (no systematic differences between the groups)
- validity of comparison of means: low

On the basis of the previous discussion we can conclude that group comparisons between the 'PES + TA' group and the 'PES only' group are rendered useless due to the different contagion problems. In other words, 'treatment' differences between the two groups are considered to be too small to allow for any meaningful interpretation of changes attributable to the project.

Figure 3. Incremental ESI points per hectare (2003-2007), three countries – PES only versus PES + TA



Source: own calculations based on RISEMP project data, January 2008

Figure 3 confirms this. Differences between the two groups in terms of incremental ESI points are very small in Nicaragua. This is also the case for the other two countries. However, in Costa Rica and Colombia, incremental change is higher in the 'PES + TA' groups. This is a more plausible result as it suggests that the combination of PES and technical assistance is more useful than PES only (see also section 3.5.).

3.4. Validity of comparison of means: 'PES 2 years' group versus 'PES 4 years' group

3.4.1. Selection bias

The decision to assign a particular farmer to a group receiving either 2 years or 4 years of payments was done more or less at random. The groups were rather similar in terms of observable characteristics (see Table 2) and (given the random nature of assigning farmers to either of two groups) most likely also for unobservable characteristics influencing adoption behavior. There is a slight bias in terms of farmers in the 4 years payment group having more livestock and a higher percentage of farmers having received TA.

3.4.2. Treatment diffusion

There were no systematic differences between the two groups in terms of access to TA or knowledge on silvopastoral practices in general. Both groups included

farmers with access to TA from the project and in addition, knowledge on silvopastoral practices was also widely available to farmers from both groups through the three mechanisms described earlier. There was no treatment diffusion with respect to PES modality (i.e. the differentiation between 2 and 4 years of payments was correctly implemented).

3.4.3. Unintended behavioral responses

Although the experiment was severely affected by unintended behavioral responses, the effect was primarily on the 'PES only' group versus CG comparison. There are no reasons to assume that there were any systematic differences in unintended behavioral responses between the 'PES 2 years' 'PES 4 years' groups.

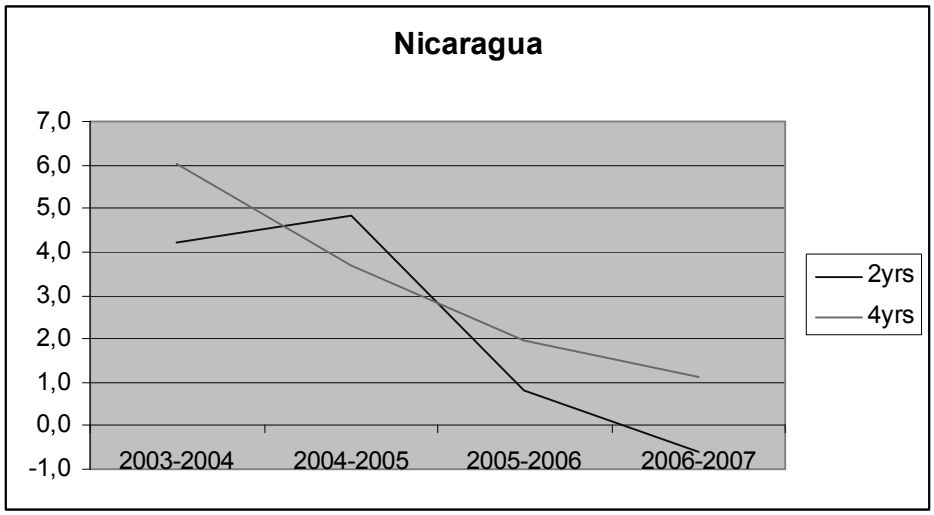
3.4.4. Conclusion on validity of comparison of means

Brief summary on threats to validity:

- selection bias: not problematic
- treatment diffusion: not problematic
- unintended behavioral responses: not problematic (no systematic differences between the groups)
- validity of comparison of means: high

The comparison between farmers receiving only 2 years of payments and farmers receiving 4 years of payments is considered to be quite valid. There were no severe problems of selection bias. In addition, treatment diffusion of PES modality did not occur. Finally, there is no indication that there were systematic differences between the two groups of farmers in terms of unintended behavioral responses.

Figure 4. Adoption behavior of the PES2yrs vs. PES4yrs group (average incremental points per farmer per year, in relation to previous year), Nicaraguan pilot site



Source: own calculations based on RISEMP project data, January 2008

Interesting findings that can be deducted from Figure 4 are the following. First, the trend appears to confirm hypothesis 4 (see section 1.3.) that adoption behavior was initially higher in the 2 years group, whereas it substantially declined in the last two years of the project (trends in other countries confirm this, see Annex 6). In the 2 years group, the incremental ESI even became negative in the last year of the project, indicating a decline in the ecological value of the farm. This had not happened (yet) in the other two countries (see Annex 6), although similar patterns can be discerned there.

The declining trend can be explained by several factors. First, once payments for land use changes from the project stopped, an important incentive for adoption was taken away. Many farmers wanted to get as close as possible to the maximum amount of PES that could be earned (6,000 US$ per farm) given the constraints they were facing. After payments ceased, adoption slowed down as the incentive to earn money had disappeared. Second, resource constraints in combination with the fact that farmers were not (yet) convinced of the (economic) advantages of some of the practices led to a decline in investments in new practices and possibly a decline in maintenance of (some) existing practices (resulting in a declining incremental ESI value (per farm) or even a negative one). The comparison depicted in Figure 4 can be useful for predicting the sustainability of land use changes in the near future. It is likely that some land use practices once they have demonstrated their economic value will be sustained and even expanded into the future (e.g. improved pastures). The initial steep decline and subsequent flattening of the curve appears to support this (see also Annex 6). However, at the same time resource constraints and the lack

of economic pay-off for some of the practices (e.g. conservation of shrubs) might result in a decline in the ecological value of the farm (as indicated by the downward trend in Figure 4). The overall sustained gain/loss in terms of ESI points per farm is a question that warrants further inquiry. In practice, whether there will be a net decline or gain in ecological value at the farm level depends on individual-specific factors such as farm household characteristics, on-farm and off-farm income activities, market integration (and correspondingly market conditions), and the existing mix of land use practices as well as recent innovations and trends. Without further insight into the question of how incentives affect types of farmers in different ways, predictions as to the overall sustainable effect of the project on land use changes and indirect benefits are impossible to make (see section 5 for further discussion).

3.5. Other issues that affected the utility of the experimental design

Other issues affecting the usefulness of the design were the following. The first aspect is treatment change. The amount of PES per ESI point was increased during the project (after already having received a first payment). Payments per ESI point were increased to 75 US$ and 110 US$ per point respectively for farmers receiving 4 years and 2 years of payments. The behavioral effect of this increase on adoption behavior has probably been small.

Quality of the data and measurement problems. In the first years of the project, staff was still learning how to measure and characterize land use activities (for example in distinguishing improved pastures from natural pastures).[18] This particularly affected the quality of the baseline data. Later on, when more experience was gained in recognizing and measuring particular land uses, the quality of the monitoring increased substantially. As a result, baseline data are probably overestimating silvopastoral land use systems in use. The implication is that actual adoption has been higher for some practices than shown in the data.

Timing of the baseline and ex post survey. The timing of the baseline was adequate; the survey was implemented before initiating payments and technical assistance activities. Land use data were adequately collected annually allowing for dynamic comparisons between groups.

Sample size. Given the small group sizes (especially of the CG) and possible high variance among farmers, statistical power is quite low. The probability of type II errors is high (failing to find a difference between groups when in fact there is a real difference). In addition, a sample size of 30 is more or less the lower boundary for applying the central limit theorem which allows for the application of parametric statistical tests. The implication is that formal confirmatory statistical analysis in some cases is not warranted (especially the comparison between CG and 'PES only'). In those cases, group comparisons should be limited to descriptive comparisons.[19]

18 Interviews with Nitlapán field staff.

19 Non-parametric tests can in principle be used to test for statistically significant differences between groups. However, not only are group sizes relatively small, given the non-random selection of some of the groups, statistical inference is not particularly useful.

Other characteristics of the design. The design did not include a group that was targeted with technical assistance without payments. The current design if well implemented could have tested for the added value of technical assistance to farmers already receiving payments. However, it could not test for the effectiveness of technical assistance as such. Testing the effectiveness of technical assistance was not an original objective of the project. However, our interviews with field staff and farmers suggest that for many farmers knowledge rather than money is the most direct constraint for successful innovation. Monitoring land use changes of separate groups of farmers with only PES or only technical assistance could have elucidated what on average would have been the most pressing constraint for land use change in the region.[20] Additional analysis (e.g. using household survey data) would have helped to shed light on the question as to what types of farmers would require PES, technical assistance or both (see also section 5).

3.6. Final remarks on the design and implementation of the experimental design for the Nicaraguan pilot site

The basic conditions for managing a controlled experiment were not fulfilled. The different stakeholders, WB staff, project staff in Nitlapán and farmers did not share the same vision about the importance of the experiment. In case of trade-offs with the development objective of improving the environmental conditions of farms and the livelihood conditions of farmers in the project, project staff acted in favor of the latter, for example by not exercising a strict control on the technical assistance component of the project. In the management of the experiment there was little or no quality control. As a result the majority of problems in terms of selection bias, contagion and behavioral responses were not addressed or even identified.

The setup of the experimental design was described in the operational manual of the project. The manual provided clear instructions on the allocation of farmers into different groups and on group sizes. In addition, the manual specified the information to be collected in the baseline survey, covering all groups of farmers. In general, the basic logic of the experiment was well understood among staff. However, most of the project staff had no experience with managing such an experiment. The majority of project staff in Nitlapán did not have a background in research, but were experienced livestock specialists and extensionists, having worked all their professional life in rural development projects. For several staff members in Nitlapán, especially the field staff who interacted most frequently with farmers, the primary goal was to implement a project that would bring benefits to the people, not to carefully manage an experiment for research purposes. There were no clear guidelines (at least not on paper) on how to manage the experiment in the field, and staff, having had no training in the logic and management of this type of experiments, when aware of problems that could affect the validity and utility of the experiment had to improvise along the way.

20 This suggestion was raised by staff from Nitlapán who would have preferred this type of design over the current one.

Communication about the experiment to the farmers suffered one serious setback. The project started out with the message that all farmers would receive payments. In a subsequent supervision mission by the WB the scheme was changed and the idea of a control group to be monitored by the project while not receiving any benefits was introduced.[21] This caused quite some resentment among farmers. In the beginning CIPAV (Colombia) and Nitlapán (Nicaragua) and to a lesser extent CATIE (Costa Rica) did not quite agree with the idea of a control group as it was considered not to be ethical. Despite several objections, the project went through with the inclusion of a control group.

The idea of differentiating treatments to farmers was not welcomed by all farmers. The differentiation between participants and control group, the division between those that would receive technical assistance and those that would not and the division between 2 years and 4 years of payments caused substantial resentment as well as confusion among farmers.

In all, the experimental framework failed on two of the three group comparisons that were to support rigorous claims on the effects of PES and technical assistance on land use change and corresponding environmental effects. The 'PES only' versus CG comparison is rendered invalid due to severe problems of selection bias and unintended behavioral responses (especially in the CG). The 'PES only' versus 'PES TA' comparison is rendered invalid due to problems of treatment diffusion. The 'PES 2 years' versus 'PES 4 years' comparison is quite valid. The data and their subsequent interpretation illustrate the utility of the experimental design in terms of providing reliable evidence on land use behavior under different types of incentives.

4. Utility of the experimental evaluation design in the assessment of project outcomes and impacts

The previous discussions allow us to present an overall picture of the strengths and weaknesses of the project's experimental design as a basis for outcome and impact assessment. In the presentation of this picture, we will focus on the concrete situation of the project (taking into account design and implementation failures discussed above). In addition, we will highlight some aspects from the point of view of an ideal experimental setting.[22]

4.1. Strengths of the experimental design for assessing outcomes and impacts

A critical comparative advantage of (quasi-)experimental designs is that they allow for a quantitative estimation of the net effect of an intervention, 'controlling' for other external factors. In other words, they have a comparative advantage in

21 Interviews with staff from CATIE and Nitlapán.

22 We will restrict ourselves to arguments that directly relate to the type of evaluation context of projects such as RISEMP. We will not go into the general debate on the applicability of (quasi-) experimental methods in development interventions (see for example Bamberger and White, 2007).

establishing to what extent changes in target variables are brought about by an intervention vis-à-vis other factors. Both positive and negative effects can be identified, and both direct and indirect effects, depending on the set of variables that is taken into account.

In addition, although design and implementation of the experimental framework may be costly and require the necessary technical expertise, subsequent analysis and interpretation of data is fairly straightforward. Experimental design data can provide reliable and easy to interpret evidence to policy makers.

In the case of the RISEMP project, if selection bias, contagion and unintended behavioral responses would have been kept under control, the design would have allowed for unbiased estimates of the net effects of the project incentives on land use changes and environmental target variables. The utility of experimental approach has been illustrated by the valid comparative analysis between groups of farmers that received 4 years of payments versus those that only received 2 years of payments.

4.2. Weaknesses of the experimental design for assessing outcomes and impacts

Disentangling the effects of project incentives from other factors. Although the comparative advantage of the experimental approach lies precisely in its potential to isolate project effects from other effects, the analytical benefits of the design can only be realized if extensive and careful attention is paid to the different threats to validity. Problems of selection bias, contagion, and unintended behavioral responses have compromised the utility of the experimental design. Given the substantial threats to validity, especially in the Nicaraguan case, but probably also in the other two countries, the effects of the project incentives cannot be satisfactorily disentangled from the influence of other factors.

Scope of the evaluation. Experimental designs are not equipped to address the full scope of impact of projects such as the RISEMP project. We briefly highlight two aspects of scope. First, as presented in the introduction of this report, we only focus here on effects at field level. The other two dimensions, institutional effects and replicatory effects, are not amenable to being assessed by means of the experimental approach.[23] The two dimensions cannot be captured in terms of discrete 'treatments' being applied to a discrete (and sufficiently large) group of subjects. Tracing and measuring the institutional and replicatory effects of the project would require a completely different approach, starting out from a theory of impact (see below). Second, even within the confines of impact at field level, the experimental design cannot address the full range of effects of the project. It is restricted to a limited set of indicators. Ideally, as is the case in the RISEMP project, these indicators are included in the baseline study of an intervention. Unexpected and unintended effects are usually not (adequately) captured by baseline studies and therefore not

23 One might think of ways to assess for example institutional effects through randomized experiments. However, at least three constraints come to mind: the unit of analysis for randomization, the range of institutional effects to be included in the experiment, and the cost-effectiveness of doing such an analysis, especially when compared to other methods of inquiry.

taken into account in the analysis. Examples of effects which cannot be adequately assessed (mainly) due to lack of data are the following:

- Indirect environmental effects such as the displacement of ecologically destructive land uses;
- Other indirect effects: e.g. diffusion of adoption by other farmers, (price) effects on local inputs and commodity markets (through production effects), (price) effects on land markets.

Moreover, in the specific case of the RISEMP project socioeconomic impacts cannot be assessed on the basis of the experimental design. This is not a shortcoming of the design itself but of data collection efforts; socioeconomic variables were only monitored for a subgroup of farmers, not including all groups. Socioeconomic impacts can be inferred from analyses of the private profitability (based on intensive monitoring of a small group farmers) of land use practices in combination with adoption grades. The baseline survey of the project includes relevant socioeconomic data to be used in impact analysis. As to date, no follow-up socioeconomic survey covering all project farmers has been organized.

Timing. Typically, adoption processes of new technologies do not take place over night. Farmers continuously experiment and assess the attractiveness of innovations on their farms. Consequently, given the relatively short time span of the project (five years) and recent closure of the project (in January 2008), one might expect farmers to continue to expand investments in land use practices that they perceive to be attractive while at the same time ceasing to invest in other land use practices (or even undo some of the changes made). Therefore, the full long-term impact in terms of the environmental and socio-economic benefits brought about by land use changes cannot be assessed yet. Usually, an outcome and impact evaluation based on an experimental approach would rely on a few snapshots of reality, the before and after situation. In addition, experimental evaluations usually look at (short- and medium term) outcomes rather than (long-term) impacts. In fewer cases, periodic measurements of target variables (as in the case of the RISEMP project) are available. In any case, the long-term effects are usually not yet apparent at the time of outcome and impact evaluations. While this is the case for any type of outcome and impact evaluation, there are other methodological approaches, such as sustainability analysis in combination with theory-based evaluation (see below) that are helpful in developing a line of argumentation on the likely sustainability of certain effects (see for example, GEF 2009a). In the case of the RISEMP project, trends in adoption data and group comparisons between the two groups receiving payments for respectively 2 and 4 years are useful for complementing such alternative analyses in assessing the likelihood of sustainability of land use changes and corresponding environmental and socioeconomic effects.

Measurement issues. Technically, this is not just about the weaknesses of the experimental design as such, but applies to a broader range of techniques for statistical analysis and beyond. However, measurement issues are especially important in this context as (quasi-) experiments focus on a narrow set of indicators only, which are used to refer to broader phenomena and processes of change. We briefly high-

light two issues: the issue of construct validity, whether a variable correctly captures the phenomenon it refers to, and the type of data.

Any type of statistical analysis (whether descriptive or inferential) is based on a succinct abstraction of reality, not only in terms of the relationship or the model that is the focus of the analysis, but also in terms of the number and choice of variables that are to represent reality. An example from the RISEMP project is the following. Using the variable quantity of land use changes both as a proxy and a basis for assessing environmental effects has its limitations. A distinguishing element not entirely adequately captured by the project's monitoring system concerns the quality of land use changes.[24] When looking at the effectiveness of TA, the distinguishing effect of TA delivered by the project on land use changes will probably not be visible in the quantity of land use changes but rather in the quality of land use changes as well as the sustainability of these changes. The quality of land use change has not been expressed in measurable indicators and consequently is not included in the experimental design-based comparisons between groups.

Regarding the second aspect, type of data, it should be highlighted that there is a substantial difference in working with data that are based on direct measurement (e.g. areas with a certain land use directly measured by project staff using GIS tools) and data generated from survey questionnaires. In general, survey data are more liable to suffer from measurement errors. In the case of the RISEMP project, despite the fact that information-sharing was part of the contract, farmers often felt reluctant to share confidential information within the framework of (socioeconomic) surveys. This was especially problematic in Colombia. The land use changes used in the group comparisons (based on direct measurement) are less likely to suffer from measurement error than survey data used in many other settings.

The rationale underlying changes. Experimental designs are equipped for determining the outcomes and (to a lesser extent) impacts of interventions on target variables while 'controlling' for known (and unknown) external factors. In reality, the changes in target variables are the result of an interplay of factors, of which an intervention is but one of many. Experimental approaches directly relate target variables to 'treatments' and do not address the underlying issue of how changes have come about. What are the causal pathways underlying processes of change influenced by the intervention? Under what circumstances do project incentives positively (or negatively) affect target variables? How do project incentives affect particular types of farmers in different ways? What are potential unintended, indirect or long-term results of interventions? Are there other instruments that might have achieved similar results more effectively? These and other questions regarding the nature of processes of change influenced by interventions and the resulting effects over time are best addressed using a theory-based evaluation approach (see below).

24 According to project staff this aspect is especially important in practices such as fodder banks or live fences. Two fodder banks of the same size (in land area) might differ substantially in terms of the diversity of species of grasses and other plants, with implications for biodiversity in terms of the fodder bank's role in supporting insects and bird species.

5. Reinforcing the experimental design-based analysis with other methods

Promising methodological options to reinforce the outcome and impact evaluation in the case of the current project as well as in other similar contexts are the following:

- Additional statistical analysis;
- Theory-based evaluation;
- Sustainability analysis.

5.1. Additional statistical analysis

Despite the problems found in the experimental design, further statistical analysis will be possible if another effort is made to collect ex post data at farm (household) level. This will open up new possibilities for using matching techniques (creating better control groups) or regression-based approaches using statistical controls to reduce observable selection bias problems (and to some extent, if measured) contagion problems. Nevertheless, some of the validity threats to attribution analysis (unobservable selection bias, contagion problems, unintended behavioral responses) cannot be corrected by further quantitative analysis.

Another option for further quantitative analysis would be to focus less on attribution of changes to project incentives and more on the general question of associations between levels and patterns of adoption on the one hand and different incentives, farm (household) characteristics and contextual variables on the other. To some extent, these data are available in the project's baseline survey. The problem is that important variables like land sales or purchases, access to credit and/or technical assistance in the past few years, are crucial explanatory variables without which further explanatory analysis would be markedly incomplete. This provides another reason for implementing an ex post survey covering all PES and CG farmers.

5.2. Theory-based evaluation

Theory-based evaluation focuses on the underlying assumptions of how an intervention is supposed to work (see for example Weiss, 1997; Rogers et al., 2000; Leeuw, 2003). A distinction can be made between process theory and impact theory, the latter focusing on the causal assumptions connecting project outputs (and some process variables) with outcomes and impacts. Several pieces of evidence can be used for reconstructing the intervention theory, for example:

- an intervention's existing logical framework (see Annex 1) provides a useful starting point for mapping causal assumptions linked to objectives; other written documents produced within the framework of an intervention are also useful in this respect;
- insights provided by as well as expectations harbored by policy makers and staff (and other stakeholders) on how they think the intervention will affect/is affecting/has affected target groups;

- (written) evidence on past experiences of similar interventions (including those implemented by other organizations);
- research literature on mechanisms and processes of change in certain institutional contexts, for particular social problems, in specific sectors, etc.

Figure 5. Basic impact theory of the RISEMP project at field level

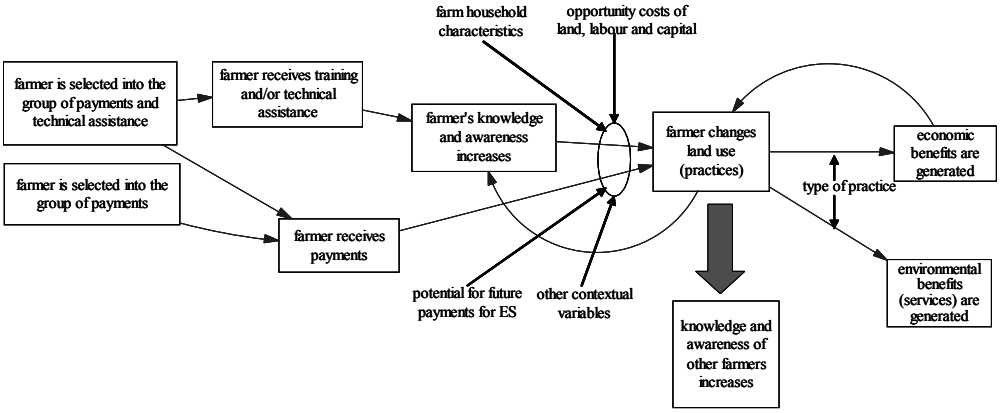


Figure 5 presents a graphical representation of the impact theory (restricted to effects at field level) of the RISEMP project. It shows the main causal assumptions linking project incentives (TA and PES) to final impact variables.

In general, the impact theory can serve multiple purposes within the framework of an outcome and impact evaluation. We briefly mention three principal purposes:

- Explanatory-analytical function: what are the main causal pathways through which the project influences processes of change leading to outcomes and impacts? Once the evaluators have reconstructed a workable theory that adequately reflects the existing knowledge on processes of change for the evaluation context at hand, the theory can be used as a basis for argumentation, supporting the analysis on impact when needed. In addition, the theory can be helpful in explaining differences in effects among farmers, i.e. conditioned by contextual variables (e.g. farm gate prices of milk, agro-physical conditions) and characteristics of farmers (e.g. age, wealth).
- Methodological design function: the initial impact theory provides a useful framework for determining what type of evidence is needed for testing and refining the theory and responding to the principal questions of the outcome and impact evaluation (see Box 2).
- Predictive function: the impact theory can be used to support predictions on what is likely to happen. The theory can serve as guidance for assessing the likelihood that certain changes will come about in the future or will be sustained into the future. While these effects are unknown at the moment the outcome and impact evaluation is undertaken, certain conditions for sustainable impact can be identified which increase the likelihood of sustainable change

(see below). This analysis will be strengthened when combined with an analysis of extrapolation of current trends into the near future.[25] Current work by the GEF EO on reviews of outcome to impact using a theory-based approach focus on these conditions (see GEF, 2009a).

It is beyond the scope of this report to provide a comprehensive theory-based analysis of project effects. To illustrate the value of an impact theory we briefly present a few corroborated aspects of the impact theory regarding the determinants of adoption behavior of silvopastoral practices:

- An important determinant of adoption behavior was the level of organization. In the Nicaraguan pilot site in two of the seven communities (San Ignacio and Paiwita) farmers were well organized in an association called 'Asociación de San Jose'. Group solidarity and shared norms (in terms of mutual assistance during workshops or in the field, extensive knowledge exchange, group discussions, etc.) positively influenced adoption behavior.
- Many farmers while receiving PES payments also had access to credit. For example, in order to effectively utilize the fodder banks, one would need a 'picadora' (equipment for processing grass into fodder). Nitlapán (in collaboration with a rural bank called the FDL) facilitated access to this equipment, respectively by offering a low interest loan (a so-called 'green loan' provided for these purposes) or directly delivering the equipment to the farmer to be repaid without interest over a period of two years. This to a large extent explains the high adoption rate of fodder banks in Nicaragua in comparison with the Costa Rican and Colombian site where access to this type of assistance was more limited.
- High milk prices had a positive effect on farmers' willingness and capacity to invest in silvopastoral land use practices. The additional incentive of these prices (next to the ex post payments of the environmental services generated by the silvopastoral land uses) reinforced farmers' preferences to invest in silvopastoral practices that generated a combination of low or intermediate environmental benefits (and corresponding relatively low amounts of PES) with direct economic benefits (e.g. improved pastures with trees), over land use practices that offered no productive benefits but high environmental benefits and corresponding high amounts of PES (e.g. reforestation).
- Small farmers are more likely to maximize value per unit of land, which explains why they had higher incremental ESI points per hectare than big farmers, who tend to maximize value per unit of labor.

25 See for example the previous discussion on the comparison between those farmers receiving 2 years and those receiving 4 years of payments.

Box 2. An example of a mixed-method design for impact evaluation

E.g. a randomized experiment could be used to assess the effectiveness of different incentives (PES and TA) on land use change and subsequent environmental and socio-economic effects of these changes (potentially strengthens the internal validity of findings);

E.g. survey data and case studies could tell us how incentives have different effects on particular types of farm households (potentially strengthens internal validity and increases external validity of findings);

E.g. semi-structured interviews and focus group conversations could tell us more about the nature of effects in terms of production, environment, poverty, etc. (potentially enhances construct validity of findings).

5.3. Sustainability analysis

An impact theory provides a useful basis for developing an argumentation on the sustainability of changes brought about by an intervention. As the long-term results and the sustainability of the results that are visible at the moment of the outcome and impact evaluation are unknown and cannot be observed, evaluators will need to assess sustainability in an indirect manner. Outcome and impact evaluations can focus on other results that are observable in the short term, such as the institutionalization of practices and the development of organizational capacity, which are likely to contribute to the sustainability of outcomes and impacts for participants and communities in the longer term.

Mog (2004) describes different dimensions of sustainability relevant to projects such as the RISEMP project. The GEF Evaluation Office has developed a framework of five dimensions of sustainability that should be considered in outcome and impact evaluation (GEF, 2009b; see also GEF, 2009a). For each dimension we provide a brief example from the perspective of the RISEMP project. These examples should not be regarded as comprehensive and conclusive results on sustainability. They merely serve the purpose of illustration.

Socioeconomic sustainability: The extent to which project activities lead to long term improvements in the social and economic situation where the project is found and where such changes are essential to ensure improved environmental management. In case of the RISEMP project, researchers at CATIE found that several of the promoted silvopastoral practices such as improved pastures with trees and fodder banks are privately profitable in the medium term. It is very likely that farmers will maintain and even increase those practices that are privately profitable.

Programmatic Sustainability: The extent to which the actions that are taken during the life of the project continue after the formal project ends. At pilot site level

the provision of PES and TA ceased by the end of the project. In different guises, the three local implementing organizations continue to support (a part) of the farmers that participated in the RISEMP project with such activities as environmental education, credits for sustainable land use practices and technical assistance on topics related to silvopastoral land use.

Institutional Sustainability: The extent to which necessary institutional structures are in place and secure for the long term as a result of the project. Institutional structures such as PES administration mechanisms and research and monitoring units ceased to exist after project closure. No new institutions were created in the pilot regions. However, an important result of the project has been the strengthening of capacities in the three implementing organizations: CATIE, CIPAV and Nitlapán.

Financial Sustainability: The extent to which post-project activities can sustain themselves financially or mechanisms are in place to provide a constant flow of external financial resources. The project did not establish links with markets for environmental services such as they exist in some countries at local, national, or international level. Examples are local markets for hydrological services or international mechanisms such as the clean development mechanism (carbon sequestration).

Replication: The extent to which successful implementation of actions in one project can be repeated in other project sites. Replication at pilot site level is fairly limited. Replication (replicatory effects) at other levels has been very successful. The project has generated many useful lessons for replication at other sites. Currently, an upscaled new version of the RISEMP project is under preparation in Colombia. The ESI, improved and validated during the course of the project, is currently being used in several new projects in different countries. Research results on the relationships between silvopastoral land uses and environmental benefits have been widely published.

6. Conclusions and recommendations

6.1. Conclusions

The Nicaraguan case shows how an experimental design that is implemented without the necessary knowledge and institutional support in the field can lose its utility. It should be emphasized that the problems with the experimental design are essentially strategic and planning failures and not implementation failures as such. Project staff were not trained or in any way prepared to manage an experimental design and could not be expected to deal with the various problems that threaten the validity of the design. The analysis shows that the utility of the experimental design in terms of resolving the attribution problem is heavily compromised by several threats to validity.

In all, the experimental framework failed on two of the three group comparisons that were to support rigorous claims on the effects of PES and technical assistance on land use change and corresponding environmental effects. The ‘PES only’ ver-

sus CG comparison is rendered invalid due to severe problems of selection bias and unintended behavioral responses (especially in the CG). The 'PES only' versus 'PES TA' comparison is rendered invalid due to problems of treatment diffusion. The 'PES 2 years' versus 'PES 4 years' comparison is quite valid. The data and their subsequent interpretation illustrate the utility of the experimental design in terms of providing reliable evidence on land use behavior under different types of incentives.

The fundamental question of the cost-benefit ratio of using an experimental design should be raised. Implementing such a design involves substantial costs:

- implementation costs: designing the experiment, selecting the farmers, managing and controlling the quality of the experiment, etc.
- costs in terms of facing ethical dilemmas or possible resistance from farmers or other stakeholders;
- foregone benefits to farmers (withholding benefits to certain groups of farmers, less outreach than without an experimental approach).

These costs can only be justified if the experiment is done carefully, thereby delivering its analytical potential. In the Nicaraguan case (and possibly the other two sites), the costs of implementing the experiment, without the necessary quality control and supervision clearly outweighed the analytical benefits of doing an experiment.

6.2. Recommendations

Despite the limited utility of the experimental design in Nicaragua and potential unidentified problems of the design in the other two countries, the logic of experimentation potentially provides a powerful tool to test the effectiveness of particular incentives on outcomes and impacts, controlling for other factors. Experiments can be especially useful in the following cases:

- when knowledge on attribution (and effectiveness) is important; for example in the case of innovative instruments when little is known about their effectiveness; in case there is a lot of existing evidence about the effectiveness of a particular approach or instrument then the benefits of an experimental design might not outweigh the costs;
- when there is an interest in the magnitude of effects (caused by the project).

However, they should only be applied:

- if sufficient attention and resources are dedicated to training and quality control of the experimental design in practice;
- if attention is paid to possible combinations of experimental approaches with other methods, which would reinforce each other and together would allow for a more comprehensive coverage of the outcome and impact dimensions of an intervention (as well as address more adequately questions of both average effects attributable to the intervention as well as heterogeneity in effects).

More specifically, we propose the following recommendations for future implementation of experimental designs in similar projects:

- Implementing an experimental design outside a laboratory in complex social environments requires a clear protocol and a shared vision among those actors involved in implementing the design on what the design is about and how it should be managed in practice. The different threats to the validity of the design should be considered before implementation.
- An experimental design (based on randomization) fundamentally affects the way an intervention is implemented. Therefore the Project Appraisal Document (PAD) should provide sufficient information on:
    - what the central hypotheses are to be tested with the experiment as well as other knowledge that is expected to be gained from the experiment;
    - the basic characteristics of the experiment;
    - how the experiment will be implemented;
    - the likely threats to validity and how they will be addressed;
    - a budget for staff training as well as for the incremental costs of managing the experiment;
    - an assessment weighing the costs against the benefits of experimentation.
- Provide training to project staff on management and quality control of experimental designs.
- Select a control group in a different region than participants/beneficiaries in order to avoid treatment diffusion; this avoids treatment diffusion effects, possible resentment and other unintended behavioral responses.
- In the communication to participants/beneficiaries try to avoid using references to the experimental design, as the idea of being part of an experiment might trigger a range of unintended responses. This might become easier if a control group is selected and monitored in a different (but similar) region.
- At national level, try to coordinate efforts by different institutional actors planning outcome and impact evaluations within (and possibly beyond) the agricultural and environmental sector. This might have several advantages:
    - Outcome and impact evaluation activities might be pooled and budgets and capacities shared, improving the prospects of high-quality evaluations;
    - Periodic surveys on target groups might incorporate variables relevant to particular interventions. As a result, individual outcome and impact evaluations can make better use of existing surveys, which is especially relevant in the case of quasi-experiments and regression-based quantitative outcome and impact evaluations.

References

Baker, J.L. (2000) Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners, The World Bank, Washington D.C.

Bamberger, M. (2006), Conducting quality impact evaluations under budget, time and data constraints, The World Bank, Washington D.C.

Bamberger, M. and H. White (2007) "Using strong evaluation designs in developing countries: experience and challenges", Journal of Multidisciplinary Evaluation, 4(8), 58-73.

Cook, T.D. and D.T. Campbell (1979) Quasi-experimentation: design and analysis issues for field settings, Houghton Mifflin, Boston.

GEF (2007a) GEF Annual Impact Report 2007, GEF Evaluation Office, Washington D.C.

GEF (2007b) Protected Areas and Avoided Deforestation: A Statistical Evaluation, prepared by K. Andam, P. Ferraro, A. Pfaff, and G. Sanchez-Azofeifa, GEF Evaluation Office, Washington D.C.

GEF (2009a) Review of Outcomes to Impacts – Guidelines and Procedures, prepared by the GEF EO in collaboration with the Conservation Development Center, GEF Evaluation Office, Washington D.C.

GEF (2009b) GEF Annual Impact Report website, http://www.thegef.org/gefevaluation.aspx?id=22444 [last consulted, June 17, 2009].

Ibrahim, M., M. Chacón, C. Cuartas, J. Naranjo, G. Ponce, P. Vega, F. Casasola and J. Rojas (2007) "Almacenamiento de carbono en el suelo y la biomasa arbórea en sistemas de usos de la tierra en paisajes ganaderos de Colombia, Costa Rica y Nicaragua", Agroforestería en las Américas, 45, CATIE, Turrialba.

Leeuw, F.L. (2003) "Reconstructing program theories: Methods available and problems to be solved", American Journal of Evaluation, 24(1), 5-20.

Mog, J. M. (2004) "Struggling with sustainability: A comparative framework for evaluating sustainable development programs", World Development, 32(12), 2139–2160.

Morgan, S.L. and C. Winship (2007) Counterfactuals and causal inference – methods and principles for social research, Cambridge University Press, Cambridge.

OECD-DAC (2002) Glossary of Key Terms in Evaluation and Results Based Management, OECD-DAC, Paris.

Pagiola, S., P. Agostini, J. Gobbi, C. De Haan, M. Ibrahim, E. Murgueitio, E. Ram'rez, M. Rosales and J.P. Ruíz (2004) "Paying for Biodiversity Conservation Services in Agricultural Landscapes", Environment Department Paper, 96, World Bank, Washington D.C.

Pagiola, S., E. Ramírez, J. Gobbi, C. De Haan, M. Ibrahim, E. Murgueitio, and. J.P. Ruíz (2007) "Paying for the environmental services of silvopastoral practices in Nicaragua", Ecological Economics, 64(2), 374-385.

Rogers P.J., T.A. Hacsi, A. Petrosino and T.A. Huebner (Eds.) (2000) Program theory in evaluation: Challenges and opportunities, New Directions for Evaluation, 87, San Francisco, Jossey-Bass.

Rossi, P.H., M.W. Lipsey and H.E. Freeman, (2004) Evaluation: A systematic approach, Thousand Oaks, Sage Publications.

Shadish, W.R., T.D. Cook, and D.T. Campbell (2002) Experimental and quasi- experimental designs for generalized causal inference, Boston, Houghton Mifflin Company.

Weiss, C.H. (1997) "Theory-based evaluation: Past, present and future", in D.J. Rog and D. Fournier (Eds.), Progress and future directions in evaluation: Perspectives on theory, practice and methods, New Directions for Evaluation, 76, San Francisco, Jossey-Bass.

White, H. (2006) Impact evaluation – The experience of the independent evaluation group of the World Bank, Independent Evaluation Group, World Bank, Washington D.C.

Wunder, S., S. Engel and S. Pagiola (2008) "Taking stock: A comparative analysis of payments for environmental services programs in developed and developing countries", Ecological Economics, 65(4), 834-852.

Annexes

Annex 1. Summary logical framework RISEMP

Project Development Objectives	**Outcome / Impact Indicators**	**Outputs**	**Output Indicators**
To obtain local environmental benefits through reduction in erosion and improvement in soil and water quality while at the same time increasing production, income and employment in rural areas. To provide global environmental benefits, through improved biodiversity and carbon sequestration services. To gain initial experience in the management of incentives required to produce global environmental benefits. To develop recommendations for sector and environmental policies in terms of land use, environmental services and socio-economic development provided by the introduction of silvopastoral systems to rehabilitate degraded pastures.	Sustainable silvopastoral systems established in three Latin American countries and improved water quality in six watersheds in Latin America. Improved habitat for diverse types of biodiversity provided and stable carbon sequestered in the soil and in commercial wood under silvopastoral systems in six watersheds in three countries. Improved resource monitoring methodologies developed for measuring carbon sequestration , biodiversity conservation developed and sustainable funding mechanism established which provide appropriate incentives to induce farmers to provide global environmental benefits. Increased awareness of the potential in environmental services provided by integrated ecosystem management and experience gained for future development of the integrated ecosystem management approaches to restore degraded pasture. Guidance for future	Increase in area with improved ecosystems functioning of 12,000 ha. currently degraded pasture land, as demonstrated by specific indicators for soil and water quality and biodiversity The increase in numbers of livestock producers, community leaders, and policy decision makers at the local, regional and national level that are familiar with the ecological and economic benefits of more intensive silvopastoral systems in livestock production. The extent of dissemination of improved monitoring methodologies developed for measuring carbon sequestration, biodiversity conservation, water quality in watersheds and socio-economic aspects. Understanding of farmer reactions to incentive systems for global environmental benefits obtained. The availability of policy guidelines to promote silvopastoral systems and	1.1.1 About 4000 ha silvopastoral systems, established, improving the eco-system in at least 12,000 ha to demonstrate the benefits of silvopastures for carbon sequestration and biodiversity in three countries. 1.1.2 Increased biodiversity conservation (at least 50 bird species/production system): 1.1.3 Increased carbon sequestration (about 25,000 ton carbon sequestered per year). 1.1.4 Increased water quality in watersheds (reduction on Biochemical Oxygen Demand (BOD) and suspended total solids (mg/l). 1.1.5 Increased socio-economic impact: Farm income to increase by 10 percent per year. 1.2.1 Local stakeholders trained in 3 countries (about 30,000 farmer days of training over 5 year period). 1.2.2 Local organization's capacity strengthened (20 organizations in 3 countries). 2.1 Methodologies to assess biodiversity, carbon sequestration, water quality on farm, watershed and community level and socio economic impact developed and tested. 2.2 Monitoring systems for biodiversity conservation, carbon sequestration, water quality using biological indicators and socio-economic impact established (monitoring systems in 3 countries). 3.1 Eco-Services payment systems implemented in each of the target countries. 3.2 Certification of ecological services conferred (results of monitoring analyzed at farm and landscape level, and environmental

	funding, lessons for replication/best practice, and policy requirements for environmental services in livestock production defined.	establish sustainable benefits sharing mechanisms related to global and local environmental services provided by integrated ecosystem management.	services paid to the farmers). 3.3 Farmers and community reaction to environmental services incentives and change of attitude and perception to local and global environment measured (measured by changes on land use, in particular in area set aside for forest regeneration). 4.1 Socioeconomic data available on key factors affecting farmer adoption of silvopastoral systems. 4.2 Alternative sources of funding for payment for eco-services, and alternative measures to promote silvopastoral systems identified and secured. 4.3. Specific recommendations for best ranching practices and land use that improve habitat heterogeneity to sustain higher biodiversity, and increase ranch yield disseminated among minimum 1200 farmers 12 NGO's and/or community-based groups, policy-makers and regional networks.

Source: PAD

Annex 2. Environmental Services Index (ESI) used by the RISEMP project

Land use	Biodiversity index	Carbon index	Total index
1 Crops (annual, grains and tubers)	0	0	0
2 Degraded Pasture	0	0	0
3 Natural Pasture without Trees	0,1	0,1	0,2
4 Improved Pasture without Trees	0,1	0,4	0,5
5 Semi-Permanent Crops	0,3	0,2	0,5
6 Natural Pasture + Low Tree Density (<30/ha)	0,3	0,3	0,6
7 Natural Pasture enriched with low tree density	0,3	0,3	0,6
8 Living Fences with new trees	0,3	0,3	0,6
9 Improved Pasture Low tree density	0,3	0,4	0,7
10 Fruit Crops (Monocrop)	0,3	0,4	0,7
11 Graminous Fodder Banks	0,3	0,5	0,8
12 Improved Pasture Low Tree Density	0,3	0,6	0,9
13 Ligneous Fodder Banks	0,4	0,5	0,9
14 Natural Pasture High Tree Density	0,5	0,5	1
15 Fruit Crops (Diverse)	0,6	0,5	1,1
16 Multistrata living fences	0,6	0,5	1,1
17 Diversified Fodder Banks	0,6	0,6	1,2
18 Commercial Tree Plantations (Monocultivation)	0,4	0,8	1,2
19 Shaded Coffee	0,6	0,7	1,3
20 Improved Pasture with High Tree Density	0,6	0,7	1,3
21 Guadua (bamboo) forest	0,5	0,8	1,3
22 Diversified Commercial Tree Plantations	0,7	0,7	1,4
23 Shrub habitats (tacotal)	0,6	0,8	1,4
24 Riparian Forest	0,8	0,7	1,5
25 Intensive Silvopastoral Areas	0,6	1	1,6
26 Secondary Forest (intervened)	0,8	0,9	1,7
27 Secondary Forest	0,9	1	1,9
28 Primary Forest	1	1	2

Source: Based on RISEMP data

Annex 3. Persons interviewed

Nicaragua (March, April and July, 2008)

Farmers		
Name	Community	Group
Absalon Guerrero	Paiwita	PSA+AT (4)
Agusto Robles	Limas Abajo	PSA+AT (4)
Albertina Jarquín	Patastule/El Gavilan	PSA+AT (4)
Alberto Saravia	San Ignacio	PSA+AT (4)
Angela Alvarado	San Ignacio	PSA+AT (4)
Carlos Urbina Luna	Limas Arriba	PSA+AT (4)
Eusebio Mendoza Dias	San Ignacio	PSA+AT (4)
Guillermo Garcia Polanco	Patastule	PSA+AT (4)
José Andrés Amador Martinez	San Ignacio	PSA (4)
José Rolando Castillo Ramirez	Patastule	PSA+AT (4)
Juan José Jarquín Jarquín	Limas Abajo	PSA (4)
Julia Gadea Amador (Jaime Robles)	Limas Abajo	PSA+AT (2)
Julio Gutierrez Obando	San Ignacio	CONTROL
Kairo Torres	Patastule	/
Orlando Urbina	El Gavilan	/
Pastor Flores Rodriguez	Paiwita	CONTROL
Pedro Reyes Urbina	Patastule	CONTROL
Pedro Talavera Valle	Paiwita	PSA+AT (4)
Hector René Zeledon Alvarado	Patastule	CONTROL
Richard José Robles Ortega	Limas Abajo	PSA (4)
Roberto Urbina	El Gavilan	PSA+AT (4)
Rosario Ramirez García	El Gavilan	PSA (4)
Santos Genaro Sevilla Suarez	El Gavilan	PSA+AT (4)
Severino Vega Martinez (Fermin del Socorro Vega Vega)	Limas Abajo	PSA+AT (4)
Tomas Castro Torres	Limas Abajo	CONTROL
Tomas Soza Morales	Limas Abajo	PSA+AT (2)
Trinidad Lanzas (Ramona del Socorro Garcia)	Limas Arriba	PSA+AT (4)
Victorina Ortega Mondoy	Limas Abajo	/
Zoyla Martinez Rubio	Patastule/El Gavilan	CONTROL

Project staff	
Alfredo Arguëllo	Project staff, Nitlapan
Omar Davila	Project staff, Nitlapan
Yuri Marin	Socio-economic analyst, Nitlapan
Guillermo Ponce	Carbon sequestration analyst, Nitlapan
Elias Ramirez	Coordinator Nicaraguan pilot site, Nitlapan
Bismark Reyes	Water analyst, Nitlapan

Costa Rica (July, 2008)

Project staff and other stakeholders	
Oliver Bach	Rainforest Alliance
Francisco Casasola	Coordinator Costa Rican pilot site, CATIE
Omar Davila	Project staff, Nitlapan
Leonardo Guerra	Consultant carbon sequestration
Muhammad Ibrahim	Project leader, CATIE
Jose Ney Rios	Water analyst, CATIE
Joel Saenz	Biodiversity specialist
Claudia Sepulveda	Project staff, CATIE
Diego Tobar	Project staff, CATIE
Cristobal Villanueva	Socio-economic analyst, CATIE

World Bank

Gunars Platais	Environmental economist
Cees De Haan (2007)	Silvopastoral specialist – team RISEMP
Stefano Pagiola (2007)	Environmental economist – team RISEMP

Annex 4. RISEMP project experimental design (situation 2006)

Group	Colombia	Nicaragua	Costa Rica	Total
Group A	29	27	28	84
Group B	50	75	69	194
Group C	25	29	27	81
Total	104	131	124	359
According to type of payment scheme*				
Payment Scheme 1 (4 years)	36	75	50	151
Payment Scheme 2 (2 years)	39	39	46	124
Total	75	104	96	275

Group A = control (without payments nor technical assistance); Group B = (payments and technical assistance); Group C = only payments.

* Only groups B and C.

Source: progress report RISEMP, 2006

Annex 5. Comparing the three sites: a broad impression

The other two cases were not analyzed in detail in this study. We restrict ourselves to some broad impressions about the quality and implementation of the design.

On the basis of interviews with staff from CATIE and looking at the data, our impression is that the experimental design in the Costa Rican site has been the least affected by the different threats to validity. The Nicaraguan case has been the most problematic one. The Colombian design can be positioned in between the others.
Notwithstanding the design-related problems in Nicaragua, the longstanding relationship between Nitlapán and many of the farmers had a positive effect on data quality. In contrast, data quality has been particularly problematic in Colombia, especially in the control group. One of the reasons for this has been the history of conflict and drug trafficking in the country, fueling an atmosphere of distrust between farmers and institutions. In terms of data quality, Costa Rica would be the intermediate case, with the Nicaraguan and the Colombian sites respectively having had the least and the most problems with data quality.

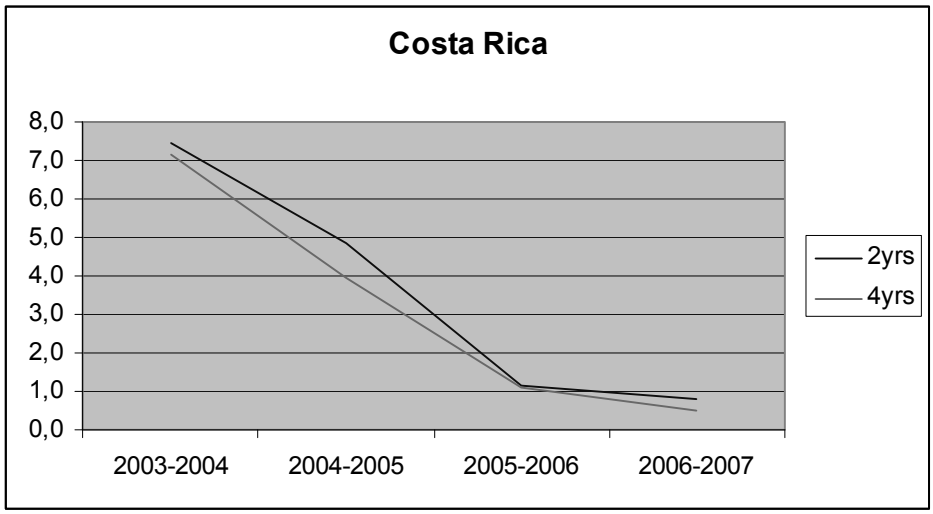
As for selection bias, the general impression is that randomization principles were more successfully implemented in both Costa Rica and Colombia, reducing the probability of selection bias.

With respect to contagion, in Costa Rica geographical proximity between farmers has probably facilitated processes of farmer to farmer diffusion. In Colombia, there was a clearer geographical separation between PES and CG farmers, resulting probably in less farmer to farmer diffusion (contagion). In addition, apart from CIPAV no other institutions in the region were providing technical assistance on silvopastoral practices or related issues. In the Costa Rican site there were a few other institutional actors, but not as many as in the Nicaraguan case.

To sum up, group comparisons for the Costa Rican and Colombian case are likely to provide a clearer picture of the effect of project incentives on adoption behavior of silvopastoral practices than in the Nicaraguan case. However, given the evidence on design implementation problems in the Nicaraguan case, using the experimental design for group comparisons in the other two countries without further insight into possible design implementation issues would seriously undermine the credibility of findings.

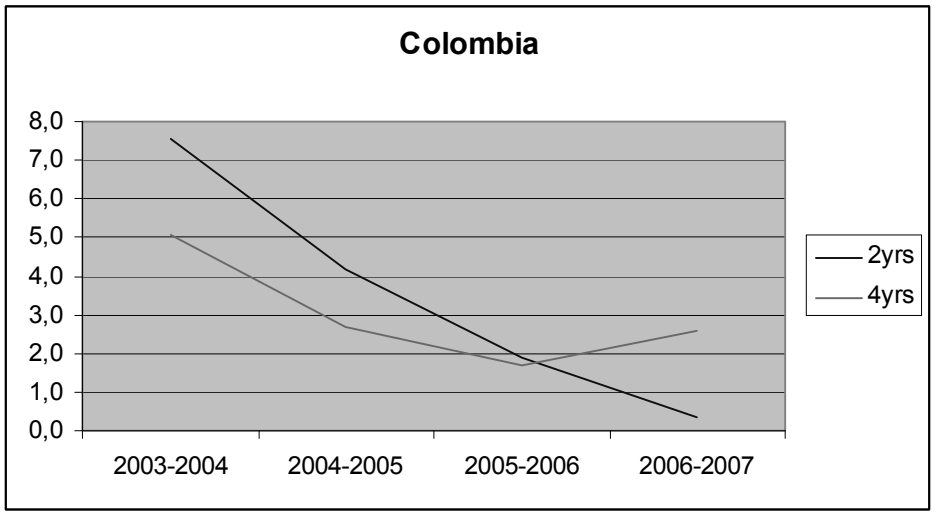
Annex 6. Group comparisons between farmers receiving 2 years versus 4 years of PES, Costa Rica and Colombia

Figure A6.1. Costa Rica



Source: own calculations based on RISEMP project data, January 2008

Figure A6.2. Colombia



Source: own calculations based on RISEMP project data, January 2008

CHAPTER 5

Vaessen, J. and J. De Groot (2004) “Evaluating training projects on low external input agriculture: lessons from Guatemala”, *Agricultural Research and Extension Network Papers*, 139, Overseas Development Institute.

odi
Agricultural Research & Extension Network

AgREN

Network Paper No.139

July 2004

EVALUATING TRAINING PROJECTS ON LOW EXTERNAL INPUT AGRICULTURE: LESSONS FROM GUATEMALA

Jos Vaessen and Jan de Groot

Abstract

Despite the popularity of the promotional activities of low external input agriculture (LEIA), systematic evaluations of the impact of these activities are scarce. This paper discusses an impact evaluation study of a training project on LEIA in Guatemala. The evaluation design is based on a simple quasi-experimental design and complemented by qualitative methods of data collection. The paper illustrates the utility of this kind of mixed-method evaluation for studying the outcome and impact of small-scale development interventions given their specific constraints of money, expertise and time. In addition, a number of specific lessons regarding the role of the evaluation study, the training project and the adoption of LEIA practices are highlighted.

Research findings

- *The study shows how a basic quasi-experimental design, complemented by qualitative research methods, and without relying on sophisticated statistical techniques, can be very useful to determine the outcome and impact of a training project at the farm level, controlling for the influence of external variables. This type of mixed evaluation method would be quite adequate for the evaluation of similar relatively small-scale interventions.*
- *Farmers' adoption behaviour after the termination of the project can be characterised as selective and partial. Given the particular circumstances of small farmers (e.g. risk aversion, high opportunity costs of labour) it is not realistic to assume that a training project as described in the paper will bring about a complete transformation from a conventional farming system to a LEIA farming system.*
- *In line with the literature, the most popular practices (in this case for example organic fertilisers, medicinal plants) are those that offer a clear short-term return while not requiring significant investments in terms of labour or capital.*

Policy implications

- *The lessons produced by the baseline study could not be used to make mid-course corrections in the design and implementation of the project, with negative consequences for its eventual outcome and impact. It is suggested that a formal appraisal study be carried out to identify potential constraints before selecting the implementing organisation and defining the terms of reference of the project.*
- *Project outreach was concentrated in a limited number of communities and social networks connected to project extensionists who were themselves farmers in the region. When working with farmer-to-farmer extension and education models, close attention should be paid to possible biases in beneficiary selection and indirect effects on local power dynamics. If necessary, corrective action to ensure a broader and more equitable outreach should be taken.*
- *An ideological faith in the absolute supremacy of LEIA practices is not in the best interests of the farmer. Projects promoting LEIA should focus on the complementary effects of LEIA practices and conventional farming techniques, encouraging each farmer to choose the best balance for his/her needs.*

Contact details

Jos Vaessen is a researcher and PhD student at the Institute of Development Policy and Management University of Antwerp, Venusstraat 35 2000, Antwerp, BELGIUM Tel. 32 3 220 4973 Fax 32 3 220 4481, *Email*: jos.vaessen@ua.ac.be

Jan de Groot is a retired Senior Research Fellow of the Department of Development Economics of the Free University of Amsterdam. Until recently, he has been working as European co-director in an EU-financed rural development programme in Totonicapán, Guatemala.

Acknowledgements

This paper is based on the reports of an evaluation study commissioned by the European Union and carried out by Halcrow Rural Management Ltd. We would like to thank both institutions for their permission to use these reports. The content of this article is the exclusive responsibility of the authors and does not commit the European Union or Halcrow Rural Management Ltd. to the authors' views

Acronyms

IRDP	Integrated Rural Development Programme
LEIA	Low External Input Agriculture
NGO	Non-Governmental Organisation

ISBN 0 85003 723 9

EVALUATING TRAINING PROJECTS ON LOW EXTERNAL INPUT AGRICULTURE: LESSONS FROM GUATEMALA

1 INTRODUCTION

Development interventions promoting low external input agriculture (LEIA) have become increasingly popular in order to tackle resource degradation and poverty in agricultural communities all over the developing world. However, as stated recently, '[t]he effectiveness and impacts of these approaches have been subject of debate' (De Jager et al., 2004: 206). Systematic evaluations of the impact of the promotional activities of LEIA practices are scarce. In practice, many of the small- and medium-scale initiatives run by non-governmental organisations (NGOs), community-based organisations, and local and regional governments face a number of constraints in terms of a lack of expertise and the financial means to carry out studies to evaluate the socio-economic and ecological effects of LEIA.

In this paper we discuss an evaluation study of a training project in LEIA in the Department of Totonicapán, Guatemala. This project was carried out within the framework of an integrated rural development programme (IRDP) implemented by the European Union (EU) in cooperation with the Guatemalan government. The main objective of the study was to assess the outcome and impact of the project by showing the presence or absence of plausible effects of the project on participants and an indication of the magnitude of these effects. Based on the results of this study IRDP would decide whether or not to extend financial support to ORGANIC[1], the implementing organisation of the project. As ORGANIC has also worked on other EU-financed rural development projects the evaluation had a wider relevance than the Totonicapán project.

The primary focus of this paper is to describe and explain the advantages and disadvantages of the particular evaluation methodology applied in this study. The methodology comprises elements of both quantitative and qualitative research. We want to illustrate that this type of mixed method evaluation (Greene and Caracelli, 1997) is very useful for studying the outcome and impact of small-scale development interventions given their specific constraints of money, expertise and time. The basis of the evaluation methodology is a simple quasi-experiment. Quasi-experiments are research designs that involve comparisons between groups affected by a certain intervention and control groups. Participation in either category is not random. Specific statistical adjustments can be made in order to make the two types of groups equivalent in terms of outcome- and impact-related variables (Cook and Campbell, 1979). The quasi-experimental data are complemented by qualitative methods of research (e.g. field visits, semi-structured stakeholder interviews) to allow for triangulation and a richer interpretation of the quantitative data.

As part of the evaluation methodology, a baseline (ex ante) study was carried out to map the situation of the participants at the beginning of the project and to identify potential constraints of the training project. The paper will show how the timing of the study and the deficient relationship between IRDP and ORGANIC reduced the policy impact of the ex ante study in terms of improving the design and implementation of the project. In fact, some of the disappointing outcomes and low impact of the project had already been anticipated in the ex ante study, but had been largely ignored by ORGANIC at that time. This evidently raises the question of how to ameliorate the connection between evaluation and improving practice, requiring a reflection on the role of evaluation in this type of project.

The paper starts with a brief description of the characteristics of the region in which the project was implemented, followed by an outline of the training project itself. The subsequent section deals with the issue of adopting LEIA practices and briefly discusses the main factors influencing the decision-making process of small farmers to do so. This section is followed by a comprehensive treatment of the methodology employed in the evaluation study. To assess the utility of the evaluation methodology in terms of analysing outcome and impact we illustrate a number of results. The paper concludes with a discussion of lessons learned regarding the applied evaluation methodology and the role of the evaluation study.

2 CONTEXT

The region

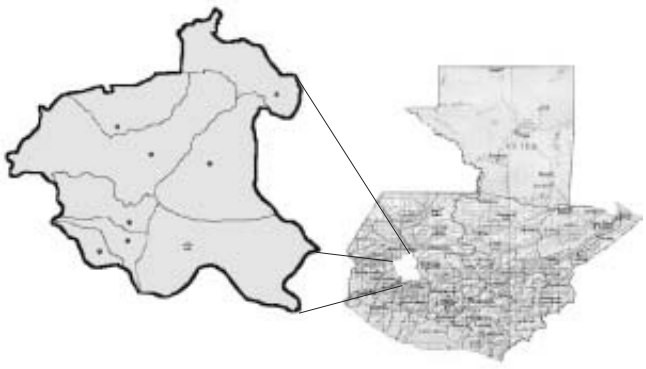
The Department of Totonicapán is situated in the western highlands of Guatemala. It is one of the smallest provinces in the country, consisting of eight municipalities. The population is predominantly indigenous (Maya-Quiché). The training project was aimed at small farmers in the four northern municipalities of the province (Santa Lucía La Reforma, San Bartolo de Aguas Calientes, Momostenango and Santa Maria Chiquimula[2]) where agriculture constitutes an important subsistence and income activity. Agricultural activities are complemented by forestry and non-agricultural activities such as weaving, tailoring, pottery and commerce. In the southern municipalities the situation is the opposite: non-

agricultural activities predominate, with agriculture of secondary importance as an income-generating activity.

Limited farm sizes, mountainous areas, mediocre soils, relative isolation and limited access to markets are important factors making agricultural production in the northern municipalities less attractive. Altitudes in the province vary between 3500 and 1700 metres, decreasing roughly from south to north. These differences in altitude imply significant differences in climatic circumstances. In the north, micro climates are more suited to subtropical crops and fruit trees, the most important crops being maize and beans. Crop diversification is limited. One of the main reasons for this is the fact that forced labour systems for coffee cultivation in the 19th and part of the 20th century extracted so much labour from the indigenous highland communities that traditional diversified Mayan crop and livestock systems degenerated into a system of mono-cropping of maize and beans (McCreery, 1994; Carmack, 1995). Some horticultural crops (tomato, peppers) are also cultivated. Among the most important fruit trees to be found are avocado, peach and citrus, a substantial part of the fruit harvest being sold in local markets. Livestock production in the region is mainly limited to pigs, sheep and chickens. These are kept chiefly for subsistence purposes with the exception of sheep, which are kept mainly to produce wool for clothes.

From the beginning of the 1960s until the mid-1990s Guatemala suffered from internal conflict, leading to

Figure 1 The province of Totonicapán



numerous victims, in particular among the rural indigenous population. This applied especially to the municipality of Santa Lucía La Reforma in the project area where, among other things, it destroyed existing local organisations. One of the consequences of the conflict was a further deepening of distrust by the indigenous population of external organisations which, until 1998, were mostly governmental. In 1998, the agricultural and livestock services of the government were dismantled. As was often the case in other countries as well, the basic consideration was that these services should be financed but not directly implemented by the state. Sadly, after abolishing the official extension services, government funds for outreach activities implemented by NGOs never became available. Nowadays, NGOs partly fill the gap with their resources but cannot maintain the same level of outreach in quantitative and qualitative terms as their governmental predecessors. Moreover, outreach is less comprehensive than before as the NGOs normally only work with existing local organisations and groups.

The training project

In 1996, an integrated rural development programme (IRDP) financed by the EU and co-implemented by the Guatemalan government was established in the province of Totonicapán in western Guatemala. The programme comprised several components, including support for agriculture, basic infrastructure and small enterprise development. In 1998, in order to support small-scale agriculture in the relatively isolated northern municipalities of Totonicapán, IRDP decided to finance a training project in LEIA.

The main features of the project that was to be implemented by ORGANIC were:

- The project was to cover a period of three years (1998–2000).
- A total of 18 courses, each consisting of two to three days of practical training, would be imparted to participating farmers in the region. These would take place at an experimental farm run by ORGANIC.
- After each course, the participants would be given 'homework' and ORGANIC extensionists would provide follow-up at the farm level.
- The teachers of the courses, who were also to act as extensionists in the field, would be, like the participants, Mayan farmers and graduates of former courses by ORGANIC. There would be only one teacher, the proposed coordinator of the project, with a formal technical degree in agronomy.
- The methodology implemented by ORGANIC (teaching and follow-up) could be characterised as a form of farmer-to-farmer extension.[3]

ORGANIC's training project offered a wide range of practices and technologies adapted to the history and culture of the Mayan population in the region. Broadly, the courses and follow-up by ORGANIC were centred around the following themes:

- soil conservation measures (e.g. barriers, ridges);
- cultivation practices (e.g. refraining from burning crop residues, contour ploughing, zero tillage);
- organic fertilisers (e.g. manure, leaves, crop residues);
- organic pesticides (e.g. onion, human urine);
- crop diversification (e.g. mixed cropping, nitrogen fixation with legumes, herbs, fruits);
- farm infrastructure (e.g. traditional ovens, special latrines for processing human manure, corrals);
- family nutrition (e.g. food preparation, composing healthy diets);
- rural organisation (e.g. group building, diffusion of knowledge to neighbouring farmers).

The courses and practices were based on the premise of a more efficient and integrated use of existing resources on the farm. Many of the proposed practices, as indicated, have been in existence in Mayan production systems for centuries but have withered over the course of time in many areas. In this sense, the project performed the role of catalyst, collecting bits of local knowledge and practices in one region and imparting them elsewhere. The peasant-to-peasant extension model is especially useful in this regard because of the tight links with local farming systems.

Over the period of three years, it was anticipated that the participants would gradually abandon conventional farming practices and have moved towards a reliance on LEIA practices by the end of 2000. The principal aim of the training project was to have achieved by its end 120 'transformed' LEIA farms. In addition, participants would be trained to become teachers in their communities. It was contemplated that each graduate would teach at least one or more practices to 10 neighbouring farmers. Moreover, graduates were expected to organise themselves into local groups which would form the basis for learning processes among graduates and neighbouring farmers.

It was assumed that the transformation from conventional farming to LEIA would lead to the following beneficial effects by the year 2000:

- a higher percentage of the harvest being sold in the market;
- higher yields (especially in maize, beans and potato);
- better soils (higher percentage of organic matter);
- improved managerial and organisational capabilities among participants, hence empowerment of the participants and their families;
- higher farm income;
- improved nutrition and health status of the participant and his/her family.

The potential outcome and impact of the training project presented by ORGANIC was rather over-ambitious, which apart from a certain marketing zeal for their services, can be explained by an inadequate attention paid to adoption processes.

3 ADOPTION OF LEIA PRACTICES

The adoption of 'sustainable' farming practices continues to be a popular topic for research and debate among practitioners and researchers (e.g. Neill and Lee, 2001; Moser and Barrett, 2003; De Jager et al., 2004). Sustainable agriculture still remains a somewhat confusing and fuzzy concept. However, as argued by Pretty, what is important is not the exact definition, but clarifying 'what is being sustained, for how long,

for whose benefit and at whose cost, over what area and measured by what criteria' (Pretty, 1995: 11). In our evaluation study this interpretation became an important guideline in our efforts to systematise the effects of the project in terms of outcome and impact. LEIA can be regarded as a form of sustainable agriculture. At the farm level it refers to an integral use of a wide range of technologies and practices that can be characterised by a low use of external resources, local regeneration and reproduction, and an intensive use of local knowledge. Sustainable agriculture, and more specifically LEIA, includes aspects such as integral pest and disease management, local nutrient management and soil and water conservation (ibid.).

While different household and farm characteristics have been identified in relation to explaining adoption behaviour (e.g. Feder et al., 1985; Pomp, 1994), the evidence is mixed. Factors such as motivation (e.g. Pannell, 1998) and perceived profitability (e.g. Cary and Wilkinson, 1997) of the practices are important determinants of adoption behaviour which are often not highly correlated with household and farm characteristics. Indeed, as argued recently by Jones (2002), many studies that approach the topic with a checklist of deterministic factors to explain adoption processes fall short of explaining the nature of the process of adoption.

In this paper it is not our aim to arrive at a thorough explanation of the adoption process. We will briefly focus on a few important factors that are expected to influence adoption of LEIA practices in Totonicapán. In general, one can state that the adoption of LEIA practices implies a substitution of knowledge and labour for external inputs (Pretty, 1995). While in the case of ORGANIC the knowledge constraint is addressed by the training project, the labour constraint must be met by the participating farm household. Time availability of the different household members is the essential resource of the farm household in developing countries (Low, 1986). Therefore, the opportunity costs of labour in relationship to the marginal returns to farm labour input is a crucial variable for farm household members in deciding whether to adopt a certain practice (Feder et al., 1985; Stocking and Abel, 1992).

In Totonicapán, opportunity costs of labour are relatively high, given the prevalence of several non-agricultural activities such as weaving and tailoring. In addition, returns to land in the case of the traditional staple crops (maize, beans) are quite low, making agriculture foremost a subsistence activity. While this situation might hamper any investment in agriculture, farmers are willing to invest in new practices that are perceived to offer a return in the short term. Some of the practices imparted by ORGANIC entail a clear return in the short term (e.g. organic fertilisers), whereas others such as (physical) soil conservation practices (e.g. ridges, barriers) require significant labour inputs in the short term while benefits occur in the long term. The perceived unattractive pay-off of the latter, compounded by the short time horizon of small farmers, substantially reduces their willingness to adopt these practices (Lutz et al., 1994).

Besides knowledge and labour as critical inputs for LEIA agriculture, lack of capital may in some cases restrict adoption of certain practices (Pomp, 1994; Ruben and Vaessen, 2000). At first, this might sound somewhat paradoxical since the reduced reliance on external inputs liberates capital that was formerly used for purchases in the market. However, the reduced reliance on purchased inputs does not rule out the possibility of not being able to finance the high initial costs associated with some practices such as the construction of stables or latrines. Offering the opportunity to apply for credit, under the right conditions and selection procedures, might take away the barrier that is keeping willing and motivated farmers from adopting certain practices.

4 METHODOLOGY OF THE EVALUATION STUDY[4]

The main objective of the study was to determine the outcome and impact of the training project, serving two underlying purposes. The principal purpose was to establish if the ORGANIC training project had proven to be a worthwhile investment. In addition, the study would help determine whether or not to extend further funding to ORGANIC after the project was ended. The utility of the evaluation study went beyond Totonicapán as the results were discussed with other EU-funded IRDPs working with similar training projects in the country, many of which were implemented by ORGANIC.

The evaluation study was designed in 1998 in collaboration with IRDP and ORGANIC staff. It was carried out by an external evaluator and a small fieldwork team in close collaboration with IRDP staff. Given the main objective, a simple quasi-experimental design was defined as a basis for measuring the outcome and impact of the project (Cook and Campbell, 1979). However, because of the size of the project (and the corresponding budget constraint for evaluation) and the size of the population to be studied, the implementation of a formal quasi-experimental study with specific matching techniques and sufficient statistical power would be too costly. In addition, the small population size meant that a good qualitative assessment would offset the need for sophisticated statistical analysis. Nevertheless, the basic framework of the quasi-experimental design was deemed essential, since an alternative approach based purely on for example farm visits, stakeholder interviews and secondary data would not sufficiently uncover the heterogeneity in patterns of adoption, the scale of adoption and the subsequent impact on farm households. The final study design could best be characterised as a kind of mixed method evaluation (Greene and Caracelli, 1997), the basis being formed by a simplified quasi-experimental design which would be thoroughly supplemented with information from field visits and semi-structured stakeholder interviews (IRDP and ORGANIC staff, participant and non-participant farmers).

The quasi-experimental design constituted a baseline survey among participating farm households in 1998,

an ex post survey covering the same sample in 2001, and, in the same year, a control group survey consisting of non-participant farm households. Ideally, a control group should be included in the baseline study as well. This would imply finding a stable control group that would be available in 1998 and in 2001. One of the reasons that this was difficult was the historical basic distrust felt by Mayan farmers for formal institutions (like IRDP),[5] which has been exacerbated by the experiences of the civil war. Whereas participants, because of their obvious links with the training project, were more prone to cooperate with the survey, non-participants were more reluctant. It was considered too costly and inefficient to cover a control group in 1998 big enough to leave a sufficiently large number of farmers willing and able to assist in the survey of 2001. In any case, the small size of the total population of participant farmers and the subsequent sample size limited the prospects and rationale for sophisticated statistical analysis while enhancing the scope for additional 'qualitative' methods of data collection and observation. Moreover, the study's objective was not to prove output and impact with a certain level of statistical accuracy but 'merely' to show the presence or absence of plausible relationships between intervention and effect with an approximate indication of magnitude. Hence, a control group in the baseline survey was not considered crucial.[6]

In 1998, 56 farmers were selected at random and interviewed, representing almost 50% of the approximately 120 farmers who volunteered to participate in the project. In 2001, 48 of that initial group could be covered by the ex post survey.[7] In the same year, a control group of 38 farm households with similar characteristics (see Table 1) was established by means of geographical sampling. The distribution of the control group sample over the territory was proportional to the participant distribution over the territory. To avoid contamination of the control group by spill-over effects from the project, each potential member of the control group was asked if he/she had had contact with the ORGANIC project. In addition, farmers were asked if they had made any changes in their production systems as a result of advice from neighbouring farmers possibly related to the ORGANIC project.

Table 1 General comparison of participants and control group

Variable	Participants (n = 48)	Control group (n = 38)
Education participant/ household head (years)	2.4 (2.3)	2.6 (2.8)
Family size	6.4 (3.1)	7.1 (3.2)
Off-farm activity participant/ household head (%)	71%	71%
Artisanal activity participant/ household head (%)	31%	21%
Remittances (%)	23%	18%
Land owned (*manzanas*)	5.1 (9.4)	4.4 (10.3)
Cultivated area (*manzanas*)	0.9 (0.9)	0.7 (0.5)
Organisational membership (%)	71%	61%
Received loan in last 3 years (%)	27%	18%
Received technical assistance in last 3 years (%)	46% [a] *	13% *
Received technical assistance (2001) (%)	6% [a]	13%

[a] Excluding participation in the ORGANIC project.
Note 1: The variables for the participants reflect the year 1998, whereas the control group data are from 2001. The values for the participant group in 2001 are almost identical to those in 1998 with the exception of the variable technical assistance (which is shown in the table for that reason).
Note 2: Variables expressed in percentages are dichotomous variables; the value refers to the percentage responding 'yes'.
Note 3: x (y) represents mean (standard deviation).
Note 4: One *manzana* is approximately 0.7 hectare.
Note 5: * $p < 0.05$, ** $p < 0.01$; depending on the measurement scale t-tests and chi-square tests were applied

Table 1 compares participants and control group farm households with regard to a number of diagnostic variables in order to check for possible differences. A first look at the table reveals that the participant group and the control group are quite similar regarding a number of diagnostic variables that were used to check for similarity. The close similarity was the result of consistent geographical sampling and application of the selection rule. No ex post matching was applied. An important difference between participants and the control group was the use of technical assistance. In 1998, almost half of the participants had been receiving such assistance. By 2001, participation in the ORGANIC project had largely substituted for the old sources of technical assistance. In contrast, the incidence of technical assistance in the control group was much lower. The high proportion of participant farmers receiving technical assistance prior to ORGANIC underlines the importance of the baseline study in recording pre-project adoption rates of several practices that other institutions in the region had already been teaching before 1998.

The difference between the two groups in terms of technical assistance received from institutions other than ORGANIC is partly due to the reduction in governmental extension services since 1998, but in part suggests a certain selection bias (see Mosley, 1997). To clarify, had we measured use of technical assistance for the control group in the year 1998, we would have come up with a higher percentage than in 2001, but still lower than the use of technical assistance by participant farmers in 1998. As suggested by the data and confirmed in farm visits, participating farmers were on average more motivated towards agricultural innovations and had had more experience with other institutions in the past than the control group. In the design no attempt was made to change the control group in order to correct for this bias. Rather, it was noted that this bias would lead to a slight 'over-estimation' of project outcome and impact.[8]

Because of the applied sampling method and the relatively small differences between participants and control group, we assumed that external factors such as market access, climatic conditions and institutional environment were similar between the two groups and therefore did not affect further analysis.

To complement the analysis from the simple quasi-experimental design, other data and information sources were used to assess the impact of the project. The most important elements that contributed to the quality of the surveys as well as constituting additional sources of information for the evaluator were the following: First, IRDP field staff intensively cooperated in the design and implementation of the surveys. This support and the collaboration of ORGANIC staff guaranteed a sound local embeddedness of the study. Moreover, it was easy to conduct interviews and informal talks with staff from both organisations during the study process. Second, the surveys were sufficiently small for the evaluator to be directly involved in all the operational tasks of the survey work (i.e. interviewer training, coordination, quality control, data processing). In this way, the evaluator was able to develop a good understanding of the field while being able to conduct more efficiently a relatively large number of farm visits and farmer interviews parallel to the formal survey.

In order to structure the different causal relationships between the project and the participating farmers we devised the following framework as depicted in Table 2.

Table 2 Main variables to be included in the evaluation framework

Output	Outcome	Impact
Course implementation	Adoption of practices	Soil quality
Course participation	Diffusion of practices	Yields
Content of courses field assistance		% of harvest sold
		Farm income
		Nutritional and health status[a]
		Organisational and managerial capacities

[a] Covered by the study on nutrition

The objective of the evaluation study was to focus on outcome and impact. A brief assessment of project output was incorporated in the project, but the principal tool for evaluating output consisted of an ongoing process of monitoring the courses and field assistance by IRDP field staff during the project implementation period. The study as a whole was best suited to measuring the outcome of the project, given the close link between project output and outcome in terms of adoption and diffusion processes. Moreover, these processes are manifest in the short and medium term, hence were identifiable at the time of the ex post evaluation study. On the other hand, the link between project output and impact is typically more indirect and of a medium- to long-term nature. The ex post study was implemented just months after the end of the project, at a time when its full impact had still to emerge.[9] Nevertheless, impact effects were measured to get an indication of potential impact, especially if it was clear that practices taught in the training project had been adopted on the participants' farms.

In order to make the evaluation study more manageable given time and language restrictions and a limited interview length per respondent,[10] it was decided to submit a selection of practices imparted by ORGANIC to the evaluation process. The selection covered more than half of all the elements that constituted the courses. In the case of the impact effects it was decided to leave out a formal measurement of farm income and instead incorporate a proxy variable for family well-being, asking the respondents whether they perceived their situation as having improved over the last five years. In addition, organisational and managerial capacities were left out of the evaluation exercise. The former was monitored by IRDP field staff while managerial capacity was considered too difficult to measure in simple terms. Moreover, it was assumed that this variable was highly correlated with other impact variables.

The baseline study was executed after the first two courses had already been implemented. The reason for this timing was the high initial fluctuation in attendance at the courses that normally occurs in the first sessions of a training project.[11] In order to select the sample for the baseline and ex post survey, some degree of certainty as to the composition of the participant population was necessary. After two sessions some 120 farmers were enlisted as participants. Given the potential restrictions that may constrain farmers from adopting LEIA practices, a substantial number of dropouts was expected. Therefore, in the initial talks between IRDP and ORGANIC and once again after the baseline study, ORGANIC was strongly recommended to select substantially more farmers, such that the target of 120 'fully trained' farmers by the end of the project could be met.

However, ORGANIC did not heed IRDP's advice, the latter having little effective influence over the former given ORGANIC's strong mandate in project implementation. On the basis of the survey and course attendance data, the dropout level over the three years was estimated to be in the neighbourhood of 45%. The most important reason for desertion was a lack of time, which points to the relatively high opportunity costs of labour in the region.

5 RESULTS

Adoption and diffusion

Table 3 shows the adoption levels of the selected practices for the three groups: participants in 1998, participants in 2001, and control group in 2001. For most practices the percentage of farmers applying a certain practice was used as an indicator for outcome. In some cases it was relatively easy and also more relevant to use quantity per farm household as an indicator of outcome. Significant differences between participants in 1998 and 2001 and between participants in 2001 and the control group provide strong evidence of an adoption effect caused by the project.

A first important observation from the table is that at the beginning of the project most practices were already known and applied by many farmers in the region. A second important observation is the fact that the participant group in 1998 and the control group are quite similar as to their adoption behaviour. Adoption rates in the participant group in 1998 are

slightly better because of the higher rates of previous technical assistance from NGOs and government organisations and the fact that two courses had already taken place in the project, resulting immediately in experimental application of the practices on the farms.

The first group of practices concerns land preparation and fertilisation. Both the burning of crop residues and the application of 'chemical' or purchased fertilisers (e.g. NPK fertilisers) were discouraged by the project. We can see that the project has had a clear effect in reducing both these practices, although the majority of participant farmers continued to apply purchased fertilisers. This ran counter to the ideological message delivered by ORGANIC's field coordinator (which in fact went further than the general philosophy of ORGANIC), who advocated a total substitution of organic fertilisers[12] for purchased fertilisers.

In some cases, the real adoption effect is not adequately reported by the table because, by the time of the study, the application of organic fertilisers and some soil conservation measures had already been dealt with in the first two courses (see Methodology section). However, other information sources suggest that the project had a significant effect on the adoption of these practices. Minimum tillage was a technique unknown in any form in the region. In contrast to their normal practice of preparing and clearing a whole plot before sowing and planting, ORGANIC taught the farmers to restrict land preparation to just the tiny area around the spot where each plant was to be sown or planted. In this way, the soil was better protected against the potential effects of wind and water. The adoption effect (among more than half of the participants) was solely due to the training project.

The data on nurseries and furnaces suggest that participants had already benefited from other organisations. The added value of the training project was less evident in these cases. The special latrines were quite popular among participants, because they are connected to one of the processes of creating organic manure on the farm. In the case of pigsties no significant effect can be noted. The same goes for other investments in livestock production (not in the table). Perhaps the most important adoption effects can be found in the area of crop and fruit tree production. First, as a special category of crops, there was a significant increase in the cultivation of medicinal plants which were highly popular among participants. In addition, crop diversification and especially fruit tree diversification increased significantly.

The favourite and most widely applied practices were the use of medicinal plants and organic fertilisers. Both practices have in common that they do not require significant investments in terms of capital or labour, and both have a clear short-term payoff which is in line with most small farmers' planning horizons and levels of risk aversion. Medicinal plants were used to cure minor illnesses and improve the quality of the diet (e.g. elderberry, rosemary, camomile and several other mostly local species). Organic fertilisers partly substituted for purchased fertilisers without incurring negative yield effects (see section on impact).

Table 3 does not tell the whole story. Findings from field visits and farmer interviews indicated that knowledge of LEIA practices had significantly increased as a result of the project. In addition, the care with which farmers implemented the practices and the diversity in modes of application had improved

Table 3 Project outcome in terms of adoption

Practice	Participants 1998	Participants 2001	Control group 2001
burning crop residues (%)	27 % **	2 %	29 % **
applying green material (crop residues, leaves,…) (%)	25 % **	63 %	18 % **
'chemical' fertilisers (%)	96 % *	79 %	97 % *
'organic' fertilisers (%)	79 % [a]	83 %	18 % **
ditches (%)	56 % [a]	73 %	24 % **
barriers (%)	44 % [a]	58 %	21 % **
minimum tillage (%)	nihil [b]	54 %	nihil [b]
latrines (%)	15 % **	56 %	8 % **
furnaces (%)	60 %	69 %	34 % **
pigsties (%)	42 %	60 %	45 %
nurseries (%)	33 %	44 %	3 % **
medicinal plants (no. plants)	3.2 (5.3) **	8.7 (7.0)	3.2 (3.5) **
crop diversity (no. crops)	4.3 (1.7) *	4.9 (2.4)	3.2 (1.4) **
fruit tree diversity (no. trees)	4.8 (2.9) *	6.2 (3.2)	4.6 (2.3) **

[a] At the time of the baseline survey, a course on the topic had already taken place.
[b] Not known by respondent.
Note1: Variables expressed in percentages are dichotomous variables; the value refers to the percentage responding 'yes'.
Note2: x (y) represents mean (standard deviation).
Note3: * $p < 0.05$, ** $p < 0.01$. Comparisons are always between the 1st and the 2nd column and the 3rd and the 2nd column; depending on the measurement scale t-tests and chi-square tests were applied.

noticeably. Also not shown in the table is the degree of adoption regarding land preparation, fertilisation and soil conservation practices, which was mostly between 25% and 55% of the total cultivated land on the farm. In reality, given the small farm sizes, this often meant that application was restricted to one or two experimental plots. We received no indication that many farmers were expanding their practices to other plots.

The literature on adoption (see Section 3) suggests that there are a number of constraints, depending on the type of practice, which can inhibit adoption processes. ORGANIC supplied valuable knowledge to participant farmers in the field of LEIA. However, capital and labour constraints were not addressed by the project. This partly explains why, for example, soil conservation measures that require significant investments in terms of labour (e.g. barriers, ditches) were not applied beyond experimental plots.[13] It also explains why relatively costly investments in terms of capital, such as the construction of pigsties, were not carried out by all the participants despite the fact that most of them owned some pigs. Other factors explaining the selective and partial adoption of the practices are first of all a lack of trust between farmers and institutions like ORGANIC. In addition, farmers in the region are reluctant to take risks especially if there is no clear (short-term) return on some of the practices. For example, those that introduced LEIA practices on their main subsistence plot (a crucial plot for the household's food security), did so gradually and selectively. Evidently, this picture undermines the somewhat ideological assumption of a complete linear transformation of conventional farming to LEIA farming as posed by ORGANIC.

The importance of these factors had been acknowledged by IRDP by the time the ex ante study had been carried out. However, although IRDP employed ORGANIC, the latter refused to cooperate with them in trying to offer solutions to the labour and capital constraints, a resistance which to a large extent can be explained by ORGANIC's (ideological) faith in its own philosophy. Short of withdrawing its financial support, IRDP did not have much leverage in terms of influencing ORGANIC's operations in the field.

Although a substantial number of participants said they had shared their knowledge with other farmers, no substantial diffusion effect took place. The ambitious target of 1200 indirect beneficiary families had not been reached, nor would it be reached in the near future. The sometimes strong social divisions within the region may in a sense have obstructed the diffusion processes. The attempts by ORGANIC to organise participant farmers in structured groups to increase knowledge sharing and stimulate diffusion to other farmers were largely unsuccessful. An important lapse in the training project was the fact that ORGANIC failed to point out the complementary benefits of combining LEIA practices with conventional techniques (see Ruben and Lee, 2000; De Jager et al., 2004). Instead, in the field it assumed a relatively extreme ideological stance, condemning conventional techniques, which was not very beneficial in terms of meeting the needs of the local farmers.

Impact at the farm level

In comparison to the causal relationship between participation in the project and the adoption of LEIA practices, the relationship between the adoption of LEIA practices and the impact variables specified in Table 4 is less straightforward. Besides the adoption of LEIA practices, various other external variables significantly influence the specified impact variables. In a formal framework, multiple regression can be used to isolate the effect of adoption rates and control for other potentially influential variables (e.g. Rossi et al., 1999). Our small sample sizes did not allow for such an analysis. Another complicating factor was the timing of the study. At the time of the ex-post analysis, it was still too early to assess the full range and magnitude of the impact effects brought about by the intervention. Therefore, even more than in the case of assessing the adoption effects, our quasi-experimental design needed to be complemented by sufficient qualitative information stemming from interviews and field visits to enable meaningful interpretation.

Table 4 shows the values on different impact variables for the three samples and forms the basis for our interpretation. In the first variable no significant improvement had occurred in the percentage of harvested fruit sold in the market. Given the fact that fruit yields in 2001 had not declined in relationship to the base year, the lack of improvement here cannot be

Table 4 Project impact

Variable	Participants 1998	Participants 2001	Control group 2001
% of fruit harvest sold	0.25 (0.37)	0.24 (0.35)	0.27 (0.35)
Soil quality (% organic matter)	2.69 (1.23) **	3.37 (1.57)	3.36 (1.70)
Yields maize (*qq* / *cuerda*)	1.85 (1.05)	2.28 (2.25)	2.30 (1.79)
Do you think your situation has improved over the last 5 years?	46 % **	88 %	55 % **

Note 1: x (y) represents mean (standard deviation).
Note 2: Variables expressed in percentages are dichotomous variables; the value refers to the percentage responding 'yes'.
Note 3: Notwithstanding some local variation, one hectare comprises approximately 23 *cuerdas*.
Note 4: *qq* refers to *quintales*. One *quintal* is approximately 50 kilograms.
Note 5: * $p < 0.05$, ** $p < 0.01$. Comparisons are always between the 1st and the 2nd column and the 3rd and the 2nd column; depending on the measurement scale t-tests and chi-square tests were applied.

attributed to bad harvests. Some diversification in fruit trees had already occurred under the influence of the project (see Table 3), though many of these new trees were not yet bearing. Probably the first harvests from these newly planted trees would have a minor positive effect on fruit sales. At the time of the evaluation study, one of the basic factors behind the lack of improvement in fruit sales was the fact that the project course module on marketing skills for farm households had not been properly implemented.

The variable soil quality requires some attention. Soil samples had been collected in 1998 from the farms of all those participating in the study. The samples were taken from those plots where the farmers would (and had already started to) practise their newly acquired knowledge from the training project. With the help of GPS (Global Positioning System), the coordinates of the plots were stored and, in 2001, soil samples were collected from the same experimental plots. Although the actual method of taking soil samples in the field allows for some variation,[14] the comparison between 'before' and 'after' was considered to be quite reliable. For the control group the rule for plot selection (to assure some level of homogeneity as a basis for comparison) was to choose the main plot for maize, a crop cultivated by all farmers in the region.

Table 4 shows a significant increase in the percentage of organic matter in the soil within the participant group. However, differences with control group farmers are not significant. This lack of difference could be explained by the fact that the main maize plots have a higher level of soil fertility than the plots where participant farmers started applying their practices (which in some cases had not been cultivated regularly). Hence, the comparison with the control group is not very reliable. In addition, caution should be taken in interpreting the increase in organic matter in the participant plots. As explained before, a combination of factors made farmers reluctant to shift fully to LEIA farming on their main subsistence plots. Where the experimental plot (from which the samples were taken) coincided with the main subsistence plot, participant farmers often applied organic fertilisers in combination with purchased fertilisers (see Table 3). Sometimes experimental plots were plots that were formerly not used intensively. The shift from little or no attention to more attention to crop cultivation on a given plot probably contributed to the increase in organic matter. Given these and other influential factors (e.g. soil type), a more controlled experimental setting would have been preferable. However, not only would such an experiment have been very costly, it would have to be determined which effect was to be isolated. Establishing the increase in soil organic matter given different combinations of purchased and organic fertiliser use would have resulted in a different experiment from isolating the effect of organic fertilisers on soil organic matter. All the same, despite the limited level of precision, we can infer from this exercise that the combined use of organic and purchased fertilisers and the extra dedication to the plots had a positive effect on soil organic matter.

In the case of yields, a slight increase over time (though not statistically significant) in maize yields was recorded. Although in this analysis one is faced with the limits of a two-year comparison of yields and the effect of specific yearly climatic conditions, different sources of information permitted some conclusions to be drawn in this regard. First of all, weather conditions, if anything, were worse in 2001 than in 1998. Hence, in a normal situation one would have expected a decline in yields. This might explain why the participants' yields in 2001, despite extra attention and in most cases the application of both organic and purchased fertilisers, were not significantly different from 1998. When asking participant farmers if they thought that their yields had improved, a majority (despite adverse climatic conditions) answered affirmatively. Although their positive assessment could have been discounted as an effort to simply speak positively (or politely) about the project, the other sources of evidence support the occurrence of better yields arising from the adoption of the LEIA practices. However, it should be stressed that in most cases the combined use of both purchased and organic fertilisers, and not a complete transformation of conventional farming to LEIA, was probably the main cause for yield improvements. As in the case of organic matter, the lack of difference between participating farmers and the control group can be explained by the sometimes structural differences in soil quality between the main plots of the control group farmers and the experimental plots of the participant farmers.

The last variable represents the respondents' perceptions of any improvement in their general situation over the last five years. In principle, this is not a pure impact variable, since the perception of general improvement might be influenced not only by the 'real' effects of the project on the livelihood of the participant household, but simply by the respondents' sentiments about participating in it. The fact that the variable was measured at two different moments in time, allows for a general interpretation of the role of 'real' effects in the perception of the participant.[15] Table 4 shows that participants were more positive about improvements in their situation than control group farmers (while starting from similar levels), suggesting that participation in the project and the adoption of LEIA practices have had an overall positive effect on the livelihoods of the participants and their families, a conclusion that was confirmed by impressions from individual interviews.

6 LESSONS LEARNED

Strengths and weaknesses of the evaluation methodology

Given the objective of the evaluation study, the applied methodology proved to be quite useful. Evaluation studies of the type discussed in this paper, based on a quasi-experimental design, are not very common in development projects of a comparable (small) size to our training project. We have shown that the formal method of comparing participants with control groups

can constitute an important framework on which to build the analysis of outcome and impact. Without complex matching procedures and with limited statistical power, the strength of a simple quasi-experiment relies heavily on additional qualitative information. This shift in emphasis should not give the impression of a lack of rigour. Problems such as the influence of selection bias still need to be addressed carefully, even if not done in a formal statistical way.

The research design has proven to be quite reliable for analysing outcome. The clear causal relationship between participation in the project and adoption of LEIA practices facilitates the interpretation of the quasi-experimental data. Controlling for starting levels before joining the project and adoption levels of similar non-participating farmers, a plausible indication of the range of practices adopted and their magnitude that may be attributed to the project was established. The findings, e.g. the logic behind the high popularity of practices such as organic fertilisers and medicinal plants, are supported by findings from other studies.

In the case of impact variables such as yields or soil quality, the causal influence from the project is weaker and more indirect. To isolate the effect of the intervention on impact variables from the influence of other factors would ideally require a more controlled experiment. Since this would raise both the budget as well as the required level of expertise, such experiments do not represent real options for many smaller projects. For these projects, a simple quasi-experimental design can be used to establish some trends in impact variables in relationship to the intervention. If prepared carefully, such a design already controls for a substantial part of the exogenous effects on impact variables. The next step would be to uncover some of the complexity in underlying causal relationships between different exogenous variables and project intervention on the one part and impact variables on the other. In practice, a wide range of techniques such as field visits, semi-structured interviews and other more participatory research techniques are available to incorporate in a structured approach to study this complexity. Our study did not entirely succeed in studying these underlying relationships. Nevertheless, despite the fact that the full impact of the project had yet to materialise at the time of the ex post study, the evaluation study was able to establish some important relationships between the project and changes in impact variables at the level of the farm and the farm household.

The role of the evaluation study: measuring effectiveness versus improving practice

In contrast to the study's successful compliance with the main objective of the evaluation study, i.e. assessing the outcome and impact of the training project, the evaluation study (and especially the ex ante study) was quite ineffective in terms of influencing the design and implementation of the project. This lack of influence is first of all due to the timing of the study which was quite adequate for establishing the baseline picture of the project participants. Nevertheless, the ex ante study was implemented at a time when the basic structure of the project was already in place and the contract and terms of reference between IRDP and ORGANIC had already been finalised. Short of withdrawing its financial support, IRDP did not have much leverage to influence the project's design and implementation, reducing the utility of the conclusions of the ex ante study in terms of its potential to influence practice. In fact, ORGANIC did not heed IRDP's 'advice' nor did it respond to their offers of assistance in the field.

As a future remedy for this kind of problem, part of the ex ante study could be implemented as a formal appraisal study before the selection of the implementing organisation takes place. Such an appraisal would include the same kind of succinct literature review as carried out in the ex ante study and the same type of general assessment of a number of potential weaknesses. On the basis of the literature on adoption and innovation processes, key constraints could be identified and posed as requisites for selecting an implementing organisation and drawing up the terms of reference of the project.

In this study some of the key constraints identified in the ex ante study but with little impact on improving project design and implementation include the following: first of all, ORGANIC failed to recruit a representative and sufficiently large group of motivated farmers from a wide variety of communities in the territory, as contemplated beforehand. Instead of investing in simple advertising campaigns (e.g. pamphlets, community meetings), communication about the project was spread through social networks. As a result, outreach was concentrated in communities and specific social networks connected to ORGANIC staff, who are themselves farmers and sometimes community leaders in the region. This is a particular danger for projects based on the model of farmer-to-farmer extension. Not only did outreach fail to cut through social divisions in and between communities, the specific outreach mechanisms through these social networks also reinforced the local power positions of community leaders involved in the project.

Second, a number of generally known constraints to adoption processes could have been acknowledged and more effectively dealt with. Problems of labour availability for investment in LEIA practices (given the farmer's opportunity costs of labour and the perceived return on investment in the practices) could have been foreseen. In addition, the training project should have focused on the complementary benefits of using both LEIA practices and conventional techniques. This would have facilitated processes of innovation and enhanced the attractiveness of LEIA agriculture, where each farmer would be motivated to choose the best balance between conventional and LEIA practices tailored to his or her personal circumstances. ORGANIC's ideological bias favouring an exclusive reliance on locally reproduced practices not only reduced the attractiveness of many LEIA practices, it also embodied a solution that was less than optimal for the farmers in the region.

REFERENCES

Cary, J.W. and Wilkinson, R.L. (1997) 'Perceived profitability and farmers' conservation behaviour'. *Journal of Agricultural Economics*, Vol. 48, No. 1, pp. 13–21.

Carmack, R.M. (1995) *Rebels of Highland Guatemala: The Quiche-Mayas of Momostenango*. Norman and London: University of Oklahoma Press.

Cook, T.D. and Campbell, D.T. (1979) *Quasi experimentation: Design & analysis issues for field settings*. Boston, MD: Houghton Mifflin.

De Jager, A., Onduru, D. and Walaga, C. (2004) 'Facilitated learning in soil fertility management: Assessing potentials of low-external-input technologies in East African farming systems'. *Agricultural Systems*, Vol. 79, No. 2, pp. 205–23.

Feder, G., Just, R.E. and Zilberman, D. (1985) 'Adoption of agricultural innovations in developing countries: A survey'. *Economic Development and Cultural Change*, Vol. 33, No. 2, pp. 254–97.

Greene, J. and Caracelli, V. (1997) 'Defining and describing the paradigm issue in mixed-method evaluation'. *New Directions for Evaluation*, Vol. 74, pp. 5–17.

Hocdé, H., Vasquez, J.I., Holt, E. and Braun, A.R. (2000) 'Towards a social movement of farmer innovation: Campesino a campesino'. *ILEIA Newsletter*, Vol. 7, pp. 26–7.

Jones, S. (2002) 'A framework for understanding on-farm environmental degradation and constraints to the adoption of soil conservation measures: Case studies from Highland Tanzania and Thailand'. *World Development*, Vol. 30, No. 9, pp. 1607–20.

Low, A. (1986) 'On-farm research on household economics' in J.L. Moock (ed.) *Understanding Africa's rural households and farming systems*. Boulder, CO: Westview Press.

Lutz, E., Pagiola, S. and Reiche, C. (eds) (1994) 'Economic and institutional analysis of soil conservation projects in Central America and the Caribbean'. *World Bank Environment Paper* No. 8. Washington D.C.: World Bank.

McCreery, D. (1994) *Rural Guatemala 1760–1940* Stanford, CA: Stanford University Press.

Moser, C.M. and Barrett, C.B. (2003) 'The disappointing adoption dynamics of a yield-increasing, low external-input technology: The case of SRI in Madagascar'. *Agricultural Systems*, Vol. 76, No. 3, pp. 1085–1100.

Mosley, P. (1997) 'The use of control groups in impact assessments for microfinance'. *ILO Working Paper* No.19. Geneva: ILO.

Neill, S.P. and Lee, D.R. (2001) 'Explaining the adoption and disadoption of sustainable agriculture: The case of cover crops in Northern Honduras'. *Economic Development and Cultural Change*, Vol. 49, No. 4, pp. 793–820.

Pannell, D.J. (1998) 'Economics, extension and the adoption of land conservation innovations in agriculture'. *International Journal of Social Economics*, Vol. 26, (7/8/9), pp. 999–1012.

Pomp, M. (1994) *Smallholders and innovation adoption: Cocoa in Sulawesi, Indonesia*. PhD thesis. Amsterdam: Free University.

Pretty, J.N. (1995) *Regenerating agriculture: Policies and practice for sustainability and self-reliance*. London: Earthscan Publications.

Rossi, P.H., Freeman, H.E. and Lipsey, M.W. (1999) *Evaluation: A systematic approach*. London: Sage Publications.

Ruben, R. and Lee, D.R. (2000) 'Combining internal and external inputs for sustainable intensification'. *2020 Brief* No. 65 Washington DC: International Food Policy Research Institute.

Ruben, R. and Vaessen, J. (2000) 'Soil conservation practices and farmers' adoption strategies in Costa Rica' in W. Pelupessy and R. Ruben (eds) *Agrarian policies in Central America*. London: Macmillan.

Stocking, M. and Abel, N. (1992) 'Labour costs: A critical element in soil conservation' in W. Hiemstra, C. Reijntjes and E. van der Werf (eds) *Let farmers judge: Experiences in assessing the sustainability of agriculture*. London: Intermediate Technology Publications.

ENDNOTES

1 ORGANIC is a fictitious name.

2 Marked in grey in Figure 1. Farmers from Xepón Grande and Xepón Pequeño in the neighbouring province of Huehuetenango also participated in the project.

3 This methodology, although criticised for its lack of formal technical expertise, has become popular in Latin America (e.g. Hocdé et al., 2000).

4 The evaluation study in fact comprised two independent complementary studies. The first focused comprehensively on project outcome and impact, mostly in agriculture, while the second dealt exclusively and in more detail with the effects of the project on the nutrition and health status of participants and their families. In this paper the focus is exclusively on the first study.

5 The Mayans have traditionally been dominated by white and *ladino* population groups.

6 Apart from the small sample size and the lack of a baseline control group that distinguishes the applied design from formal 'Best Practices', there are a number of techniques (e.g. matching techniques, use of instrumental variables in two-stage regression analyses) to isolate the effect of a certain intervention on outcome and impact variables in a more rigorous manner (see for example Rossi et al., 1999; Mosley, 1997).

7 Eight of the original 56 farmers could not be located mainly because of temporal or permanent migration to other regions. Only one of these eight participants had graduated from the project. This dropout rate was much larger than for the sample as a whole. Using only the 48 cases for comparison between the baseline and the ex post survey would result in a slight 'overestimation' of the project outcome and impact.

8 This is because the counterfactual, i.e. what would have happened with the group of relatively

motivated farmers without the project, is not entirely accurately captured by the control group (see Mosley, 1997).

9 However, it was important that the ex post evaluation study be carried out soon after the termination of the project for at least two reasons: First, the budgetary planning and limited time horizon of IRDP as a whole (the programme ended in 2002) made any delay in the timing of the study difficult. Second, the decision about further financial support for ORGANIC and follow-up by IRDP staff with the participant farmers depended in part on the results of the evaluation study.

10 In practice, a lot of time was needed for careful explanation and formulation of the survey questions. Sometimes this required a mix of Spanish and Quiché, the local Mayan language.

11 There is always some degree of adverse selection, i.e. farmers who enrol for dubious reasons or with unrealistic expectations and who come to the conclusion that the project does not serve their purposes.

12 The application of green material (crop residues, leaves, etc.) was also encouraged by ORGANIC. The green material was simply gathered and distributed on the plots. It was also incorporated into the process of preparing organic fertiliser. One such process involved composting green material in combination with animal (and sometimes human) manure and lime.

13 This was despite the fact that the steep slopes on the majority of the farms required protection from erosion.

14 Soil samples were collected according to a standardised procedure in which the person collecting the samples walked in a zigzag through a plot, taking small samples from different parts of the plot, then mixing all the samples together.

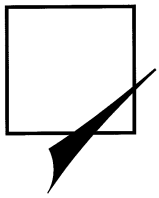
15 This assumes that the feeling of optimism due to participating in the project did not increase significantly during the course of the project.

CHAPTER 6

Vaessen, J. (2006) “Programme theory evaluation, multicriteria decision aid and stakeholder values: a methodological framework”, *Evaluation*, 12(4), 397-417.

Evaluation

SAGE Publications (London,
Thousand Oaks and New Delhi)
DOI: 10.1177/1356389006071217
Vol 12(4): 397–417



Programme Theory Evaluation, Multicriteria Decision Aid and Stakeholder Values

A Methodological Framework

JOS VAESSEN
University of Antwerp, Belgium

Nowadays there are a number of evaluation approaches that specifically focus on the elicitation of stakeholder values. Multicriteria decision aid enables evaluators to go one step further by systematically showing how different stakeholder values would affect evaluative outcomes and subsequent policy decisions about a programme's future. This type of technique can greatly benefit from being embedded in a programme theory evaluation methodology. Taking into account lessons from the literature on stakeholder values in evaluation, a methodological framework combining elements of programme theory evaluation and multicriteria decision aid is presented. The framework is illustrated by means of an example.

KEYWORDS: evaluation methodology; multicriteria decision aid; programme theory evaluation; rural development; stakeholder values

Introduction

Over the past decades, evaluators of social programmes have developed a number of evaluation approaches which start out from some kind of 'theory' of how a programme works or should work. The common element that unites these 'theory-oriented' approaches (Stame, 2004)[1] is the reconstruction of a causal model (the programme theory) on the basis of different sources of information in order to arrive at an understanding of how programmes bring about intended and unintended outcomes. The term 'programme theory evaluation' (PTE) as used in this article refers to this process of reconstruction of the theory as well as an assessment of the validity of the reconstructed theory (vis-à-vis multiple benchmarks).

Leeuw (2003) argues that, notwithstanding a renewed attention in the literature to programme theory in evaluation, there continues to be a lack of systematic

methods for reconstructing programme theories. An important element that is not sufficiently addressed in most of the methodological discussions on programme theory reconstruction and evaluation is the question of how to deal with multiple stakeholder perspectives on what a programme is (or should be) about. While this question has received relatively little attention in the literature on PTE, elsewhere the significance of the topic is reflected in the myriad of approaches in evaluation dealing with the issue of stakeholder participation (e.g. Cousins and Whitmore, 1998; Stufflebeam, 2001).

Notwithstanding the diversity in evaluation approaches explicitly dealing with participation and stakeholder values, both the participatory process and taking into account all the preferences of stakeholders in the evaluation design can be costly. In the light of the foregoing, a recent approach called 'values inquiry' (Mark et al., 1999) is useful in helping to determine the most important stakeholder values to be taken into account in the evaluation design. While values inquiry might be meaningfully integrated with PTE (see e.g. Renger and Bourdeau, 2004), the approach is limited to the determination of the most important stakeholder value positions, e.g. by determining which criteria to take into account in the evaluation design.

A range of techniques that fall under the banner of multicriteria decision aid (MCDA) enable the evaluator to go one step further by showing how different stakeholder values would affect evaluative outcomes and subsequent policy decisions about a programme's future. MCDA serves the purpose of helping decision makers to make informed choices about (future) policies. The core of MCDA constitutes the evaluation of different alternatives (e.g. programme strategy scenarios) based on a number of criteria, each receiving a specific weight in the evaluation process (Belton, 1990).

While MCDA as a technique and a process stands perfectly well on its own, the articulation of MCDA with PTE has clear advantages for the MCDA process. In turn, from the point of view of PTE, MCDA offers a useful procedure for systematically addressing stakeholder values in contexts of formative (programme theory) evaluations that serve the purpose of programme improvement in the future.

This article presents a methodological framework that combines elements of PTE and MCDA. The benefits of integrating these two methodological approaches are optimally present in evaluations that serve the double objective of retrospective assessment of merit and worth and programme improvement in the future. In such cases, a summative programme theory evaluation exercise can provide the adequate basis for a subsequent formative evaluation exercise that is structured by the MCDA technique. I will especially focus on the latter type of exercise, which (with some changes) can also be implemented independently.

The structure of the article is the following. The next section will briefly review some of the principal elements in the general debate on stakeholder values in evaluation. Subsequently, the issue of stakeholder values will be considered within the realm of PTE. These two sections provide the basis for the third section which discusses the benefits of combining elements of programme theory evaluation

with multicriteria decision aid especially with regard to the issue of addressing stakeholder values in the evaluation. PTE and MCDA are then brought together in a methodological framework. The framework is illustrated on the basis of an example regarding the evaluation of a training programme in organic agriculture. Finally, some limitations and advantages of the framework are discussed.

Programme Evaluation and Stakeholder Values

In a comprehensive review of participatory approaches in evaluation, Cousins and Whitmore (1998) distinguish between transformative participation and practical participation. In the latter approach the main principle underlying participation is that it enhances stakeholder ownership and therefore the relevance of the study and utilization of the findings. In contrast, transformative participatory evaluation aims at strengthening the capacities of stakeholder groups by enhancing their control over the programmes that affect them. Both approaches can be characterized by a rather intensive form of participation (i.e. dialogue and joint deliberation among evaluators and stakeholders as opposed to 'mere' consultation) as well as being comprehensive (i.e. including stakeholder participation in all phases of the evaluation study).

A recent approach called 'deliberative democratic evaluation' (House and Howe, 1999) is rather similar to the participatory approaches just mentioned in the sense that it presupposes an intensive interaction between evaluators and different stakeholder groups. House and Howe (1999) argue that eliciting the interests of different stakeholders is not an easy and straightforward task. Stakeholders do not have a clear viewpoint about a given programme but need to be involved in a process of extensive interaction in order to be able to clarify and express their views. The authors advocate a three-tiered approach for such a process: inclusion of all major stakeholder groups, extensive dialogue and finally a phase of deliberation in order to arrive at a judgement of a programme's merit and worth.

House and Howe acknowledge that their model for evaluation in a democratic society is rather ideal typical. More specifically, there are a number of constraints to a successful implementation of this approach in practice. These constraints also apply (to a certain extent) to the two approaches of practical and transformative participatory evaluation already mentioned. First, there is the difficulty and high cost of *physically* incorporating all major stakeholder groups in a process of dialogue and deliberation. Second, even if one might be able to organize such processes (e.g. in smaller programmes), there are several potential problems of group processes that inhibit the realization of an open dialogue that reflects the views of all participants (see e.g. Cooke, 2001).[2] For example, power imbalances might crowd out the viewpoints of less powerful and under-represented groups.

Mark et al.'s approach – values inquiry – has the potential to circumvent these problems. 'Values inquiry refers to a variety of methods that can be applied to the systematic assessment of the value positions surrounding the existence, activities, and outcomes of a social policy and program' (Mark et al., 1999: 183). Values

inquiry is in fact a rather generic approach to incorporating stakeholder values into the evaluation process. In contrast to the deliberative democratic evaluation approach which also focuses on the elicitation of stakeholder values, the degree of participation is in general much lower. In principle it is not a method-specific type of inquiry. Henry (2002) for example used surveys to elicit the values of four different stakeholder groups in the evaluation of a preschool programme. Alternatively, focus groups or other more qualitative techniques of inquiry can be used to capture stakeholder values.

A central element of values inquiry is to determine which criteria (according to different stakeholder groups) should be used to determine a programme's worth and merit. In this sense, values inquiry can be perceived as being an elaboration of the earlier approach of stakeholder-based evaluation (Bryk, 1983). A core premise hereby is that 'the choice of criteria for program success should be justified by the process used to obtain them' (Henry, 2002: 183). In values inquiry, the mapped value positions are important findings that can be presented to different evaluation audiences (Mark et al., 1999). MCDA is a particular approach to take the analysis a step further by showing in a systematic manner how these different value positions would affect evaluative conclusions and subsequent policy choices.

Programme Theory Evaluation and Stakeholder Values

In PTE, much of the discussion on values has focused on the question of whose assumptions make up the programme theory (Chen, 1990; Weiss, 1998). This discussion has largely revolved around the relative roles of the evaluator and programme staff in determining the programme theory. Who is/are the principal determinant(s) of the programme theory? Some authors (e.g. Wholey, 1987) have argued for a focus on programme staff and other key stakeholders (e.g. policy makers, interest groups) as principal determinants of the programme theory, on the grounds that they are the major actors who continuously shape the implementation and outcome of the programme. Others have emphasized the role of the evaluator backed by her experience and knowledge of social science theory (e.g. Chen and Rossi, 1980). In more recent work, authors tend to favour an integrative approach (e.g. Chen, 2004; Pawson and Tilley, 1997).

Most recent approaches to programme theory reconstruction imply a high degree of stakeholder participation. Examples are Christie and Alkin's (2003) iterative process of Delphi inquiry, or the so-called strategic assessment approach based on four stages of group discussion and deliberation described by Leeuw (2003). Other approaches used in policy analysis and decision support also serve the purpose of programme theory reconstruction. Techniques like concept mapping (Trochim, 1989) and cognitive mapping (Eden and Ackermann, 1998; Leeuw, 2003) rely on a high degree of participation of different stakeholders. Most of these approaches are applied in collaboration with programme staff only (or a slightly wider group of stakeholders), e.g. not including the views of programme beneficiaries. Incorporation of a wider audience of stakeholders might prove too costly (and/or difficult) given the required level of intensity of the interaction between evaluators and stakeholders.

In the discussion of stakeholder values in programme theory evaluation, it is important to distinguish between the purpose of (ex post) evaluation of merit and worth and programme improvement (strategy clarification). In the former case, the question of whose assumptions make up the theory is likely to be determined by a condition that the evaluator should reconstruct a theory that best reflects programme practice and major assumptions of primary users of the evaluation (e.g. programme staff, commissioners) and subsequently assess the theory on the basis of further data collection and analysis. In the assessment phase, different stakeholders can then be consulted regarding which assumptions to test and on the basis of which criteria (e.g. following a values inquiry procedure). If the purpose is programme improvement, the evaluation takes on a more formative character and the issue of stakeholder values becomes closely linked to the redefinition of the programme strategy. MCDA is one of the possible approaches for structuring such a formative evaluation process while systematically taking into account divergent stakeholder values.

Programme Theory Evaluation and Multicriteria Decision Aid

Multicriteria decision aid (MCDA) is a decision support approach that has been widely used in ex ante product, project and programme evaluations in both corporate and public domains of decision making. Nowadays, there are many different methods in MCDA and the number is growing (see e.g. Belton and Stewart, 2002; Roy, 1996). The essence of an MCDA approach is to support decision makers in making informed choices regarding a number of alternatives based on a number of criteria. In addition, the relative importance (weight) of the different evaluation criteria in the final decision can vary according to the preferences of different stakeholders involved the decision-making process (Belton, 1990).

MCDA is mostly used for formative purposes, i.e. to facilitate a decision-making process regarding the choice between different policy options. While MCDA as a technique and a process stands perfectly well on its own, there are definite advantages for applying MCDA within the framework of a PTE. First, PTE can be very helpful in the identification of potential policy alternatives to improve a situation of ineffective policy intervention in a certain area. A PTE of the existing policy intervention will consist of a systematic reconstruction of the assumptions regarding the implementation, generation of outputs and processes of change. Subsequent assessment of these assumptions will provide evaluators with insight into the questions of *why* and (very importantly) *where* an intervention went wrong. Subsequently, the weaker links in the programme theory constitute the basis for distilling lessons for improvement, which become the building blocks for a number of programme alternatives designed to remedy (some of) these weaknesses. Second, when assessing programme alternatives on multiple criteria, a programme theory can be a very useful framework for guidance. The programme theory of the existing intervention can be adapted to represent the logic of each of the different alternatives, i.e. an overview of the causal links that reflect the

major assumptions of each of the alternatives in combination with major external influences that affect processes of implementation and change. As a result, these (slightly) different programme theories facilitate the analysis and reasoning by the evaluator in order to arrive at an assessment of an alternative's expected performance on a particular criterion. This works in two ways: the programme theory constitutes a structure for argumentation as well as for efficiently selecting relevant bodies of evidence to analyse specific links in the theory (which are related to particular criteria). While PTE has the potential to enhance the quality of the MCDA process, from the point of view of PTE the combination is also advantageous. MCDA provides a sound framework for systematically dealing with stakeholder values in a PTE in the context of a formative process geared towards the definition of a new programme strategy.

Presenting the Methodological Framework

A comprehensive methodological framework built on elements of PTE and MCDA can be divided into two 'steps' as shown in Box 1. The first step serves the main purpose of assessing the strengths and weaknesses of an existing programme under review. The second step constitutes a trajectory for programme improvement towards the future. In principle, both steps can be implemented separately but, as argued earlier, the full advantages of a combination of PTE and MCDA are most visible in an evaluation study that covers both steps. I will focus mostly on step 2 of the framework as the particular combination of MCDA and PTE in a formative evaluation context has received little attention in the literature. In contrast, step 1 largely resembles existing programme theory evaluations as described in the literature (for recent examples see e.g. Christie and Alkin, 2003; Ehren et al., 2005).

Box 1. Methodological Framework for Retrospective Programme Evaluation and Programme Improvement Towards the Future

STEP 1: Retrospective programme evaluation

- Reconstruction of the programme theory
- (Elicitation of criteria and relative importance of criteria)
- Evaluation of the programme theory on the basis of the criteria

STEP 2: Programme improvement towards the future

- Definition of alternatives
- (Elicitation of criteria and relative importance of criteria)
- Assessment of the alternatives on the basis of the criteria
- Application of an MCDA technique as a basis for a process of deliberation towards an improved programme strategy

Step 1: Retrospective Programme Evaluation

The main element that distinguishes this step from other summative PTE approaches is the explicit dimension of stakeholder value elicitation. In an evaluation study covering both steps of the framework, in the first step the evaluator would pose two sets of questions to stakeholders. The first set of questions refers to how stakeholders think that a programme works and is expected to achieve certain outcomes. This information (together with other information like programme documents, staff interviews, etc.) will provide the building blocks for a (descriptive) programme theory. The second set refers to the issue of what stakeholders deem important regarding the implementation and effects of a programme. This is akin to the normative part of a programme theory (see Chen, 1990) and can be captured by a values inquiry exercise (e.g. based on surveys). By identifying different sets of criteria and differences in the relative importance of these criteria (the weights), the evaluator can determine the major stakeholder value positions regarding a programme. These criteria would be the basis for the evaluation of the descriptive programme theory. In addition, the different value positions would constitute the basis for the assessment of and choice between programme alternatives under step 2. If step 2 is implemented without step 1 as a basis then the evaluator would need to acquire some of the essential information that would have been provided by the first step through other means (perhaps in a more rapid and superficial way). One of these pieces of information is the inventory of stakeholder value positions. In addition, as there would be no evaluative results (as provided by step 1) available about the programme in question, a different way to define alternative programme strategies would need to be developed (see Keeney, 1992, for discussion on how to do this).

Step 2: Programme Improvement towards the Future

Definition of alternatives In the PTE under step 1, each (major) assumption in the programme theory of the existing programme is assessed separately. As a result, the evaluator is able to identify where (in what part of the programme logic) major weaknesses are to be found. The detected weaknesses in the original programme theory provide the basis for defining:

- a set of general directives to improve the program;
- a number of alternative strategies that provide partial solutions to the detected weaknesses in the existing programme (see also Keeney, 1992).

Given the trade-offs between the different criteria,[3] it is highly unlikely that there will be one alternative that can (potentially) solve all the weaknesses in the original programme, hence the definition of multiple alternatives. These alternatives represent partial solutions to the weaknesses identified in the original programme, each alternative as it were emphasizing particular evaluation criteria. Given the fact that stakeholder groups differ in terms of the values they attach to particular criteria, the list of alternatives to some extent already represents the diversity in stakeholder value positions.

Assessment of the alternatives on the basis of the criteria The original programme theory provides essential support for appraising the different alternative strategies that have been defined, since all the alternatives are in essence (slightly) different versions of the existing programme. The evaluator starts with a review of the original material (interview transcripts from stakeholders and experts, survey data, academic literature, other written sources) collected for the assessment of the original programme theory. On the basis of these information sources, (slightly) adjusted programme theories reflecting the logic of each alternative, and consultations with key stakeholders and experts, the evaluator is able to appraise the alternatives on the different criteria. The appraisal is recorded by means of a score expressing the expected performance of each alternative on each criterion. Such an evaluation score can be measured on a metric scale (e.g. expressing monetary costs) or on a non-metric scale (e.g. an ordinal semantic scale ranging from 'very low' to 'very high'). A special type of non-metric score concerns the procedure of ranking alternatives according to their expected performance on a criterion. In this procedure the alternative with the best expected performance receives the rank 1, the second-best alternative receives rank 2, etc. Ranking is often used in cases when the available information is imprecise and/or the expected effects are uncertain and depend on a complex interplay of (external) factors. All the evaluation scores are summarized in a performance matrix of m alternatives by n criteria.

Application of an MCDA technique as a basis for a process of deliberation towards an improved programme strategy MCDA is an approach that helps to structure the kind of intuitive deliberation that always takes place when decision makers need to choose between alternative courses of action. It is especially useful when the number of alternatives and/or criteria is high. In such cases, without a decision support technique (MCDA), the human mind alone simply cannot grasp the complexity of advantages and disadvantages of different alternatives as expressed in the different criteria. Moreover, without such a technique it would be nearly impossible for decision makers to gauge how different preferences of stakeholders regarding programme dimensions would translate into preferences for alternative strategies.

It is not the purpose of this article to present a comprehensive technical discussion of MCDA. There is a rich literature on MCDA covering a multitude of different techniques and their corresponding technical underpinnings. These techniques mainly differ in the way they address issues like:

- the way criteria are assessed;
- the application and computation of weights;
- the mathematical algorithm used to derive the overall ranking of alternatives;
- the model to describe individual preferences (compensatory versus non-compensatory criteria,[4] linear versus non-linear preferences);
- the uncertainty embedded in the data;
- the ability for stakeholders to participate in the process.

For guidance on these issues and choosing appropriate MCDA techniques for different situations see e.g. Dodgson et al. (2001), Belton and Stewart (2002), Roy (1996) or more classical texts like Keeney and Raiffa (1976).

MCDA techniques vary from very complicated to very simple. One of the most simple and widely applied MCDA techniques is the so-called linear additive model.[5] In the next section I will illustrate an example which is based on a specific version of this model. In the evaluation literature a few authors have referred to this type of technique in the context of aggregation of judgements on different criteria and also programme strategy clarification (see e.g. Chen, 1990; House, 1995; Scriven, 1991). Among these, Scriven has been rather critical of the technique. Scriven's critique is centred on four points (1991: 380–1), two of them specifically related to the linear additive model. The most important critique is probably the aspect of linear preferences and compensation. In the linear additive model one unit of loss in performance on one criterion can be fully compensated by a fixed amount of units in another criterion which is not a very realistic assumption. A more sophisticated critique of this model developed by proponents of the so-called outranking school in MCDA (see Roy, 1996) also highlights this problem of linear preferences.

Most of the proponents of the linear additive model acknowledge its limitations. Still, the technique continues to be widely used, mainly because of the fact that it is fairly simple, relatively cheap to implement and easy to understand for decision makers. Belton and Stewart (2002), very much aware of the shortcomings of the method, downplay the critique such as developed by Scriven (1991) in the context of programme design or improvement. They argue that in such contexts simple models like variations of the linear additive model very often generate the same global insights as more elaborate models that (among other things) more realistically model the preferences of decision makers. The key lies in sensitivity analysis, i.e. an extensive discussion and variation of the assumptions of the model.

Within the light of the proposed framework, sensitivity analysis would entail the following. After generating an overall ranking of alternatives, the evaluator should undertake a sensitivity analysis first on performance scores and subsequently on the choice and relative importance of criteria. The latter analysis constitutes the basis for a deliberation process towards a new programme strategy. The evaluator starts with producing different overall rankings of the alternatives, each ranking being based on a particular set of criteria and weights of one stakeholder group. Subsequently, contrasting these different rankings will fuel a process of reflection and deliberation, finally resulting in a choice about which strategy to undertake. In principle, there are three broad scenarios for organizing this type of deliberation process:

- the evaluator facilitates a discussion between representatives of different stakeholder groups where the initial viewpoint of each group is represented by a ranking of alternatives that is based on each group's value position;

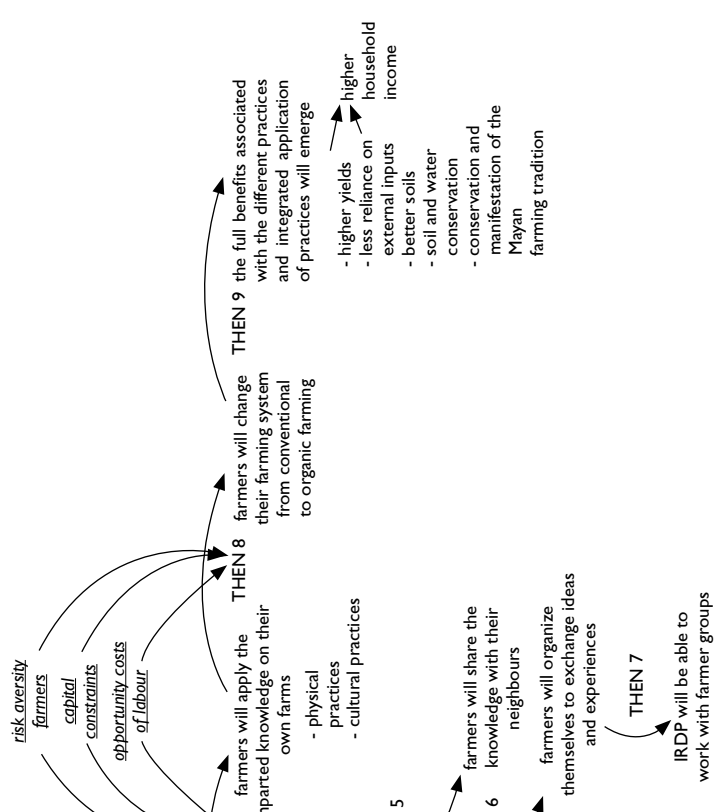
- the evaluator facilitates a discussion between programme decision makers only; it is the task of the evaluator to confront programme decision makers with the consequences of different stakeholder values on the ranking of the programme alternatives;
- the evaluator does not participate in the decision-making process but describes in her report how different value positions would affect choices between alternative strategies and where possible compromises between stakeholders might be found.

An Example: The Evaluation of a Training Programme in Organic Agriculture in Guatemala

Short Description of the Programme and the Programme Evaluation

In this section I will illustrate the potential of the methodological framework starting from an evaluation study of a training programme on organic agriculture in the Western Highlands of Guatemala (Vaessen and De Groot, 2004). The study was implemented in 2001 and served the main purpose of evaluating programme outcome and impact. For the present, I have modified some of its features in order to increase its relevance for the discussion at hand. I assume that, instead of being only a summative ex post study, the evaluation study would serve the additional purpose of assisting decision makers in redefining the programme for a second phase of implementation. I will limit my discussion to the latter purpose (step 2).

The three most important stakeholder groups involved in the training programme are: indigenous (Mayan) farmers participating in the programme, the managers and implementing staff from ORGANIC (a pseudonym, the organization which implemented the original programme) and finally the managers of an integrated rural development programme (IRDP) which financed ORGANIC. The training programme involved two main components. The first component concerned the organization of training workshops of two to three days every two months (for a period of three years) on an experimental farm in the central part of the territory. Participants were selected on the basis of their commitment to attend the courses and apply the knowledge on their own farms. The second component was the provision of technical assistance by ORGANIC staff to participant farmers in between the courses to assist them in the application of organic practices in the field.[6] Besides participating in the courses and applying the acquired knowledge in their own farms, farmers were stimulated to organize themselves in groups for future exchange of ideas and cooperation on farming techniques. In addition, they were stimulated to share their knowledge with neighbouring farmers in their communities, thereby creating a diffusion effect of the innovations. The main objective of the programme was to improve agricultural production and hence the living conditions of participating farm households. The programme theory is depicted in Figure 1.



Note: Underlined variables (and in italic) are examples of external variables that influence specific links in the programme theory.

Figure 1. Programme Theory Training Programme in Organic Agriculture

STEP 2: Programme Improvement towards the Future

Definition of alternatives The findings of the summative evaluation study (Vaessen and De Groot, 2004) provide the basis for defining the alternative programme strategies. In principle, the alternatives are to be determined by the evaluator (to safeguard that the alternatives cover the array of weaknesses detected in the original strategy) in consultation with other stakeholder groups. In this case, five alternatives have been defined which adequately represent different (partial) remedies to the flaws detected in the original programme. For reasons of scope, I will not discuss in detail the relationship between detected weaknesses and the definition of alternatives. Looking at Figure 1, some of the logic underlying the alternatives can be easily traced to specific parts of the programme theory. For example, alternatives 5 and 6 are assumed to reinforce the link 'THEN 7' in the programme theory. The list of alternatives (including the original programme) is the following:

1. original programme;
2. maintaining more or less the same management structure and content but with more emphasis on labour-saving techniques;
3. abandoning the original management structure and content; the new programme focuses on a balance between conventional (labour-saving) and organic techniques;
4. 3 plus credit provision;
5. 3 plus assistance in organization building;
6. 3 plus assistance in organization building and credit provision.

Elicitation of criteria and relative importance of criteria Per stakeholder group, the criteria for evaluating the six programme alternatives (including the original programme) are reported in Table 1. In addition, Table 1 shows the relative importance of the different criteria. To keep things simple, I will restrict our analysis to contrasting the preferences of IRDP decision makers with those of farmers (assuming that the preferences of ORGANIC staff are adequately represented by the original programme, which they designed). In addition, I assume that there are no significant differences within the stakeholder groups with respect to how the programme is perceived and valued. In reality, differences in values regarding a programme can also be used to determine the definition of groups rather than the generic categories of programme beneficiaries, programme managers, etc.

Given the difficulties surrounding the determination of the relative importance of criteria, a ranking approach supported by a simple semantic scale is used. As shown in the table, for IRDP decision makers, total costs, the effect on yields and the creation of farmer organizations are the most important criteria. In contrast, for farmers labour input is the primary criterion. Programme costs and the creation of farmer organizations are considered not to be important at all.

Table 1. Evaluation Criteria and Relative Importance of Evaluation Criteria for Two Stakeholder Groups: IRDP Decision Makers and Farmers

	Total costs	*Yields per unit land*	*Soil quality*	*Soil and water conservation*	*Labour use*	*Reliance on external inputs*	*Conservation of Mayan farming tradition*	*Working with farmer organizations*
IRDP	1	1	2	2	3	2	2	1
farmers	n.r.	2	3	3	1	3	3	n.r.

1 = most important; 2 = relevant but not very important; 3 = least important; n.r.= not relevant (in the case of farmers).

Assessment of the alternatives on the basis of the criteria Table 2 presents the rankings of the alternatives per criterion. Given the imprecise nature of many of the expected effects of the alternatives on the criteria, an ordinal ranking approach is used to represent the expected performance of the alternatives on the criteria. Quite elaborate reasoning (guided by the programme theory) may lie behind a certain ranking. For example, the ranking regarding the criterion 'conservation of the Mayan farming tradition' depends on explicit assumptions about expected adoption levels, expected diffusion levels and specific content of the training programme.[7] This explains why the original programme strategy has such a low ranking. Although the original programme implies the richest variety in traditional Mayan techniques, expected adoption and diffusion levels are much lower than for the other alternatives.

Application of an MCDA technique as a basis for a process of deliberation towards an improved programme strategy In this subsection I will illustrate how the differences in values between IRDP decision makers and farmers will lead to different rankings of programme alternatives. Subsequently, these different rankings provide the basis for a process of deliberation, either among IRDP decision makers or including representatives from the farmers.

There are a number of MCDA techniques available that can be used to transform the ordinal rankings of criteria (Table 1) and expected performances of alternatives on criteria (Table 2) into overall rankings of programme alternatives. In this example, for illustrative purposes I have used a variation of the linear model (see Belton and Stewart, 2002: chs 5 and 6). Basically, the approach entails a transformation of ordinal rankings to cardinal scores. In the case of performance scores Belton and Stewart argue

> . . . that if the number of alternatives is small, then it may be possible to rank order all alternatives in terms of the criterion under consideration. Each rank position might then in this case be represented as a 'category' . . . [on an ordinal scale], and the estimation of values corresponding to each category becomes in effect a direct rating of alternatives. (2002: 168)

Table 2. Expected Performance of the Alternatives per Criterion

A	*Costs*	*Yields*	*Soil quality*	*Soil and water conservation*	*Labour use*	*Reliance on external inputs*	*Conservation of Mayan farming tradition*	*Working with farmer organizations*
1	1	4	5	6	3	2	5	4
2	2	3	4	5	2	1	4	4
3	3	2	3	4	1	3	3	3
4	5	1	2	2	2	4	2	3
5	4	2	3	3	1	3	3	2
6	6	1	1	1	2	4	1	1

The alternatives are ranked according to their perceived performance on the different criteria. 1 = best performance; 2 = second-best, etc. The rankings already take into account that some criteria (costs, labour use and reliance on external inputs) are criteria to be minimized (e.g. the alternative that implies the lowest costs receives the highest ranking = 1) while the rest of the criteria are to be maximized.

In other words, on the basis of Table 2 one additional step in the analysis is needed to determine the 'distances' between the rank positions and to generate more precise estimations of the expected performance of the alternatives (for example, on the basis of expert interviews). Subsequently, the scores are normalized and ready for being matched with the scores on weights. In the case of weights we used the simple rank sum approach to create cardinal weights (Stillwell et al., 1987). After normalization, these scores, which are crude estimates of stakeholder value positions, constitute the starting points for further sensitivity analysis (Belton and Stewart, 2002: 142). By combining the normalized scores on weight and performance, aggregate preference scores for each alternative can be generated which will result in an overall ranking of alternatives. After having produced the overall ranking the evaluator will perform a sensitivity analysis on performance scores to test to what extent changes in performance scores would affect the overall ranking. Those scores that have a relatively high influence on the overall ranking require further scrutiny by the evaluator (i.e. by reviewing the assessment pertaining to a particular performance score).

While this approach is useful for illustrative purposes (and might also be considered for structuring stakeholder choice processes in practice) there are specialized techniques available which in many cases are better suited to deal with this type of data. An example of an alternative technique is MACBETH, which has a standardized procedure for transforming ordinal rankings into cardinal ratings (see Bana e Costa and Vansnick, 1994). Another approach called ARGUS (De Keyser and Peeters, 1994) maintains the ordinal nature of both weights and evaluation scores up until the final ranking. Both approaches are quite accessible and the main ideas underlying these approaches can be quite easily explained to decision makers. In addition, software is available on the market to facilitate calculations and provide graphic displays to facilitate the discussion.

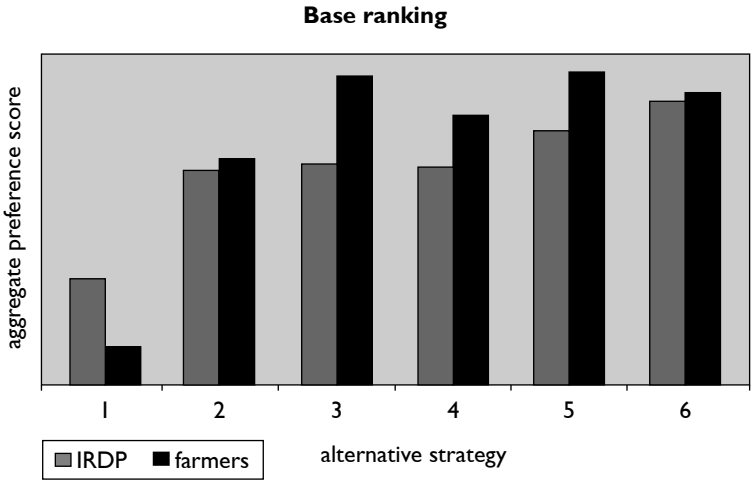


Figure 2. Base Ranking

Returning to our example, Figure 2 shows the base ranking of alternatives for the two stakeholder groups. This ranking by no means represents the end of the evaluation process. In fact it is the beginning of a process of deliberation which involves a combination of human judgement and adaptations of the model. The ideal situation would involve some kind of workshop involving decision makers (and other stakeholders) guided by the evaluator, in which the assumptions (e.g. regarding stakeholder preferences) underlying the model would change as the deliberation moves along. In this way, one would create a constructive process of deliberation, progressively moving towards a shared conviction of the best strategy to be undertaken. Alternatively, if the evaluator is barred from the decision-making process, or if for some other reason such a deliberation process cannot take place, it is the evaluator's task to inform decision makers as best as possible of the consequences of different assumptions regarding preferences (and expected performance) on the ranking of programme strategies.

Let me briefly illustrate the kind of deliberation that could arise after presenting the base ranking (Figure 2) and changing the underlying assumptions of the model (Figure 3). In Figure 2 we can see that for both groups the original strategy (alternative 1) represents the least desirable option. For IRDP alternative 6 is the most desirable option given the high potential for working with farmer organizations, the high adoption rates and high yield effects. These benefits apparently sufficiently offset the high costs associated with alternative 6 (as long as the high costs do not pose too high an obstacle in the eyes of IRDP decision makers). For the farmers, alternatives 3 to 6 all represent a significant improvement in comparison to the original strategy or its close derivative (alternative 2). Alternatives 3 and 5 are the most desirable options. The provision of credit (implied in alternatives 4 and 6) tied to the adoption of (physical) practices only pays off if farmers are sufficiently willing to invest heavily in their farm. However, in reality labour is an important

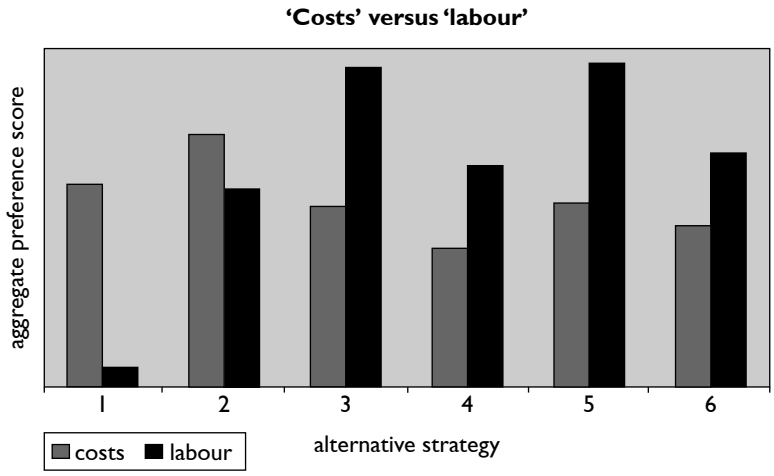


Figure 3. Sensitivity Analysis: Enhanced Weight of Costs (IRDP) and Labour Use (Farmers)

bottleneck. Given the overall importance of non-farm activities, often making up more than half of the household income, for the majority of farmers there are clear limits in terms of how much labour they are willing to invest in agriculture without endangering other income activities. Given current preferences and assumptions regarding expected performance perhaps alternative 5 would constitute a good compromise between farmers and IRDP.

What if we change the assumptions regarding stakeholder preferences? Let us suppose that the cost criterion for IRDP and the labour use criterion for the farmers become more important vis-à-vis other criteria. With respect to the farmers, Figure 3 shows that the ranking does not change significantly but the preference for alternatives 3 and 5 is now more marked. In contrast, from the point of view of IRDP the ranking substantially changes. IRDP decision makers now face a different trade-off than before. Alternative 2 becomes the most attractive one, with alternative 1 (the original strategy and the cheapest option) ending second. If IRDP chooses alternative 2, it is able to offer the farmers a substantially better alternative than the original strategy, though not exactly the most attractive option from the farmers' point of view. On the other hand, alternative 1, the second most attractive strategy for IRDP, is out of the question since it is less attractive for both IRDP and the farmers in comparison to alternative 2. Perhaps alternative 2 would constitute the best compromise solution for the two parties. Choosing alternatives 3 or 5 would content the farmers but at a substantial cost for IRDP.

This reasoning supported by the MCDA model in principle could be extended to include additional value positions. The discussion illustrates the supportive nature of the MCDA model, not substituting but complementing the deliberation process among decision makers.

Conclusions

The purpose of this article has been to illustrate the utility of a methodological framework encompassing elements of PTE and MCDA. The attractiveness of PTE approaches lies (among other things) in the aspect of making explicit the consecutive assumptions underlying social programmes. Such a reconstruction of the programme theory is quite useful as a structure to further evaluation activities in order to determine (retrospectively) the programme's merit and worth as well as processes of strategy clarification, consensus building and improving the programme towards the future. Regarding the latter purpose, MCDA can structure the process of programme strategy improvement by means of systematically showing the advantages and disadvantages of alternative strategies. PTE reinforces MCDA in two ways, i.e. by providing a basis for the definition of meaningful alternatives to the original programme and by strengthening the assessment of alternatives on criteria.

Stakeholder value positions can be systematically addressed in such a process. Participatory evaluation approaches that rely on intensive (group) interactions between evaluators and stakeholders can be criticized because of their costliness and the potential distortions of group processes. Rather than presupposing a participatory process based on intensive interaction between all stakeholders, in the framework that was presented a systematic values inquiry (e.g. based on surveys) can be used to elicit stakeholder preferences, focusing on the criteria that are deemed important in the different valuations of the programme. In such a way, the values of large groups of stakeholders can be incorporated in the evaluation process. Subsequently, MCDA can be used to show how different stakeholder value positions translate into different preferences for programme alternatives.

To some extent the framework chooses scope (i.e. the ability to incorporate the values of many stakeholders) over degree of participation (i.e. consultation rather than joint deliberation). This implies a certain 'distance' between programme management and other stakeholders (e.g. staff and beneficiaries) which can be advantageous to the former. In practice, programme decision makers are often reluctant to share their decision-making power with other stakeholders. The creation of an 'artificial' decision-making forum with multiple stakeholders to decide on a programme's future risks ending up being disarticulated from the decision makers' real concerns, and as a result might be ignored at the moment of taking decisions. Alternatively, in the proposed framework evaluators contrast the evaluative conclusions drawn from the point of view of decision makers' values with the conclusions drawn on the basis of other stakeholders' values, which would probably have quite an illuminating effect on decision makers. In practice, they often do not have at their disposal this type of information. As discussed earlier, evaluators and decision makers are left with the option to include or exclude other stakeholders in the final deliberation process which is structured by the different rankings of alternatives reflecting stakeholder value positions.

Notwithstanding the advantages of the proposed framework, the most notable shortcoming of the framework is that it improves the integration of stakeholder values into the evaluation process in a way that is beneficial to decision makers but

not necessarily to other groups of stakeholders. Apart from the final deliberation process which might include representatives from other groups, participation is limited to different forms of stakeholder consultation (e.g. to capture preferences regarding criteria). Other potential disadvantages concern the level of acceptance of the framework by decision makers. The evaluator should devote some attention to explaining that the MCDA process (step 2 of the framework) should be perceived as a supportive and not as a prescriptive tool, as in some cases it could be perceived as taking the decision out of the hands of the decision makers. In addition, the evaluator should spend some time on choosing the appropriate MCDA technique. A very complicated model, while being more realistic than a simple one, might prove to be too costly and more easily rejected by decision makers. The particular MCDA technique in this article was chosen for illustrative purposes. In practice, the choice of a particular MCDA technique should be based on a balance between realism (e.g. in terms of the assumptions regarding stakeholder preferences, being able to show 'real' trade-offs) and applicability, taking into account aspects like the amount and type of data available, stakeholder willingness to cooperate, available computer hardware and software, time and resource constraints and conditions posed by programme decision makers. If both issues are treated carefully then the framework might be quite successful in meeting the needs of decision makers and possibly other stakeholders in an evaluation process that is both enlightening and inclusive, bringing together the evaluation study, the decision-making process and stakeholder participation.

Notes

The author would like to thank Frans Leeuw, Robrecht Renard and three anonymous reviewers for their comments on the paper and Johan Springael for the interesting discussions on multicriteria techniques. This article is based on a paper presented at the 6th Conference of the European Evaluation Society held in Berlin, in October 2004. The author would like to thank participants for their useful feedback. Any remaining errors are the responsibility of the author.

1. The three most important evaluation approaches that fall under the banner of theory-oriented evaluation are: theory-driven evaluation (Chen and Rossi), theory-based evaluation (Weiss) and realistic evaluation (Pawson and Tilley).
2. Cooke (2001) discusses four of the most widely cited problems of group processes: risky shift, Abilene paradox, groupthink and coercive persuasion.
3. For example, there might be a trade-off between the costs of a programme (which need to be controlled) and its potential impact (which should be optimized), i.e. a higher positive impact implying higher costs.
4. In case of the former, compensation of a low score on a criterion can be compensated by a higher score on another. In contrast, low scores on non-compensatory criteria cannot be compensated by higher scores on other criteria. In such cases, alternatives have to meet certain threshold levels in order to be still considered as viable alternatives. In practice, non-compensatory criteria could constitute the basis for a set of general directives which should be met by each alternative in order to be considered for further analysis and ranking based on the (remaining) compensatory criteria.

5. The generic form of the linear additive model is the following:

$$V(a) = \sum_{i=1}^{m} w_i v_i(a)$$

where $V(a)$ is the overall value associated with alternative a, w_i is the weight of criterion i and $v_i(a)$ is the expected performance of alternative a on criterion i.

6. The main topics that were covered by the programme can be roughly divided into two categories: physical practices and cultural practices. Physical practices involve the use of knowledge, labour and sometimes capital in order to be implemented (and maintained). Examples are the construction of sties, latrines (for human manure collection) and soil conservation measures like windshields, ditches and terraces. Cultural practices basically concern a change of habit or technique. The only essential input is knowledge, though sometimes additional labour might be required. Examples include the substitution of organic 'homemade' fertilizers for 'chemical' purchased fertilizers, ploughing along the contour lines of the plot, crop diversification and collecting instead of burning crop residues.
7. Alternatives 1 and 2 are not so different from each other in terms of content of the training programme. In contrast, alternatives 3 to 6 are substantially different from the first two.

References

Bana e Costa, C. A. and J. C. Vansnick (1994) 'MACBETH: An Interactive Path towards the Construction of Cardinal Value Functions', *International Transactions in Operational Research* 1(4): 489–500.

Belton, V. (1990) 'Multiple Criteria Decision Analysis: Practically the Only Way to Choose', in L. C. Hendry and R. W. Eglese (eds) *Operational Research Tutorial Papers 1990*, pp. 53–101. Birmingham: Operational Research Society.

Belton, V. and T. J. Stewart (2002) *Multiple Criteria Decision Analysis: An Integrated Approach*. Dordrecht: Kluwer Academic Publishers.

Bryk, A. S., ed. (1983) *Stakeholder-Based Evaluation.* New Directions for Program Evaluation, 17. San Francisco, CA: Jossey-Bass.

Chen, H. T. (1990) *Theory-Driven Evaluations*. Newbury Park, CA: SAGE.

Chen, H. T. (2004) *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness.* Thousand Oaks, CA: SAGE.

Chen, H. T. and P. H. Rossi (1980) 'The Multi-Goal, Theory-Driven Approach to Evaluation: A Model Linking Basic and Applied Social Science', *Social Forces* 59 (Sept.): 106–22.

Christie, C. A. and M. C. Alkin (2003) 'The User-Oriented Evaluator's Role in Formulating a Program Theory: Using a Theory-Driven Approach', *American Journal of Evaluation* 24(3): 373–85.

Cooke, B. (2001) 'The Social Psychological Limits of Participation?', in B. Cooke and U. Kothari (eds) *Participation: The New Tyranny?*, pp. 102–21. London: Zed Books.

Cousins, J. B. and E. Whitmore (1998) 'Framing Participatory Evaluation', in E. Whitmore (ed.) *Understanding and Practicing Participatory Evaluation*, pp. 5–23. New Directions for Evaluation, 80. San Francisco, CA: Jossey-Bass.

De Keyser, W. S. M. and P. H. M. Peeters (1994) 'ARGUS: A New Multiple Criteria Method Based on the General Idea of Outranking', in M. Paruccini (ed.) *Applying Multiple Criteria Aid for Decision to Environmental Management*, pp. 263–78. Dordrecht: Kluwer Academic Publishers.

Dodgson, J., M. Spackman, A. Pearman and L. Phillips (2001) *Multi Criteria Analysis: A Manual*. London: Department for Transport, Local Government and the Regions.

Eden, C. and F. Ackermann (1998) *Making Strategy: The Journey of Strategic Management.* London: SAGE.

Ehren, M. C. M., F. L. Leeuw and J. Scheerens (2005) 'On the Impact of the Dutch Educational Supervision Act: Analyzing Assumptions Concerning the Inspection of Primary Education', *American Journal of Evaluation* 26(1): 60–76.

Henry, G. T. (2002) 'Choosing Criteria to Judge Program Success: A Values Inquiry', *Evaluation* 8(2): 182–204.

House, E. R. (1995) 'Putting Things Together Coherently: Logic and Justice', in D. M. Fournier (ed.) *Reasoning in Evaluation: Inferential Links and Leaps*, pp. 33–48. New Directions for Evaluation, 68. San Francisco, CA: Jossey-Bass.

House, E. R. and K. R. Howe (1999) *Values in Evaluation and Social Research*. Thousand Oaks, CA: SAGE.

Keeney, R. L. (1992) *Value-Focused Thinking: A Path to Creative Decision Making*. Cambridge, MA: Harvard University Press.

Keeney, R. L. and H. Raiffa (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

Leeuw, F. L. (2003) 'Reconstructing Program Theories: Methods Available and Problems to be Solved', *American Journal of Evaluation* 24(1): 5–20.

Mark, M. M., G. T. Henry and G. Julnes (1999) 'Toward an Integrative Framework for Evaluation Practice', *American Journal of Evaluation* 20(2): 177–98.

Pawson, R. and N. Tilley (1997) *Realistic Evaluation.* Thousand Oaks, CA: SAGE.

Renger, R. and B. Bourdeau (2004) 'Strategies for Values Inquiry: An Exploratory Case Study', *American Journal of Evaluation* 25(1): 39–49.

Roy, B. (1996) *Multicriteria Methodology for Decision Aiding*. Dordrecht: Kluwer Academic Publishers.

Scriven, M. (1991) *Evaluation Thesaurus*. Newbury Park, CA: SAGE.

Stame, N. (2004) 'Theory-Based Evaluation and Types of Complexity', *Evaluation* 10(1): 58–76.

Stillwell, W. G., D. Von Winterfeldt and R. S. John (1987) 'Comparing Hierarchical and Nonhierarchical Weighting Methods for Eliciting Multiattribute Value Models', *Management Science* 33(4): 442–50.

Stufflebeam, D. L. (2001) *Evaluation Models*. New Directions for Evaluation, 89. San Francisco, CA: Jossey-Bass.

Trochim, W. M. K. (1989) 'An Introduction to Concept Mapping for Planning and Evaluation', *Evaluation and Program Planning* 12(1): 1–16.

Vaessen, J. and J. De Groot (2004) 'Evaluating Training Projects on Low External Input Agriculture: Lessons from Guatemala', *Agricultural Research and Extension Network Papers* 139. London: Overseas Development Institute.

Weiss, C. H. (1998) *Evaluation: Methods for Studying Programs and Policies*. Upper Saddle River, NJ: Prentice Hall.

Wholey, J. S. (1987) 'Evaluability Assessment: Developing Program Theory', in L. Bickman (ed.) *Using Program Theory in Evaluation*, pp. 77–92. New Directions for Program Evaluation, 33. San Francisco, CA: Jossey-Bass.

CHAPTER 7

Reflections on the practice and methodology of impact evaluation in development

7.1. Introduction

This dissertation has illustrated several important conceptual and methodological challenges in impact evaluation, which can roughly be classified under the banners of delimitation, attribution versus explanation and implementation challenges. As was done in Chapter 1, in this Chapter I will continue using REs as a focal point to reflect on the practice and methods of impact evaluation.

The controversy surrounding the promotion and application of REs in development has led to a sense of polarization in the development policy and evaluation community. As some proponents claim epistemological supremacy of REs (with respect to attribution) the counter reaction among others has been rejection. Needless to say, such extreme positions are counterproductive to reaching a goal that is commonly endorsed: to learn more about what works and why in development. Polarization leads to 'argument mining', with proponents bringing up (arguably valid) arguments in defense of REs while adversaries pick their favorite (and arguably valid) arguments against REs. Clearly, this is not the way forward on the path to knowledge growth. If one explores the growing body of evidence in development generated through REs, one cannot deny the positive (though divergent) direct and indirect benefits to knowledge generation about what works. At the same time, as acknowledged by most scholars in the literature, the applicability of REs is limited to certain contexts (see Chapter 1). Moreover, as illustrated in this dissertation, the range of potential challenges (including different questions) to be addressed in an empirical exercise of assessing what works and why in development intervention is too broad to be adequately addressed by REs *only*. By presenting a diverse array of methodological and conceptual perspectives on impact evaluation this dissertation has illustrated potential benefits but also limitations as well as complementary perspectives to REs. The implicit lesson is twofold. First of all, the question 'to randomize or not to randomize' is overrated in the current debate. Other challenges demand the attention of policymakers and evaluation researchers alike. Second, 'do not throw out the baby with the bath water'. There is a risk that the current popularity of REs in certain policy circles might lead to a backlash. Too high expectations of REs may quicken its demise.[1] There are several reasons why it is important to avoid this.

1 A famous example of the rise and fall of REs in evaluation concerns the 1960s and 1970s in the US, sometimes referred to as the 'golden age of evaluation'. The application of REs rose in 1960s as Lyndon B. Johnson's 'War on Poverty' and 'Great Society' programs were to be evaluated systematically using REs (and quasi-experiments). However, due to several factors, by the end of the 1970s REs had lost much ground. In part, the demise was due to methodological problems (see Oakley, 2000; Leeuw, 2009).

7.2. Some lessons for REs and impact evaluation from the perspective of a non-randomista

Starting with the latter issue in the previous paragraph, I will gradually introduce some of the challenges in impact evaluation which are currently insufficiently on the radar of policymakers and/or researchers and evaluators, (in part) due to the randomization debate.

First of all, it is important to get the applicability picture in order. Bamberger and White (2007) are explicit in pointing out the limitations in applicability of REs in development. Reasons are manifold. Proponents of REs have convincingly argued that barriers of costs, ethical concerns, practical organization and quality control can be overcome in some cases. For example, designs with clear geographical separations between participant groups and control groups can reduce unintended behavioral responses,[2] and allocation principles such as lotteries can enhance collaboration. However, whether solutions work depends on context. Therefore, context should (in part) determine methodological design choices. More serious restrictions to applicability of REs remain. Many of the interventions funded by bilateral and multilateral donors are not amenable to REs. New aid instruments such as General Budget Support or Sector Budget Support, various forms of institutional support, administrative reform procedures, full-coverage interventions such as laws and macroeconomic policies are by and large not amenable to REs. In addition, macro- or sector-level strategies or any type of comprehensive program, as a whole are not suitable to be evaluated with REs, although specific intervention components may be (Cohen and Easterly, 2009). An important barrier to RE application is that policy-driven impact evaluations, i.e. impact evaluations commissioned by funding or implementing agencies, often favor scope over depth. With a limited budget, evaluators are forced to develop plausible statements on impact over a range of intervention activities in divergent contexts. In such cases, a tension between accountability (implying a coverage of most or all of the funded activities) and learning (giving more attention to one particular intervention or type of intervention) may exist.[3] Even if the applicability range of REs broadens in the future due to new experiences (e.g. experiments at the institutional level), aspects such as implementation challenges (e.g. the ethics, logistics of doing REs but also the willingness among policy makers, implementing agencies and other stakeholders to transform intervention formats into randomized experiments), budget allocation priorities (e.g. scope versus depth) and other limitations will inevitably restrict RE-type evaluations to being a minority practice in the impact evaluation business (see below).

The realization that there are inherent limitations in applicability of REs should not be an argument against the further promotion of REs. Nor should the range of threats to validity that can affect the analytical strength of an RE (see Chapter 4)

2 Although in such a case one has to check whether this might lead to selection effects due to observable or unobservable differences between participants and control groups (i.e. does the assumption that populations from different geographical regions are similar hold?).

3 See CGD (2006) for a broader discussion on the incentive problems that explain the limited investments of donors in REs.

constitute an argument against REs.[4] I briefly discuss a few points of interest here. First, experimentation should not be perceived as an anomaly in development intervention. The policy field of education in the Netherlands is in this aspect not much different from the policy field of agricultural technologies in Latin America. From the former perspective, Borghans (2009) argues that teachers continuously experiment with teaching methods, whereas the target group, students, are continuously receiving different 'treatments', for example simply by going to different schools. Similarly, in Latin America (or elsewhere), farmers and development agencies alike continuously experiment with new techniques and new intervention activities to achieve particular desired objectives.[5] If one considers a specific sample of farmers or agencies, at any given point in time, one can identify a range of similar yet different practices aimed at boosting productivity per unit of land given particular resource constraints. *Observation over time* in combination with a *systematic variation* in respectively the application or promotion of certain techniques, are common behavior among farmers, cooperatives, NGOs or state agencies. A RE is a more systematic approach to experimentation than what most stakeholders are used to. As argued by Borghans (2009) from the perspective of education in the Netherlands, if we take this experimenting attitude as a given, then one would expect it to become more acceptable and desirable for decision makers and target groups to organize and participate in more systematic experiments, such as REs. In the context of development interventions Banerjee and Duflo (2008) talk about developing long-term working relationships between researchers and development agencies. In such conditions it is possible to show that a RE is much more efficient in proving whether something works than much of the haphazard experimentation that goes on in the daily practice of target groups and implementing agencies. And it is in such cases that the comparative advantage of a RE can truly blossom: it magnifies heterogeneity in 'treatment' by introducing a clear comparative perspective between those with and without an intervention and it reduces bias in estimation of effects through the principle of randomization. These are two powerful features that help us to identify efficiently what works for a particular group in a particular context.

Now, in order to go from a specific setting to a more generalizable conclusion about effectiveness we need theory. To illustrate this, consider the adoption of certain organic farming practices. We know from research that knowledge and labor substitute for capital. In other words, less inputs are bought on the market in exchange for an increased input of labor and knowledge. Peasant economics (e.g. Ellis, 1993) and the diffusion of innovation literature (e.g. Rogers, 2003) teach us that the opportunity costs of labor (next to a range of other variables, and contingent

4 In fact, the systematic identification and discussion in the literature of threats to the (internal) validity of REs strengthens the scientific basis of the methodology. As such, it should not be considered as an argument against REs. However, caution is in order when talking about the possibility of mainstreaming REs in the intervention cycle of an agency. I remember a recent debate in a multilateral donor organization where there was talk of mainstreaming REs in the design of projects. It is very likely that such experiments will not benefit from the same level of quality control present in most (research-driven) RE studies which feature so prominently in the literature. Instead, situations such as illustrated in Chapter 4 may occur where the analytical potential of REs is severely damaged by implementation issues and lack of quality control.

5 However, smallholder farmers are often risk averse and tend to experiment on a small scale. Once a particular new practice has demonstrated its pay-off, experimentation may lead to broader adoption.

upon among other things the type of crop) is an important explanatory variable of smallholder adoption behavior. With this information in mind we can test whether this theory holds for a specific farmer population or region. For example, we may set up a clustered RE with different samples in several regions which mainly differ in the opportunity costs of labor yet are similar in other potential explanatory variables (as derived from theory). While this example may not be watertight, it can prove to be very informative on the effectiveness of certain policy instruments (e.g. subsidies, training) in promoting the adoption of organic practices and at the same time test whether the opportunity costs of labor is a decisive factor in adoption behavior. Of course, the stratification may be a little bit more complex as more explanatory variables (derived from theory) may determine the setup of the series of REs. The core idea is that in this way REs may more effectively contribute to existing theories on the determinants of adoption processes. In general, it shows that theory may be rigorously tested by means of REs and that theory itself can be a guideline for determining how to set up a series of REs (see also Cohen and Easterly, 2009, for a discussion on the theory-testing potential of REs). Limitations to external validity remain, especially with respect to scaling-up effects. Nevertheless, a theory-driven RE is potentially stronger on external validity than a 'theory-empty' RE. As commented by Deaton (2009) and also Banerjee and Duflo (2008) the number of theory-informed and theory-testing REs is on the increase. This is important as REs without a basis in theory are prone to be weaker not only in terms of external validity, but also the internal validity and construct validity of findings.

The biggest gains from this growing attention for rigorous impact evaluation, this 'new push for objective evidence' as one may call it, are not to be found in the growing body of REs but rather in its spin-offs. Whereas development evaluations used to be largely on process issues or output delivery, the paradigm shift in development policy towards more attention for results has strengthened the belief across the developing world that interventions should not be based on hunches, intuitions or ideologies but on evidence on whether an intervention is likely to make a difference in terms of the desired objectives. A randomized experiment in a way is one of the elegant flagships of this new evolution. As a methodological design it has drawn a lot of attention yet it is unlikely to win the war on its own. Several other trends and opportunities can be noted which in part have been strengthened by the 'randomista' movement. These (partial) 'spin-offs'[6] point at other challenges in impact evaluation. First of all, a RE is just one of the designs based on explicit counterfactual analysis. The number of quasi-experimental studies in development has increased sharply over the last ten years or so. For example, where randomization is not possible or appropriate, regression discontinuity analysis may be used instead, as it relies on a different principle for the definition of groups.[7] In addition, as data gathering efforts have increased under the influence of results-based thinking and

6 In fact, it is better to think in terms of an association between growth in REs and other quantitative methodological designs.

7 Comparing over time a participant group beneath a certain threshold value of a particular targeting variable (e.g. distance to road) with a control group just above the threshold value.

new M&E systems, opportunities for ex post statistical matching or multivariate analyses with statistical controls have also markedly increased.

This dissertation has focused on design and implementation issues in impact evaluation, scarcely touching upon the large and expanding body of statistical impact evaluations. Statistical techniques are often used within the framework of (quasi-)experimental designs but more often than not in non-experimental settings. Comparative advantages of (non-experimental) statistical impact evaluations are among other things their broad coverage in terms of the number of individuals, households, districts, and regions encompassed by the data sets. The increased availability of data and the growing capacities (in terms of expertise and technological support) to process and analyze these data have enhanced the opportunities for impact evaluation at aggregate (e.g. regional) levels. The issue of the comparative advantage of REs vis-à-vis non-experimental statistical analyses has been briefly raised in Chapter 1. One particular argument all too often is ignored. As a methodological design requiring active manipulation of an intervention, an RE relies on original data, collected specifically for the purposes of the study. In most cases, the evaluation researchers analyzing the data are also involved in the data collection. In qualitative research one also commonly finds a strong link between data collection and analysis. Not necessarily so in statistical impact evaluation exercises. All too often data analyses are based on data sets constructed by others for other purposes. Econometricians and economists involved in statistical impact evaluations using (mostly) non-experimental data tend to overemphasize 'threats to validity' in the data analysis phase (e.g. specifying the right selection model) and often blatantly ignore (or are ignorant of) any problems or biases that may have arisen in the data collection phase. The severed link between data collection and analysis in many impact evaluation exercises is distressing and merits a higher profile in methodological debates.

Data quality is also a concern in the literature on mixed methods in impact evaluation.[8] Previously, I already underlined the importance of mixed methods from the perspective of comparative advantages of methods. Two other reasons make this area of research particularly relevant to policymakers and evaluation researchers. First of all, practically all evaluation work is multi-method in nature. This is most clearly visible in large program or portfolio evaluations where approach papers and evaluation matrices specify the range of data collection and analysis tools used to analyze specific questions. A second reason is that the majority of impact evaluations take place under less than ideal circumstances (see Bamberger et al., 2009). Often evaluation researchers are not present in the design phase of interventions; they are called in when an intervention is already in the implementation phase or after completion. Consequently, evaluators often have to resort to baseline reconstruction, secondary data or ex post data only. In addition, budgets often do not permit large sample sizes or elaborate designs with multiple group comparisons. Pressures on scope further limit the chances of evaluators to set up for example a solid quasi-experiment. Time pressures may force evaluators to do the work in

8 For examples see Leeuw and Vaessen (2009) or Bamberger et al. (2009). Karlan (2009) argues the case for more mixed method research in an RE context.

'quicker and dirtier' ways than what is needed to appropriately address the attribution issue. Methodological options for mixed method evaluations under less than ideal circumstances are existent yet, as argued by one of the principal authors on this subject, Michael Bamberger, much remains to be done in terms of developing new methodologies and standards for mixed method evaluations. Chapter 5 is an example of a mixed method evaluation which tries to incorporate elements of theory, quasi-experimentation and qualitative methods. It is exactly this promising mix of ingredients which constitutes the basis for 'good practice' mixed method evaluations (see for example Bamberger et al., 2004, 2009).

In this dissertation I talked about theory from two perspectives: intervention theory as a framework of guidance and interpretation or, in other terms, as a systematic reconstruction of causal assumptions; and explanatory theory from the social and behavioral sciences which can feed into the former type of theory. It is important to realize that not every intervention can be subject to an elaborate impact evaluation exercise. Yet, in principle every intervention can be assessed on the basis of a reasonably acceptable intervention theory using relatively few resources. Chapter 6 shows how an intervention theory of an existing intervention can be a useful framework for assessing different useful policy alternatives. In practice, the reconstruction and assessment of intervention theories can be organized in a 'resource-light' manner or in a more elaborate expensive manner.[9]

If decision makers want answers to their questions on what works across interventions and across contexts, they may need to change part of their focus. Instead of seeing administrative categories such as projects or country portfolios as the only principal units of analysis (often mainly for accountability purposes) they need to look more at particular policy instruments or intervention types which recur throughout projects (see Chapter 2).[10] Consequently, decision makers may learn to appreciate and invest more in rigorous evaluations of particular intervention types (e.g. using REs) and see how these pieces of evidence connect to the macro picture, i.e. looking at the importance of a particular intervention type and the divergent contexts in which it is implemented. Such a focus on particular intervention types and policy instruments can perfectly coexist with more comprehensive approaches to impact evaluation which start out from programs and portfolios, including a variety of strategies and interventions.

Chapters 2 and 3 provide some discussion on higher-level 'theories' on particular types of interventions at portfolio, country or agency level. We know there are no laws in social sciences (Elster, 2007), yet we also know there patterns of regularity, or demi-regularities, in individual, social and institutional behavior (Pawson, 2009). Identifying and refining such patterns of regularity require a particular way of theorizing about interventions, deconstructing or unpacking interventions into their active ingredients: policy instruments linked to contexts linked to behavioral

9 Recent work on theory-based evaluations based on different methodological modalities (from cheap and relatively superficial but with a broad scope to expensive, in-depth but with a narrower scope) can be found in GEF (2009).

10 For example, even though microcredit may only be a small component of a particular project or country portfolio of a donor agency, throughout its entire portfolio a sizeable portion of the budget may be allocated to supporting microcredit activities (at different levels).

mechanisms (see Chapter 3; see also Pawson, 2006, 2009). It is our mission as development researchers and evaluators to uncover these patterns of (demi-) regularities as they are the building stones of knowledge about what works and why across interventions. Intervention theories with a certain degree of external validity are becoming more and more important in the context of review and synthesis work as well. 3IE is one of the organizations which is currently investing in this type of work (for an example on microcredit see Vaessen et al., 2009).[11] We still have a long way to go. One thing is for sure, we need good empirical impact evaluations - which rely on intervention theory, explanatory theory and multiple methods tailored to a specific context- in order to succeed, eventually . . .

References

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

Bamberger, M., J. Rugh, M. Church and L. Fort (2004) "Shoestring evaluation: Designing impact evaluations under budget, time and data constraints", *American Journal of Evaluation* 25(1), 5-37.

Bamberger, M., V. Rao and M. Woolcock (2009) "Using mixed methods in monitoring and evaluation: Experiences from international development", MIMEO, World Bank, Washington D.C.

Banerjee, A.V. and E. Duflo (2008) "The experimental approach to development economics", *NBER Working Paper* 14467, Cambridge.

Borghans, L. (2009) "Leren over leren", in: R. Rouw, D. Satijn and T. Schokker (eds.) *Bewezen beleid in het onderwijs*, Ministerie van Onderwijs, Cultuur en Wetenschap, The Netherlands.

CGD (2006) *When will we ever learn? Improving lives through impact evaluation*, Report of the Evaluation Gap Working Group, Center for Global Development, Washington, DC.

Cohen, J. and W. Easterly (eds.) (2009) *What works in development? Thinking big and thinking small*, Brookings Institution press, Washington D.C.

Deaton, A. (2009) "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development", *NBER Working Paper* 14690, Cambridge.

Ellis, F. (1988) *Peasant economics: Farm households and agrarian development*, Cambridge University Press, Cambridge.

Elster, J. (2007) *Explaining social behavior - More nuts and bolts for the social sciences*, Cambridge University Press, Cambridge.

GEF (2009) *Review of outcomes to impacts: Practitioners' handbook*, GEF Evaluation Office, Washington D.C.

Karlan, D. (2009) "Thoughts on randomised trials for evaluation of development: presentation to the Cairo evaluation clinic, *Journal of Development Effectiveness* 1(3), 237-242.

Leeuw, F.L. (2009) "On the contemporary history of experimental evaluations and its relevance for policy making", in: O. Rieper, F.L. Leeuw and T. Ling (eds.) *The evidence book: concepts, generation, and use of evidence*, Transaction Publishers, New Brunswick.

Leeuw, F.L. and J. Vaessen (2009) *Impact evaluations and development – NONIE guidance on impact evaluation*, Network of Networks on Impact Evaluation, Washington D.C.

Oakley, A. (2000) *Experiments in knowing: Gender and method in the social sciences*, Polity Press, Cambridge.

Pawson, R. (2006) *Evidence-based policy: A realist perspective*, Sage Publications, London.

Pawson, R. (2009) "Middle range theory and program theory: from practice to provenance", in : J. Vaessen and F.L. Leeuw (eds.) *Mind the gap: perspectives on policy evaluation and the social sciences*, Transaction Publishers, New Brunswick.

11 Comparable to the work by institutions such as the Campbell and the Cochrane Collaboration on health, education, crime and justice and social work, based on empirical work from (mostly) OECD countries.

Rogers, E.M. (2003) *Diffusion of Innovations*, Free Press, New York.
Vaessen, J., F. Leeuw, S. Bonilla, R. Lukach and J. Bastiaensen (2009) "Protocol for synthetic review of the impact of microcredit", *Journal of development effectiveness*, 1(3), 285-294.

SAMENVATTING

Inleiding

In de afgelopen tien jaar hebben verschillende ontwikkelingen in de internationale gemeenschap geleid tot een toenemende aandacht voor de effecten van de ontwikkelingsinterventies. Zowel binnen de internationale gemeenschap van ontwikkelingssamenwerking als daarbuiten heerste het gevoel dat de resultaten van ontwikkelingsinterventies tegenvielen in vergelijking tot de verwachtingen. Bovenal was het gevoel dat de bewijsvoering omtrent de effecten van ontwikkelingsinterventies zeer zwak was. Als gevolg hiervan zien we in de tweede helft van het eerste decennium van de 21e eeuw een sterke toename in de aandacht en financiering voor impactevaluatie. Impactevaluatie kan in brede zin gedefinieerd worden als het geheel van evaluatieve praktijken dat zich richt op het vaststellen van beoogde en niet-beoogde resultaten van beleidsinterventies.

Een bijzondere vorm van impactevaluatie die veel aandacht krijgt betreft zogenaamde gerandomiseerde experimenten. Deze vorm van effectonderzoek wordt door velen beschouwd als de meest rigoureuze techniek om de effecten van beleidsinterventies vast te stellen. De discussie rond de toepassing van gerandomiseerde experimenten in ontwikkelingssamenwerking heeft in de afgelopen paar jaren geleid tot een grote controverse tussen de zogenaamde 'randomistas', voorstanders van gerandomiseerde experimenten, en tegenstanders. Een van de basisprobleemstellingen van dit proefschrift is dat deze controverse een averechts effect heeft op het bereiken van een doelstelling waar iedereen het wel over eens is: het accumuleren van kennis omtrent wat werkt en wat niet, en onder welke omstandigheden.
Hoofdstuk 1 beschrijft het overkoepelende kader van het proefschrift. In dit hoofdstuk worden drie essentiële (categorieën van) uitdagingen in impactevaluatie beschreven. Deze zijn achtereenvolgens: afbakening, attributie versus verklaring, en praktische uitdagingen in de opzet en uitvoering van impactevaluaties. Dit kader is ook de kapstok waaraan alle hoofdstukken van dit proefschrift aan kunnen worden opgehangen.

In de bespreking van de drie uitdagingen wordt voortdurend de koppeling gemaakt met gerandomiseerde experimenten. Dit proefschrift toont aan dat gerandomiseerde experimenten niet geschikt zijn om op al deze uitdagingen een bevredigend antwoord te bieden. Tegelijkertijd wordt ook geargumenteerd dat gerandomiseerde experimenten zeker een prominente plaats verdienen binnen het gamma van methodologieën van effectonderzoek. Het doel van dit proefschrift is het illustreren en bekrachtigen van de volgende stellingen:

- De vraag van wat werkt en waarom binnen het beleidsveld van ontwikkelingssamenwerking kan niet bevredigend worden beantwoord op basis van een enkele onderzoeksopzet (i.e. gerandomiseerde experimenten);
- Er zijn een aantal factoren die de toepassing van gerandomiseerde experimenteren beperken;

- De kwaliteit van impactevaluaties kan worden verbeterd door deze te kaderen in een beleidstheoretische benadering.

Hoewel de titel van dit proefschrift een algemene beschouwing omtrent bovenstaande punten impliceert, zijn omwille van praktische redenen een aantal keuzes gemaakt omtrent de interventies, beleidsvelden en contexten die worden behandeld in dit boek. Een drietal elementen kunnen worden onderscheiden die een verdere afbakening van dit proefschrift vormgeven. Ten eerste zijn de hoofdstukken 4, 5 en 6 gebaseerd op empirische analyses van beleidsinterventies op het gebied van (duurzame) landbouw gericht op kleine en middelgrote boeren. In hoofdstukken 2 en 3 wordt ook gesproken over interventies op het gebied van landbouw en mileu. Ten tweede is er ook een geografische focus in de empirische analyses. Hoofdstuk 4 is gebaseerd op veldwerk in Nicaragua (en Costa Rica) en hoofdstukken 5 en 6 komen voort uit veldwerk in Guatemala. Het derde element betreft de institutionele context van beleidsevaluatie. Hoofdstukken 2 en 4 gaan over interventies gefinancierd door de *Global Environment Facility*, een globaal fonds dat projecten steunt in ontwikkelingslanden op het gebied van biodiversiteit, klimaatsverandering en andere milieudoelstellingen.

Uitdaging 1: afbakening

Hetgeen wel of niet wordt bestudeerd binnen het kader van een impactevaluatie kan heel erg verschillen naar gelang de aard van het evaluandum, de aard van de effecten die kunnen optreden, als ook verdere keuzes die worden gemaakt omtrent de aspecten die meegenomen moeten worden. In hoofdstuk 1 komen de volgende aspecten aan de orde: het omgaan met uiteenlopende waarden en prioriteiten van de verschillende groepen van betrokkenen bij een evaluatie, de vraag van ‘effecten van wat?’, en de vraag van ‘effecten op wat?’ Deze aspecten komen in verschillende hoofdstukken aan de orde.

Hoofdstuk 2 behandelt het thema van methodologische uitdagingen in het evalueren van de effecten van een portfolio van interventies op het gebied van biodiversiteit. Gegeven de heterogeniteit binnen de portfolio en budgettaire restricties is het belangrijk om te reflecteren over het spanningsveld tussen ‘breedte’ en ‘diepte’ van de analyse. Ter ondersteuning worden de volgende twee vragen gesteld: de ‘effecten van wat’ en de ‘effecten op wat’? Verschillende aspecten van afbakening komen aan de orde. Een belangrijk aspect is niveau van analyse. In de keuze van niveau van analyse kan men een onderscheid maken tussen administratieve categorieën zoals projecten of programma’s en analytische categorieën zoals strategische prioriteiten en beleidsinstrumenten die teruggevonden kunnen worden in meerdere interventies. Niveau van analyse heeft ook betrekking op de effecten. Het maakt nogal een verschil in termen van onderzoeksopzet of men kijkt naar veranderingen op institutioneel en individueel niveau of op het niveau van milieudoelstellingen. De causaliteit tussen beleidsinterventies en dit laatste niveau is veel indirecter en diffuser dan in het geval van individueel en institutioneel gedrag. Hoofdstuk 2 be-

handelt ook de beleidstheoretische evaluatiebenadering als kader voor afbakening en analyse op verschillende niveaus.

Hoofdstuk 3 bespreekt in detail de beleidstheoretische evaluatiebenadering. Met betrekking tot het thema van afbakening gaat het hoofdstuk in op een specifiek kader voor het deconstrueren van veranderingsprocessen naar verschillende typen van mechanismen. Deze mechanismen betreffen achtereenvolgens macro-micro-, micro-micro-, en micro-macro-veranderingsprocessen. Het hoofdstuk laat zien hoe een dergelijke deconstructie nuttig kan zijn in het structureren van de verzameling en analyse van data binnen een interpretatiekader van hoe veranderingsprocessen kunnen verlopen.

Hoofdstuk 4 betreft een analyse van een gerandomiseerd experiment in het kader van een ontwikkelingsinterventie in Nicaragua, Costa Rica en Colombia. Op systematische wijze wordt het potentieel van het experiment om antwoord te geven op vragen omtrent de effectiviteit van monetaire incentives en technische assistentie op landgebruiksveranderingen onderzocht. Gerandomiseerde experimenten zijn normaliter uitermate geschikt om directe effecten ('outcomes') vast te stellen. In dit hoofdstuk wordt ook bekeken in hoeverre het experiment iets kan zeggen over impact in brede zin zoals gedefinieerd door de *OESO*. In het kader van deze oefening worden verschillende beperkingen vastgesteld en concrete suggesties voor additionele vormen van analyse worden besproken. Voorbeelden van effecten buiten het bereik van gerandomiseerde experimenten zijn zogenaamde 'verplaatsingseffecten' (effecten op landgebruik elders buiten de zone van interventie maar wel als gevolg van de interventie) en de duurzaamheid van landgebruiksveranderingen en dieperliggende effecten op milieudoelstellingen.

Hoofdstuk 6 presenteert een methodologisch kader voor impactevaluatie vanuit het oogpunt van ex ante effectonderzoek van beleidsalternatieven. Een belangrijke veronderstelling in dit onderzoek is dat enkel de bewijsvoering rond potentiële effecten onvoldoende is als basis voor toekomstig beleid. Er moet ook rekening worden gehouden met de prioriteiten van verschillende groepen van betrokkenen. Het methodologische kader dat wordt uitgewerkt in dit hoofdstuk maakt het mogelijk om beleidsalternatieven met elkaar te vergelijken op basis van potentiële effecten, daarbij rekening houdend met verschillen in het relatieve belang van effecten voor verschillende groepen van betrokkenen.

Uitdaging 2 : attributie versus verklaring

Gerandomiseerde experimenten maken het mogelijk om de netto-effecten van een interventie vast te stellen met een hoge graad van interne validiteit. Echter, tegelijkertijd zijn er een aantal beperkingen in termen van het helpen verklaren van veranderingsprocessen. Een beleidstheoretische benadering kan een gerandomiseerd experiment op verschillende punten versterken. Hoofdstuk 1 behandelt een aantal aspecten onder dit thema, waaronder een korte bespreking van het attributievraagstuk en de rol van gerandomiseerde experimenten, de mogelijke wisselwerking tussen interne en externe validiteit in effectonderzoek, het belang van beleids-

theorieën, en de 'wet van de comparatieve voordelen' van verschillende onderzoeksmethodologieën.

De belangrijkste doelstelling van hoofdstuk 4 is het bestuderen van het potentieel van een gerandomiseerd experiment in de vaststelling van bepaalde effecten van een concrete interventie in drie landen in Latijns-Amerika. De analyse laat zien hoe verschillende tekortkomingen dit analytisch potentieel in gevaar brengen. Het experiment bestaat uit verschillende facetten. Twee van de drie vergelijkingen tussen interventie- en controlegroepen zijn door de verschillende problemen in het experiment niet meer betrouwbaar. De derde optie, de vergelijking tussen groepen met verschillende modaliteiten van monetaire incentives, biedt daarentegen een betrouwbaar kader voor het analyseren van netto-effecten. Daarnaast laat dit hoofdstuk ook zien hoe het analytische potentieel van een gerandomiseerd experiment versterkt kan worden met additionele methoden en de toepassing van een beleidstheoretische benadering.

Hoofdstukken 3 en 4 laten beiden op hun eigen wijze zien hoe een beleidstheoretische benadering in evaluatie de interne, externe en constructvaliditeit van bevindingen in effectonderzoek kan versterken. In hoofdstuk 3 wordt ingegaan op de vragen van wat een beleidstheorie precies is, hoe deze kan worden gereconstrueerd en hoe deze kan worden toegepast in evaluatie-onderzoek. Het hoofdstuk bespreekt ook hoe een specifieke versie van een beleidstheoretisch kader, gebaseerd op Coleman's *Theory of Social Action*, van nut kan zijn om veranderingsprocessen beter kunnen begrijpen. Dit kader biedt niet alleen een bruikbare een structuur voor de verzameling en analyse van data maar ook een kader van interpretatie ingebed in een theorie van individueel en collectief gedrag.

Hoofdstuk 5 presenteert een impactevaluatie gebaseerd op kwantitatieve en kwalitatieve methoden en ingebed in een theoretisch kader. Het is een voorbeeld van een zogenaamde *shoestring approach*; de evaluatie is gericht op het genereren van maximaal betrouwbare gegevens en bevindingen binnen een context van beperkingen in budget, tijd en data.

Het principe dat verschillende methoden comparatieve voordelen hebben bij bepaalde aspecten of vragen binnen een impactevaluatie wordt ook geïllustreerd in hoofdstuk 6. In het geval van gerandomiseerde experimenten is de zogenaamde 'counterfactual' de situatie waarin er geen interventie plaatsvindt. Echter, in de praktijk zijn beleidsmakers geïnteresseerd in andere 'counterfactuals', namelijk een andere interventie (in plaats van geen interventie). Het methodologische kader dat wordt ontwikkeld in dit hoofdstuk bestaat uit drie bouwstenen: multicriteria-analyse, een inventaris van beleidsprioriteiten onder betrokkenen, en een beleidstheorie rond veranderingsprocessen. Door deze elementen te combineren kunnen beleidsalternatieven systematisch met elkaar worden vergeleken, wat uiteindelijk leidt tot een rangorde van alternatieven. De beleidstheorie speelt hierbij een cruciale rol in de ondersteuning van de ex ante evaluatie van beleidsalternatieven.

Uitdaging 3: praktische uitdagingen in de opzet en uitvoering van impactevaluaties

Kwaliteitsvolle impactevaluatie impliceert goed onderzoek. Huidige debatten rond impactevaluatie benadrukken de rol en keuze van onderzoeksopzet, i.e. gerandomiseerde experimenten. Echter, het genereren van goed onderzoek, in de context van gerandomiseerde experimenten en daarbuiten, hangt vooral af van de mate waarin een aantal praktische uitdagingen in de opzet en uitvoering van onderzoek op een bevredigende wijze kunnen worden aangepakt. In hoofdstuk 1 worden verschillende mogelijke tekortkomingen (of 'dreigingen') besproken die de interne validiteit van bevindingen van gerandomiseerde experimenten in gevaar kunnen brengen. Daarnaast worden een aantal concrete uitdagingen in de opzet en uitvoering van impactevaluaties besproken.

Hoofdstuk 4 analyseert de mogelijke tekortkomingen van een gerandomiseerd experiment die de interne validiteit van de bevindingen in gevaar kunnen brengen. De belangrijkste problemen die systematisch worden onderzocht zijn: bias in selectie, contaminatie, en verschillende onbedoelde gedragsveranderingen binnen de doelgroep. In het algemeen kan men stellen dat differentiatie tussen groepen van boeren binnen de experimentele opzet nogal wat onrust veroorzaakte binnen de doelgroep. Dit valt deels te wijten aan de communicatiestrategie in de beginfase van het project. De problemen bij de opzet en uitvoering van het gerandomiseerde experiment zijn in wezen terug te voeren naar een aantal basisoorzaken die betrekking hebben op de organisatie en kwaliteitscontrole van het experiment. Het project had een zeer goede professionele staf, met o.a. experten in ecologie, zoölogie en agronomie. De stafleden in het veld waren echter geen ervaren onderzoekers. Onderzoekers waren slechts in beperkte mate op bepaalde tijdstippen aanwezig. De staf in het veld had noch de kennis noch de incentives om een gerandomiseerd experiment tot een goed einde te brengen. In dit hoofdstuk worden verschillende praktische aanbevelingen verstrekt ter verbetering van de opzet en uitvoering van gerandomiseerde experimenten in soortgelijke projecten.

Hoofdstuk 5 betreft een impactevaluatie van een trainingsprogramma in organische landbouw in de hooglanden van Guatemala ten tijde (1998-2001) van een herstelperiode na een langdurige burgeroorlog. Gegeven structurele problemen van wantrouwen en taalbarrières (Spaans versus lokale Maya-talen) tussen onderzoekers (inclusief lokale onderzoekers) en boeren, is het onderzoek in sterke mate gebaseerd op het principe van triangulatie van bevindingen uit enquête-onderzoek, semigestructureerde interviews en veldbezoeken. Daarnaast is het onderzoek ingebed in een theoretisch kader gebaseerd op literatuur rond adoptie-gedrag en economisch gedrag van kleine boeren. Het onderzoek is gebaseerd op een simpele quasi-experimentele opzet bestaande uit drie gekoppelde steekproeven: participerende boeren ex ante en ex post en een controlegroep ex post. De reden waarom een controlegroep ex ante ontbreekt heeft te maken met de beperkte omvang van de steekproeven. Gegeven de hoge migratiecijfers (permanent en seizoensgebonden) en andere factoren zou het te moeilijk zijn geweest om in 2001 voldoende boeren uit de controlegroep te vinden die bereid zouden zijn om mee te werken aan het onder-

zoek. Dit probleem deed zich niet voor in de deelnemersgroep gegeven hun band met het project. Hoofdstuk 5 is een illustratief voorbeeld van hoe men bij de opzet en uitvoering van een impactevaluatie zou kunnen inspelen op verschillende praktische problemen teneinde zo betrouwbaar mogelijke resultaten te kunnen genereren in sub-optimale onderzoeksomstandigheden.

Tot slot

In het slothoofdstuk worden de initiële stellingen van dit proefschrift nogmaals kort onder de loep gehouden. Een aantal aspecten komen hierbij naar voren. Allereerst kan men stellen dat de vraag van 'wel of niet gerandomiseerde experimenten' teveel aandacht heeft gekregen in het debat. Tegelijkertijd moet ook niet het tegenovergestelde pad worden bewandeld door deze waardevolle methodologie aan de kant te schuiven. Het slothoofdstuk gaat verder in op de problematiek van beperkte toepasbaarheid van gerandomiseerde experimenten enerzijds en de kracht van een gerandomiseerd experiment anderszijds. In combinatie met een goed theoretisch kader en voortbouwend op de in vele contexten al bestaande interesse voor experimenteren met beleid (in velerlei vormen) verdienen gerandomiseerde experimenten een plaats in impactevaluatie. Een belangrijk indirect effect van het hele debat omtrent gerandomiseerde experimenten betreft de toegenomen aandacht voor kwaliteit in effectonderzoek. In het stijgende aantal impactevaluaties gebaseerd op een gamma van onderzoeksmethodologieën is er duidelijk aandacht voor kernuitdagingen zoals het attributie-vraagstuk en de externe validiteit van bevindingen. Andere belangrijke aspecten zoals de kwaliteit van dataverzameling worden echter onterecht nog steeds onvoldoende belicht.

CURRICULUM VITAE

Jos Vaessen was born on April 19, 1974 in Maasbree, The Netherlands. He studied Rural Development Studies with a specialization in Agrarian Development Economics, at Wageningen University, The Netherlands, where he obtained his Master degree in 1997 (with honors, top 5%). Part of 1997 and 1998 Jos worked in Guatemala and Nicaragua, undertaking household survey work. From 1998 to 2001 he worked at the University of Antwerp as a researcher on rural development issues, with a strong link to Central America. After a year of coordinating survey work at a Dutch consulting company, in 2002 he returned to the University of Antwerp, taking up a research position at the Institute of Development Policy and Management.

Currently, Jos Vaessen is lecturer at the University of Maastricht (since 2008) and researcher at the Institute of Development Policy and Management of the University of Antwerp (since 2002). Over the last twelve years he has worked on research activities in the field rural development, mostly in Central America. In addition, he has been involved in teaching research and evaluation methods, at the University of Antwerp, Maastricht University and other institutions. From the very beginning of his professional career, Jos Vaessen has been involved in policy evaluation work, an orientation which has become more marked over time. His primary research interest has been the study of evaluation theory and methods in the context of (rural) development interventions. Over the years, he has become increasingly involved in external policy advisory work, having worked on a variety of evaluation-related assignments for several bilateral and multilateral development organizations such as the Dutch Ministry of Foreign Affairs, the World Bank and others. Apart from the abovementioned activities, Jos is also active in the wider professional community revolving around policy evaluation.