NONIE

DAC Evaluation Network | Evaluation Cooperation Group | International Organization for Cooperation in Evaluation | UN Evaluation Group

Impact Evaluations and Development

NONIE Guidance on Impact Evaluation

Inputs
»
Activities
»
Outputs
»
Outcomes
»
Impacts

What Is NONIE?

NONIE is a Network of Networks for Impact Evaluation comprised of the Organisation for Economic Co-operation and Development's Development Assistance Committee (OECD/DAC) Evaluation Network, the United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and the International Organization for Cooperation in Evaluation (IOCE)—a network drawn from the regional evaluation associations.

NONIE was formed to promote quality impact evaluation. NONIE fosters a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation. NONIE focuses on impact evaluation and does not attempt to address wider monitoring and evaluation issues.

To this end NONIE aims to—

- Build an international collaborative research effort for high-quality and useful impact evaluations as a means of improving development effectiveness.
- Provide its members with opportunities for learning, collaboration, guidance, and support, leading to commissioning and carrying out impact evaluations.
- Develop a platform of resources to support impact evaluation by member organizations.

www.worldbank.org/ieg/nonie

NONIE

DAC Evaluation Network | Evaluation Cooperation Group | International Organization for Cooperation in Evaluation | UN Evaluation Group

Impact Evaluations and Development

NONIE Guidance on Impact Evaluation

Frans Leeuw
Maastricht University

Jos Vaessen
Maastricht University and University of Antwerp

IOCE

ECG
Evaluation Cooperation Group

DAC Network on Development Evaluation

UNEG
United Nations Evaluation Group

c/o Independent Evaluation Group
1818 H Street, NW
Washington, DC 20433
Internet: www.worldbank.org/ieg/nonie/

This volume is a product of the volume's authors, Frans Leeuw and Jos Vaessen, who were commissioned by NONIE. The findings, interpretations, and conclusions expressed in this volume are those of the authors and do not necessarily reflect the views of NONIE, its members, or other participating agencies. NONIE does not guarantee the accuracy of the data included in this work and accepts no responsibility for any consequence of their use.

Cover: Pakistani girl reading. Photo by Curt Carnemark, courtesy of World Bank Photo Library.

ISBN-10: 1-60244-120-0
ISBN-13: 978-1-60244-120-0


Printed on recycled paper

Contents

vii **Acknowledgments**

ix **Executive Summary**

xix **Introduction**

1 **PART I – METHODOLOGICAL AND CONCEPTUAL ISSUES IN IMPACT EVALUATION**

3 **1 Identify the (type and scope of the) intervention**

3 1.1. The impact evaluation landscape and the scope of impact evaluation

4 1.2. Impact of what?

7 1.3. Impact on what?

9 Key message

11 **2 Agree on what is valued**

11 2.1. Stakeholder values in impact evaluation

12 2.2. Intended versus unintended effects

12 2.3. Short-term versus long-term effects

12 2.4. The sustainability of effects

13 Key message

15 **3 Carefully articulate the theories linking interventions to outcomes**

15 3.1. Seeing interventions as theories: The black box and the contribution problem

15 3.2. Articulating intervention theories on impact

17 3.3. Testing intervention theories on impact

19 Key message

21 **4 Address the attribution problem**

21 4.1. The attribution problem

23 4.2. Quantitative methods addressing the attribution problem

29 4.3. Applicability of quantitative methods for addressing the attribution problem

31 4.4. Other approaches

34 Key message

35 **5 Use a mixed-methods approach: The logic of the comparative advantages of methods**

35 5.1. Different methodologies have comparative advantages in addressing particular concerns and needs

36 5.2. Advantages of combining different methods and sources of evidence

38 5.3. Average effect versus distribution of costs and benefits

39 Key message

41 **6 Build on existing knowledge relevant to the impact of interventions**
43 Key message

45 **PART II – MANAGING IMPACT EVALUATIONS**

47 **7 Determine if an impact evaluation is feasible and worth the cost**
48 Key message

49 **8 Start collecting data early**
49 8.1. Timing of data collection
49 8.2. Data availability
51 8.3. Quality of the data
51 8.4. Dealing with data constraints
52 Key message

53 **9 Front-end planning is important**
53 9.1. Planning tools
53 9.2. Staffing and resources
54 9.3. The balance between independence and collaboration between evaluators and stakeholders
54 9.4. Ethical issues
55 9.5. Norms and standards
56 9.6. Ownership and capacity building
56 Key message

57 **Appendices**
59 1. Examples of diversity in impact evaluation
61 2. The General Elimination Methodology as a basis for causal analysis
63 3. Overview of quantitative techniques of impact evaluation
65 4. Technical aspects of quantitative impact evaluation techniques
69 5. Evaluations using quantitative impact evaluation approaches
71 6. Decision tree for selecting quantitative evaluation designs to deal with selection bias
73 7. Hierarchical modeling and other statistical approaches
75 8. Multi-site evaluation approaches
77 9. Methodological frameworks for assessing the effects of interventions, mainly based on quantitative methods
79 10. Where to find reviews and synthesis studies on mechanisms underlying processes of change
81 11. Evaluations based on qualitative and quantitative descriptive methods
101 12. Further information on review and synthesis approaches in impact evaluation
105 13. Basic education in Ghana
109 14. Hierarchy of quasi-experimental designs
111 15. International experts who contributed to the subgroup documents

113 **Endnotes**

117 **References**

Boxes
7 1.1. "Unpacking" the aid chain
16 3.1. Social funds and government capacity: Competing theories

18 3.2. Social and behavioral mechanisms as heuristics for understanding processes of change and impact
25 4.1. Using propensity scores to select a matched comparison group—The Vietnam Rural Roads Project
33 4.2. Participatory impact monitoring in the context of the poverty reduction strategy process
39 5.1. Brief illustration of the logic of comparative advantages
42 6.1. Narrative review and synthesis study: Targeting and impact of community-based development initiatives
73 A7.1. Impact of the Indonesian financial crisis on the poor: Partial equilibrium modeling and CGE modeling with microsimulation

Figures

xi ES.1. Levels of intervention, programs, and policies and types of impact
xii ES.2. Simple graphic of net impact of an intervention
8 1.1. Levels of intervention, programs, and policies and types of impact
17 3.1. Basic intervention theory of a fictitious small business support project
22 4.1. Graphic display of the net impact of an intervention
29 4.2. Regression discontinuity analysis
66 A4.1. Estimation of the effect of class size with and without the inclusion of a variable correlated with class size
87 A11.1. Final impact assessment triangulation
92 A11.2. Generic representation of a project's theory of change
93 A11.3. Components of impact evaluation framework
96 A11.4. Project outputs and outcomes
99 A11.5. Framework to establish contribution
99 A11.6. Model linking outcome to impact

Tables

5 1.1. Aspects of complication in interventions
6 1.2. Aspects of complexity in interventions
26 4.1. Double difference and other designs
52 8.1. Evaluation scenarios with time, data, and budget constraints
96 A11.1. Project outcome
97 A11.2. Change in key ecological attributes over time
97 A11.3. Current threats to the global environment benefits

Acknowledgments

This Guidance document could not have existed without the numerous contributions of Network of Networks on Impact Evaluation (NONIE) members and others in terms of papers, PowerPoint® presentations, and suggestions.

In particular, this Guidance document builds on two existing draft guidance documents, a document on experimental and quasi-experimental approaches to impact evaluation (NONIE subgroup 1, May 17, 2007) and a document on qualitative approaches to impact evaluation (NONIE subgroup 2, January 9, 2008). A third draft document prepared by NONIE members on the impact evaluation of macroeconomic policies and new aid modalities such as budget support is outside the scope of this Guidance document. The subgroup 1 document was prepared mainly by Howard White and Antonie De Kemp. The subgroup 2 document, which was somewhat broader in content than methodology, was coordinated by Sukai Prom-Jackson. The primary authors were Patricia Rogers, Zenda Ofir, Sukai Prom-Jackson, and Christine Obester. Case studies were prepared by Jocelyn Delarue, Fabrizio Felloni, Divya Nair, Christine Obester, Lee Risby, Patricia Rogers, David Todd, and Rob van den Berg. The development of this document benefited extensively from a reference group of international evaluators.

Whereas the two subgroup documents provided the basis for the current Guidance document, the purpose of the current document was to develop a new structure that could accommodate some of the diversity in perspectives on impact evaluation. In addition, within this new structure, new content was added where necessary to support key points. The process of developing this Guidance was supervised by a steering committee of NONIE members. An external peer reviewer critically assessed the first draft of this document.

The Guidance document represents the views of the authors, who were commissioned by NONIE. Given the fact that perspectives on the definition, scope, and appropriate methods of impact evaluation differ widely among practitioners and other stakeholders, the document should not be taken to represent the agreed positions of all of the individual NONIE members. The network membership and the authors recognize that there is scope to develop the arguments further in several key areas.

We would like to thank all of the above people for their contributions to the process of writing the Guidance document. First, we thank the authors of the subgroup documents for providing building blocks for this document. In addition, we would like to thank the steering committee of this project, Andrew Warner, David Todd, Zenda Ofir, and Henri Jorritsma, for their pertinent suggestions. We also would like to thank Antonie De Kemp for exchanging ideas on design questions. We are grateful to Patricia Rogers, the external peer reviewer, for providing valuable input to this document. Our thanks also go to Victoria Gunnarsson and Andrew Warner from the NONIE secretariat for accompanying us throughout the whole process and providing excellent feedback. Nick York, Howard White, David Todd, Indran Nadoo, and John Mayne provided helpful insights in the final phase of this project. We thank Arup Banerji for drafting the executive summary. Comments from NONIE members were received at the Lisbon European Evaluation Society Conference (October, 2008) and the Cairo Conference on Impact Evaluation (March, 2009). Networks within NONIE, such as the International Organization for Cooperation in Evaluation and the European Evaluation Society, contributed by submitting written comments. Moreover, many individual NONIE members also

sent in their feedback through email. We would like to thank all NONIE members for the stimulating discussions and inputs on impact evaluation.

Finally, within the restricted time available for writing this document, we have tried to combine different complementary perspectives on impact evaluation into an overall framework, in line with our own views on these topics and feedback from the steering committee and others. Though we have not included all perspectives on impact evaluation, an important and quite diverse selection of the thinking and practice on the subject has been incorporated. The result, we hope, represents a balance between coherence, a comprehensive structure of key issues, and diversity. Any remaining errors are our own.

Frans Leeuw
frans.leeuw@maastrichtuniversity.nl
Jos Vaessen
jos.vaessen@maastrichtuniversity.nl

Executive Summary

In international development, impact evaluation is principally concerned with final results of interventions (programs, projects, policy measures, reforms) on the welfare of communities, households, and individuals, including taxpayers and voters. Impact evaluation is one tool within the larger toolkit of monitoring and evaluation (including broad program evaluations, process evaluations, ex ante studies, etc.).

The Network of Networks for Impact Evaluation (NONIE) was established in 2006 to foster more and better impact evaluations by its membership—the evaluation networks of bilateral and multilateral organizations focusing on development issues, as well as networks of developing country evaluators. NONIE's member networks conduct a broad set of evaluations, examining issues such as project and strategy performance, institutional development, and aid effectiveness. But the focus of NONIE is narrower. By sharing methodological approaches and promoting learning by doing on impact evaluations, NONIE aims to promote the use of this more specific approach by its members within their larger portfolio of evaluations. This document, by Frans Leeuw and Jos Vaessen, has been developed to support this focus.[1]

The Guidance document was written by and represents the views of the authors. Given the fact that perspectives on the definition, scope, and appropriate methods of impact evaluation differ widely among practitioners and other stakeholders, the document should not be taken to represent the agreed positions of all of the individual NONIE members.

Why promote impact evaluations? For development practitioners, impact evaluations play a key role in the drive for better evidence on results and development effectiveness. They are particularly well suited to answer important questions about whether development interventions do or do not work, whether they make a difference, and how cost-effective they are. Consequently, they can help ensure that scarce resources are allocated where they can have the most developmental impact.

Although there is debate within the profession about the precise definition of impact evaluation, NONIE's use of the term proceeds from its adoption of the Development Assistance Committee of the Organisation for Economic Co-operation and Development (DAC) definition of impact, as "*the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.*"[2]

Adopting the DAC definition of impact leads to a focus on two underlying premises for impact evaluations:

- *Attribution:* The words *"effects produced by"* in the DAC definition imply an approach to impact evaluation that is about attributing impacts to interventions, rather than just assessing what happened.
- *Counterfactual:* It follows that in most contexts, knowledge about the impacts produced by an intervention requires an attempt to gauge what *would have occurred* in the absence of the intervention and a comparison with *what has occurred* with the intervention implemented.

These two premises do not, however, lead to a determination of a set of analytical methods that is above all others in all situations. In fact, this Guidance note underlines that—

- *No single method is best* for addressing the variety of questions and aspects that might be part of impact evaluations.
- However, depending on the specific questions or objectives of a given impact evaluation, some methods have a *comparative advantage* over others in analyzing a particular question or objective.
- Particular methods or perspectives *complement each other* in providing a more complete "picture" of impact.

The document is structured around nine key issues that provide guidance on conceptualizing, designing, and implementing an impact evaluation:

Methodological guidance:

1. Identify the type and scope of the intervention.
2. Agree on what is valued.
3. Carefully articulate the theories linking interventions to outcomes.
4. Address the attribution problem.
5. Use a mixed-methods approach—the logic of the comparative advantages of methods.
6. Build on existing knowledge relevant to the impact of interventions.

Guidance on managing impact evaluations:

7. Determine if an impact evaluation is feasible and worth the cost.
8. Start collecting data early.
9. Front-end planning is important.

1. Identify the (type and scope of the) intervention

Interventions range along a continuum from single-"strand" initiatives with explicit objectives to complex institutional policies, and the particular type of impact evaluation would be affected by the type and scope of the intervention.

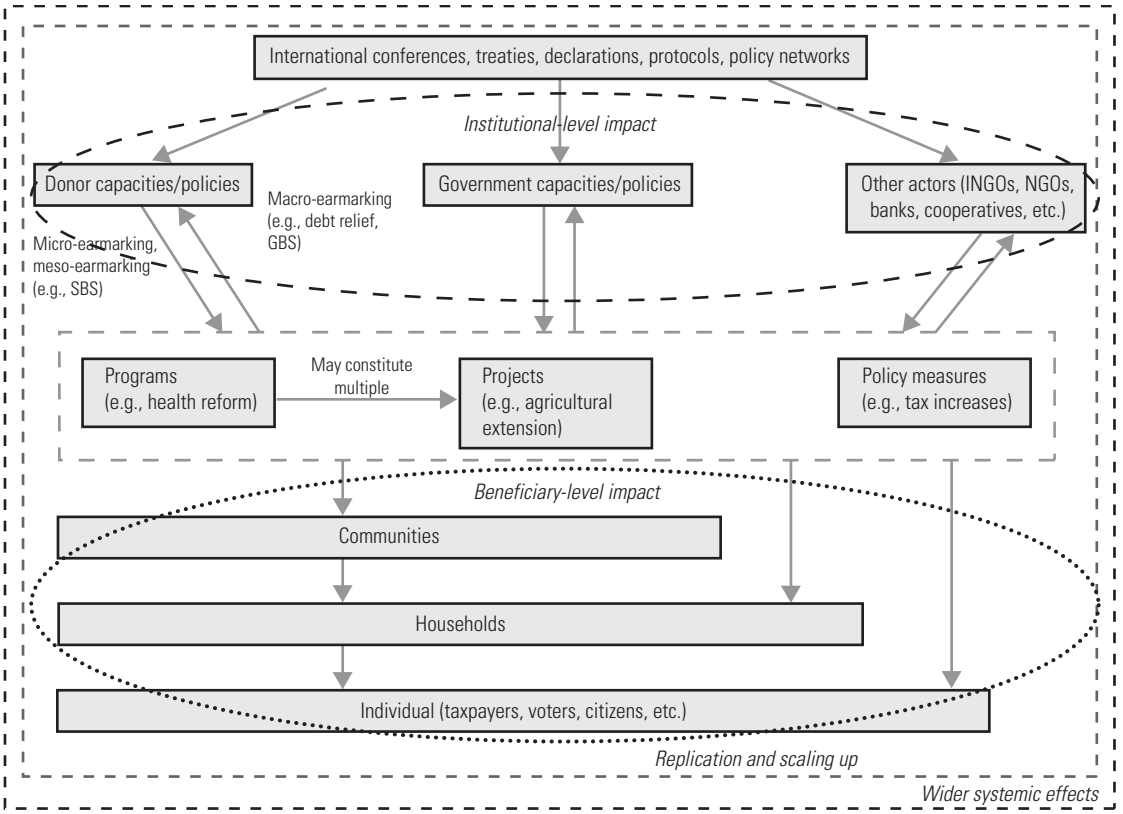
Yet across this continuum, the scope of an impact evaluation can be identified through the lens of two questions: the impact *of what* and the impact *on what*?

When asking the "of what" question, it is useful to differentiate among intervention characteristics. Take single-strand initiatives with explicit objectives—for example, the change in crop yield after introduction of a new technology, or reduction in malaria prevalence after the introduction of bed nets. Such interventions can be isolated, manipulated, and measured, and experimental and quasi-experimental designs may be appropriate for assessing causal relationships between these single-strand initiatives and their effects.

At the other end of the continuum are programs with an extensive range and scope that have activities that cut across sectors, themes, and geographic areas. These can be complicated—multiple agencies, multiple simultaneous causes for the outcomes, and causal mechanisms differing across contexts and complex (recursive, with feedback loops, and with emergent outcomes) (Rogers, 2008). In such cases, impact evaluations have to proceed systematically—first, through *locating and prioritizing key program components* through a comprehensive mapping of the potential influences shaping the program, including possible feedback loops and emerging outcomes; second, *evaluating program components by subsets* of this prioritized program mapping.

When asking the "on what" question, impact evaluations have to unpack interventions that affect multiple institutions, groups, individuals and sites. For tractability, this guidance distinguishes between two principal levels of impact: impact at the *institutional level* and impact at the *beneficiary level* (figure ES1). Examples of the former are policy dialogues, training programs, and strategic support to institutional actors such as governmental and civil society institutions or private corporations and public-private partnerships.

Figure ES1: Levels of intervention, programs, and policies and types of impact



Most policy makers and stakeholders are, however, primarily interested in beneficiary-level interventions that directly affect communities, households, and individuals—whether they be trade liberalization measures, technical assistance programs, antiretroviral treatments, cash transfer programs, construction of schools, etc. This Guidance document accordingly focuses on this level. But it should be recognized that policy interventions primarily geared at inducing sustainable changes at institutional levels can also have indirect effects at the beneficiary level.

2. Agree on what is valued

When conducting impact evaluations, evaluators also need to ask a third question—not only the impact *of what* and *on what*, but impact *for whom*. The fundamental principles to follow here are to agree on the most important, and most valued, objectives of the intervention, and then as much as possible to try to translate these objectives into measurable indicators while keeping track of important aspects that are difficult to measure.

The "for whom" question is inherently a question about stakeholder values—which impacts and processes are judged as significant or valuable, and whose values are used to judge the distribution of costs and benefits? The first and most important reference source to answer this question is the objectives of an intervention, as stated in the official documents. However, interventions evolve and objectives might be implicit or may change. To bring stakeholder values to the surface, evaluators may need to have informal or structured (e.g., "values inquiry") consultations with representatives from different stakeholder groups or use a participatory evalua-

tion approach to include stakeholder values directly in the evaluation.

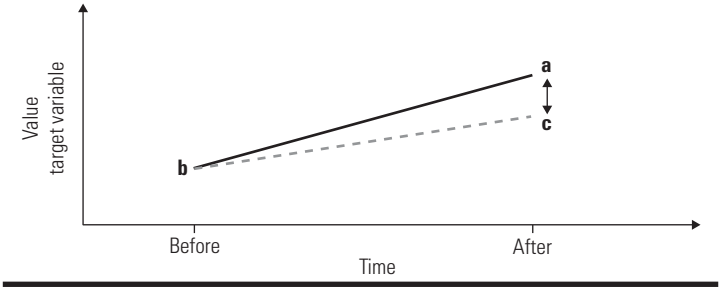
Three other issues are critical to creating measurable indicators to capture the effects of an intervention. First, the evaluation has to consider the possibility of *unintended effects* that go beyond those envisaged in the program theory of the intervention—for example, governments reducing spending on a village targeted by an aid intervention. Second, there may be *long-term effects* of an intervention (such as environmental changes, or changes in social impacts on subsequent generations) or time lags not captured in an impact evaluation that occur relatively soon after the intervention period. Third, and related, is evidence on the *sustainability* of effects, which few impact evaluations will be able to directly capture. Impact evaluations therefore need to identify shorter-term impacts and, where possible, indicate whether longer-term impacts are likely to occur.

3. Carefully articulate the theories linking interventions to outcomes

Development policies and interventions are typically aimed at changing the behavior or knowledge of households, individuals, and organizations. Underlying the design of the intervention is a "theory"—explicit or implicit—with social, behavioral, and institutional assumptions indicating why a particular policy intervention will work to address a given development challenge.

For evaluating the nature and direction of an impact, understanding this theory is critical. But often, these theories are partly "hidden" and require reconstruction and articulation. This articulation can use one or more pieces of evidence—ranging from the intervention's existing logical framework, to insights and expectations of policy makers and other stakeholders on the expected way target groups are affected, to theoretical and empirical research on processes of change or past experiences of similar interventions. However, it is important to critically look for, and articulate, plausible explanations for the changes.

Figure ES2: Simple graphic of net impact of an intervention



After articulating the assumptions on the effect of an intervention on outcomes and impacts, these assumptions will need to be tested. This can be done in two ways—by carefully constructing the causal "story" about the way the intervention has produced results (as by using "causal contribution analysis") or by formally testing the causal assumptions using appropriate methods.

4. Address the attribution problem

The steps above are important to identify the "factual"—the observed outcome that is a result of the intervention. But given that multiple factors can affect the outcomes pertaining to individuals and institutions, the unique point of an impact evaluation is to go beyond the factual—to know the *added value* of the policy intervention under consideration, separate from these other factors.

Any observed changes will be, in general, only partly caused by the intervention of interest. Other interventions inside or outside the core area will often interact and strengthen/reduce the effects of the intervention of interest for the evaluation. Therefore, addressing this "attribution problem" implies both isolating and accurately measuring the particular contribution of an intervention and ensuring that causality runs from the intervention to the outcome.

Analysis of the attribution problem compares the situation "with" an intervention to what would have happened in the absence of an intervention, the "without" situation (the counterfactual, figure ES2). The impact is not

measured by either the value of a target variable (point a) or even the difference between the before and after situation (a–b, measured on the vertical axis). The net impact is the difference between the target variable's value after the intervention and the value the variable would have had in case the intervention had not taken place (a–c).

In doing impact evaluations, there is no "gold standard" (in the sense of a single method that is best in all cases). However, depending on factors such as the scope, objectives, and design of the intervention, as well as data availability, some methods can be better than others in specific cases.

Quantitative techniques can be broadly categorized into experimental, quasi-experimental, and regression-based techniques. These, if well done, have a comparative advantage in addressing the issue of attribution. In each case, the counterfactual is simulated by examining the situation of a participant group (receiving benefits from or affected by an intervention, the "treatment" group) with the situation of an equivalent comparison or "control" group that is not affected by the intervention. A key issue these techniques aim to tackle is *selection bias*—when those in the treatment group are different in some way from those in the control group.

Experimental techniques avoid selection effects by randomly selecting treatment and control groups from the same eligible population, before the intervention starts.

- In a *randomized controlled trial (RCT)*, both groups are expected to have similar average characteristics, with the single exception that the treatment group received the intervention. Thus, a simple comparison of average outcomes in the two groups solves the attribution problem and yields accurate estimates of the impact of the intervention. *But*, despite the clean design, RCTs have to be managed carefully to ensure that the two groups do not have different rates of attrition and that there is a minimum of "contamination," when the control group ends up being exposed to the intervention (either because of geographic proximity or because of the presence of similar parallel interventions affecting the control group).

Quasi-experimental techniques can simulate comparable intervention and comparison groups.

- A *pipeline approach* takes advantage of projects that are rolled out gradually and compares outcomes for households or communities that have already experienced the intervention (the treatment group) with households or communities that are selected but that have not yet participated (the control group). *But* for pipeline approaches to be valid, it is critical that both the treatment and control groups have similar characteristics. Self-selection (due to earlier participation by those eager to receive the intervention) or geographical biases (such as moving from rural to urban areas) do introduce selection biases.
- In *propensity score matching*, a control group is created ex post by selecting its members on the basis of observed and relevant characteristics that are similar to those of members of the treatment group. The pairs are formed not by matching every characteristic exactly, but by selecting groups that have similar probabilities of being included in the sample as the treatment group on the basis of observable characteristics. *But* the technique does not solve the potential bias that results from the omission of unobserved differences between the groups and may require a large sample for the selection of the comparison group. This is usually accounted for through the added use of *double difference* or *difference-in-difference*, which measures differences between the two groups, before and after the intervention, thus netting out the unobservables (as long as they remain constant over time).
- *Judgmental matching* is a less precise method using descriptive information to construct comparison groups—first consulting with clients and other knowledgeable

persons to identify relevant matching characteristics, and then combining geographic information, secondary data (such as household surveys), interviews, and key informants to select comparison areas or individuals/households with the best match of characteristics. *But* the element of subjectivity may induce biases, and further qualitative work is essential to tease out unobserved differences.

Regression-based techniques are more flexible tools for ex post impact evaluation, which can flexibly deal with a range of issues—heterogeneity of treatment, multiple interventions, heterogeneity of participant characteristics, interactions between interventions, and interactions between interventions and specific characteristics. With a regression approach, it may be possible to estimate the contribution of a separate intervention to the total effect or to estimate the effect of the interaction between two interventions.

- Dealing with *unobservables* and *endogeneity*: "Difference-in-difference" approaches in a regression model, by examining the changes within groups over time, can have unobserved (time invariant) variables drop from the equation. The approach is similar to a fixed-effects regression model. "Instrumental variables" can help with endogeneity, as a good instrument correlates with the original endogenous variable in the equation, but not with the error term. *But* the difference-in-difference method is more vulnerable than others to the presence of measurement error in the data, and good instruments are not always possible to find, given the available data.
- *Regression discontinuity* takes advantage of programs that have a cut-off point regarding who receives the treatment (for example, geographic boundaries or income thresholds). It compares the treatment group just within the cut-off point with a control group of those just beyond. At that point, it is unlikely that there are unobserved differences between the two groups. Estimating the impact can now be done by comparing the mean difference between the regression line of treatment outcomes before the intervention with the regression line after. *But* this method assesses the marginal impact of the program only around the cut-off point for eligibility and not across the whole spectrum of the people affected by the intervention. Moreover, care must be taken that individuals were not able to manipulate the selection process or threshold.

Quantitative techniques are not foolproof and can have limitations that go beyond the technical constraints identified above. Narrow counterfactual estimation is not applicable in full-coverage interventions such as price policies or regulation on land use, which affect everybody (although to different degrees)—so regression-based techniques that focus on the variability in exposure/participation are called for. There are also some pragmatic constraints—such as ethical objections to randomization or lack of data representing the baseline situation of intervention target groups. And simple quantitative approaches may not be appropriate in "complex" contexts—though the methodological difficulties of evaluating complicated interventions can to some extent be "neutralized" by deconstructing them into their "active ingredients."

Nonquantitative techniques are often less effective in many cases in addressing attribution, though they can have the comparative advantage when addressing issues of contribution in complex settings. But they can be useful in impact evaluations to both obtain information about the scope, objectives, and theory of change and to generate or supplement data and evidence.

Participatory approaches are a central nonquantitative tool and are built on the principle that stakeholders should be involved in some or all stages of the evaluation. In the case of impact evaluation, this includes aspects such as the determination of objectives, indicators to be taken into account, and stakeholder participation in data collection and analysis. The various methodologies under this umbrella rely on different degrees of participation, ranging from consultation to collaboration to joint decision making. Participatory approaches can be valuable in identifying a

more comprehensive and/or more appropriate set of valued impacts, greater ownership and a better level of understanding among stakeholders, and a better understanding of processes of change and the ways in which interventions affect people. *But* the higher the degree of participation, the more costly and difficult it is to set up an impact evaluation—and thus these may be inappropriate for large-scale comprehensive interventions such as sector programs. Also, there are serious limitations to the validity of information based only on stakeholder perceptions. Finally, strategic responses, manipulation, or advocacy by stakeholders can also influence the validity of the data collection and analysis.

Overall, for impact evaluations, well-designed quantitative methods are usually preferable for addressing attribution and should be pursued when possible. Qualitative techniques cannot quantify the changes attributable to interventions, but should be used to evaluate important issues for which quantification is not feasible or practical and to develop complementary and in-depth perspectives on processes of change induced by interventions.

5. Use a mixed-methods approach

Each different methodology mentioned above has comparative advantages in addressing particular concerns and needs in impact evaluation. A lens to examine these comparative advantages is the four different types of validity:

- *Internal validity*: Establishing the causal relationship between intervention outputs and processes of change leading to outcomes and impacts
- *Construct validity*: Ensuring that the variables measured adequately represent the underlying realities of development interventions linked to processes of change
- *External validity*: Establishing the generalizability of findings to other settings
- *Statistical conclusion validity*: For quantitative techniques, ensuring the degree of confidence about the existence of a relationship between intervention and impact variable and the magnitude of change.

For example, RCTs are arguably better than most other methods in terms of *internal* validity, because if well designed, the counterfactual can be cleanly identified—the randomized project benefits (within a relatively homogenous population) would ensure that there are no systematic differences between those that receive benefits and those that do not. But RCTs control for differences between groups within the *particular setting* that is covered by the evaluation; other settings have different characteristics that are not controlled, so the *external validity* of such RCTs may be limited—unless there has been a systematic and large set of RCTs undertaken that test the intervention across the range of settings and policy options found in reality.

Again, in-depth qualitative methods that attempt to capture complexity and diversity of institutional and social change can have a comparative advantage in *construct validity* in assessing the contribution of complex and multidimensional interventions or impacts. Take the example of impacts on poverty or governance—these may be difficult to fully capture in terms of the distinct, quantifiable indicators usually employed by RCTs and some quasi-experimental methods and may be better addressed through qualitative techniques. Yet these methods also may be lacking in terms of external validity. In such cases, methods having comparative advantages are those large sample quantitative approaches that cover substantial diversity in context and people.

A mix of methods—"triangulating" information from different approaches—can be used to assess different facets of complex outcomes or impacts, yielding greater validity than from one method alone. For example, if looking at the impact of incentives on farmers' labor utilization and livelihoods, a randomized experiment can test the effectiveness of different individual incentives on labor and income effects (testing internal validity); survey data and case studies can deepen the analysis by looking at the distribution of these effects among different types of farm households (triangulating with the RCT evidence on internal validity and increasing external validity); and semistructured interviews

and focus group conversations can broaden the information about the nature of effects in terms of production, consumption, poverty, and so on (establishing construct validity).

Finally, important to note is that an analysis of the distribution of costs and benefits as a result of an intervention—distinguishing between coverage, effects on those that are directly affected, and indirect effects—cannot be addressed with one particular method. If one is interested in all these questions, then inevitably one needs a framework of multiple methods and sources of evidence.

6. Build on existing knowledge relevant to the impact of interventions

Review and synthesis methods can play a pivotal role in marshalling existing evidence to deepen the power and validity of an impact evaluation, to contribute to future knowledge building, and to meet the information needs of stakeholders. Specifically, these methods can serve two major purposes:

- They strengthen external validity by evaluating comparable interventions across different countries and regions—thus assessing the relative effectiveness of alternative interventions in different contexts.
- Because many interventions rely on similar mechanisms of change, they help refine the hypotheses or expected results chain to help greater selectivity for the impact evaluation.

There are several methods that fall into this category:

- ***Systematic reviews*** are syntheses of primary studies that, from an initial explicit statement of objectives, follow a transparent, systematic, and replicable methodology of literature search, inclusion and exclusion of studies according to clear criteria, and extracting and synthesizing of information from the resulting body of knowledge.
- ***Meta-analyses,*** a common type of systematic review, quantitatively synthesize "scores" for the impact of a similar set of interventions from a number of individual studies across different environments. They follow a strict procedure to search for and select appropriate evidence, typically using a hierarchy of methods, with more quantitatively rigorous (experimental) studies being ranked higher as sources of evidence.
- ***Narrative reviews*** are descriptive accounts of intervention processes and/or results covering a series of interventions, relying on a common analytical framework and template to extract data from the individual studies and summarizing the main findings in a narrative account and/or tables and matrices representing key aspects of the interventions.
- ***Realist syntheses*** are theory based and do not use a hierarchy of methods. They collect earlier research findings by placing the policy instrument or intervention that is evaluated in the context of other similar instruments and describe the intervention in terms of its context, social and behavioral mechanisms (what makes the intervention work), and outcomes (the deliverables).

7. Determine if an impact evaluation is feasible and worth the cost

Impact evaluations can be costly exercises in terms of their need for human, financial, and often political resources. They complement rather than replace other types of monitoring and evaluation activities and should therefore be seen as one of several in a cycle of potentially useful evaluations in the lifetime of an intervention. Thus, at each juncture of deciding whether to set up an impact evaluation, it is useful to examine its objectives, benefits, and feasibility and to weigh these against the cost.

Impact evaluations are *feasible* when they have a clearly defined purpose and design, adequate resources, support from influential stakeholders, and data availability and when they are appropriate, given the nature and context of the intervention. They provide the *greatest value* when there is an articulated *need* to obtain the information from them—either to know whether a specific intervention worked, to learn from the intervention, to increase transparency of the intervention, or to know its "value for

money." If they are feasible, their value can then be weighed against the expected *costs*—including the costs of establishing a credible counterfactual, or what would have happened without the intervention.

8. Start collecting data early

As good baseline data are essential to understanding and estimating impact, starting early is critical to the success of the eventual evaluation. When working with secondary data, a lack of information on the quality of data collection can restrict data analysis options and validity of findings. Those managing an impact evaluation have to take notice of and deal effectively with the constraints—of time, data, and resources—under which an impact evaluation has to be carried out.

Depending on the type of intervention, the collection of baseline data and the setup of other aspects of the impact evaluation require an efficient relationship between the impact evaluators and the implementers of the intervention—thus policy makers and commissioners need to involve experts in impact evaluation as early as possible in the intervention to design high-quality impact evaluations.

9. Front-end planning is important

For every impact evaluation, front-end planning is important to help manage the study, its reception, and its use.

When managing the evaluation, it is critical to manage costs and staffing and to make essential and transparent decisions on ethical issues and levels of independence (of the evaluating team vis-à-vis the stakeholders with whom they are collaborating).

To ensure that the evaluation is used, it is also important, at the beginning, to pay attention to country and regional ownership of the impact evaluation and to build capacity to understand and use it. Providing a space for consultation and agreement on impact evaluation priorities among the different stakeholders of an intervention will help enhance utilization and ownership.

Introduction

Over the last 15–20 years, governments and other (public sector) organizations have been paying much more attention to evaluation. It has become a growth industry in which systems of evaluation exist, with their methodologies, organizational infrastructures, textbooks, and professional societies (Leeuw and Furubo, 2008).

In the development world, the growth of monitoring and evaluation (M&E) in particular has been acknowledged as crucial. Kusek and Rist (2004) have articulated its underlying philosophy. M&E stimulates capacity development within countries and organizations to do their "own" evaluations and to produce their "own" performance data. M&E is not focused on *one type of evaluation,* but concerns all of them, including, for example, ex ante studies, rapid appraisals, process evaluations, cost-benefit analyses, and impact evaluations.

Part of the philosophy of evaluation and therefore also M&E is *to put questions first*. Different questions raise a need for different approaches. If the question an evaluator is confronted with is directed toward understanding what a program or policy is about, what the underlying theory of change or logic is, and what the risk factors are when implementing the program, an evaluability assessment or an ex ante evaluation will be an appropriate route to follow. If the question is focused on the implementation of the program or policy or on the role agencies play, then an implementation analysis or a review of the performance of agencies can be appropriate. This can include an audit or inspection. However, if the question is about whether and to what extent the policy intervention made a significant difference (compared with the status quo, compared with other factors and interventions and with or without side effects), then an impact evaluation is the appropriate answer. This Guidance document looks at the latter type of question and corresponding evaluative inquiry, *impact evaluation*. This document discusses questions of what impact evaluation is about, when it is appropriate, and how to do it.

The Network of Networks for Impact Evaluation (NONIE) was established in 2006 to foster more and better impact evaluations by its membership. NONIE uses the definition of the Organisation for Economic Co-operation and Development's Development Assistance Committee (DAC), defining impacts as "[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).

The impact evaluations that NONIE pursues are expected to reinforce and complement the broader evaluation work by NONIE members. The DAC definition refers to the *"effects produced by,"* stressing the attribution aspect. This implies an approach to impact evaluation that is about attributing impacts rather than assessing what happened. 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 and a comparison with *what has occurred* with the intervention implemented.

Following this line of argument, this document subscribes to a somewhat more comprehensive view on impact than the DAC definition does.

Much of the work on impact evaluation that stresses the attribution problem is in fact about attributing short- and medium-term outcomes (to interventions). In practice, this type of attribution analysis is also referred to as impact evaluation, although (in a strict sense) not within the scope of the DAC evaluation. This document includes a discussion on the latter type of analysis as well as the more long-term effects emphasized in the DAC definition (for further discussion of these issues, see White, 2009).

The purpose of NONIE is to promote more and better impact evaluations among its members. Issues relating to evaluations in general are more effectively dealt with within the parent networks and are thus not the primary focus of NONIE. NONIE will focus on sharing methods and learning by doing to promote the practice of impact evaluation. This Guidance document was developed to support those purposes.

The Guidance document was written by and represents the views of the authors, Frans Leeuw and Jos Vaessen, who were commissioned by NONIE. In writing the document, the authors included previous work by NONIE members and took account of their comments in finalizing the document. Given the fact that perspectives on the definition, scope, and appropriate methods of impact evaluation differ widely among practitioners and other stakeholders, the document should not be taken to represent the agreed positions of all of the individual NONIE members. The current Guidance document, highlighting key conceptual and methodological issues in impact evaluation, provides ample coverage of such topics as delimitation, intervention theory, attribution, and combining methods in impact evaluation. It also presents an introduction to such topics as participatory approaches to impact evaluation and assessing impact for complex interventions. These and other topics, such as the evaluation of new aid modalities and country perspectives to impact evaluation, should be developed further in the future.

Impact evaluation in development assistance has received considerable attention over the last few years. The major reason is that many outside of development agencies believe that achievement of results has been poor, or at best not convincingly established. Many development interventions appear to leave no trace of sustained positive change after they have been terminated, and it is hard to determine the extent to which interventions are making a difference. However, the development world is not "alone" in attaching increasing importance to impact evaluations. In fields such as crime and justice, education, and social welfare, impact evaluations have over the last decade become more and more important.[1] Evidence-based (sometimes "evidence-informed") policies are high on the (political) agenda, and some even refer to the "Evidence Movement" (Rieper et al., 2009). This includes the development of *knowledge repositories,* where results of impact evaluations are summarized. In some fields such as criminology and in some professional associations such as the Campbell Collaboration, methodological standards and scales are used to *grade* impact evaluations,[2] although not without discussion (Leeuw, 2005; Worral, 2002, 2007).

Important reasons for doing impact evaluations are the following:

- Impact evaluations provide evidence on "what works and what doesn't" (under what circumstances) and how large the impact is. As the Independent Evaluation Group (IEG) of the World Bank (IEG, 2005) puts it: measuring outcomes and impacts of an activity and distinguishing these from the influence of other, external factors is one of the rationales behind impact evaluation.
- Measuring impacts and relating the changes in dependent variables to development policies and programs is not something that can be done "from an armchair." Impact evaluation is *the* instrument for these tasks.
- Impact evaluation can gather evidence on the sustainability of effects of interventions.
- Impact evaluations produce information that is relevant from an accountability perspective; they disclose knowledge about the (societal) effects of programs that can be linked to the (fi-

nancial) resources used to reach these effects.
- Individual and organizational learning can be stimulated by doing impact evaluations. This is true for organizations in developing countries but also for donor organizations. Informing decision makers on whether to expand, modify, or eliminate projects, programs, and policies is linked to this point, as is IEG's (2005) argument that impact evaluations enable sponsors, partners, and recipients to compare the effectiveness of alternative interventions.

The authors of this Guidance document believe that the ultimate reason for promoting impact evaluations is to learn about "what works and what doesn't and why" and thus to contribute to the effectiveness of (future) development interventions. In addition to this fundamental motive, impact evaluations have a key role to play in the international drive for better evidence on results and development effectiveness. They are particularly well suited to answering important questions about whether development interventions made a difference (and how cost-effective they were). Well-designed impact evaluations also shed light on why an intervention did or did not work, which can vary across time and space.

Decision makers need better evidence on impact and its causes to ensure that resources are allocated where they can have most impact and to maintain future public funding for international development. The pressures for this are already strong and will increase as resources are scaled up for international development. Without such evidence there is a risk of the case for aid and future funding sources being undermined.

Using the word "effects" and "effectiveness" implies that the changes in the "dependent variable[s]" that are measured within the context of an impact evaluation are *caused* by the intervention under study. The concept of "goal achievement" is used when causality is *not* necessarily present. Goals can also be achieved *independent* of the intervention. Changes in financial or economic situations in the world of health and agriculture or in other social conditions can help realize goal achievement, even in a situation where the "believed-to-be-effective" intervention under review is *not* working.

The question of whether impact evaluation should always attempt to measure all possible impacts is not easy to answer. Impact evaluation involves finding the appropriate balance between the desire to understand and measure the full range of effects in the most rigorous manner possible and the practical need to delimit and prioritize on the basis of interests of stakeholders as well as resource constraints.

Key issues addressed in this document

The guidance is structured around nine key issues in impact evaluation:

1. Identify the (type and scope of the) intervention.
2. Agree on what is valued.
3. Carefully articulate the theories linking interventions to outcomes.
4. Address the attribution problem.
5. Use a mixed-methods approach: the logic of the comparative advantages of methods.
6. Build on existing knowledge relevant to the impact of interventions.
7. Determine if an impact evaluation is feasible and worth the cost.
8. Start collecting the data early.
9. Front-end planning is important.

The discussion of these nine issues constitutes the structure of this Guidance document. The first part, comprising the first six issues, deals with methodological and conceptual issues in impact evaluation and constitutes the core of the document. In addition, a shorter second part focuses on managing impact evaluation and addresses aspects of evaluability, benefits, and costs of impact evaluation and planning.

There is no universally accepted definition of "rigorous" impact evaluation. There are some who equate rigorous impact evaluation with particular methods and designs. Given the diversity in

thinking and practice on the topic and the variety in terms of interventions and contexts in which impact evaluation is being applied, the writing of this document has been guided by three basic premises:

- *No single method is best* for addressing the variety of questions and aspects that might be part of impact evaluations.
- However, depending on the specific questions or objectives of a given impact evaluation, some methods have a *comparative advantage* over others in analyzing a particular question or objective.
- Particular methods or perspectives *complement each other* in providing a more complete "picture" of impact.

Moreover, in our view, rigorous impact evaluation is more than methodological design. Rigorous impact evaluation requires addressing the issues described above in an appropriate manner, especially the core methodological and conceptual issues described in Part I.

Part I
Methodological and Conceptual Issues in Impact Evaluation

Chapter 1

Identify the (type and scope of the) intervention

In international development, impact evaluation is principally concerned with final results of interventions (programs, projects, policy measures, reforms) on the welfare of communities, households, and individuals.

1.1. The impact evaluation landscape and the scope of impact evaluation

Impact is often associated with progress at the level of the Millennium Development Goals, which primarily comprise indicators of welfare of these households and individuals. The renewed attention on results- and evidence-based thinking and ensuing interest in impact evaluation provides new momentum for applying rigorous methods and techniques in assessing the impact of interventions.

There is today more than ever 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, along with attention to the other tasks of impact evaluation. At the other end of the continuum are comprehensive programs with an extensive range and scope (increasingly at the country, regional, or global level), with a variety of activities that cut across sectors, themes, 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.

Some evidence of this continuum is provided in appendix 1, in which two examples of impact evaluations are presented, implemented at different (institutional) levels, and based on divergent methodologies with different timeframes (see also figure 1.1.).

The endorsement in 2000 of the Millennium Development Goals by all heads of state, together with other defining events and occurrences, has propelled new action that challenges development evaluation to enter new arenas. There is a shift away from fragmented, top-down, and asymmetrical approaches. Increasingly, ideals such as "harmonization," "partnership,"

"participation," "ownership," and "empowerment" are being emphasized by stakeholders.

However, this trend in policy is not yet reflected in evaluative practices, including impact evaluation. Institutional policies such as anticorruption policies—but also regional and global policy networks and public-private partnerships with their different forms and structures[1]—appear to be less often a part of the goal of impact evaluations, when compared with (top-down) small programs for specific groups of beneficiaries. Ravallion (2008: 6) is of the opinion that there is "a 'myopia bias' in our knowledge, favoring *development projects that yield quick results.*"[2] In the promotion of more rigorous impact evaluation, development agencies, national governments, civil society organizations, and other stakeholders in development should be aware of this bias in focus, keeping in mind the full range of policy interventions that (eventually) affect the welfare of developing societies.

Evaluating the impact of policies—with their own settings and levels—requires appropriate methodological responses. These can be usefully discussed under the banner of two key issues: the impact *of what* and the impact *on what*. These two issues point to a key challenge in impact evaluation: *the scope of the impact evaluation.*

1.2. Impact of what?

What is the independent variable (intervention) we are looking at? In recent years, we have seen a broadening in the range of policy interventions that should/could be subject to impact evaluation.

One of the trends in development is that donors are moving up the *aid chain*. In the past, donors were very much involved in "micro-managing" their own projects and (sometimes) bypassing government systems. In contrast, nowadays a sizeable chunk of aid is allocated to national support for recipient governments. Conditionality 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 being 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 and sector or country strategies, often comprising multiple instruments, stakeholders, sites of intervention, and target groups.

There is a growing demand for assessing the impact of new instruments and modalities, such as—

- International treaties governing the actions of multiple stakeholders (e.g., the Paris Declaration, the Kyoto Protocol)
- New aid modalities such as sector budget support or general budget support
- Instruments such as institutional capacity building, institutional reform, partnership development, and stakeholder dialogues at national or regional levels.

In most countries donor organizations are (still) the main promoters of impact evaluation. The shift of the unit of analysis to the macro and (government) institutional level requires that impact evaluators pay more attention to complicated and more complex interventions at the 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 "project level" remains urgent. The majority of aid money is (still) micro-earmarked money for particular projects managed by donors in collaboration with national institutions. Furthermore, the ongoing efforts in capacity building on national M&E systems (see Kusek and Rist, 2004) and the promotion of country-led evaluation efforts stress the need for further guidance on impact evaluation at the "single" intervention level.

Earlier we referred to 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 relatively easy isolated, manipulated, and measured. Examples of these kinds of interventions include building new roads, repairing roads, reducing the price of fertilizer for farmers, providing clean drinking water at lower cost, etc. It is important to be precise in what the interventions are and what they focus on. In the case of new roads or the rehabilitation of existing ones, the goal often is a reduction in journey time and therefore reduction of societal transaction costs.

At the other end of the continuum are comprehensive programs with an extensive range and scope (increasingly at the country, regional, or global level), with a variety of activities that cut across sectors, themes, and geographic areas and emergent specific activities. Rogers (2008) has outlined several aspects of what constitutes complicated interventions (multiple agencies, alternative and multiple causal strands) and complex interventions[3] (recursive causality, and emergent outcomes; see tables 1.1 and 1.2).

Rogers (2008: 40) recently argued that "the greatest challenge [for the evaluator] comes when interventions have both complicated aspects (multi-level and multi-site) and complex aspects (emergent outcomes)." These aspects often converge in interventions in the context of public-private partnerships or new aid modalities, which have become more important in the development world. Demands for accountability and learning about results at the country, agency, sector, or program and strategy levels are also increasing, which has made the need for appropriate methodological frameworks to assess their impact more pressing.

Pawson (2005) has distinguished five principles on complicated programs that can be helpful when designing impact evaluations of aid:

1. Locate key program components. Evaluation should begin with a comprehensive scoping study, mapping out the potential conjectures and influences that appear to shape the program under investigation. One can envisage stage-one mapping as the hypothesis generator. It should alert the evaluator to the array of decisions that constitute a program, as well as providing some initial deliberation on their intended and wayward outcomes.
2. Prioritize among program components. The general rule here is to concentrate on (i) those components of the program (intervention) theory that seem likely to have the most significant bearing on overall outcomes, and (ii) those segments of program theory about which the least is known.
3. Evaluate program components by subsets. This principle is about when and where to locate evaluation efforts in relation to a program. The evaluation should take on *subsets* of program theory. Evaluation should

Table 1.1: Aspects of complication in interventions

Aspect of complication	Simple intervention	Complicated intervention
Governance and location	Single organization	Multiple agencies, often interdisciplinary and cross-jurisdictional
Simultaneous causal strands	Single causal strand	Multiple simultaneous causal strands
Alternative causal strands	Universal mechanism	Different causal mechanisms operating in different contexts

Source: Rogers (2008).

Table 1.2: Aspects of complexity in interventions

Aspect of complexity	Simple intervention	Complex intervention
Recursive causality and disproportionate effect	Linear, constant dose-response relationship	Recursive, with feedback loops, including reinforcing loops; disproportionate effects at critical limits
Emergent outcomes	Pre-identified outcomes	Emergent outcomes

Source: Rogers (2008).

occur in ongoing portfolios rather than one-off projects. Suites of evaluations and reviews should track program theories as and wherever they unfold.

4. Identify bottlenecks in the program network. "Theories of Change" analysis perceives programs as implementation chains and asks, "What are the flows and blockages as we put a program into action?" The basic strategy is to investigate how the implementation details sustain or hinder program outputs. The main analytic effort is directed at configurations made up of selected segments of the implementation chains across a limited range of program locations.
5. Provide feedback on the conceptual framework. What the theory-based approach initiates is a process of *thinking through* the pathways along which a successful program has to travel. What would be described are the main series of decision points through which an initiative has proceeded, and the findings would be used in alerting stakeholders to the caveats and considerations that should inform those decisions. The most durable and practical recommendations that evaluators can offer come from research that begins with a theory and ends with a refined theory.

If interventions are complicated, in that they have multiple active components, it is helpful to state these separately and treat the intervention as a package of components. Depending on the context, the impact of intervention components can be analyzed separately and/or as part of a package.[4]

To a large extent interventions can be identified and categorized on the basis of the main theme addressed. Examples of thematic areas of interventions are roads and railroads, protected area management, alternative livelihoods, and research on innovative practices.

A second way to identify interventions is to find out which *generic policy instruments and their combinations* constitute the intervention: economic incentives (e.g., tax reductions, subsidies), regulations (e.g. , laws or restrictions), or information (e.g., education or technical assistance). As argued by authors such as Pawson (2006), Salamon (1981), and Vedung (1998), using this relatively simple classification helps identify the interventions. ''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: 256). Acknowledging the central role of policy instruments enables evaluators to take into account lessons from the application of particular (combinations of) policy interventions elsewhere (see Bemelmans-Videc and Rist, 1998).

Third, the separate analysis of intervention components implies interventions being *unpacked* in such a way that the most important social and behavioral mechanisms believed to make the "package" work are spelled out (see chapter 3).

Box 1.1: "Unpacking" the aid chain

The importance of distinguishing among different levels of impact is also discussed by Bourguignon and Sundberg (2007), who "unpack" the aid effectiveness box by differentiating among three essential links between aid and final policy outcomes:

- Policies to outcomes: How do policies, programs and projects affect investment, production, growth, social welfare, and poverty levels? (beneficiary level impact)
- Policy makers to policies: How does the policy-making process at national and local levels lead to "good policies"? This is about governance (institutional capacities, checks and balances mechanisms, etc.) and is likely to be affected by donor policies and aid. (institutional level impact)
- External donors and international financial institutions to policy makers: How do external institutions influence the policy-making process through financial resources, dialogue, technical assistance, conditionalities, etc.? (institutional-level impact)

The above links can be perceived as channels through which aid eventually affects beneficiary-level impact. At the same time, the processes triggered by aid generate lasting impacts at institutional levels.

Source: Bourguignon and Sundberg (2007).

Although complicated interventions are becoming more important and therefore should be subject to impact evaluation, this evolution should not imply a reduction of interest in evaluating the impact of *relatively simple, single-strand interventions*. The sheer number of these interventions makes doing robust impact evaluations of great importance.

1.3. Impact on what?

This topic concerns the "dependent variable problem." Interventions often affect multiple institutions, groups, and individuals. What level of impact should we be 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 might be "fairly" straightforward[5] (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 on 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, we think it is useful for purposes of "scoping" to distinguish between two principal levels of impact: *at the institutional level* and *at the beneficiary level.*[6] It broadens impact evaluation beyond either simply 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. In terms of a simple logic model, there can be multiple intermediate (short- and medium-term) outcomes over time that eventually lead to impact—some or all of which may be included in an evaluation of impact at a specific moment in time.

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 design, manage, and implement better policies for communities, households, and individuals).

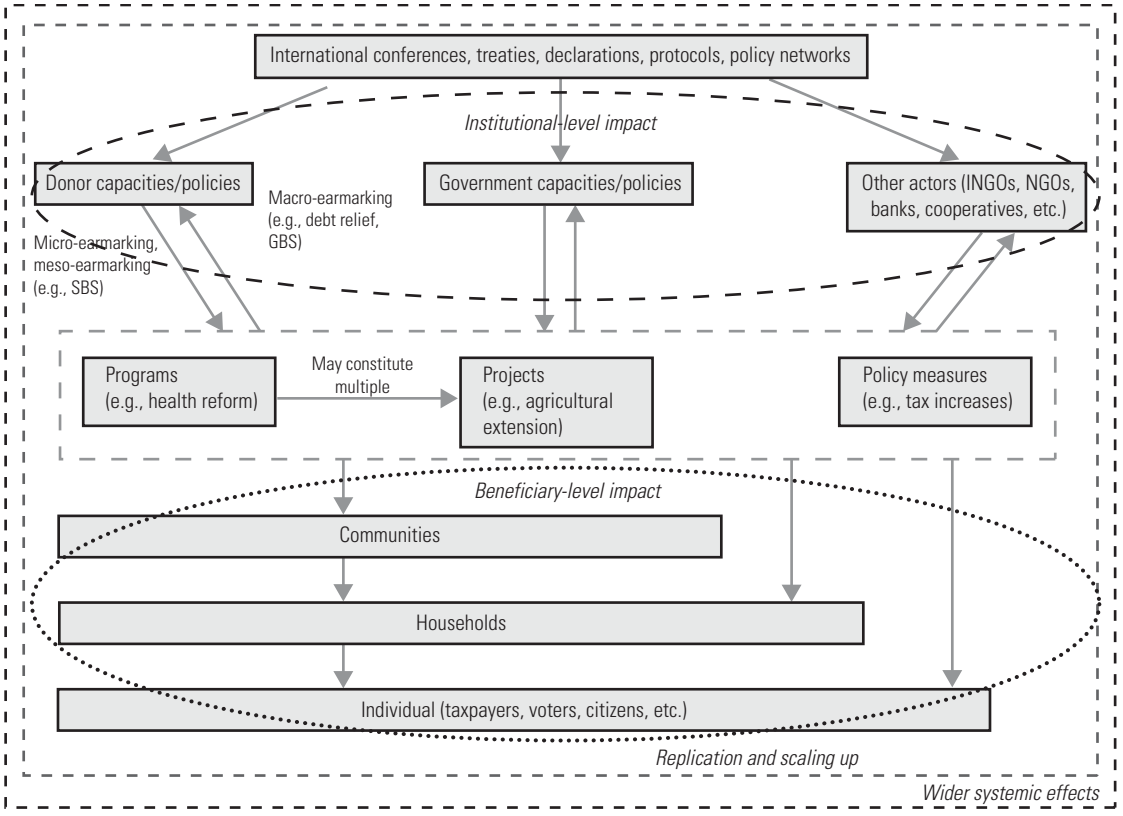
Examples are policy dialogues, policy networks, training programs, institutional reforms, and strategic support to institutional actors (i.e., governmental and civil society institutions, private corporations, and hybrids) and public-private partnerships.

Other types of interventions directly aim at/affect *communities, households, and individuals,* including voters and taxpayers. Examples are fiscal reforms, trade liberalization measures, technical assistance programs, cash transfer programs, construction of schools, etc.

Figure 1.1. graphically presents different levels of intervention and levels of impact. The differentiation between impact at the institutional level and impact at the beneficiary level[7] can be useful in the discussion on choice of scope and method in impact evaluation.

Having illustrated this differentiation, it is important to note that for many in the development community, impact assessment is essentially about impact at the beneficiary level. The *main* concern is how (sets of) policy interventions directly or indirectly affect the welfare of beneficiaries and to what extent changes in welfare can be attributed to these interventions. In line with this interpretation of impact evaluation,[8] *throughout this document we will focus on impact assessment at the beneficiary level* (see the *dotted oval* in figure 1.1.), addressing key methodological concerns and methodological approaches as well as the choice of methodological approach in a particular evaluation context.

Figure 1.1: Levels of intervention, programs, and policies and types of impact



Where necessary, other levels and settings of impact will be addressed (see the *dashed oval* in figure 1.1.). The implication is that with respect to the impact evaluation of, for example, new aid modalities (e.g., general budget support or sector budget support), this will only be discussed as far as interventions financed through these modalities (aim to) affect the lives of households and individuals.[9] We do *not* address the question of how to do impact evaluations of new aid modalities as such (see Lister and Carter, 2006; Elbers et al., 2008).

Key message

Identify the scope and type of the intervention. Interventions range from single-strand initiatives with explicit objectives to complicated institutional policies. Across this continuum, the scope of an impact evaluation can be identified by answering two questions: the impact *of what* and *on what*? Look closely at the nature of the intervention, for example, on the basis of the main theme addressed or by the generic policy instruments used. If interventions are complicated in that they have multiple active components, state these separately and treat the intervention as a package of components that should be unpacked.

Although complicated interventions, sometimes of an institutional nature, are becoming more important and therefore should be subject to impact evaluation, this should not imply a reduction of interest in evaluating the impact of relatively simple, single-strand interventions. The sheer number of these interventions makes doing robust impact evaluations of great importance. In addition, one should be clear about the level of impact to be evaluated. Although most policy makers and stakeholders are primarily interested in beneficiary-level impact (e.g., impact on poverty), specific policy interventions are primarily geared at inducing sustainable changes at the institutional (government) level ("second-order" effects), with only indirect effects at the beneficiary level.

Chapter 2

Agree on what is valued

Impact evaluation requires finding a balance between taking into account the values of stakeholders and paying appropriate attention to the empirical complexity of processes of change induced by an intervention. Some of this complexity has been unpacked in the discussion on the topic of scope of the impact evaluation, where we distinguished between levels of impact that neatly capture the often complex and diffuse causal pathways from intervention to different outcomes and impact: institutional or beneficiary level and replicatory impact. It is best to—as much as possible—translate objectives into measurable indicators, but at the same time not lose track of important aspects that are difficult to measure.

After addressing the issue of stakeholder values, we briefly discuss three dimensions that are particularly important and at the same time challenging to capture in terms of measurable indicators: intended versus unintended effects, short-term versus long-term effects, and the sustainability of effects.

2.1. Stakeholder values in impact evaluation

Impact evaluation needs to assess the value of the results derived from an intervention. This is not only an empirical question but inherently a question about values—which impacts are judged as significant (whether positive or negative), what types of processes are valued in themselves (either positive or negative), and what and whose values are used to judge the distribution of the costs and benefits of interventions.

First, stakeholder values are reflected in the objectives of an intervention, as stated in the official documents produced by an intervention. However, interventions evolve and objectives might change. In addition, stakeholder groups, besides funding and implementing agencies, might harbor expectations not adequately covered by official documents. Impact evaluations need to answer questions related to "for whom" the impacts have been intended and how context influences impacts of interest. Some of the main tasks of an impact evaluation are, therefore, to be clear about who decides what the right aims are and to ensure that the legitimate different perspectives of different stakeholders are given adequate weight. Where there are multiple aims, there must be agreement about the standards of performance required in the weighting of these—for example, can an intervention be

considered a success overall if it fails to meet some of the targets but does well in terms of the main intended outcome?

Depending on the evaluation context, there are different ways for evaluators to address stakeholder values:

- Informal consultation with representatives from different stakeholder groups
- Using values inquiry[1] (Henry, 2002) as a basis for more systematic stakeholder consultation
- Using a participatory evaluation approach to include stakeholder values in the evaluation (see, e.g., Cousins and Whitmore, 1998).

2.2. Intended versus unintended effects

In development programs and projects, intended effects are often translated into measurable indicators as early as the design phase. Impact evaluation should go beyond assessing the expected effects, given an intervention's logical framework and objectives. Interventions often change over time, with consequences for how they affect institutional and people's realities. Moreover, effects are sometimes context specific, where different contexts trigger particular processes of change. Finally, in most cases, the full scope of an intervention's effects is not known in advance. A well-articulated intervention theory can help anticipate some of the unintended effects of an intervention (see chapter 3).

Classic impact evaluations assume that there are no impacts for nonparticipants, but this is unlikely to be true for most development interventions. Spillover effects or replicatory effects (see chapter 1) can stem from market responses (given that participants and nonparticipants trade in the same markets), the (nonmarket) behavior of participants/nonparticipants or the behavior of intervening agents (governmental/nongovernmental organization). 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).

2.3. Short-term versus long-term effects

In some types of interventions, impacts emerge quickly. In others, impact may take much longer 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 more useful and immediate information for policy and decision making. 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 changes in social impacts on subsequent generations. Searching for evidence of such impacts too early might mistakenly lead to the conclusion that they have failed.

In this context, the *exposure time* of an intervention in making 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) and the diffusion effect (other farmers want to see evidence of results before they copy any new behavior). In such gradual, nonlinear processes of change with cascading effects, the timing of the ex post measurement (of land use) is crucial. Ex post measurements that occur 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).

2.4. The sustainability of effects

Focusing on short- or intermediate-term outcomes may underestimate the importance of designs that are able to measure effects (positive or negative) in the long term. One example is an effective strategy to reduce child malnutrition

in a certain population that may quite quickly produce impressive results, yet fail soon after in the absence of systems, resources, and capacities to maintain the work—or follow-up work—after termination of the intervention.

Few impact evaluations will probably provide direct evidence of long-term impacts, and in any case results are needed before these impacts become evident to inform decisions on continuation, next phases, and scaling-up. Impact evaluations therefore need to identify short-term impacts and, where possible, indicate whether longer-term impacts are likely to occur.

To detect negative impacts in the long term, early warning indicators are important to include. A well-articulated intervention theory (see chapter 3) that also addresses the time horizons over which different types of outcomes and impacts could reasonably be expected to occur can help to identify impacts that can and should be explored in an evaluation. The sustainability of positive impacts is also likely to be evident only in the longer term. Impact evaluations therefore can focus on other impacts that will be observable in the short term, such as the institutionalization of practices and the development of organizational capacity, that are likely to contribute to the sustainability of impacts for participants and communities in the longer term.[2]

Key message

Agree on what is valued. Select objectives that are important to the stakeholders' values. Do not be afraid of selecting one objective; focus and clarity are virtues, not vices. As much as possible try to translate objectives into measurable indicators, but at the same time do not lose track of important aspects that are difficult to measure. In addition, keep in mind the dimensions of exposure time and the sustainability of changes.

Chapter 3
Carefully articulate the theories linking interventions to outcomes

When evaluators talk about the black box "problem," they are usually referring to the practice of viewing interventions primarily in terms of effects, with little attention paid to how and why those effects are produced. The common thread underlying the various versions of theory-based evaluation is the argument that "interventions are theories incarnate" and evaluation constitutes a test of intervention theory or theories.

3.1. Seeing interventions as theories: The black box and the contribution problem

Interventions are embodiments of theories in at least two ways. First, they comprise an expectation that the introduction of a program or policy intervention will help ameliorate a recurring social problem. Second, they involve an assumption or set of assumptions about how and why program activities and resources will bring about changes for the better. The underlying theory of a program often remains hidden, typically in the minds of policy architects and staff. Policies—be they relatively small-scale direct interventions like information campaigns, training programs, or subsidization; meso-level interventions such as public-private partnerships and social funds, or macro-level interventions such as "general budget support"—rest on social, behavioral and institutional assumptions indicating why "this" policy intervention will work, which at first view are difficult to uncover.

By seeing interventions as theories and by using insights from theory-based evaluations, it is possible to open up the *black box*. Development policies and interventions, in one way or another, have to do with changing behavior/intentions/knowledge of households, individuals, and organizations (grass roots, private, and public sector). Crucial for understanding what can change behavior is information on *behavioral and social mechanisms*. An important insight from theory-based evaluations is that policy interventions are (often) believed to address and trigger certain social and behavioral responses among people and organizations; in reality this may not be the case.

3.2. Articulating intervention theories on impact

Program theory (or intervention theory) can be identified (articulated) and expressed in many ways—a graphic display of boxes and arrows, a table, a narrative description, and so on. The methodology for constructing intervention theory, as well as the level of detail and complexity, also varies significantly (e.g., Connell et al., 1995; Leeuw, 2003; Lipsey, 1993; McClintock, 1990; Rogers et al., 2000; Trochim, 1989; Wholey, 1987).

Too often the role of methodology is neglected, and it is assumed that "intervention theories" are like *manna* falling out of the sky. That is not the case. Often the underlying theory has to be dug up. Moreover, much of what passes as theory-based evaluation today is simply a form of "analytic evaluation [which] involves no theory in anything like a proper use of that term" (Scriven, 1998: 59).

The intervention theory provides an overall framework for making sense of potential processes of change induced by an intervention. Several pieces of evidence can be used for articulating the intervention theory:

- An intervention's existing logical framework as a starting point for mapping causal assumptions linked to objectives and other written documents produced within the framework of an intervention
- Insights provided by and 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)
- Literature on mechanisms and processes of change in certain institutional contexts, for particular social problems, in specific sectors, etc.

Sometimes stakeholders have contrasting assumptions and expectations about an intervention's impact that has implications for reconstructing the intervention theory. Basically, there are two ways to address this issue. The first is to try to combine the perspectives of different people (for example, program managers and target group members) into an overarching intervention theory that consists of (parts of) arguments from these different sources. The overall theory might be created through an itinerary process of dialogue and refinement and as such might contribute to a shared vision among stakeholders (see, e.g., Pawson and Tilley, 1997). Second, when differences are substantial, several competing intervention theories have to be reconstructed. Carvalho and White (2004) give an example of a "theory" and an "anti-theory" dealing with the assumed impact of social funds (see box 3.1).

For an example of what an impact theory might look like, consider the case of a small business development project that provides training to young managers who have started a business. The direct goal is to help make small businesses financially sustainable and the indirect goal is to generate more employment in the region. Closer scrutiny reveals that the project might have a positive influence on the viability of small businesses in two ways: First, by training young

Box 3.1: Social funds and government capacity: Competing theories

Proponents of social funds argue they will develop government capacity in several ways. Principle among these are that the social fund will develop superior means of resource allocation and monitoring, which will be transferred to the government either directly through collaborative work or indirectly by copying the procedures shown to be successful by the social fund. But critics argue that social funds bypass normal government channels and so undermine government capacity, an effect reinforced by drawing away the government's best people by paying a project premium. Hence, these are rather different theories of how social funds affect government capacity. Carvalho and White (2004) refer to both sets of assumptions in terms of "theory" and "anti-theory." Their study found that well-functioning, decentralized social funds, such as the Zambia Social Investment Fund in Zambia, worked through—rather than parallel to—existing structures and that the social fund procedures were indeed adopted more generally by district staff. But at national level there was generally little evidence of either positive or negative effects on capacity—with some exceptions, such as the promotion of poverty mapping in some countries.

Source: Carvalho and White (2004).

people in basic management and accounting skills, the project intends to have a positive effect on financial viability and ultimately on the growth and sustainability of the business; second, by supporting the writing of a business plan, the project aims to increase the number of successful applications for credit with the local bank, which previously excluded the project's target group because of the small loan sizes (high transaction costs) and high risks involved. Following this second causal strand, efficient and effective spending of the loan is also expected to contribute to the strength of the business. Outputs are measured in terms of the number of people trained by the project and the number of loans the bank extends (see figure 3.1.).

Any further empirical analysis of the impact of the project requires insight into the different factors—besides the project itself—that affect small business development and employment generation. Even in this rather simple example, the number of external variables that affect the impact variables either directly or by moderating the causal relations specified in figure 3.1. is manifold. Some examples are the following:

- Short-term demands on the labor efforts of business owners in other activities may lead to suboptimal strategic choices, jeopardizing the sustainability of the business.
- Inefficient or ineffective use of loans because of short-term demands for cash for other expenditures might jeopardize repayment and the financial viability of the business.
- Deteriorating market conditions (in input or output markets) may jeopardize the future of the business.
- The availability and quality of infrastructure or skilled labor at any point may become constraining factors on business development prospects.
- The efforts of other institutions promoting small business development or any particular aspect of it might positively (or negatively) affect businesses.

Methods for reconstructing the underlying assumptions of project/program/policy theories are the following (see Leeuw, 2003):

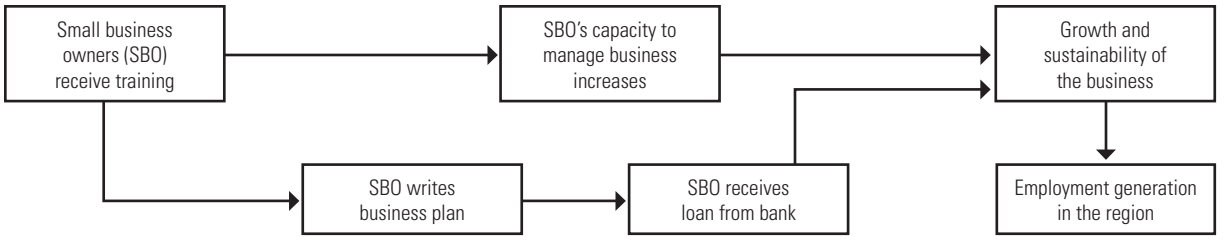
- A policy-scientific method, which focuses on interviews, documents, and argumentation analysis
- A strategic assessment method, which focuses on group dynamics and dialogue
- An elicitation method, which focuses on cognitive and organizational psychology.

Central in all three approaches is the search for mechanisms that are believed to be "at work" when a policy is implemented. Box 3.2 discusses social and behavioral mechanisms for understanding impact.

3.3. Testing intervention theories on impact

After articulating the assumptions on how an intervention is expected to affect outcomes and impacts, the question arises as to what extent these assumptions are valid. In practice,

Figure 3.1: Basic intervention theory of a fictitious small business support project



Box 3.2: Social and behavioral mechanisms as heuristics for understanding processes of change and impact

Hedström (2005: 25) has defined the concept of social mechanisms as "a constellation of entities and activities that are organized such that they regularly bring about a particular type of outcome." Mechanisms form the "nuts and bolts" (Elster, 1989) or the "engines" (Leeuw, 2003) of interventions (policies and programs), making them work, given certain contexts (Pawson and Tilley, 1997). Hedström and Swedberg (1998: 296–98), building on the work of Coleman (1990), discuss three types of mechanisms: situational mechanisms, action formation mechanisms, and transformational mechanisms.

Examples of *situational mechanisms* are self-fulfilling and self-denying prophecies and crowding-out (e.g., by striving to force people who are already largely compliant with laws and regulations into full compliance, the opposite is realized, because due to the extra focus on laws and regulation, the internal motivation of people to comply is reduced).

Action-formation mechanisms are the heuristics that people develop to deal with their bounded rationality, such as—

- Framing and the endowment effect—"The fact that people often demand much more to give up an object than they would be willing to pay to acquire it," but also the tendency for people to have a stronger preference for more immediate payoffs than for later payoffs, the closer to the present both payoffs are
- Types of learning (social learning, vicarious learning)
- "Game-theoretical" mechanisms, such as the "grim strategy" (to repeatedly refuse to cooperate with another party as a punishment for the other party's failure to cooperate previously) and the shadow of the future /shadow of the past mechanism
- Mechanisms such as the "fight-or-flight-response" to stress and the "tend-and-befriend mechanism" are other examples.

Transformational mechanisms illuminate how processes and results of interacting individuals and groups are "transformed" into collective outcomes. Examples are the following:

- Cascading is a process by which people influence one another, so much so that participants ignore their private knowledge and rely instead on the publicly stated judgments of others. The bandwagon phenomenon (the tendency to do [or believe] things because many other people do [or believe] the same) is related to this, as are group think, the common knowledge effect, and herd behavior.
- "Tipping points," "where a small additional effort can have a disproportionately large effect, can be created through virtuous circles, or be a result of achieving certain critical levels" (Rogers, 2008: 35).

Relevance of mechanisms for impact evaluations

Development policies and interventions, in one way or another, have to do with changing behavior/intentions/knowledge of households, individuals, and organizations (grass roots, private, and public sector). Crucial for understanding what can change behavior is information about these mechanisms. *The mechanisms underlying processes of change might not be necessarily those that are assumed to be at work by policy makers, programs designers, and staff.* Creating awareness on the basis of (public) information campaigns does not always lead to behavioral change. Subsidies and other financial incentives run the risk of causing unintended side effects, such as benefit snatching, but also create the "Mitnahme-effect" (people already tended to behave in a way the incentive wanted them to behave before the incentive existed). Mentoring dropouts in education might cause "learned helplessness" and therefore increase dropout rates. Many other examples are available in the literature. The relevance of knowing which social and behavioral mechanisms are believed to do the work increases as the complication and complexity of interventions increases.

A focus on mechanisms helps evaluators and managers open up and test the theory underlying an intervention. Spending time and money on programs based on "pet theories" of policy makers or implementation agents that are not corroborated by relevant research should probably not be high on the agenda. If a policy intervention is based on mechanisms that are known not to work (in a given context or in general), that is a signal that the intervention probably will not be very effective. This can be found out on the basis of desk research as a first test of the relevance and validity of an intervention theory, that is, by confronting the theory with existing knowledge about mechanisms. That knowledge stems from synthesis and review studies (see chapter 6). Further empirical impact evaluation can generate more contextualized and precise tests of the intervention theory.

evaluators have at their disposal a wide range of methods and techniques to test the intervention theory. We can distinguish between two broad approaches. The first is that the theory constitutes the basis for constructing a "causal story" about how and to what extent the intervention has produced results. Usually different methods and sources of evidence are used to further refine the theory in an iterative manner until a credible and reliable causal story has been generated. The second approach is to use the theory as an explicit benchmark for testing (some of) the assumptions in a formal manner. Besides providing a benchmark, the theory provides the template for method choice, variable selection, and other data collection and analysis issues. This approach is typically applied in statistical analysis but is not in any way restricted to this type of method. In short, theory-based methodological designs can be situated anywhere in between "telling the causal story" and "formally testing causal assumptions."

The systematic development and corroboration of the causal story can be achieved through *causal contribution analysis* (Mayne, 2001), which aims to demonstrate whether the evaluated intervention is one of the causes of observed change. Contribution analysis relies on chains of logical arguments that are verified through careful analysis. Rigor in causal contribution analysis involves systematically identifying and investigating alternative explanations for observed impacts. This includes being able to rule out implementation failure as an explanation for lack of results and developing testable hypotheses and predictions to identify the conditions under which interventions contribute to specific impacts.

The causal story is inferred from the following evidence:

- There is a reasoned theory of change for the intervention: it makes sense, is plausible, and is agreed to by key players.
- The activities of the intervention were implemented.
- The theory of change—or key elements thereof—is verified by evidence: the chain of expected results occurred.
- Other influencing factors have been assessed and either shown not to have made a significant contribution or their relative role in contributing to the desired result has been recognized.

The analysis is best done iteratively, building up a more robust assessment of causal contribution. The overall aim is to reduce the uncertainty about the contribution the intervention is making to the observed results through an increased understanding of why the observed results have occurred (or not) and the roles played by the intervention and other factors. At the impact level this is the most challenging, and a "contribution story" has to be developed for each major strategy that is part of an intervention, at different levels of analysis. They would be linked, as each would treat the other strategies as influencing factors.

One of the key challenges in the foregoing analysis is to pinpoint the exact causal effect from intervention to its impact. Despite the potential strength of the causal argumentation on the links between the intervention and impact, and despite the possible availability of data on indicators, as well as data on contributing factors, etc., there remains uncertainty about the *magnitude* of the impact as well as *the extent* to which the changes in impact variables are really due to the intervention or to other influential variables. This is called the attribution problem and is discussed in chapter 4.

Key message

Carefully articulate the assumptions behind the theories linking interventions to outcomes. What are the causal pathways linking intervention outputs to processes of change and impact? Be critical if an "intervention theory" appears to assert or assume changes without much explanation. The focus should be on dissecting the causal (social, behavioral, and institutional) mechanisms that make interventions "work."

Chapter 4

Address the attribution problem

Multiple factors can affect the livelihoods of individuals or the capacities of institutions. For policy makers as well as stakeholders it is important to know what the added value of the policy intervention is, apart from these other factors.

4.1. The attribution problem

The attribution problem is often referred to as the central problem in impact evaluation. The central question is to what extent changes in outcomes of interest can be *attributed* to a particular intervention. Attribution refers to both isolating and estimating accurately the particular contribution of an intervention and ensuring that causality runs from the intervention to the outcome.

The changes in welfare for a particular group of people can be observed by undertaking before and after studies, but these rarely accurately measure impact. *Baseline* data (before the intervention) and *end-line* data (after the intervention) give facts about the development over time and describe "the factual" for the treatment group (not the counterfactual). But changes observed by comparing before-after (or pre-post) data are rarely caused by the intervention alone, as other interventions and processes influence developments, both in time and space. There are some exceptions in which before versus after will suffice to determine impact. For example, supplying village water pumps reduces time spent fetching water. If nothing else of importance happened during the period under study, attribution is so clear that there is no need to resort to anything other than before versus after to determine this impact.

In general, the observed changes are only partly caused by the intervention of interest. Other interventions inside or outside the core area will often interact and strengthen/reduce the effects of the intervention of interest for the evaluation. In addition, other unplanned events or general change processes will often influence development, such as natural catastrophes, urbanization, growing economies, business cycles, war, or long-term climate change. For example, in evaluating the impact of microfinance on poverty, we have to control for the influences of changing market conditions, infrastructure developments, or climate shocks such as droughts, and so on.

A discussion that often comes up in impact evaluation is the issue of *attribution of what*. This issue is complementary to the independent variable question discussed in chapter 1.

How the impact of the intervention is measured may be stated in several ways:

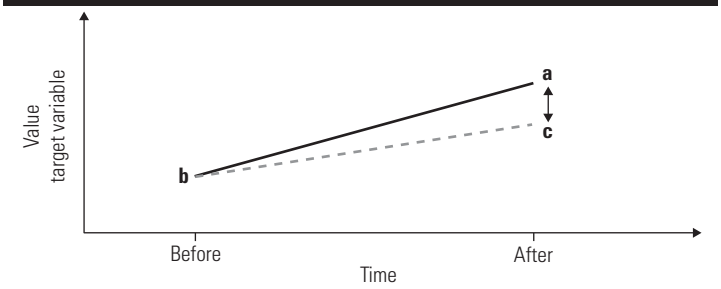
- What is the impact of an additional dollar of funding to program X?[1]
- What is the impact of country Y's contribution to a particular intervention?
- What is the impact of intervention Z?

In this guidance we will focus on the third level of attribution: *What is the impact of a particular policy intervention (from very simple to complex), independent of the specific monetary and nonmonetary contributions of the (institutional) actors involved?*

The issue of attributing impact to a particular intervention can be a quite complicated issue in itself (especially when talking about complicated interventions such as sector strategies or programs). Additional levels of attribution, such as tracing impact back from interventions to specific (financial) contributions of different donors, are either meaningless or too complicated to achieve in a pragmatic and cost-effective manner.

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 because it is not possible to observe how the situation would have been without the intervention, so that has to be constructed by the evaluator. The counterfactual is illustrated in figure 4.1.

Figure 4.1: Graphic display of the net impact of an intervention



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 net impact (at a given point in time) is the difference between the target variable's value after the intervention and the value the variable would have had if the intervention had not taken place (a–c).

The starting point for an evaluation is a good account of the factual—what happened in terms of the outputs/outcomes targeted by the intervention? A good account of the factual requires articulating the intervention theory (or theories) and connecting the different causal assumptions from intervention outputs to outcomes and impacts, as discussed earlier in chapter 3. This guidance will discuss several options for measuring the counterfactual.

Evaluations can either be experimental, as when the evaluator purposely collects data and designs evaluations in advance, or quasi-experimental, as when data are collected to mimic an experimental situation. Multiple regression analysis is an all-purpose technique that can be used in virtually all settings (provided that data are available); when the experiment is organized in such a way that no controls are needed, a simple comparison of means can be used instead of a regression, because both will give the same answer. (Experimental and quasi-experimental approaches will be discussed in § 4.2.) We briefly introduce the general principles and the most common 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 comparison group that is not affected by the intervention.

Several designs exist of combinations of ex ante and ex post measurements of participant and control groups (see § 4.2.). Randomization of intervention participation is considered the best way to create equivalent groups. Random assignment to the participant and control group leads

to groups with similar average characteristics[2] for both observables and non-observables, except for the intervention. 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).

4.2. Quantitative methods addressing the attribution problem[3]

In this section we discuss experimental (e.g., randomized controlled trials), quasi-experimental (e.g., propensity score matching), and regression-based techniques.[4, 5]

Three related problems that quantitative impact evaluation methods attempt to address are the following:

- The establishment of a *counterfactual*: What would have happened in the absence of the intervention(s)?
- The elimination of *selection effects*, leading to differences between the intervention group (or treatment group) and the control group
- A solution for the problem of *unobservables*: The omission of one or more unobserved variables, leading to biased estimates.

Selection effects occur, for example, when those in the intervention group are more or less motivated than those in the control group. It is particularly a problem when the variable in question, in this case motivation, is not easily observable. As long as selection is based on *observable* characteristics and these are measured in the evaluation, they may be included—and thus controlled for—in the regression analysis. However, not all relevant characteristics are observed or measured. This problem of *selection of unobservables* is one of the main problems in impact evaluation.

In the following sections we will discuss different techniques of quantitative impact evaluation, thereby mainly focusing our discussion on the selection bias issue. In trying to deal systematically with selection effects, (quasi-) experimental design-based approaches such as the randomized controlled trial (RCT) or the pipeline approach can be compromised by two sets of problems: contamination and unintended behavioral responses.

Contamination: Contamination (or contagion, treatment diffusion) refers to the problem of groups of people that are not supposed to be exposed to certain project benefits but in fact are benefiting from them. Contamination comes from two possible sources. The first is from the intervention itself, as a result of *spill-over effects*. Interventions are most often planned and implemented within a *delimited space* (a village, district, nation, region, or institution). The *influence zone* of an intervention may, however, be larger than the *core area* where the intervention takes place or is intended to generate results (geographical spill-over effects). To avoid contamination, control and comparison groups must be located outside the influence zone. Second, the selected comparison group may be subject to similar interventions implemented by *different agencies*, or even somewhat dissimilar interventions that affect the same outcomes. The counterfactual is thus a *different type of intervention* rather than *no intervention*. This problem is often overlooked. A good intervention theory as a basis for designing a measurement instrument that records the different potential problems of contamination is a good way to address this problem.

Unintended behavioral responses: In any experiment people may behave differently when they know that they are part of the intervention or treatment. Consequently, this will affect data. The resulting bias is even more pronounced when the researcher has to rely on recall data or self-reported effects. Several unintended behavioral responses not caused by the intervention or by "normal" conditions might therefore disrupt the validity of comparisons between groups and hence the ability to attribute changes to project incentives. Important possible effects are the following (see Shadish et al., 2002; Rossi et al., 2004):

- *Expected behavior or compliance behavior*: Participants react in accordance with intervention staff expectations for reasons such

as compliance with the established contract or certain expectations about future benefits from the organization (not necessarily the project).

- *Compensatory equalization*: Discontent among staff or recipients with inequality between incentives might result in compensation of groups that receive less than other groups.
- *Compensatory rivalry*: Differentiation of incentives to groups of people might result in social competition between those receiving (many) intervention benefits and those receiving fewer or no benefits.
- *Hawthorne effect*: The fact of being part of an experiment rather than the intervention as such causes people to change their behavior.
- *Placebo effect*: The behavioral effect is not the result of the incentives provided by the intervention but of people's perception of the incentives and the subsequent anticipatory behavior.

These problems are relevant in most experimental and quasi-experimental design approaches that are based on ex ante participant and control/comparison group designs.[6] They are less relevant in regression-based approaches that use statistical matching procedures or that do not rely on the participant-control group comparison for counterfactual analysis.[7]

4.2.1. Randomized controlled trial

The safest way to avoid selection effects is a *randomized selection* of the intervention and control groups *before* the experiment starts. When the experimental group and the control group are selected randomly from the same eligible population, 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 RCT, a simple comparison of average outcomes in the two groups can adequately resolve the attribution problem and yield accurate estimates of the impact of the intervention on a variable of interest; by design, the only difference between the two groups was the intervention.

To determine if the intervention had a statistically significant impact, one simply performs a test of equality between the mean outcomes in the experiment and control group. Statistical analysis will tell you if the impact is statistically significant and how large it is. Of course, with larger samples, the statistical inferences will be increasingly precise; but if the impact of an intervention really is large, it can be detected and measured even with a relatively small sample.

A proper RCT addresses many attribution issues, but has to be planned and managed carefully to avoid contamination and other risks. Risks of a RCT are (i) different rates of attrition in the two groups, possibly caused by a high dropout in one of the two groups, (ii) spillover effects (contamination) resulting in the control group receiving some of the treatment, and (iii) unintended behavioral responses.

4.2.2. Pipeline approach

One of the problems for the evaluation of development projects or programs is that evaluators rarely get involved early enough to design a good evaluation (although this is changing). Often, households or individuals are selected for a specific project, but not everybody participates (directly) in the project. A reason may be a gradual implementation of the project. Large projects (such as in housing or construction of schools) normally have a phased implementation.

In such a case, it may be possible to exploit this phase of the project by comparing the outcomes of households or communities that actually participate (the experiment group) with households or communities that are selected but do not participate (the comparison group). A specific project (school building) may start, for instance, in a number of villages and be implemented later in other villages. This creates the possibility of evaluating the effect of school building on enrollment. One has to be certain, of course, that the second selection—the actual inclusion in the project—does not introduce a selection bias. If, for instance, at the start of the project a choice is made to start construction in a number of specific villages, the

(relevant) characteristics of these villages must be similar to other villages that are eligible for new schools. Self-selection (of villages that are eager to participate) or other selection criteria (starting in remote areas or in urban areas) may introduce a selection bias.

4.2.3. Propensity score matching

When no comparison group has been created at the start of the project or program, a comparison group may be created ex post through a *matching* procedure: for every member of the treatment group, one or more members in a control group are selected on the basis of similar *observed* (and relevant) characteristics.

Suppose there are two groups, one a relatively small intervention group of 100 pupils who will receive a specific reading program. If we want to analyze the effects of this program, we must compare the results of the pupils in the program with other pupils who were not included in the program. We cannot select just any control group, because the intervention group may have been self-selected on the basis of specific characteristics (pupils with relatively good results or relatively bad results, pupils from rural areas, from private schools or public schools, boys, girls, orphans, etc.). Therefore, we need to select a group with similar characteristics. One way of doing this would be to find for every boy age 10 years from a small rural school with a high pupil:teacher ratio in a poor district another boy with the same observed characteristics. This would be a time-consuming procedure, especially for 100 pupils.

An alternative way to create a control group for this case is the method of *propensity score matching*. This technique involves forming pairs, not by matching every characteristic exactly, but by selecting groups that have similar *probabilities* of being included in the sample as the treatment group. The technique uses all *available* information to construct a control group (see box 4.1.).[8] Rosenbaum and Rubin (1983) showed that this method makes it possible to create a control group ex post with characteristics that are similar to the intervention group that would have been created had its members been selected randomly before the beginning of the project.

It should be noted that the technique only deals with selection bias on observables and does not solve potential endogeneity bias (see appendix 4), which results from the omission of unobserved variables. Nevertheless, propensity score matching may be combined with the technique of double differencing to correct for the influence of time-invariant unobservables (see below). Moreover, the technique may require a large sample for the selection of the comparison

Box 4.1: Using propensity scores to select a matched comparison group—The Vietnam Rural Roads Project

The survey sample included 100 project communes and 100 non-project communes in the same districts. Using the same districts simplified survey logistics and reduced costs, but communes were still far enough apart to avoid "contamination" (control areas being affected by the project). A logit model of the probability of participating in the project was used to calculate the propensity score for each project and non-project commune. *Comparison communes* were then selected with *propensity scores* similar to the project communes. The evaluation was also able to draw on commune-level data collected for administrative purposes that cover infrastructure, employment, education, health care, agriculture, and community organization. These data will be used for contextual analysis, to construct commune-level indicators of welfare, and to test program impacts over time. The administrative data will also be used to model the process of project selection and to assess whether there are any selection biases.

Sources: Van De Walle and Cratty (2005); Bamberger (2006).

group, which might pose a problem if secondary data are not available (see chapter 8).

4.2.4. Judgmental matching[9]

A less precise method for selecting control groups uses descriptive information from, for example, survey data to construct comparison groups.

Matching areas on observables. In consultation with clients and other knowledgeable persons, the researcher identifies characteristics that should be matched (e.g., access to services, type or quality of house construction, economic level, location, or types of agricultural production). Information from maps (sometimes including geographic information system data and/or aerial photographs), observation, secondary data (e.g., censuses, household surveys, school records), and key informants are then combined to select comparison areas with the best match of characteristics. Operating under real-world constraints means that it will often be necessary to rely on easily observable or identifiable characteristics (e.g., types of housing and infrastructure). Although this may expedite matters, there may also be unobservable differences; the researcher must address these as much as possible through qualitative research and attach the appropriate caveats to any results.

Matching individuals or households on observables. Similar procedures as those noted above can be used to match individuals and households. Sample selection can sometimes draw on existing survey data or ongoing household surveys; however, in many cases researchers must find their own ways to select the sample. Sometimes the selection is based on physical characteristics that can be observed (type of housing, distance from water and other services, type of crops or area cultivated), whereas in other cases selection is based on characteristics that require screening interviews (e.g., economic status, labor market activity, school attendance). In these latter cases, the interviewer must conduct quota sampling.

4.2.5. Double difference (difference in difference)

Differences between the intervention group and the control group may be unobserved and therefore problematic. Nevertheless, even though such differences cannot be measured, the technique of double difference (or difference-in-difference) deals with these differences as long as they are time invariant. The technique measures differences between the two groups, before and after the intervention (hence the name double difference).

Suppose there are two groups, an intervention group I and a control group C. One measures, for instance, enrollment rates before (0) and after (1) the intervention. According to this method, the effect is

$$(I_1 - I_0) - (C_1 - C_0) \text{ or } (I_1 - C_1) - (I_0 - C_0).$$

For example, if enrolment rates at t = 0 would be 80% (for the intervention group) and 70% for the control group and at t = 1, these rates would be, respectively, 90% and 75%, then the effect of

Table 4.1: Double difference and other designs

	Intervention group	Control group	Difference across groups
Baseline	I_0	C_0	$I_0 - C_0$
Follow-up	I_1	C_1	$I_1 - C_1$
Difference across time	$I_1 - I_0$	$C_1 - C_0$	*Double-difference:* $(I_1 - C_1) - (I_0 - C_0) =$ $(I_1 - I_0) - (C_1 - C_0)$

Source: Adapted from Maluccio and Flores (2005).

the intervention would be (90% – 80%) – (75% – 70%) = 5%.

The techniques of propensity score matching (see above) and double difference may be combined. Propensity score matching increases the likelihood that the treatment and control groups have similar characteristics, but cannot guarantee that all relevant characteristics are included in the selection procedure. The double difference technique can eliminate the effects of an unobserved selection bias, but this technique may work better when differences between the intervention group and the control group are eliminated as much as possible. The approach eliminates *initial* differences between the two groups (e.g., differences in enrollment rates) and therefore gives an unbiased estimate of the effects of the intervention, as long as these differences are time invariant. When an unobserved variable is time variant (changes over time), the measured effect will still be biased.

4.2.6. Regression analysis and double difference

In some programs the interventions are all or nothing (a household or individual is subjected to the intervention or not); in others they vary continuously over a range, as when programs vary the type of benefit offered to target groups. One example is a cash transfer program or a microfinance facility where the amount transferred or loaned may depend on the income of the participant; improved drinking water facilities are another example. These facilities differ in capacity and are implemented in different circumstances with beneficiaries living at different distances to these facilities.

In addition to the need to deal with both discrete and continuous interventions, we also need to control for other factors that affect the outcome other than the magnitude of the intervention. The standard methodology for such an approach is a regression analysis. One of the reasons for the popularity of regression-based approaches is their flexibility: they may deal with the heterogeneity of treatment, multiple interventions, heterogeneity of characteristics of participants, interactions between interventions, and interactions between interventions and specific characteristics, as long as the treatment (or intervention) and the characteristics of the subjects in the sample are observed (can be measured). With a regression approach, it may be possible to estimate the contribution of a specific intervention to the total effect or to estimate the effect of the interaction between two interventions. The analysis may include an explicit control group.

We must go beyond a standard regression-based approach when there are *unobserved* selection effects or endogeneity (see next section). A way to deal with unobserved selection effects is the application of the "difference-in-difference" approach in a regression model (see appendix 4). In such a model we do not analyze the (cross-section) effects between groups, but the changes (within groups) over time. Instead of taking the specific values of a variable in a specific year, we analyze the *changes* in these variables over time. In such an analysis, unobserved time-invariant variables drop from the equation.[10]

Again, the quality of this method as a solution depends on the validity of the assumption that unobservables are time invariant. Moreover, the quality of the method also depends on the quality of the underlying data. The method of double differencing is more vulnerable than some other methods to the presence of measurement error in the data.

4.2.7. Instrumental variables

An important problem when analyzing the impact of an intervention is the problem of *endogeneity*. The most common example of endogeneity is when a third variable causes two other variables to correlate without there being any causality. For example, doctors are observed to be frequently in the presence of people with fevers, but doctors do not cause the fevers; it is the third variable (the illness) that causes the two other variables to correlate (people with fevers and the presence of doctors). In econometric language, when there is endogeneity an explanatory variable will be correlated with the error term in a mathematical model (see appendix 4). When an explanatory

variable is endogenous, it is not possible to give an unbiased estimate of the causal effect of this variable.

Selection effects also give rise to bias. Consider the following example. Various studies in the field of education find that repeaters produce lower test results than non-repeaters. A preliminary and false conclusion would be that repetition does not have a positive effect on student performance and that it is simply a waste of resources. But such a conclusion neglects the endogeneity of repetition: intelligent children with well-educated parents are more likely to perform well and therefore not repeat. Less intelligent children, on the other hand, will probably not achieve good results and are therefore more likely to repeat. So, both groups of pupils (i.e., repeaters and non-repeaters) have different characteristics, which at first view makes it impossible to draw conclusions based on a comparison between them.

The technique of instrumental variables is used to address the endogeneity problem. An instrumental variable (or instrument) is a third variable that is used to get an unbiased estimate of the effect of the original endogenous variable (see appendix 4). A good instrument correlates with the original endogenous variable in the equation, but not with the error term. Suppose a researcher is interested in the effect of a training program. Actual participation in the program may be endogenous, because, for instance, the most motivated employees may subscribe to the training. Therefore, one cannot compare employees who had the training with employees who did not without incurring bias. The effect of the training may be determined if a subset were assigned to the training by accident or through some process unrelated to personal motivation. In this case, the instrumental variables procedure essentially only uses data from that subset to estimate the impact of training.

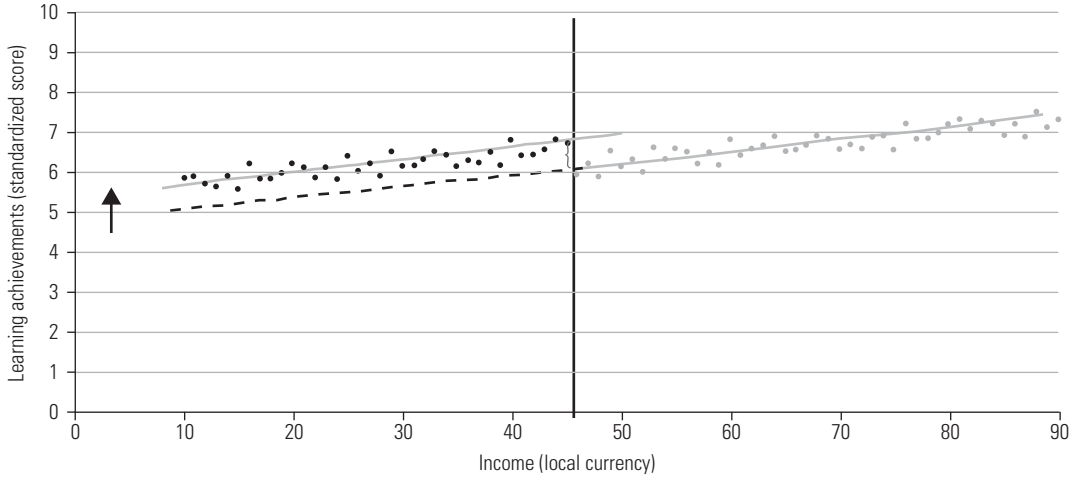
4.2.8. Regression discontinuity analysis

The basic idea of regression discontinuity analysis is simple. Suppose program participation depends on income. On the left side of the *cut-off point,* people (or households) have an income that is just low enough to be eligible for participation; on the right side of the cut-off point, people are no longer allowed to participate, even though their income is just slightly higher. There may be more criteria that define the threshold, and these criteria may be explicit or implicit. Regression discontinuity analysis compares the treatment group with the control group at the cut-off point. At that point, it is unlikely that there are unobserved differences between the two groups.

Suppose we want to analyze the effect of a specific program to improve learning achievements. This program focuses on the poorest households: the program includes only households with an income below a certain level. We know that learning achievements are correlated with income,[11] and therefore we cannot compare households participating in the program with households that do not participate. Other factors may induce an endogeneity bias (such as differences in the educational background of parents or the distance to the school). Nevertheless, at the cut-off point, there is no reason to assume that there are systematic differences between the two groups of households (apart from small differences in income). Estimating the impact can now be done, for example, by comparing the mean difference between the regression line of learning achievements in function of income *before* the intervention with the regression line *after* (see figure 4.2).

A major disadvantage of a regression discontinuity design is that the method assesses the marginal impact of the program only around the cut-off point for eligibility. Moreover, it must be possible to construct a specific threshold, and individuals should not be able to manipulate the selection process (ADB, 2006: 14). Many researchers prefer regression discontinuity analysis above propensity score matching, because the technique generates a higher likelihood that estimates will not be biased by unobserved variables.[12]

Figure 4.2: Regression discontinuity analysis



4.3. Applicability of quantitative methods for addressing the attribution problem

There are some limitations to the applicability of the techniques discussed in the previous section. We briefly highlight some of the more important ones (for a more comprehensive discussion see, e.g., Bamberger and White, 2007). First, in general, counterfactual estimation is not applicable in full-coverage interventions such as price policies or regulation on land use, which affect everybody (although to different degrees). In this case there are still possibilities to use statistical "counterfactual-like" analyses, such as those that focus on the variability in exposure/participation in relation to changes in an outcome variable (see, e.g., Rossi et al., 2004). Second, there are several pragmatic constraints to applying this type of analysis, especially with respect to randomization and other design-based techniques. For example, there might be ethical objections to randomization or lack of data representing the baseline situation of intervention target groups (see chapter 8). Third, applicability of quantitative approaches (experimental and non-experimental) also largely depends on the number of observations (n) available for analysis. Quantitative analysis is only meaningful if n is reasonably large: statistically based approaches are not applicable if there is a small n. The small n problem can arise either because the intervention was applied to a single unit (e.g., capacity building in a single ministry or a national policy change) or a small number of units or because there is heterogeneity in the intervention so that only a small number of units received support of a specific type. Where this is a small n, then a variety of other approaches can be used (see § 4.4.).

An important critique of the applicability of these methods refers to the nature of the intervention and the complexity of the context in which the intervention is embedded. The methodological difficulties of evaluating complicated interventions to some extent can be "neutralized" by deconstructing them into their "active ingredients" (see, e.g., Vaessen and Todd, 2008).[13] Consider the example of school reform in Kenya as described by Duflo and Kremer (2005). School reform constitutes a set of different simultaneous interventions at different levels, ranging from revisions in and decentralization of the budget allocation process, to addressing links between teacher pay and performance, to vouchers and school choice. Although the total package of interventions constituting school reform represents an impressive landscape of causal pathways of change at different levels,

directly and indirectly affecting individual school, teacher, and student welfare in different ways, it can be unpacked into different (workable) components, such as teacher incentives and their effects on student performance indicators or school vouchers and their effects on student performance.

True experimental designs have been relatively rare in development settings (but not rare in developing countries, as medical tests routinely use a randomized approach). Alternatively, quasi-experiments using non-random assignment to participant and control groups are more widely applicable. Preferably, double difference (participant-control group comparisons over time) designs should be used. However, it is more usual that impact assessments are based on less rigorous—and reliable—designs, where—

- Baseline data are reconstructed or collected late during the implementation phase.
- Baseline data are collected only for the treatment group.
- There are no baseline data for the treatment or control group.

If no baseline data exist, then the impact of the intervention is measured by comparing the situation afterward between the treatment and control groups. This comparison of end-line data is measured by a single difference (see also appendix 14).

Some impact evaluations are based on pure "before and after" comparisons of change only for the treatment group, with no comparison group at all. The measure in such cases is also a single difference, but the lack of a proxy for the counterfactual makes conclusions based on this design less robust. This design gives a valid measure of impacts only in the rare situations when no other factors can explain the observed change, or when the intervention of interest is the only factor influencing the conditions. In other words, all other factors are stable, or there are no other cause-effect relationships than between the intervention and the observed change. A systematic control of the influence of other factors can significantly increase the reliability of findings (see also chapter 8).

Some final remarks on attribution are in order. Given the centrality of the attribution issue in impact evaluation, we concur with many of our colleagues that there is scope for more quantitative impact evaluation, as these techniques offer a comparative advantage of formally addressing the counterfactual. Therefore, with a relatively large *n,* a quantitative approach is usually preferred. However, at the same time it is admitted that, given the limitations discussed above, the application of experimental and quasi-experimental design-based approaches will necessarily be limited to only a part of the total amount of interventions in development.[14]

The combination of theory-based evaluation and quantitative impact evaluation provides a powerful methodological basis for rigorous impact evaluation for several reasons:

- The intervention theory will help indicate which of the intervention components are amenable to quantitative counterfactual analysis through, for example, quasi-experimental evaluation and how this part of the analysis relates to other elements of the theory.[15]
- The intervention theory approach will help identify key determinants of impact variables to be taken into account in a quantitative impact evaluation.
- The intervention theory approach can provide a basis for analyzing how an intervention affects particular individuals or groups in different ways; although quantitative impact evaluation methods typically result in quantitative measures of average net effects of an intervention, an intervention theory can help to support the analysis of distribution of costs and benefits (see chapter 5).
- The intervention theory can help strengthen the interpretation of findings generated by quantitative impact evaluation techniques.

This symbiosis between theory-based evaluation and quantitative impact evaluation has been

acknowledged by a growing number of authors in both the general impact evaluation literature (e.g., Cook, 2000; Shadish et al., 2002; Rossi et al., 2004; Morgan and Winship, 2007) as well as in the literature on development impact evaluation (e.g., Bamberger et al., 2004; Bourguignon and Sundberg, 2007; Ravallion, 2008). When this combination is not feasible, alternative methods embedded in a theory-based evaluation framework should be applied.

4.4. Other approaches

In this section we introduce a range of methodological approaches that can be used to address the attribution problem or particular aspects of the impact evaluation.[16]

4.4.1. Alternative approaches for addressing the attribution problem

The methods discussed in the previous sections have the advantage of allowing for an estimation of the magnitude of change attributable to a particular intervention using counterfactual analysis. There are also other (qualitative) methods that can be useful in addressing the issue of attribution. However, *these methods as such do not quantify effects attributable to an intervention.*[17]

A first example of an alternative approach is the so-called General Elimination Methodology (GEM). This approach is epistemologically related to Popper's falsification principle. Michael Scriven added it to the methodology of (impact) evaluations. Although in some papers he suggested that the GEM approach was particularly relevant for dissecting causality chains within case studies, both in his earlier work and in a more recent paper (Scriven, 1998), he makes clear that the GEM approach is relevant for every type of expert practice, including RCTs and case studies (see appendix 2 for a more detailed discussion).

What is the *relevance of this approach for impact evaluation?* Given the complexity of solving the attribution problem, GEM can help "test" different counterfactuals that have been put forward in a theoretical way. When doing (quasi-)experiments, using GEM can be an extra check on the validity of the conclusions and can help one understand why the results are as they are. Pawson and Tilley (1997) criticized experimentalists by highlighting what they perceive as a lack of attention to explanatory questions in (quasi-) experiments. Consequently, GEM can be helpful by involving the evaluator in setting up a "competition" between the conclusions from the evaluation and possible *other* hypotheses.

A second example is causal contribution analysis (see Mayne, 2001; described in chapter 3). Contribution analysis relies on chains of logical arguments that are verified through careful analysis. Rigor in this type of causal analysis involves systematically identifying and investigating alternative explanations for observed impacts. This includes being able to rule out implementation failure as an explanation for lack of results and developing testable hypotheses and predictions to identify the conditions under which interventions contribute to specific impacts.

Some of these hypotheses can be tested using the quantitative methods discussed previously. In this sense, contribution analysis, and other variants of theory-based analysis, provide a framework in which quantitative methods of impact evaluation could be used to test particular causal assumptions. If the latter is not possible, the verification and refinement of the causal story should exclusively rely on other (multiple) methods of inquiry (see chapter 5).

4.4.2. Participatory approaches[18]

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. Participatory evaluation approaches (see, e.g., Cousins and Whitmore, 1998) are built on the principle that stakeholders should be involved in some or all stages of the evaluation. As Greene (2006: 127ff) illustrates, "[P]articipatory approaches to evaluation directly engage the micropolitics of power by involving stakeholders in important decision-making roles within

the evaluation process itself. Multiple, diverse stakeholders collaborate as *co-evaluators*, often as members of an evaluation team." Participatory evaluation can be perceived as a developmental process in itself, largely because it is "the process that counts" (Whitmore, 1991). 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).

Participatory approaches to impact evaluation can be important for several reasons. First, one could ask the legitimate question of impact "according to whom." Participatory approaches can be helpful in engaging stakeholders on the issue of what is to be valued in a particular impact evaluation. By engaging a range of stakeholders, a more comprehensive and/or appropriate set of *valued* impacts is likely to be identified (see the second key issue of this Guidance document). When identifying the (type and scope of the) intervention to be evaluated (see first chapter), participatory methods might be of particular use; aspects that might be "hidden" behind official language and political jargon (in documents) can be revealed by narrative analyses and by consulting stakeholders. More generally, the process of participation in some cases can enhance stakeholder ownership, the level of understanding of a problem among stakeholders, and utilization of impact evaluation results.

Within the light of the attribution issue, stakeholder perspectives can help improve an evaluator's understanding of the complex reality surrounding causal relationships among interventions and outcomes and impacts. In addition, insight into the multiple and (potentially) contrasting assumptions about causal relationships between an intervention and processes of change can help enrich an evaluator's perspective on the attribution issue. As discussed in chapter 3, stakeholder perspectives can be an important source for reconstructing an intervention theory or multiple theories,[19] which subsequently can be refined or put to the test during further analysis.

Some of the latter benefits can also be realized by using qualitative methods that are nonparticipatory (see Mikkelsen, 2005; see also appendix 9). This brings us to an important point. There is a common misperception that there is a finite and clearly defined set of so-called "participatory" evaluation methods. Although certain methods are often (justifiably) classified under the banner of participatory methods because stakeholder participation is a defining feature, many methods not commonly associated with stakeholder participation (including, for example, (quasi-) experimental methods) can also be used in more or less participatory ways, with or without stakeholder involvement. The participatory aspect of methodology is largely determined by the issues of who is involved and who does or decides on what and how. For example, the methodology for testing water quality to ascertain the impact of treatment facilities can become participatory if community-level water users are involved in deciding, for example, what aspects of water quality to measure and how to collect the data and report the results.

Methodologies commonly found under the umbrella of participatory (impact) evaluation include appreciative inquiry; beneficiary assessment; participatory impact pathway analysis; participatory impact monitoring (see box 4.2.); poverty and social impact analysis; social return on investment; systematic client consultation; self-esteem, associative strength, resourcefulness, action planning and responsibility; citizen report cards; community score cards; and the Participatory Learning and Action toolbox[20] (see, for example, IFAD, 2002; Mikkelsen, 2005; Pretty et al., 1995; Salmen and Kane, 2006).

These methods rely on different degrees of participation, ranging from consultation to collaboration to joint decision making. In general, the higher the degree of participation, the more costly and difficult it is to set up the impact evaluation. In addition, a high degree of participation might be difficult to realize in

large-scale comprehensive interventions such as sector programs.[21]

Apart from the previously discussed potential benefits of an impact evaluation involving some element of stakeholder participation, disadvantages of participatory approaches include the following:

- Limitations to the validity of information based on stakeholder perceptions (only); this problem is related to the general issue of shortcomings in individual and group perceptional data.
- The risk of strategic responses, manipulation, or advocacy by stakeholders can influence the validity of the data collection and analysis.[22]
- Limitations to the applicability of impact evaluation with a high degree of participation especially in large-scale, comprehensive, multi-site interventions (aspects of time and cost).

4.4.3. Useful methods for data collection and analysis that are often part of impact evaluation designs[23]

In this section we distinguish a set of methods that are useful:

- For testing/refining particular parts (i.e., assumptions) of the impact theory but not specifically focused on impact assessment as such
- For strengthening particular lines of argumentation with additional/detailed knowledge, useful for triangulation with other sources of evidence
- For deepening the understanding of the nature of particular relationships between intervention and processes of change.

The literature on (impact) evaluation methodology, as in any other field of methodology, is riddled with labels representing different (and sometimes not so different) methodological approaches. In essence however, methodologies are built upon specific methods. Survey data collection and (descriptive) analysis, semi-structured interviews, and focus-group interviews are but a few of the specific methods that are found throughout the landscape of methodological approaches to impact evaluation.

Evaluators, commissioners, and other stakeholders in impact evaluation should have a basic knowledge about the more common research techniques:[24]

Box 4.2: Participatory impact monitoring in the context of the poverty reduction strategy process

Participatory impact monitoring builds on the *voiced perceptions and assessments* of the poor and aims to strengthen these as relevant factors in decision making at national and subnational levels. In the context of poverty reduction strategy monitoring it will provide systematic and fast feedback on the implementation progress, early indications of outcomes, impact, and the unintended effects of policies and programs.

The purposes are as follows:

- Increase the voice and the agency of poor people through participatory monitoring and evaluation
- Enhance the effectiveness of poverty oriented policies and programs in countries with poverty reduction strategies
- Contribute to methodology development, strengthen the knowledge base, and facilitate cross-country learning on the effective use of participatory monitoring at the policy level, and in the context of poverty reduction strategy processes in particular.

Conceptually, the proposed project impact monitoring approach combines (1) the analysis of relevant policies and programs at the national level, leading to an inventory of "impact hypotheses," with (2) extensive consultations at the district/local government level, and (3) joint analysis and consultations with poor communities on their perceptions of change, their attributions to causal factors, and their contextualized assessments of how policies and programs affect their situation.

Source: Booth and Lucas (2002).

Descriptive statistical techniques (e.g., of survey or registry data): The statistician Tukey (e.g., Tukey, 1977) argued for more attention to exploratory data analysis techniques as powerful and relatively simple ways to understand patterns in data. Examples include univariate and bivariate statistical analysis of primary or secondary data using graphical analysis and simple statistical summaries (e.g., for univariate analysis: mean, standard deviation, median, interquartile range; for bivariate analysis: series of boxplots, scatterplots, odds ratios).

Inferential statistical techniques (e.g., of survey or registry data): Univariate analysis (e.g., confidence intervals around the mean; t-test of the mean), bivariate analysis (e.g., t-test for difference in means), and multivariate analysis (e.g., cluster analysis, multiple regression) can be rather useful in estimating impact effects or testing particular causal assumptions of the intervention theory. These techniques (including the first bullet point) are also used in the (quasi-) experimental and regression-based approaches described in § 4.2. For more information, see Agresti and Finlay (1997) or Hair et al. (2005) or, more specifically for development contexts, see Casley and Lury (1987) or Mukherjee et al. (1998).

"*Qualitative methods*" include widely used methods, such as semi-structured interviews, open interviews, focus group interviews, participant observation, and discourse analysis—but also less conventional approaches such as mystery guests, unobtrusive measures (e.g., through observation; see Webb et al., 2000), etc. For more information, see Patton (2002) or, more specifically for development contexts, see Mikkelsen (2005) or Roche (1999).[25]

Key message

Address the attribution problem. Although there is no single method that is best in all cases (a gold standard), some methods are indeed best in specific cases. When empirically addressing the attribution problem, experimental and quasi-experimental designs embedded in a theory-based evaluation framework have clear advantages over other designs. If addressing the attribution problem can only be achieved by doing a contribution analysis, be clear about that and specify the limits and opportunities of this approach. Overall, for impact evaluations, well-designed quantitative methods may better address the attribution problem. Baseline data are critical when using quantitative methods. Qualitative techniques cannot quantify the changes attributable to interventions but should be used to evaluate important issues for which quantification is not feasible or practical, and to develop complementary and in-depth perspectives on processes of change induced by interventions (see next section). Evaluators need a good basic knowledge about all techniques before determining what method to use to address this problem.

Chapter 5
Use a mixed-methods approach: The logic of the comparative advantages of methods

The work by Campbell and others on validity and threats to validity within experiments and other types of evaluations have left deep marks on the way researchers and evaluators have addressed methodological challenges in impact evaluation (see Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish et al., 2002).

5.1. Different methodologies have comparative advantages in addressing particular concerns and needs

Validity can be broadly defined as the "truth of, or correctness of, or degree of support for an inference" (Shadish et al., 2002: 513). Campbell distinguished among four types of validity, which can be explained in a concise manner by looking at the questions underlying the four types:

- *Internal validity:* How do we establish that there is a causal relationship between intervention outputs and processes of change leading to outcomes and impacts?
- *Construct validity:* How do we make sure that the variables we are measuring adequately represent the underlying realities of development interventions linked to processes of change?
- *External validity:* How do we (and to what extent can we) generalize about findings to other settings (interventions, regions, target groups, etc.)?
- *Statistical conclusion validity:* How do we make sure that our conclusion about the existence of a relationship between intervention and impact variable is in fact true? How can we be sure about the magnitude of change?[1]

Applying the logic of comparative advantages makes it possible for evaluators to compare methods on the basis of their relative merits in addressing particular aspects of validity. This provides a useful basis for methodological design choice; given the evaluation's priorities, methods that better address particular aspects of validity are selected in favor of others. In addition, the logic of comparative advantages can support decisions on combining methods to be able to simultaneously address multiple aspects of validity.

We will illustrate this logic using the example of RCTs. Internal validity usually receives (and justifiably so) a lot of attention in impact evaluation, as it lies at the heart of the attribution problem; is there a causal link between intervention outputs and outcomes and impacts? Arguably, RCTs (see § 4.2.) are viewed by many as the best method for addressing the attribution problem *from*

the point of view of internal validity. Random allocation of project benefits reduces the likelihood that there are systematic (observable and unobservable) differences between those that receive benefits and those that do not. However, this does not make it necessarily the best method *overall.* For example, RCTs control for differences between groups within the particular setting that is covered by the study. Other settings have other characteristics that are not controlled, hence there are limitations of *external validity* here.

To resolve this issue, Duflo and Kremer (2005) propose to undertake series of RCTs on the same type of instrument in different settings. However, as argued by Ravallion, "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).

Another limitation of RCTs (also valid for other approaches discussed in § 4.2.) lies in the realm of construct validity. Does the limited set of indicators adequately represent the impact of a policy on a complex phenomenon such as poverty? In-depth qualitative methods can more adequately capture the complexity and diversity of aspects that define (and determine) poverty than the singular or limited set of impact indicators taken into account in RCTs. Consequently, the latter have a *comparative advantage* in addressing *construct validity* concerns. However, a downside of most qualitative approaches is that the focus is local and findings are very context specific, with limited external validity. *External validity* can be adequately addressed by, for example, quantitative quasi- and non-experimental approaches that are based on large samples covering substantial diversity in context and people.

Theory-based evaluation provides the *basis* for combining different methodological approaches that have comparative advantages in addressing validity concerns. In addition, the intervention theory as a structure for making explicit causal assumptions, generalizing findings, and making in-depth analysis of specific assumptions can help strengthen internal, external, and construct validity claims.

To conclude:

- There is no single best method in impact evaluation that can always address the different aspects of validity better than others.
- Methods have particular advantages in dealing with particular validity concerns; this provides a strong rationale for combining methods.

5.2. Advantages of combining different methods and sources of evidence

In principle, each impact evaluation is in some way supported by different methods and sources of evidence. For example, even the quite technical quantitative approaches described in § 4.2 include other modes of inquiry, such as the research review to identify key variables that should be controlled for in, for example, a quasi-experimental setting. Nevertheless, there is a growing literature on the explicit use of *multiple methods* to strengthen the quality of the analysis.[2] At the same time the discordance between the practice and "theory" of mixed-methods research (Bryman, 2006) suggests that mixed-methods research is often more an art than a science.

Triangulation is a key concept that embodies much of the rationale behind doing mixed-methods research and represents a set of principles to fortify the design, analysis, and interpretation of findings in impact evaluation.[3] Triangulation is about looking at things from multiple points of view, a method "to overcome the problems that stem from studies relying upon a single theory, a single method, a single set of data … and from a single investigator" (Mikkelsen, 2005: 96). As can be deduced from the definition, there are different types of triangulation. Broadly, these are the following (Mikkelsen, 2005):

- Data triangulation—To study a problem using different types of data, different points in time, or different units of analysis
- Investigator triangulation—Multiple researchers looking at the same problem
- Discipline triangulation—Researchers trained in different disciplines looking at the same problem
- Theory triangulation—Using multiple competing theories to explain and analyze a problem
- Methodological triangulation—Using different methods, or the same method over time, to study a problem.

As can be observed from this list, particular methodologies already embody aspects of triangulation. Quantitative double-difference impact evaluation (see § 4.2.), for example, embodies aspects of methodological and data triangulation. Participatory impact evaluation approaches are often used to seek out and reconstruct multiple (sometimes contrasting) perspectives on processes of change and impact using diverse methods, often relying on teams of researchers with different disciplinary backgrounds (that may include members of target groups). Theory-based evaluation often involves theory triangulation (see chapter 3; see also Carvalho and White [2004], who refer to competing theories in their study on social funds). Moreover, it also allows for methodological and data triangulation by relying on different methods and sources of evidence to test particular causal assumptions.

Discipline triangulation and theory triangulation both point to the need for more diversity in perspectives for understanding processes of change in impact evaluation. Strong pleas have recently been made for development evaluators to recognize and make full use of the wide spectrum of frameworks and methodologies that have emerged from different disciplines and that provide evaluation with a rich arsenal of possibilities (Kanbur, 2003; White, 2002; Bamberger and White, 2007). For example, when doing impact evaluations, evaluators can benefit from approaches developed in different disciplines and subdisciplines. Neo-institutionalist economists have shown ways to study the impact of institutions as "rules of the game" (see North, 1990), and interventions such as policies can be considered as attempts to establish specific rules with the expectation (through a "theory of change") of generating certain impacts (Picciotto and Wiesner, 1997). In addition, the literature on behavioral and social mechanisms (see appendix 10; see also chapter 6) provides a wealth of explanatory insights that help evaluators better understand and frame processes of change triggered by interventions.

A good methodological practice in impact evaluation is to encourage applying these principles of triangulation as much as possible.

Advantages of mixed-methods approaches to impact evaluation are the following:

- A mix of methods can be used to assess important outcomes or impacts of the intervention being studied. If the results from different methods converge, then inferences about the nature and magnitude of these impacts will be stronger. For example, triangulation of standardized indicators of children's educational attainments with results from an analysis of samples of children's academic work yields stronger confidence in the educational impacts observed than either method alone (especially if the methods employed have offsetting biases).
- A mix of methods can be used to assess different facets of complex outcomes or impacts, yielding a broader, richer portrait than one method alone can. For example, standardized indicators of health status could be mixed with onsite observations of practices related to nutrition, water quality, environmental risks, or other contributors to health, jointly yielding a richer understanding of the intervention's impacts on targeted health behaviors. In a more general sense, quantitative impact evaluation techniques work well for a limited set of pre-established variables (preferably determined and measured ex ante) but less well for capturing unintended, less expected (indirect) effects of interventions. Qualitative methods or descriptive (secondary) data

analysis can be helpful in better understanding the latter.
- One set of methods could be used to assess outcomes or impacts and another set to assess the quality and character of program implementation, including program integrity and the experiences during the implementation phase.
- Multiple methods can help ensure that the sampling frame and the sample selection strategies cover the whole of the target intervention and comparison populations. Many sampling frames leave out important sectors of the population (usually the most vulnerable groups or people who have recently moved into the community), while respondent selection procedures often under-represent women, youth, the elderly, or ethnic minorities. This is critical because important positive or negative impacts on vulnerable groups (or other important sectors) are completely ignored if they do not even get included in the sample. This is particularly important (and frequently ignored) where the evaluation uses secondary data sets, as the evaluator often does not have access to information on how the sample was selected.
- Multiple methods are needed to address the complementary questions of average effect and distribution of costs and benefits of an intervention (see § 5.3.)

Appendix 11 presents four interesting examples of impact evaluations that are based on a mixed method perspective:

- Case 1: Combining qualitative and quantitative descriptive methods—Ex post impact study of the Noakhali Rural Development Project in Bangladesh
- Case 2: Combining qualitative and quantitative descriptive methods—Mixed-methods impact evaluation of International Fund for Agricultural Development projects in Gambia, Ghana, and Morocco
- Case 3: Combining qualitative and quantitative descriptive methods—Impact evaluation: agricultural development projects in Guinea
- Case 4: A theory-based approach with qualitative methods (GEF, 2007).

5.3. Average effect versus distribution of costs and benefits

Sometimes policy makers and stakeholders are concerned with the question of whether an intervention (for a specific context and group of people) has been effective overall. This is typically a question that can be addressed by using (quasi) experimental evaluation techniques. However, another important question, one that might not be easily answered with these techniques, is whether and how people are differently affected by an intervention.[4] This question can be answered by using regression analysis. A regression model can incorporate different moderator variables (e.g., through modeling interaction effects) to analyze to what extent important characteristics co-determine outcome variables. In addition, many qualitative methods such as those used for case studies can help evaluators study in detail how interventions work differently in different situations. From a methodological design perspective, a mixed-methods study combining quasi-experimental survey data with a limited number of in-depth, semistructured interviews among different types of people from the target population is an example of a potentially good framework to provide credible answers to both questions (see box 5.1.).

When talking about the issue of distribution of costs and benefits of an intervention, it is useful to distinguish between different levels or foci. First, one should consider the issue of outreach or coverage. Who are the people (individuals, households, and communities) directly affected by an intervention? Sometimes this question can be answered in a relatively straightforward manner, such as when the intervention is clearly delineated and targeted to a specific group of people (e.g., a training program). In other cases (e.g., a tax cut or construction of a road), coverage or outreach, or indeed the delineation of the group of people affected by the intervention, is not that easy to determine. In the last case, the issue of delineation is closely linked to the second level, how an intervention has different effects on

Box 5.1: Brief illustration of the logic of comparative advantages

Consider the example of an intervention that provides monetary incentives and training to farmers to promote land use changes leading to improved livelihoods conditions.

We could use the following methods in the impact evaluation:

- A randomized experiment 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)*
- Survey data and case studies could tell how incentives have different effects on particular types of farm households *(potentially strengthens internal validity and increases external validity of findings)*
- Semistructured interviews and focus group conversations could tell us more about the nature of effects in terms of production, consumption, poverty, etc. *(potentially enhances construct validity of findings).*

groups of people (e.g., how the construction of a road affects different types of businesses and households near or farther from the new road). In the case of a simple training program, the first level (who participates, who is directly affected) can be neatly separated from the second (how an intervention affects participants in different ways). A third level concerns the indirect effects of an intervention. For example, an objective of a training program may be that participants in turn become teachers for the population at large. While this is an intended indirect effect, multiple indirect effects on participants, their families, and non-participants may occur, some of which may be quite substantial. Time and scale are important dimensions here (see also chapter 2).

Often, impact evaluation is about level two—determining the effects on those that are directly targeted by/participating in the intervention. In those cases, it is often assumed that level one (targeting, outreach) is fully known and mapped. In other cases, level one—outreach and coverage or indeed the determination of the scope of direct effects of an intervention on the population at risk—is the great "unknown" and should be a first priority in an impact evaluation exercise. Level three—indirect processes of change induced by an intervention, with potentially important implications for the distribution of costs and benefits among target populations and beyond—is often outside the scope of impact evaluations (see Ravallion, 2008).

Important to note is that an analysis of the distribution of costs and benefits as a result of an intervention—distinguishing among coverage, effects on those who are directly affected, and indirect effects—cannot be addressed with one particular method. If one is interested in all these questions, then inevitably one needs a framework of multiple methods and sources of evidence. For example, descriptive analysis of survey data can help to map coverage, quasi-experiments can help to assess attribution of change among those directly affected, and case studies and survey data analysis can help to map indirect effects over time.

Key message

Use a mixed-methods design. Bear in mind the logic of the comparative advantages of designs and methods. A mix of methods can be used to assess different facets of complex outcomes or impacts, yielding more breadth, depth, and width in the portrait than one method alone can. One set of methods could be used to assess outcomes or impacts and another set to assess the quality and nature of intervention implementation, thus enhancing impact evaluation with information about program integrity and program experiences. It is important to note that an analysis of the distribution of costs and benefits of an intervention—distinguishing among coverage, effects on those directly affected, and indirect effects—cannot be addressed with one particular method. Answering these questions requires a framework of multiple methods and sources of evidence.

Chapter 6
Build on existing knowledge relevant to the impact of interventions

Review and synthesis approaches are commonly associated with systematic reviews and meta-analyses. Using these methods, comparable interventions evaluated across countries and regions can provide the empirical basis to identify "robust" performance goals and to help assess the relative effectiveness of alternative interventions under different country contexts and settings. These methods can lead to increased emphasis on the rigor of impact evaluations so they can contribute to future knowledge building as well as meet the information needs of stakeholders. These methods can also lead to a more selective approach to extensive impact evaluation, where existing knowledge is more systematically reviewed before undertaking a local impact evaluation.

"Systematic review" is a term that is used to indicate a number of methodologies that deal with synthesizing lessons from existing evidence. In general, one can define a systematic review as a synthesis of primary studies that contains an explicit statement of objectives and is conducted according to a transparent, methodical, and replicable methodology (Greenhalgh et al., 2004). Typical features of a protocol underlying a systematic review are the following (Oliver et al., 2005):

- Defining the review question(s)
- Developing the protocol
- Searching for relevant bibliographic sources
- Defining and applying criteria for including and excluding documents
- Defining and applying criteria for assessing the methodological quality of the documents
- Extracting information[1]
- Synthesizing the information into findings.

A meta-analysis is a quantitative aggregation of effect scores established in individual studies. The synthesis is often limited to a calculation of an overall effect score that expresses the impact attributable to a specific intervention or group of interventions. To arrive at such a calculation, meta-analysis involves a strict procedure to search for and select appropriate evidence of the impact of single interventions. The selection of evidence is based on an assessment of the methodology of the single-intervention impact study. In this type of assessment, usually a hierar-

chy of methods is applied in which RCTs rank highest and provide the most rigorous sources of evidence for meta-analysis. Meta-analysis differs from multicenter clinical trials in that in the former, the evaluator has no control over the single-intervention evaluations as such. As a result, despite the fact that homogeneity of implementation of similar interventions is a precondition for successful meta-analysis, inevitably meta-analysis is confronted with higher levels of variability in individual project implementation, context, and evaluation methodology than in multicenter clinical trials.

Meta-analysis is most frequently applied in professional fields such as medicine, education, and (to a lesser extent) criminal justice and social work (Clarke, 2006). Knowledge repositories such as the Campbell Collaboration and Cochrane Society rely heavily on meta-analysis as a rigorous tool for knowledge management on what works. Both from within these professional fields as well as from other fields criticism has emerged. In part, this criticism reflects a resistance to the idea of a "gold standard" underlying the practice of meta-analysis. The discussion has been useful in that it has helped define the boundaries of applicability of meta-analysis and the idea that, given the huge variability in parameters characterizing evaluations, there is no such thing as a gold standard (see Clarke, 2006).

Box 6.1: Narrative review and synthesis study: Targeting and impact of community-based development initiatives

The study was performed by Mansuri and Rao (2004), who reviewed the evidence on community-based development (CBD) projects funded by the World Bank. At the time, it was estimated that an estimated US$ 7 billion of World Bank projects were about CBD.

Review questions

1. Does community participation improve the targeting of private benefits such as welfare or relief?
2. Are the public goods created by community participation projects better targeted to the poor?
3. Are such goods of higher quality, or better managed, than similar public goods provided by the government?
4. Does participation lead to the empowerment of marginalized groups—does it lessen exclusion, increase the capacity for collective action, or reduce the possibility that locally powerful elites will capture project benefits?
5. Do the characteristics of external agents—donors, governments, nongovernmental organizations (NGOs), and project facilitators—affect the quality of participation or project success or failure?
6. Can community participation projects be sustainably scaled up?

To obtain relevant and reliable evidence on CBD projects, the reviewers decided to restrict the review process to peer-reviewed publications or studies conducted by independent researchers. This provided an exogenous rule that improved the quality and reduced the level of potential bias while casting a wide enough net to let in research from a variety of disciplinary perspectives on different types of CBD projects. The following sources of evidence were included: impact evaluations, which use statistical or econometric techniques to assess the causal impact of specific project outcomes; and ethnographic or case studies, which use anthropological methods such as participant observation, in-depth interviews, and focus group discussions.

Some conclusions

- Projects that rely on community participation have not been particularly effective at targeting the poor; there is some evidence that CBD/community-driven development projects create effective community infrastructure, but not a single study establishes a causal relationship between any outcome and participatory elements of a CBD project.
- A naïve application of complex contextual concepts like "participation," "social capital," and "empowerment" is endemic among project implementers and contributes to poor design and implementation.

Source: Mansuri and Rao (2004).

Partly as a response to the limitations in applicability of meta-analysis as a synthesis tool, more comprehensive methodologies of systematic review have been developed. One example is a systematic review of health behavior among young people in the United Kingdom that involves both quantitative and qualitative synthesis (see Oliver et al., 2005). The case shows that meta-analytic work on evidence stemming from what the authors call "intervention studies" (evaluation studies on similar interventions) can be combined with qualitative systematic review of "non-intervention studies," mainly research on relevant topics related to the problems addressed by the intervention. Regarding the latter, similar to the quantitative part, a systematic procedure for evidence search, assessment, and selection is applied. The difference lies mostly in the synthesis part, which in the latter case is a qualitative analysis of major findings. The two types of review can subsequently be used for triangulation purposes, reinforcing the overall synthesis findings.

Other examples of review and synthesis approaches are the narrative review and the realist synthesis. A narrative review is a descriptive account of intervention processes and/or results covering a series of interventions (see box 6.1.). Often, the evaluator relies on a common analytical framework, which serves as a basis for a template that is used for data extraction from the individual studies. In the end, the main findings are summarized in a narrative account and/or tables and matrices representing key aspects of the interventions.

A realist synthesis is a theory-based approach that helps synthesize findings across interventions. It focuses on the question of which mechanisms are assumed to be at work in a given intervention, taking into account the context the intervention operates in (see appendix 10). Although interventions often appear different, they often rely on strikingly similar mechanisms. Recognition of this can broaden the range of applicable evidence from other studies.

Combinations of meta-approaches are also possible. In a recent study on the impact of public policy programs designed to reduce and/or prevent violence in the public arena, Van der Knaap et al. (2008) have shown the relevance of *combining* synthesis approaches (see appendix 12).

Key message

Build on existing knowledge relevant to the impact of interventions. Review and synthesis methods can play a pivotal role in impact evaluation in synthesizing results and contributing to knowledge. Although interventions often appear different, they often may rely on strikingly similar mechanisms. Recognition of this can broaden the range of applicable evidence. As there are several approaches available, it is worthwhile to try to combine (some of) them. Review and synthesis work can provide a useful basis for empirical impact analysis of a specific intervention and in some cases may even take away the need for further in-depth impact evaluation.

Part II
Managing Impact Evaluations

Chapter 7
Determine if an impact evaluation is feasible and worth the cost

Managers and policy makers sometimes assume that impact evaluation is synonymous with any other kind of evaluation. They might request an "impact evaluation" when the real need is for a quite different kind of evaluation (e.g., to provide feedback on an implementation process or to assess the accessibility of program services to vulnerable groups). Ensuring clarity of the information needed and for what purposes is a prerequisite to defining the type of evaluation to conduct.

Moreover, impact evaluation is not "the" alternative but draws on and complements rather than replaces other types of M&E activities. It should therefore be seen as *one of several in a cycle of potentially useful evaluations in the lifetime of an intervention*. The rather traditional difference between ex ante and ex post impact evaluations remains important, where the ex ante impact assessment is, by nature, largely an activity in which "predictions" are made of any effects and side effects a particular intervention might have. Ex post impact evaluation, or simply "impact evaluation," as defined by the development community (and elsewhere), can test whether and to what extent these ex ante predictions have been correct. In fact, one of the potential uses of impact evaluations, not yet frequently applied in the field of development intervention, could be to strengthen the process of ex ante impact assessments.

When should an impact evaluation *ideally* be conducted?

- When there is an *articulated need* to obtain the information from an impact evaluation to know whether the intervention worked, to learn from it, to increase transparency of the intervention, and to know its "value for money."
- When a "readiness assessment" shows that political, technical, resource, and other practical considerations are adequate and it is feasible to do an impact evaluation. More specifically, this would include the following conditions:
  - The evaluation has a clearly defined purpose and an agreed-upon intended use, appropriate to its timing and with support of influential stakeholders.
  - There is clarity about the evaluation design. The evaluation design has to be clearly described and well justified after due consideration of alternatives and constraints.
  - The evaluation design has a chance to be credibly executed, given the nature and context of the intervention, the data, and

information needs and the availability of adequate resources and expertise to conduct the evaluation.
- When an intervention is functioning long enough to have visible effects.
- When there is sufficient scale (e.g., in terms of funding, number of people affected) to justify a thorough assessment.
- When the evaluation is likely to produce "new" knowledge, adding value to the public knowledge on the effectiveness of particular (innovative) types of interventions and the mechanisms that "do the work."

Impact evaluations may *not* be appropriate at particular times:

- When other valuable forms of evaluation will yield more useful information to support decisions to be made or other purposes
- When they move too many resources and too much attention away from the need to develop and use a rich spectrum of evaluation approaches and capacities
- When political, technical, practical, or resource considerations are likely to prevent a credible, rigorous, and useful evaluation
- When there are signs that the evaluation will not be used (or may be misused, for example, for political reasons).

Not all interventions should be submitted to elaborate and costly impact evaluation exercises. Rather, those sectors, regions, and intervention approaches about which less is known (including new, innovative ideas) should receive funding and support for impact evaluation. Ideally, organizations should pool their resources and expertise to select interventions of interest for rigorous and elaborate impact evaluation and consequently contribute jointly to the public good of knowledge on the impact of (under-evaluated) interventions.

Key message

Determine if an impact evaluation is feasible and worth the cost. Costs can be significant; what are the benefits? In what ways does the impact evaluation contribute to accountability, learning, and information about the "value for money" of what works? What is the likely added value of an impact evaluation in relation to what is already known about a particular intervention? What are the costs? What are the costs of estimating or measuring what would have happened without the intervention? Is the likelihood of getting accurate information on impact high enough to justify the cost of the evaluation?

Chapter 8

Start collecting data early

Although issues of data and data collection such as availability and quality often sound like "mere" operational issues that only need to be discussed on a technical level, it should not be forgotten that these aspects are of crucial importance for any impact evaluation (and any evaluation in general). Data are needed to test whether there have been changes in the dependent variables or to represent the *counterfactual* estimate of what the populations' situation would have been if the project had not taken place. The data issue is strongly linked to the method of evaluation.

8.1. Timing of data collection

Ideally, impact evaluations should be based on data from both before and after an intervention has been implemented.[1] An important question is if the baseline period or end-line period is representative or normal. If the baseline or end-line year (or season) is not normal, then this affects the observed change over time. If, for example, the baseline year is influenced by unusually high or low agricultural production or a natural disaster, then the observed change up to the end-line year can be strongly biased. In most cases it is the timing of the intervention, or the impact evaluation, that determines the timing of the baseline and end-line studies. This timing is not random, and evaluators need to investigate if the baseline/end-line data are representative of "normal" periods before they draw conclusions. If not, even rigorous evaluations may produce unreliable conclusions about impacts.

An additional issue concerns short-term versus long-term effects. Depending on the intervention and its context, at the time of ex post data collection some effects might not have occurred or not be visible yet, whereas others might wither over time. Evaluators should be aware of how this affects conclusions about impact.

8.2. Data availability

In practice, impact evaluation starts with an appraisal of existing data, the data that have been produced during the course of an intervention on inputs, processes, and outputs (and outcomes). This inventory is useful for several reasons:

- Available data are useful for reconstructing the intervention theory that further guides primary and secondary data collection efforts.
- Available data might affect the choice of methodological design or options for further data processing and analysis; for example, ex ante

and ex post data sets of target groups might be complemented with other data sets to construct useful control groups; the amount and type of data available might influence the choice of whether to organize additional primary data collection efforts.
- Available data from different sources allow for triangulation of findings.

In addition, evaluators can rely on a variety of data from other sources that can be used in the evaluation process:

- National census data
- General household surveys such as Living Standards Measurement Surveys
- Specialized surveys such as demographic and health surveys
- Administrative data collected by line ministries and other public agencies (e.g., on school enrolment, use of health facilities, market prices for agricultural produce)
- Studies conducted by donor agencies, NGOs, and universities
- Administrative data from agencies, ministries, or other organizations
- Mass media (newspapers, television documentaries, etc.); these can be useful, among other things, for understanding the local economic and political context of an intervention.

Appendix 13 describes an example of an impact evaluation implemented by IEG. In 1986 the government of Ghana embarked on an ambitious program of educational reform, shortening the length of pre-university education from 17 to 12 years, reducing subsidies at the secondary and tertiary levels, lengthening the school day, and taking steps to eliminate unqualified teachers from schools. There was no clearly defined "project" for this study, but the focus was World Bank support to the subsector through four large operations. These operations had supported a range of activities, from rehabilitating school buildings to assisting in the formation of community-based school management committees. The impact evaluation heavily relied on existing data sets such as the Ghana Living Standards Survey for impact analyses.

A useful stepwise approach for assessing data availability is the following:

1. Make an inventory of the availability of data and assess its quality. Sometimes secondary data can be used to carry out the whole impact study. This is especially true for national or sector-wide interventions. More usually, secondary data can be used to buttress other data.
2. Analyze, from the perspective of the intervention theory, the necessity of additional data. The process of data gathering must be based on the evaluation design which is, in turn, partly based on the intervention theory. Data must be collected across the results chain, not just on outcomes.
3. Assess the best way(s) to obtain additional data.
4. A comparison group sample must be of adequate size, and subject to the same, or virtually the same, questionnaire or other data collecting instruments. While some intervention-specific questions may not be appropriate, similar questions of a more general nature can help test for contagion.
5. It is necessary to check if other interventions, unexpected events, or other processes have influenced developments in the comparison group or the treatment group (i.e., check whether the comparison group is influenced by other processes than the treatment group).
6. Multiple instruments (e.g., household and facility level) are usually desirable and must be coded in such a way that they can be linked.
7. Baseline data must cover the relevant welfare indicators but preferably also the main determinants of the relevant welfare elements, so it will be easier to investigate later if other processes than the intervention have influenced welfare developments over time. End-line data must be collected across the results chain, not just on intended outcomes.

When there is no baseline, the option of a field survey using recall on the variables of interest may

be considered. Many commentators are critical of relying on recall. But all survey questions in the end are recall, so it is a question of degree. The evaluator needs to use his or her judgment (and knowledge about cognitive processes) as to what are credible data, given a respondent's capacity to recall.

8.3. Quality of the data

The quality of data can make or break any impact evaluation. Mixed methods and triangulation are strategies to reduce the problem of data quality. Yet in terms of the quality control that is needed to ensure that evaluation findings are not (heavily) biased because of data quality problems, they are insufficient.

The evaluator should ask several questions:

- What principles should we follow to improve the quality of data (collection)? Some examples of subquestions:
  - How to address missing data (missing observations in a data set, missing variables).
  - How to address measurement error—Does the value of a variable or the answer to a question represent the true value?
  - How to address specification error—Does the question asked or variable measured represent the concept that it was intended to cover?
- Does the quality of the data allow for (advanced) statistical analysis? New advances in and the more widespread use of quasi-experimental evaluation and multivariate data analysis are promising in light of impact evaluation. Yet often data quality is a constraining factor in terms of the quality of the findings (see Deaton, 2005).
- In the case of secondary data, what do we know about the data collection process that might strengthen or weaken the validity of our findings?[2]

De Leeuw et al. (2008) discuss data quality issues in survey data analysis. Much of their discussion on measurement error (errors resulting from respondent, interviewer, method, and question-related sources or a combination of these; examples are recall problems or the sensitivity of certain topics) is equally relevant for semistructured interviews and similar techniques in qualitative research. With respect to group processes in qualitative research, Cooke (2001) discusses three of the most widely cited problems: risky shift, groupthink, and coercive persuasion. A detailed discussion of these issues is beyond the scope of this guidance. However, they lead us to some important points:

- Data based on the perceptions, views, and opinions of people on the causes and effects of an intervention (e.g., target groups) do not necessarily adequately reflect the real causes of an intervention; data collected through observation, measurement, or counting (e.g., assets, farm size, infrastructure, profits) in general are less prone to measurement error (but are not always easy to collect or sufficient to cover all information needs).
- The quality of data is more often than not a constraining factor in the overall quality of the impact evaluation; it cannot be solved by sophisticated methods but might be solved in part through triangulation among data sources.

8.4. Dealing with data constraints

According to Bamberger et al. (2004: 8), "Frequently, funds for the evaluation were not included in the original project budget and the evaluation must be conducted with a much smaller budget than would normally be allocated for this kind of study. As a result, it may not be possible to apply the desirable data collection instruments (tracer studies or sample surveys, for example), or to apply the methods for reconstructing baseline data or creating control groups." Data problems are often correlated with or compounded by time and budget constraints. The scenarios laid out in table 8.1 can occur.

Bamberger et al. (2004) describe scenarios for working within these constraints. For example, the implications for quasi-experimental designs are that evaluators have to rely on less robust designs such as ex post comparisons only (see appendix 14).

Table 8.1: Evaluation scenarios with time, data, and budget constraints

The constraints under which the evaluation must be conducted			
Time	**Budget**	**Data**	**Typical Scenarios**
X			The evaluator is called in late in the project and is told that the evaluation must be completed by a certain date so that it can be used in a decision making process or contribute to a report. The budget may be adequate but it may be difficult to collect or analyze survey data within the time frame.
	X		The evaluation is only allocated a small budget, but there is not necessarily excessive time pressure. However, it will be difficult to collect sample survey data because of the limited budget.
		X	The evaluator is not called in until the project is well advanced. Consequently no baseline survey has been conducted either on the project population or on a control group. The evaluation does have an adequate scope, either to analyze existing household survey data or to collect additional data. In some cases the intended project impacts may also concern changes in sensitive areas such as domestic violence, community conflict, women's empowerment, community leadership styles, or corruption, on which it is difficult to collect reliable data—even when time and budget are not constraints.
X	X		The evaluator has to operate under time pressure and with a limited budget. Secondary survey data may be available but there is little time and few resources to analyze it.
X		X	The evaluator has little time and no access to baseline data or a control group. Funds are available to collect additional data but the survey design is constrained by the tight deadlines.
	X	X	The evaluator is called in late and has no access to baseline data or control groups. The budget is limited but time is not a constraint.
X	X	X	The evaluator is called in late, is given a limited budget, has no access to baseline survey data and no control group has been identified.

Source: Bamberger et al. (2004).

Key message

Start collecting data early. Good baseline data are essential to understanding and estimating impact. Depending on the type of intervention, the collection of baseline data, as well as the setup of other aspects of the impact evaluation, requires an efficient relationship between the impact evaluators and the implementers of the intervention. Policy makers and commissioners need to involve experts in impact evaluation as early as possible in the intervention design to be able to design high-quality evaluations. Ensuring high-quality data collection should be part and parcel of every impact evaluation. When working with secondary data, a lack of information on the quality of data collection can restrict data analysis options and the validity of findings. Take notice of and deal effectively with the restrictions under which an impact evaluation has to be carried out (time, data, and money).

Chapter 9

Front-end planning is important

Front-end planning refers to the initial planning and design phase of an impact evaluation. Ad hoc commissioned impact evaluations usually do not have a long planning period, thereby risking a suboptimally planned and executed evaluation process.

As good impact evaluation relies on good data, preferably including baseline data, attention to proper front-end planning should be a priority issue. Ideally, front-end planning of impact evaluations should be closely articulated to the initial design and planning phase of the policy intervention. Indeed, this articulation is most clearly visible in an RCT, in which intervention and impact evaluation are inextricably linked.

9.1. Planning tools

Clear definition of scope (chapters 1 and 2) and sound methodological design (chapters 3–6) cannot be captured in standardized frameworks. Decision trees on assessing data availability (see § 8.2.) and method choice (see appendix 6) are useful, though they provide only partial answers to methodological design choice issues. Pragmatic considerations of time, budget, and data (see § 8.4) but culture and politics also play a role. Two tools that are particularly helpful in the planning phase of an impact evaluation are the *approach paper* and the *evaluation matrix*.

The approach paper outlines what the evaluation is about and how it will be implemented. This document can be widely circulated and gives stakeholders and others a chance to comment and improve upon the intended evaluation design from an early stage. It also helps to generate broad "buy-in" or at worst to define the main grounds of potential disagreement between evaluators and practitioners. In addition, it is wise to use an evaluation matrix when planning and executing the work. This tool ensures that key questions are identified, together with the ways to address them, sources of data, role of theory, etc. This can also play an important role in stakeholder consultation, ensuring that important elements are not omitted.

9.2. Staffing and resources

Resources are important, and spending should be safeguarded up front. The longer the time horizon of a study, the more difficult this is. Resources are also important to realize the much-needed independence of an evaluator and the evaluation team. A template for assessing the independence of evaluation organizations can be downloaded from http://www.ecgnet.org/docs/ecg.doc. This document specifies a number of criteria and questions that can be asked.

Evaluation is not only a *financial resources business* but even more a *people's business.* So is the *planning of an evaluation*. As evaluation projects are usually no longer "*lonely hunter*" *activities*, *staffing* is crucial. So when starting the preparation of the study, a crucial point concerns addressing a number of questions:

- Who are the people who do the evaluation?
- Under which (contractual) conditions are they doing the job?
- What is their expertise?
- Which roles will they be carrying out?

Topics that deserve attention are the following:

- The *mix of disciplines and traditions* that are brought together in the team.
- The *competencies* the team has "in stock." Competencies range from methodological expertise to negotiating with institutional actors and stakeholders, getting involved in "hearing both sides" (those evaluated and the principal) and in the clearance of the report.
- The *structure of the evaluation team.* For the evaluation to be planned and carried out effectively, the roles of the project director, staff, and other evaluators must be made clear to all parties.
- The *responsibilities* of the team members.
- The more an evaluation is linked to a political "hot spot," the more it is necessary that at least one member of the team have a "political nose"—not primarily to deal with administrators and (local) politicians, but to understand when an evaluation project becomes too much of what is known as a *partnerial evaluation* (Pollitt, 1999).
- Also, staff should be active in realizing an *adequate documentation and evaluation trail*.

A range of skills is needed in evaluation work. The quality and eventual utility of the impact evaluation can be greatly enhanced with coordination between team members and policymakers from the outset. It is therefore important to identify team members as early as possible, agree on roles and responsibilities, and establish mechanisms for communication during key points of the evaluation.

9.3. The balance between independence and collaboration between evaluators and stakeholders

One of the questions within the world of impact evaluations is what degree of institutional separation to put between the evaluation providers and the evaluation users. There is much to be gained from the objectivity provided by having the evaluation carried out independently of the institution responsible for the project being evaluated. Pollitt (1999) warned against "partnerial" evaluations, where positions of stakeholders, commissioners, and evaluators blurred too much.[1] However, evaluations often have multiple goals, including building evaluation capacity within government agencies and sensitizing program operators to the realities of their projects once they are carried out in the field. At a minimum, the evaluation users, who can range from government agencies in client countries to bilateral and multilateral donors, international NGOs, and grass roots/civil society organizations, must remain sufficiently involved in the evaluation to ensure that the evaluation process is recognized as legitimate and that the results produced are relevant to their information needs. Otherwise, the evaluation results are less likely to be used to inform policy. The evaluation manager and his or her clients must achieve the right balance between involving the users of evaluations and maintaining the objectivity and legitimacy of the results (Baker, 2000).

9.4. Ethical issues

It is important to take the ethical objections and political sensitivities seriously. There can be ethical concerns with deliberately denying a program to those who need it and providing the program to those who do not; this applies to both experimental and non-experimental methods. For example, with too few resources, randomization may be seen as a fair solution, possibly after conditioning on observables. However, the information available to the evaluator (for conditioning) is typically a partial subset of the information available "on the ground" (includ-

ing voters/taxpayers). The idea of "intention-to-treat" helps alleviate these concerns; one has a randomized assignment, but anyone is free to not participate. Even then, the "randomized out" group may include people in great need. All these issues must be discussed openly and weighed against the (potentially large) longer-term welfare gains from better information for public decision making (Ravallion, 2008).[2]

9.5. Norms and standards

As noted before, impact evaluations are often designed, implemented, analyzed, disseminated, and used under budget, time, and data constraints while facing diverse and often competing political interests. Given these constraints, the management of a real-world evaluation is much more complicated than textbook descriptions.

Evaluations sometimes fail because the stakeholders were not involved, or the findings were not used because they did not address the stakeholders' priorities. Others fail because of administrative or political difficulties in getting access to the required data, being able to meet with all the individuals and groups that should be interviewed, or being able to ask all the questions that the evaluator feels are necessary. Many other evaluations fail because the sampling frame, often based on existing administrative data, omits important sectors of the target population—often without anyone being aware of this. In other cases the budget was insufficient, or was too unpredictable to permit an adequate evaluation to be conducted. Needless to say, evaluations also fail because of emphasizing stakeholders´ participation too much, leading to partnerial evaluations (Pollitt, 1999), and because of insufficient methodological and theoretical expertise.

Although many of these constraints are presented in the final evaluation report as being completely beyond the control of the evaluator, in fact their effects could very probably have been reduced by more effective management of the evaluation. For example, a more thorough scoping analysis could have revealed many of these problems, and the client(s) could then have been made aware of the likely limitations on the methodological rigor of the findings. The client(s) and evaluator could then strategize to either seek ways to increase the budget or extend the time, or agree to limit the scope of the evaluation and what it promises to deliver. If clients understand that the current design will not hold up under the scrutiny of critics, they can find ways to help address some of the constraints: "We have found that impact evaluations generally provide rudimentary documentation of the data being used. There is evidently a trade-off between decision makers' and bureaucrats' appeal for short and crisp reports and principles for scientific documentation, but we want to emphasise that displaying descriptive statistics improves the transparency of the methodological approach" (Jerve and Villanger, 2008: 34).

For the sake of honest commitment to development, evaluators and evaluation units should ensure that impact evaluations are designed and executed in a manner that limits manipulation of processes or results that lean toward any ideological or political agenda. They should also ensure that there are realistic expectations of what can be achieved by a single evaluation within existing time and resource constraints, and that findings from the evaluation are presented in ways that are accessible to the intended users. This includes finding a balance between simple, clear messages and properly acknowledging the complexities and limitations of the findings.

International evaluation standards (such as the OECD-DAC or the United Nations Evaluation Group Norms and Standards and/or the standards and guidelines developed by national or regional evaluation associations) should be applied where appropriate (Picciotto, 2004).

Greater emphasis on impact evaluation for evidence-based policy making can create greater risk of manipulation aimed at producing desirable results (House, 2008). Impact evaluations require an honest search for the truth and thus place high demands on the integrity of those commissioning and conducting them. For the sake of honest commitment to development, evaluators and evaluation units need to ensure that

impact evaluations are designed and executed in a manner that limits manipulation of processes or that produces results favoring any ideological or political agenda.

9.6. Ownership and capacity building

Capacity building at the level of governmental or non-governmental agencies involved should be an explicit purpose in impact evaluation. In cases where sector-wide investment programs are financed by multidonor co-financing schemes, participating donors would make natural partners for a joint evaluation (OECD-DAC, 2000).

Other factors in selecting other donors as partners in a joint evaluation work may also be relevant. Selecting donors with similar development philosophies, cultures, evaluation procedures and techniques, and regional affiliations, and that are geographically close may make working together easier. Another issue may be keeping the total number of donors "manageable." Where more donors are involved, a key group of development partners (including national actors) could assume management responsibilities and the role of others can be more limited. Once appropriate donors that have a likely stake in an evaluation topic are identified, the next step is to contact them and discern whether they are interested in participating. In some cases, an appropriate consortium or group may already exist, where the issue of a joint evaluation can be raised and expressions of interest easily solicited. The DAC Working Party on Aid Evaluation, the United Nations Evaluation Group, and the Evaluation Cooperation Group have a tradition of cooperation, shared vision on evaluation, and longstanding relationships and have fostered numerous joint evaluations.

The interaction among the international development evaluation community, the countries/regions themselves, and the academic evaluation communities should also be stimulated, as it is likely to affect the pace and quality of capacity building in impact evaluation. Capacity building will also strengthen (country and regional) *ownership* of impact evaluation. Providing a space for consultation and agreement on impact evaluation priorities among the different stakeholders of an intervention will also help enhance utilization and ownership.

Key message

Front-end planning is important. It can help manage the study, its reception, and its use. When managing the evaluation, keep a clear eye on items such as costs, staffing, ethical issues, and level of independence—of the evaluator and the team, versus the level of collaboration with stakeholders. Pay attention to country and regional ownership of impact evaluation and capacity building and promote it. Providing a space for consultation and agreement on impact evaluation priorities among the different stakeholders of an intervention will help enhance utilization and ownership

Appendixes

Example 1. Evaluating the impact of a European Union-funded training project on Low External Input Agriculture in Guatemala

Within the framework of a European Union-funded integrated rural development project, financial support was provided to a training project aimed at the promotion of Low External Input Agriculture (LEIA) as a viable agricultural livelihood approach for small farmers in the highlands of western Guatemala.

The impact evaluation design of this project was based on a quasi-experimental design and complemented by qualitative methods of data collection (Vaessen and De Groot, 2004). An intervention theory was reconstructed on the basis of field observations and relevant literature to make explicit the different causal assumptions of the project, facilitating further data collection and analysis. The quasi-experimental design included data collection on the ex ante and ex post situation of participants, complemented with ex post data collection involving a control group (based on judgmental matching using descriptive statistical techniques). Without complex matching procedures and with limited statistical power, the strength of the quasi-experiment relied heavily on additional qualitative information. This shift in emphasis should not give the impression of a lack of rigor. Problems such as the influence of selection bias were explicitly addressed, even if not done in a formal statistical way.

Farmers' adoption behavior after the termination of the project can be characterized as selective and partial. Given the particular circumstances of small farmers (e.g., risk aversion, high opportunity costs of labor), it is not realistic to assume that a training project will bring about a complete transformation from a conventional farming system to a LEIA farming system (as assumed in the objectives). In line with the literature, the most popular practices (in this case, for example, organic fertilizers and medicinal plants) were those that offer a clear short term return while not requiring significant investments in terms of labor or capital. Finally, an ideological faith in the absolute supremacy of LEIA practices is not in the best interest of the farmers. Projects promoting LEIA should focus on the complementary effects of LEIA practices and conventional farming techniques, encouraging each farmer to choose the best balance fitted to his/her needs.

Example 2. Assessing the impact of Swedish program aid

White and Dijkstra (2003) analyzed the impact of Swedish program aid. Their analysis accepted from the start that it is impossible to separate the impact of Swedish money from that of other donors' money. Therefore, the analysis focuses on all program aid with nine (country) case studies that trace how program aid has affected macro-economic aggregates (like imports and government spending) and (through these indicators) economic growth. The authors discern two channels for influencing policy: money and policy dialogue. The main evaluation questions are—

1. How has the policy dialogue affected the pattern and pace of reform (and what has been the contribution of program aid to this process)?
2. What has the impact of the program aid funds (on imports, government expenditure, investment, etc.) been?

3. What has the impact of reform programs been?

Their analytical model treats donor funds and the policy dialogue as inputs; specific economic, social, and political indicators as outputs; and the main program objectives (like economic growth, democracy, human rights and gender equality) as outcomes; and poverty reduction as the overall goal.

The analysis focuses on marginal impact and uses a combination of quantitative and qualitative approaches (interviews, questionnaires, and e-mail enquiries). The analysis of the impact of aid is largely quantitative, while the analysis of the impact of the policy dialogue is mainly qualitative.

An accounting approach is used to identify aid impact on expenditure levels and patterns using a number of ad hoc techniques, such as analyzing behavior during surges and before versus after breaks in key series and searching the data for other explanations of the patterns observed.

Moreover, the authors analyze the impact of aid on stabilization through—

a. The effect on imports
b. Its impact on the markets for domestic currency and foreign exchange
c. The reduction of inflationary financing of the government deficit.

In terms of the impact of program aid on reform, domestic political considerations are a key factor in determining reform: most countries have initiated reform without the help from donors and have carried out some measure of reform not required by them, while ignoring others that have been required.

APPENDIX 2: THE GENERAL ELIMINATION METHODOLOGY AS A BASIS FOR CAUSAL ANALYSIS

What are the core elements of the General Elimination Methodology (also known as the modus operandi approach)? We follow Scriven (2008).[1]

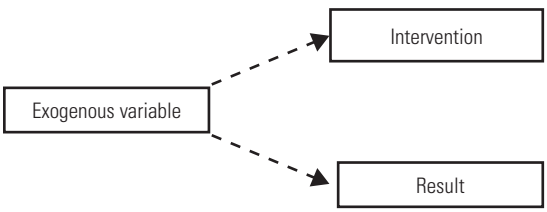
i. The general premise is the deterministic principle: all macro events (or conditions, etc.) have a cause. This is only false at the micro-level, where the uncertainty principle applies, but the latter principle has essentially no detectable effect on the truth of macro determinism (though it is easy enough to deliberately create bizarre experiments where it does).
ii. The first "premise from practice" is the list of possible causes (LOPC) of events of the type in which we are interested, e.g., learning gains, reduction of poverty, and extension of life for AIDS patients. We have used LOPCs for more than a million years, in tracking and cooking and healing and repairing, and today every detective knows the list for murder, just as every competent mechanic knows the list for a big-end rattle or a brake failure, though the knowledge is as often tacit as explicit, outside the classroom and the maintenance videos. An LOPC usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualization, and will vary depending on these parameters; of course, the context of the investigation determines the appropriate distance parameters. The distant LOPC for murder is the list of possible motives; a more proximate one, developed in a particular case by applying the general one, is the list of suspects. When dealing with new effects, we may not be certain the list is complete, but we work with the list we have and extend it when necessary.
iii. The second practical premise is the list of the modus operandi for each of the possible causes (the MOL). Each cause has a set of footprints, a short one if it's a proximate cause, a long one if it's a remote cause, but in general the modus operandus is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective. There's often a rubric for this; for example, in criminal (and most other) investigations into human agency, we use the rubric of means/motives/opportunity to get from the motives to the list of "suspects." The list of modus operandi is the magnifying lens that fleshes out the candidate causes from the LOPC so that we can start fitting them to the case or rejecting them, for which we use the next premise.
iv. The fourth premise comprises the "facts of the case," and these are now assembled selectively, by looking for the presence or absence of factors listed in the modus operandi of each of the LOPCs. Only those causes are (eventually) left standing whose modus operandi are completely present. Ideally, there will be just one of these, but sometimes more than one, which are then co-causes. (Note that there is no reference to counterfactuals.)

APPENDIX 3: OVERVIEW OF QUANTITATIVE TECHNIQUES OF IMPACT EVALUATION

		Analysis of intervention(s)	
		Explicit counterfactual (with/without)	**Analysis of multiple interventions and influences**
SELECTION EFFECTS	OBSERVED	Propensity score	Regression analysis
	UNOBSERVED	Randomized controlled trial pipeline approach Double difference (Difference in difference) Regression discontinuity	Difference in difference regression Fixed effects regression Instrumental variables

Endogeneity

The selection on unobservables is an important cause of *endogeneity*, a correlation of one of the explanatory variables with the error term in a mathematical model. This correlation occurs when an omitted variable has an effect at the same time on the dependent variable and an explanatory variable.[1]



When a third variable is not included in the model, the effect of the variable becomes part of the error term and contributes to the "unexplained variance." As long as this variable does not have an effect at the same time on one of the explanatory variables in the model, this does not lead to biased estimates. However, when this third variable has an effect on one of the explanatory variables, this explanatory variable will "pick up" part of the error and therefore will be correlated with the error. In that case, omission of the third variable leads to a biased estimate.

Suppose we have the relation

$$Y_i = a + bP_i + cX_i + e_i ,$$

where Y_i is the effect, P_i is the program or intervention, X_i is an unobserved variable, and e_i is the error term. Ignoring X we try to estimate the equation

$$Y_i = a + bP_i + e_i ,$$

while in effect we have

$$Y_i = a + bP_i + (e_i + e_x),$$

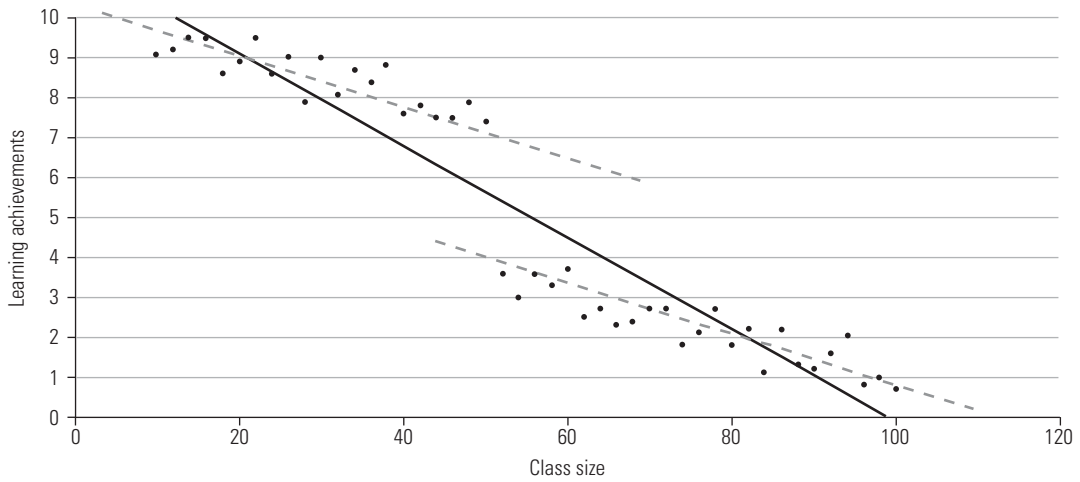
where e_i is a random error term and e_x is the effect of the unobserved variable. P and e_x are correlated and therefore P is *endogenous*. Ignoring this correlation results in a biased estimate of b. When the source of the selection bias (X) is known, inclusion of this variable (or these variables) leads to an unbiased estimate of the effect

$$Y_i = a + bP_i + cX_i + e_i .$$

An example is the effect of class size on learning achievements. The school choice of motivated (and probably well-educated) parents is probably correlated with class size, as these parents tend to send their children to schools with low pupil:teacher ratios. The neglect of the *endogeneity* of class size may lead to biased estimates (with an overestimation of the real effect of class size). When the selection effects are observable, a regression-based approach may be used to get an unbiased estimate of the effects.

Figure A4.1 gives the relation between class size and learning achievements for two groups of schools: the left side of the figure shows private schools in urban areas with pupils with relatively rich and well educated parents; the right side shows public schools with pupils from poor remote rural areas. A neglect of the differences between the two schools leads to a biased estimate, as shown by the black line. Including these effects in the equation leads to the smaller effect of the dotted lines.

Figure A4.1: Estimation of the effect of class size with and without the inclusion of a variable correlated with class size



Double difference and regression analysis

The technique of "double differencing" can also be applied in a regression analysis. Suppose that the anticipated effect (Y) is a function of participation in the project (P) and of a vector of background characteristics. In a regression equation we may estimate the effect as

$$Y_i = a + bP_i + cX_i + e_i ,$$

where e is the error term and a, b, and c the parameters to be estimated.

When we analyze changes over time, we get (taking the *first differences* of the variables in the model):

$$(Y_{i,1} - Y_{i,0}) = a + b(P_{i,1} - P_{i,0}) + c\,(X_{i,1} - X_{i,0}) + e_i.$$

When the (unobserved) variables X are time invariant, $(X_{i,1} - X_{i,0}) = 0$, and these variables drop from the equation. Suppose, for instance that a variable X denotes the "year of birth." For every individual the year of birth in year 1 = year of birth in year and therefore $(X_{i,1} - X_{i,0}) = 0$. So, if we expect that the year of birth is correlated with the probability of being included in the program and with the anticipated effect of the program, but we have no data on the year of birth, we may get an unbiased estimate by taking the first differences of the original variables. This technique helps to get rid of the problem of "unobservables."[2]

Instrumental variables

The use of instrumental variables is another technique to get rid of the endogeneity problem. A good instrument correlates with the (endogenous) intervention, but not with the error term. This instrument is used to get an unbiased estimate of the effect of the endogenous variable.

In practice, researchers often use the method of *two-stage least squares*: in the first stage an *exogenous* variable (Z) is used to give an estimate of the endogenous intervention-variable (P):

$$P_i' = a + dZ_i + e_i .$$

In the second stage this new variable is used to get an unbiased estimate of the effect of the intervention:

$$Y_i = a + bP'_i + cX_i + e_i .$$

The computation of propensity scores

The method of *propensity score matching* involves forming pairs by matching on the *probability* that subjects have been part of the treatment group. The method uses all *available* information to construct a control group. A standard way to do this is using a *probit* or *logit* regression model. In a logit specification, we get

$$ln\,(p_i / (1–p_i)) = a + bXi + cYi + dZi + e_i\,,$$

where p_i is the probability of being included in the intervention group and X, Y, and Z denote specific *observed* characteristics. In this model, the probability is a function of the observed characteristics. Rosenbaum and Rubin (1983) proved that when subjects in the control group have the same probability of being included in the treatment group as subjects who actually belong to the treatment group, the treatment and control groups will have similar characteristics.

Agriculture and rural development

Case study: Pakistan

The projects: Irrigation in Pakistan suffers from the "twin menaces" of salinity and waterlogging. These problems have been tackled through Salinity Control and Reclamation Projects (SCARPs), financed in part by the Bank. Although technically successful, SCARP tubewells imposed an unsustainable burden on the government's budget. The project was to address this problem in areas with plentiful groundwater by closing public tubewells and subsidizing farmers to construct their own wells.

Methodology: The Independent Evaluation Group (IEG) commissioned a survey in 1994 to create a panel from two earlier surveys undertaken in 1989 and 1990. The survey covered 391 farmers in project areas and 100 from comparison areas. Single and double differences of group means are reported.

Findings: The success of the project was that the public tubewells were closed without the public protests that had been expected. Coverage of private tubewells grew rapidly. However, private tubewells grew even more rapidly in the control area. This growth may be a case of contagion, though a demonstration effect. But it seems more likely that other factors (e.g., availability of cheaper tubewell technology) were behind the rapid diffusion of private water exploitation. Hence the project did not have any impact on agricultural productivity or incomes. It did, however, have a positive rate of return by virtue of the savings in government revenue.

Case study: Philippines

The project: The Second Rural Credit Projects (SRCP) operated between 1969 and 1974 with a US$12.5 million loan from the World Bank. SRCP was the continuation of a pilot credit project started in 1965 and completed in 1969. As its successful predecessor, SRCP aimed to provide credit to small and medium rice and sugar farmers for the purchase of farm machinery, power tillers, and irrigation equipment. Credits were to be channeled through 250 rural banks scattered around the country. An average financial contribution to the project of 10% was required from both rural banks and farmers. The SRCP was followed by a third loan of US$22.0 million from 1975 to 1977 and by a fourth loan of US$36.5 million that was still in operation at the time of the evaluation (1983).

Methodology: The study uses data of a survey of 738 borrowers (nearly 20% of total project beneficiaries) from seven provinces of the country. Data were collected through household questionnaires on land, production, employment, and measures of standard of living. In addition, 47 banks were surveyed to measure the impact on their profitability, liquidity, and solvency. The study uses before-and-after comparisons of means and ratios to assess the project impact on farmers. National level data are often used to validate the effects observed. Regarding the rural banks, the study compares measures of financial performance before and after the project, taking advantage of the fact that the banks surveyed joined the project at different stages.

Findings: The mechanization of farming did not produce an expansion of holding sizes (though

the effect of a contemporaneous land reform should be taken into account). Mechanization did not change cropping patterns, and most farmers were concentrating on a single crop at the time of the interviews. No change in cropping intensity was observed, but production and productivity were found to be higher at the end of the project. The project increased the demand for both family and hired labor. Farmers reported an increase in incomes and savings, and in several other welfare indicators, as a result of the project. Regarding the project impact on rural banks, the study observes an increase in the net income of the sample banks from 1969 to 1975 and a decline thereafter. Banks' liquidity and solvency position was negatively affected by poor collection and loan arrears.

Health, nutrition, and population

Case study: India

The project: The Tamil Nadu Integrated Nutrition Project (TINP) operated between 1980 and 1989, with a credit of US$32 million from the International Development Association (IDA). The overall objective of the project was to improve the nutritional and health status of pre-school children, pregnant women, and nursing mothers. The intervention consisted of a package of services including nutrition education, primary health care, supplementary feeding, administration of vitamin A, and periodic de-worming. The project was the first to employ Growth Monitoring and Promotion (GMP) on a large scale. The evaluation is concerned with the impact of the project on the nutritional status of children.

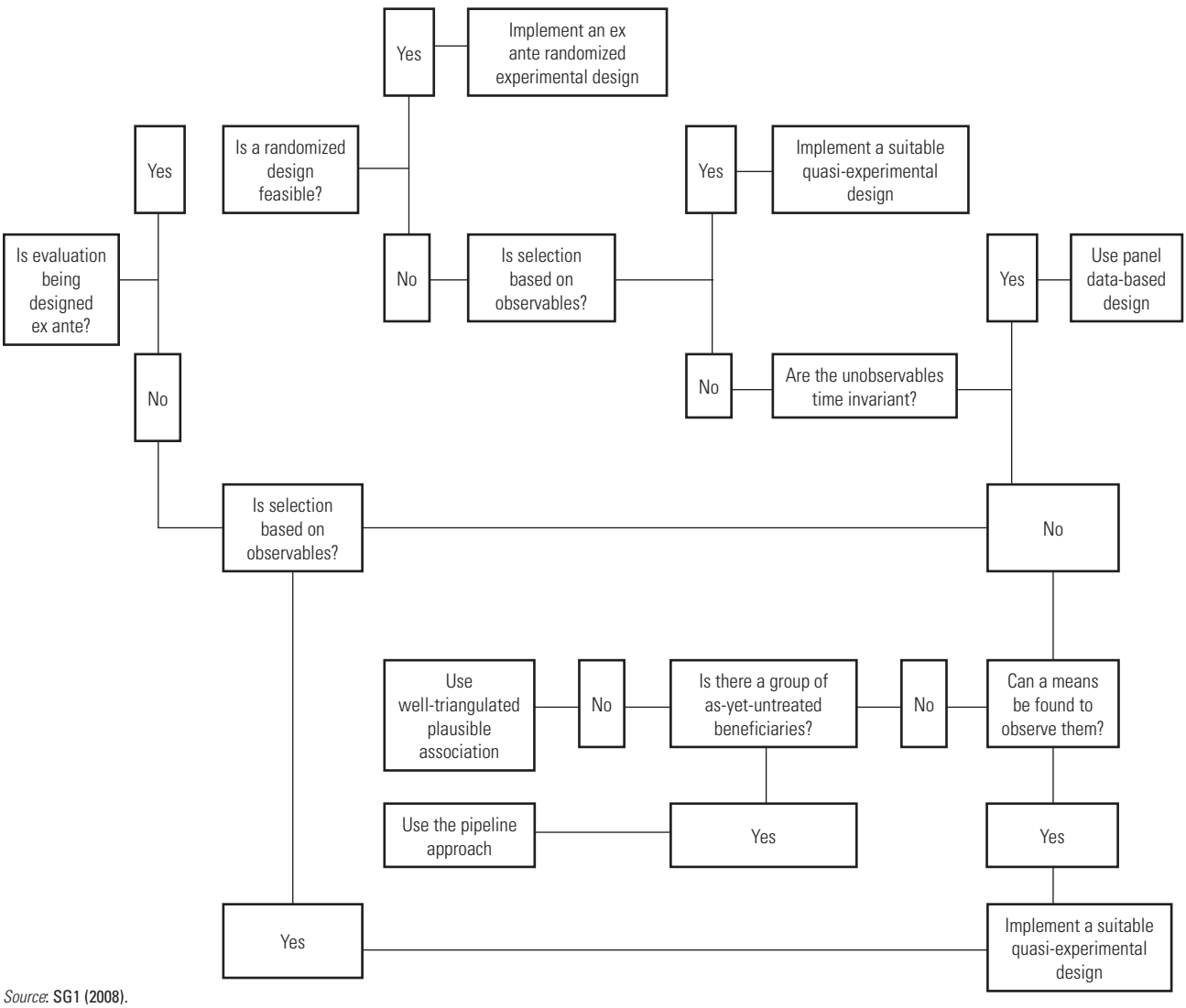
Methodology: The study uses three cross-sectional rounds of data collected by the TINP Monitoring Office. Child and household characteristics of children participating in the program were collected in 1982, 1986, and 1990, each round consisting of between 1,000 and 1,500 observations. The study uses before-and-after comparisons of means, regression analysis, and charts to provide evidence of the following: frequency of project participation, improvement in nutritional status of participating children over time, differential participation, and differential project impact across social groups. Data on the change in nutritional status in project areas are compared to secondary data on the nutritional status of children outside the project areas. With some assumptions, the use of secondary data makes the findings plausible.

Findings: The study concludes that the implementation of GMP programs on a large scale is feasible and that this had a positive impact on the nutritional status of children of Tamil Nadu. More specifically, these are the findings of the study:

- Program participation: Among children participating in GMP, all service delivery indicators (age at enrolment, regular attendance of sessions, administration of vitamin A, and de-worming) show a substantial increase between 1982 and 1986, though subsequently they declined to around their initial levels. Levels of service delivery, however, are generally high.
- Nutritional status: Mean weight and malnutrition rates of children aged between 6 and 36 months and participating in GMP have improved over time. Data on non-project areas in Tamil Nadu and all-India data show a smaller improvement over the same time period. Regression analysis of nutritional status on a set of explanatory variables, including the participation in a cotemporaneous nutrition project (the National Meal Program) shows that the latter had no additional benefit on nutritional outcomes. Positive associations are also found between nutritional status and intensive participation in the program and complete immunization.
- Targeting: Using tabulations and regression analysis, it is shown that initially girls have benefited more from the program, but that at the end of the program boys have benefited more. Children from the scheduled caste are shown to have benefited more than other groups. Nutritional status was observed to be improving at all income levels, the highest income category benefiting slightly more than the lowest.

Decision tree for impact evaluation design using quantitative impact evaluation techniques

1. If the evaluation is being designed before the intervention (ex ante), is randomization possible? If the treatment group is chosen at random, then a random sample drawn from the sample population is a valid control group and will remain so provided they are outside the influence zone and contamination is avoided.



Source: SG1 (2008).

This approach does not mean that targeting specific analytical units is not possible. The random allocation may be to a subgroup of the total population, e.g., from the poorest districts.

2. If randomization is not possible, are all selection determinants observed? If they are, then there are a number of regression-based approaches that can remove the selection bias.
3. If the selection determinants are unobserved and if they are thought to be time invariant, then using panel data will remove their influence, so a baseline is essential (or some means of substituting for a baseline).
4. If the study is done ex post so it is not possible to get information for exactly the same units (a panel of persons, households, etc.) and selection is determined by unobservables, then some means of observing the supposed unobservables should be sought. If that is not possible, then a pipeline approach can be used if there are as-yet untreated beneficiaries. For example, the Asian Development Bank's impact study of microfinance in the Philippines matched treatment areas with areas that were in the program but that had not yet received the intervention.
5. If none of the above mentioned procedures is possible, then the problem of selection bias cannot be addressed. The impact evaluation will have to rely heavily on the intervention theory and triangulation to build an argument by plausible association.

This group of approaches covers a quite diverse set of advanced modeling and statistical approaches. Detailed discussion of these technical features is beyond the scope of this document. The common element that binds these approaches is purpose modeling and estimating direct and indirect effects of interventions at various levels of aggregation (from micro to macro). At the risk of substantial oversimplification we briefly mention a few of the approaches. In hierarchical modeling, evaluators and researchers look at the interrelationships between different levels of a program. The goal is "to measure the true and often intertwined effects of the program. In a typical hierarchical linear model analysis, for example, the emphasis is on how to model the effect of variables at one level on the relations occurring at another level. Such analyses often attempt to decompose the total effect of the program into the effect across various program levels and that between program sites within a level (Dehejia, 1999)" (Yang et al., 2004: 494).

Also part of this branch of approaches is a range of statistical approaches such as nested models, models with latent variables, multi-level regression approaches, and others (see, for example, Snijders and Bosker 1999). Other examples are typical economist tools such as partial equilibrium analyses; general computable equilibrium models (CGEs) are often used to assess the impact of, for example, macroeconomic policies on markets and example, subsequently on household welfare (see box A7.1).

Box A7.1: Impact of the Indonesian financial crisis on the poor: Partial equilibrium modeling and CGE modeling with microsimulation

General equilibrium models permit the analyst to examine explicitly the indirect and second-round consequences of policy changes. These indirect consequences are often larger than the direct, immediate impact, and may have different distributional implications. General equilibrium models and partial equilibrium models may thus lead to significantly different conclusions. A comparison of conclusions reached by two sets of researchers, examining the same event using different methods, reveals the differences between the models. Levinsohn et al. (1999) and Robillard et al. (2001) both look at the impact of the Indonesian financial crisis on the poor—the former using partial equilibrium methods, the latter using a CGE model with micro-simulation. The Levinsohn study used consumption data for nearly 60,000 households from the 1993 SUSENAS survey, together with detailed information on price changes over the 1997–98 crisis period, to compute household-specific cost-of-living changes. It finds that the poorest urban households were hit hardest by the shock, experiencing a 10%–30% increase in the cost of living (depending on the method used to calculate the change). Rural households and wealthy urban households actually saw the cost of living fall.

These results suggest that the poor are just as integrated into the economy as other classes but have fewer opportunities to smooth consumption during a crisis. However, the methods used have at least three serious drawbacks. First, the consumption parameters are fixed; that is, no substitution is permitted

(*continued on next page*)

Box A7.1: Impact of the Indonesian financial crisis on the poor: Partial equilibrium modeling and CGE modeling with microsimulation (*continued*)

between more expensive and less expensive consumption items. Second, the results are exclusively *nominal*, in that the welfare changes are due entirely to changes in the price of consumption and do not account for any concomitant change in income. Third, this analysis cannot control for other exogenous events, such as the El Niño drought and resulting widespread forest fires.

Robillard et al. (2001) use a CGE model, connected to a microsimulation model. The results are obtained in two steps. First, the CGE is run to derive a set of parameters for prices, wages, and labor demand. These results are fed into a micro-simulation model to estimate the effects on each of 10,000 households in the 1996 SUSENAS survey. In the microsimulation model, workers are divided into groups according to sex, residence, and skill. Individuals earn factor income from wage labor and enterprise profits, and households accrue profits and income to factors in proportion to their endowments. Labor supply is endogenous. The micro-simulation model is constrained to conform to the aggregate levels provided by the CGE model. The Robillard team finds that poverty did increase during the crisis, although not as severely as the previous results suggest. Also, the increase in poverty was due in equal parts to the crisis and to the drought. Comparing their microsimulation results to those produced by the CGE alone, the authors find that the representative household model is likely to *underestimate* the impact of shocks on poverty. In contrast, ignoring both substitution and income effects, as Levinsohn et al. (1999) do, is likely to lead to *overestimating* the increase in poverty, since it does not permit the household to reallocate resources in response to the shock.

Source: World Bank (2003).

Multi-site evaluation approaches involve primary data collection processes and analyses at multiple sites or interventions. They usually focus on programs encompassing multiple interventions implemented in different sites (Turpin and Sinacore, 1991; Straw and Herrell, 2002). Although these approaches are often referred to as a family of methodologies, in what follows, and in line with the literature, we will use a somewhat more narrow definition of multi-site evaluations alongside several specific methodologies to address the issue of aggregation and cross-site evaluation of multiple interventions.

Straw and Herrell (2002) use the term "multi-site evaluation" both as an overarching concept, i.e., including cluster evaluation and multi-center clinical trials, as well as a particular type of multi-level evaluation distinguishable from cluster evaluation and multi-center clinical trials. Here we use the latter definition to refer to a particular (though rather flexible) methodological framework applicable to the evaluation of comprehensive multilevel programs addressing health, economic, environmental, or social issues.

The *multi-center clinical trial* is a methodology in which empirical data collection in a selection of homogenous intervention sites is systematically organized and coordinated. Basically it consists of a series of randomized controlled trials. The latter are experimental evaluations in which treatment is randomly assigned to a target group while a similar group not receiving the treatment is used as a control group. Consequently, changes in impact variables between the two groups can be traced back to the treatment, as all other variables are assumed to be similar at group level. In the multi-center clinical trial sample size is increased and multiple sites are included in the experiment in order to strengthen the external validity of the findings. Control over all aspects of the evaluation is very tight to keep as many variables constant over the different sites. Applications are mostly found in the health sector (see Kraemer, 2000).

Multi-site evaluation distinguishes itself from cluster evaluation in the sense that its primary purpose is summative. In addition, multi-site evaluations are less participatory in nature vis-à-vis intervention staff. In contrast to settings in which multi-center clinical trials are applied, multi-site evaluations address large-scale programs that, because of their (complex) underlying strategies, implementation issues, or other reasons, are not amenable to controlled experimental impact evaluation designs. Possible variations in implementation among interventions sites, and variations in terms of available data require a different, more flexible approach to data collection and analysis than in the case of the multi-center clinical trials. A common framework of questions and indicators is established to counter this variability, enabling data analysis across interventions in function of establishing generalizable findings (Straw and Herrell, 2002).

Cluster evaluation is a methodology that is especially useful for evaluating large-scale interventions that address complex societal themes such as education, social service delivery, and health promotion. Within a cluster of projects under evaluation, implementation among interventions may vary widely, but single interventions are still linked in terms of common strategies, target populations, or problems that are addressed (Worthen and Schmitz, 1997).

The approach was developed by the Kellogg Foundation in the 1990s and since then has been taken up by other institutions. Four elements characterize cluster evaluation (Kellogg Foundation, 1991):

- It focuses on a group of projects in order to identify common issues and patterns.
- It focuses on what happened as well as why.
- It is based on a collaborative process involving all relevant actors, including evaluators and individual project staff.
- Project-specific information is confidential and not reported to the higher level; evaluators only report aggregate findings; this type of confidentiality between evaluators and project staff induces a more open and collaborative environment.

Cluster evaluation is typically applied during program implementation (or during the planning stage) in close collaboration with stakeholders from all levels. Its purpose is, on the one hand, formative, as evaluators in close collaboration with stakeholders at project level try to explore common issues as well as variations between sites. At the program level the evaluation's purpose can be both formative in terms of supporting planning processes as well as summative, i.e., judging what went wrong and why. A common question at the program level would be, for example, to explore the factors that in the different sites are associated with positive impacts. In general, the objective of cluster evaluations is not so much to prove as to improve, based on a shared understanding of why things are happening the way they are (Worthen and Schmitz, 1997). It should be noted that not only cluster evaluations but also multi-site evaluations are applicable to homogenous programs with little variation in terms of implementation and context among single interventions.

APPENDIX 9: METHODOLOGICAL FRAMEWORKS FOR ASSESSING THE EFFECTS OF INTERVENTIONS, MAINLY BASED ON QUALITATIVE METHODS[1]

Outcome mapping

Outcome mapping (IDRC, 2001) is a methodology that focuses on outcomes as behavioral change. The outcomes can be logically linked to an intervention's activities, although they may not be necessarily directly caused by them. These changes are aimed at contributing to specific aspects of human and ecological well-being by providing partners with new tools, techniques, and resources to contribute to the development process. "Boundary partners" are individuals, groups, and organizations with whom the intervention interacts directly and with whom the intervention anticipates opportunities for influence; most activities will involve multiple outcomes because they have multiple boundary partners.

Success case method

The success case method (Brinkerhoff, 2003) is a widely adopted example of a mixed-method framework, drawing from several established traditions, including theory-based evaluation, organizational development, appreciative inquiry, narrative analysis, and quantitative statistical analysis of impact. It has been expanded in scope by those who combine it with realist methodologies (e.g., Dart) and soft systems methodologies (e.g., Williams). It also shares much in common with the positive deviance approach that has been applied to health interventions in many developing countries. The success case method identifies individual cases that have been particularly successful (and unsuccessful) and uses case study analytical methods to develop credible arguments about the contribution of the intervention to these.

Most significant change

The most significant change technique (Davies and Dart, 2005) is a form of participatory monitoring and evaluation. It is participatory because many intervention stakeholders are involved both in deciding the types of change to be recorded, and in analyzing the data. It is a form of monitoring because it occurs throughout the intervention cycle and provides information to help people manage the intervention. It contributes to impact evaluation in part because it provides data on impact and outcomes that can be used to help assess the performance of the intervention as a whole—but largely through providing a tool for identifying and rating the impacts that are valued by different stakeholders.

MAPP

The Method for Impact Assessment of Projects and Programs (Späth, 2004) is a methodological framework for combining a qualitative approach with participatory assessment instruments, including a quantification step. It orients itself toward principles and procedures of Participatory Rural Appraisal methodology, including triangulation, "optimal ignorance," and communal learning. A major element of this methodology is conducting workshops with representatives of relevant stakeholders. Perceived key processes are jointly reflected in structured group discussions in which at least six interlinked and logically connected steps are accomplished: (i) lifeline; (ii) trend analysis; (iii) activity list; (iv) influence matrix; (v) transect—or data cross checking; and (vi) development and impact profile.

Books on social mechanisms

Authors like Elster (1989; 2007), Farnsworth (2007), Hedström and Swedberg (1998), Swedberg (2005), Bunge (2004), and Mayntz (2004) have summarized and synthesized the research literature on different (types of) social mechanisms. Elster's explanation of social behavior (2007) summarizes insights from neurosciences to economics and political science and discusses 20-plus mechanisms. They range from motivations, emotions, and self-interest to rational choice, games and behavior and collective decision making.

Farnsworth (2007) takes legal arrangements like laws and contracts as a starting point and dissects which (types of) mechanisms play a role when one wants to understand why laws sometimes do or do not work. He combines insights from psychology, economics, and sociology and discusses mechanisms such as the "slippery slope," the endowment effect, framing effects, and public goods production.

Review journals

Since the 1970s review journals have been developed to address important developments within a discipline. An example is *Annual Reviews,* which publishes analytic reviews in 37 disciplines within the biomedical, life, physical, and social sciences.

Knowledge repositories

Hansen and Rieper (2009) have inventoried a number of second-order evidence-producing organizations within the social (and behavioral) sciences. In recent years the production of systematic reviews has been institutionalized in these institutions. There are two main international organizations: the Cochrane Society, working within the health field; and the Campbell Collaboration, working within the fields of social welfare, education, and criminology. Both organizations subscribe to the idea of producing globally valid knowledge about the effects of interventions, if possible through synthesizing the results of primary studies designed as RCTs and using meta-analysis as the form of syntheses. In many (Western) countries second-order knowledge-producing organizations have been established at the national level that not all are based on findings from RCTs. Hansen and Rieper (2009) present information about some 15 of them, including web addresses.

Knowledge repositories and development intervention impact

The Coalition for Evidence-Based Policy offers "Social Programs That Work," a Web site providing policy makers and practitioners with clear, actionable information on what works in social policy, as demonstrated in scientifically valid studies (www.evidencebasedprograms.org/).

The International Organization for Cooperation in Evaluation, a loose alliance of regional and national evaluation organizations from around the world, builds evaluation leadership and capacity in developing countries, fosters the cross-fertilization of evaluation theory and practice around the world, addresses international challenges in evaluation, and assists evaluation professionals to take a more global approach to identifying and solving problems. It offers links to other evaluation organizations; forums that network evaluators internationally; news of events and important initiatives; and opportunities to exchange ideas, practices, and

insights with evaluation associations, societies, and networks (http://ioce.net).

The Abdul Latif Jameel Poverty Action Lab (J-PAL) fights poverty by ensuring that policy decisions are based on scientific evidence. Located in the Economics Department at the Massachusetts Institute of Technology, J-PAL brings together a network of researchers at several universities who work on randomized evaluations. It works with governments, aid agencies, bilateral donors, and nongovernmental organizations to evaluate the effectiveness of antipoverty programs using randomized evaluations, disseminate findings and policy implications, and promote the use of randomized evaluations, including by training practitioners to carry them out (www.povertyactionlab.com/).

The Development Impact Evaluation Initiative (DIME) is a World Bank-led effort involving thematic networks and regional units under the guidance of the Bank's Chief Economist. Its objectives are—

- To increase the number of Bank projects with impact evaluation components
- To increase staff capacity to design and carry out such evaluations
- To build a process of systematic learning based on effective development interventions with lessons learned from completed evaluations.

Case 1: Combining qualitative and quantitative descriptive methods—Ex post impact study of the Noakhali Rural Development Project in Bangladesh[1]

1. Summary

The evaluation examined the intended and unintended socio-economic impacts of the project, with particular attention to the impact on women and to the sustainability and sustainment of these impacts. The evaluation drew on a wide range of existing evidence and also used mixed methods to generate additional evidence; because the evaluation was conducted nine years after the project had ended, it was possible to directly investigate the extent to which impacts had been sustained. Careful attention was paid to differential impacts in different contexts to interpret the significance of before/after and with/without comparisons; the intervention was only successful in contexts that provided the other necessary ingredients for success. The evaluation had significant resources and was preceded by considerable planning and review of existing evidence.

2. Summary and main characteristics

The Noakhali Rural Development Project (NRDP) was an integrated rural development project (IRDP) in Bangladesh, funded for DKK 389 million by Danida. It was implemented in two phases over a period of 14 years, 1978–92, in the greater Noakhali district, one of the poorest regions of Bangladesh, which had a population of approximately 4 million. More than 60 long-term expatriate advisers—most of them Danish—worked 2–3 years each on the project together with a Bangladeshi staff of up to 1,000 (at the peak).

During NRDP-I the project comprised activities in 14 different areas grouped under four headings:

- Infrastructure (roads, canals, market places, public facilities)
- Agriculture (credit, cooperatives, irrigation, extension, marketing)
- Other productive activities (livestock, fish ponds, cottage industries)
- Social sector (health & family planning, education).

The overarching objective of NRDP-I was to promote economic growth and social progress, in particular aiming at the poorer sections of the population. The poorer sections were to be reached through the creation of temporary employment in construction activities (infrastructure) and engaging them in income-generating activities (other productive activities). There was also an aim to create more employment in agriculture for landless laborers through intensification. Almost all the major activities started under NRDP-I continued under NRDP-II, albeit with some modifications and additions. The overarching objective was kept, with one notable addition: to promote economic growth and social progress in particular, aiming at the poorer segments of the population, including women. A special focus on women was thus included, based on the experience that most of the benefits of the project had accrued to men.

3. Purpose, intended use, and key evaluation questions

This ex post impact study was carried out nine years after the project was terminated. At the time of implementation NRDP was one of the largest projects funded by Danida, and it was

considered an excellent example of integrated rural development, which was a common type of support during the 1970s and '80s. In spite of the potential lessons to be learned from the project, it was not evaluated upon completion in 1992. This fact and an interest in the sustainability factor in Danish development assistance led to the commission of the study. What type of impact could still be traced in Noakhali nine years after Danida terminated its support to the project?

Although the study dealt with aspects of the project implementation, its main focus was on the project's socioeconomic impact in the Noakhali region. The study aimed to identify the intended as well as unintended impact of the project, in particular whether it had stimulated economic growth and social development and improved the livelihoods of the poor, including women, which the project had set out to do.

The evaluation focused on the following questions:

- What has been the short- and long-term—intended as well as unintended—impact of the project?
- Has the project stimulated economic growth and social development in the area?
- Has the project contributed to improving the livelihoods of the poorest section of the population, including women?
- Have the institutional and capacity-building activities engendered or reinforced by the project produced sustainable results?

4. Concise description of the evaluation

Identifying impacts of interest

This study focuses on the impact of NRDP, in particular the long-term impact (i.e., nine years after). But impact cannot be understood in isolation from implementation so the study analyzes various elements and problems in the way the project was designed and executed. Impact can also not be understood isolated from the context, both the natural/physical and in particular the societal—social, cultural, economic, political—context. In comparison with ordinary evaluations, this study puts a lot more emphasis on understanding the national and in particular the local context.

Gathering evidence of impacts

One of the distinguishing features of this impact study, compared to normal evaluations, is the order and kind of fieldwork. The fieldwork lasted four months and involved a team of eight researchers (three European and five Bangladeshi) and 15 assistants. The researchers spent 1.5–3.5 months in the field, the assistants 2–4 months.

The following is a list of the methods used:

- Documentary study (project documents, research reports, etc.)
- Archival work (in the Danish embassy, Dhaka)
- Questionnaire with former advisers and Danida staff members
- Stakeholder interviews (Danida staff, former advisers, Bangladeshi staff, etc.)
- Quantitative analysis of project monitoring data
- Key informant interviews
- Compilation and analysis of material about context (statistics, articles, reports, etc.)
- Institutional mapping (particularly NGOs in the area)
- Representative surveys of project components
- Assessment of buildings, roads and irrigation canals (function, maintenance, etc.)
- Questionnaire-based interviews with beneficiaries and non-beneficiaries
- Extensive and intensive village studies (surveys, interviews, etc.)
- Observation
- Focus group interviews
- In-depth interviews (issue-based and life stories).

In the history of Danish development cooperation no other project has been subject to so many studies and reports, not to speak of the vast number of newspaper articles. Most important for the impact study have been the appraisal reports and the evaluations plus the final project

completion report. But in addition to this, there exists an enormous number of reports on all aspects of the project. A catalogue from 1993 lists more than 1,500 reports produced by and for the NRDP. Both the project and the local context were, moreover, intensively studied in a research project carried out in cooperation between the Centre for Development Research and Bangladesh Institute of Development Studies.

A special effort was made to solicit the views of a number of key actors (or stakeholders) in the project and other key informants. This included numerous former NRDP and BRDB officers, expatriate former advisers as well as former key Danida staff, both based in the Danish Embassy in Dhaka and in the Ministry of Foreign Affairs in Copenhagen. They were asked about their views on strengths and weaknesses of the project and the components they know best, about their own involvement and about their judgment regarding likely impact. A questionnaire survey was carried out among the around 60 former expatriate long-term advisers and 25 former key staff members in the Danish embassy, Danida, and other key informants. In both cases about half returned the filled-in questionnaires. This was followed up by a number of individual interviews.

The main method in four of the five component studies was surveys with interviews, based on standardized questionnaires, with a random—or at least reasonably representative—sample of beneficiaries (of course combined with documentary evidence, key informant interviews, etc.). A great deal of effort was taken in ensuring that the survey samples were reasonably representative.

The infrastructure component was studied by partly different methods, because in this case the beneficiaries were less well defined. It was decided to make a survey of all the buildings that were constructed during the first phase of the project to assess their current use, maintenance standard, and benefits. In this phase the emphasis was on construction; in the second phase it shifted to maintenance. Moreover, a number of roads were selected for study, both of their current maintenance standard, their use, etc., but also the employment the road construction and maintenance generated, particularly for groups of destitute women. The study also attempted to assess the socio-economic impact of the roads on different groups (poor/better-off, men/women, etc.).

Assessing causal contribution

The impact of a development intervention is a result of the interplay of the intervention and the context. It is the matching of what the project has to offer and people's needs and capabilities that produces the outcome and impact. Moreover, the development processes engendered unfold in a setting that is often characterized by inequalities, structural constraints, and power relations. This certainly has been the case in Noakhali. As a consequence there will be differential impacts, varying between individuals and according to gender, socio-economic group and political leverage.

In addition to the documentary studies, interviews, and questionnaire survey, the actual fieldwork has employed a range of both quantitative and qualitative methods. The approach can be characterized as a contextualized, tailor-made ex post impact study. There is considerable emphasis on uncovering elements of the societal context in which the project was implemented. This covers both the national context and the local context. The approach is tailor-made in the sense that it will be made to fit the study design outlined above and apply an appropriate mix of methods.

An element in the method is the incorporation in the study of both before/after and with/without perspectives. These, however, are not seen as the ultimate test of impact (success or failure), but interpreted cautiously, bearing in mind that the area's development has also been influenced by a range of other factors (market forces, changing government policies, other development interventions, etc.), both during the 14 years the project was implemented and during the 9 years after its termination.

Considerable weight was accorded to studying what has happened in the villages that have previously been studied and for which some comparable data exist. Four villages were studied intensively in 1979 and briefly restudied in 1988 and 1994. These studies—together with a thorough restudy in the year 2001—provide a unique opportunity to compare the situation before, during, and after the project. Moreover, 10 villages were monitored under the project's village-wise impact monitoring system in the years 1988–90, some of these being with (+NRDP) and some (largely) without (–NRDP) the project. Analysis of the monitoring data combined with a restudy of a sample of these villages illuminates the impact of the project in relation to other factors. It was decided to study a total of 15 villages, 3 intensively (all +NRDP, about 3 weeks each) and 12 extensively (9 +NRDP, 3 – NRDP, 3–5 days each). As a matter of principle, this part of the study looks at impact in terms of the project as a whole. It brings in focus the project benefits as perceived by different groups and individuals and tries to study how the project has impinged on economic and social processes of development and change. At the same time it provides a picture of the considerable variety found in the local context.

In the evaluation of the mass education program, the problem of attribution was dealt with as carefully as possible. First, a parallel comparison has been made between the beneficiaries on the one hand and non-beneficiaries on the other, to identify (if any) the changes directly or indirectly related to the program. Such comparison was vital due to the absence of any reliable and comparable baseline data. Second, specific queries were made in relation to the impact of the program as perceived by the beneficiaries and other stakeholders of the program, assuming that they would be able to perceive the impact of the intervention on their own lives in a way that would not be possible for others. And finally, views of non-beneficiaries and non-stakeholders were sought to have opinions from people who do not have any valid reason for either understating or overstating the impact of the program. It was through such a cautious approach that the question of attribution was addressed. Arguably, elements of subjectivity may still have remained in the conclusions and assumptions, but that is unavoidable in a study that seeks to uncover the impact of an education project.

Managing the impact evaluation

The impact study was commissioned by Danida and carried out by Centre for Development Research, who also co-funded the study as a component of its Aid Impact Research Program. The research team comprised independent researchers from Bangladesh, Denmark, and the UK. A reference group of nine persons (former advisers, Danida officers, and researchers) followed the study from the beginning to the end. It discussed the approach paper in an initial meeting and the draft reports in a final meeting. In between it received three progress reports from the team leader and took up discussions by e-mail correspondence. The study was prepared during the year 2000 and fieldwork carried out in the period January–May 2001. The study consists of a main report and seven topical reports.

The first step in establishing a study design was the elaboration of an approach paper (study outline) by the team leader. This was followed by a two-week visit to Dhaka and the greater Noakhali area. During this visit, Bangladeshi researchers and assistants were recruited to the team, and more detailed plans for the subsequent fieldwork were drafted. Moreover, a background paper by Hasnat Abdul Hye, former Director General of BRDB and Secretary, Ministry of Local Government, was commissioned.

The fieldwork was preceded by a two-day methodology-cum-planning workshop in Dhaka. The actual fieldwork lasted four months—from mid-January to mid-May 2001. The study team comprised 23 people: 5 Bangladeshi researchers, 3 European researchers, 6 research assistants, and 9 field assistants (all from Bangladesh). The researchers spent 1.5–3.5 months in the field, the assistants 2–4 months. Most of the time the team worked 60–70 hours a week. So it takes a good

deal of resources to accomplish such a big and complex impact study.

Case 2: Combining qualitative and quantitative descriptive methods—Mixed-method impact evaluation of IFAD projects in Gambia, Ghana, and Morocco[2]

1. Summary

The evaluation included intended and unintended impacts and examined the magnitude, coverage, and targeting of changes. It used mixed methods to gather evidence of impacts and the quality of processes with cross-checking among sources. With regard to *assessing causal contribution*, it must be noted that no baseline data were available. Instead a comparison group was constructed, and analysis of other contributing factors was made to ensure appropriate comparisons. The evaluation was undertaken within significant resource constraints and was carried out by an interdisciplinary team.

2. Introduction and background

Evaluations of rural development projects and country programs are routinely conducted by the Office of Evaluation of IFAD. The ultimate objectives of these evaluations is to set a basis for accountability by assessing the development results and contribute to learning and improvement of design and implementation by providing lessons learned and practical recommendations. These evaluations follow a standardized methodology and a set of evaluation questions including the following: (i) project performance (relevance, effectiveness, and efficiency), (ii) project impact, (iii) overarching factors (sustainability, innovation, and replication) and (iv) the performance of the partners. As can be seen, impact is but one of the key evaluation questions and the resources allocated to the evaluation (budget, specialists, and time) that have to be shared for the entirety of the evaluation.

Thus, these evaluations are to be conducted under resource constraints. In addition, very limited data are available on socio-economic changes taking place in the project area that can be ascribed to an impact definition. IFAD adopts an impact definition which is similar to the DAC definition. The key feature of IFAD evaluations is that they are conducted just before or immediately after project conclusion: the effects can be observed after 4–7 years of operations and the future evolution can be estimated through an educated guess on sustainability perspectives. Several impact domains are considered, including household income and assets, human capital, social capital, food security, environment, and institutions.

3. Sequencing of the process and choice of methods

This short case study is based on evaluations conducted in Gambia, Ghana, and Morocco between 2004 and 2006. As explained above, evaluations had multiple questions to answer and impact assessment was but one of them. Moreover, impact domains were quite diverse. This meant that some questions and domains required quantitative evidence (e.g., in the case of household income and assets), whereas a more qualitative assessment would be in order for other domains (e.g., social capital). In many instances, however, more than one method would have to be used to answer the same questions to cross-check the validity of findings, identify discrepancies, and formulate hypotheses on the explanation of apparent inconsistencies.

As the final objective of the evaluation was not only to assess results but also to provide future intervention designers with adequate knowledge and insights, the evaluation design could not be confined to addressing a dichotomy between "significant impact has been observed" and "no significant impact has been observed." Findings would need to be rich enough and grounded in field experience to provide a plausible explanation that would lead, when suitable, to a solution to identified problems and to recommendations to improve the design and the execution of the operations.

Countries and projects considered in this case study were diverse. In all cases, however, the first step in the evaluation consisted of a desk review of the project documentation. This allowed the evaluation team to understand or reconstruct

the intervention theory (often implicit) and the logical framework. In turn, this would help to identify a set of hypotheses on changes that may be observed in the field as well as on intermediary steps that would lead to those changes.

In particular, the preliminary desk analysis highlighted that the results assessment would have to be supplemented with some analysis of implementation performance. The latter would include some insight into the business processes (e.g., the management and resource allocation made by the project implementation unit) and the quality of service rendered (e.g., the topics and the communication quality of an extension service or the construction quality of a feeder road or of a drinking water scheme).

The second step was to conduct a preparatory mission. This mission was instrumental in fine-tuning our hypotheses on project results and designing the methods and instruments. Given the special emphasis of the IFAD interventions on the rural poor, impact evaluation would need to shed light, to the extent possible, on the following dimensions of impact: (i) magnitude of changes, (ii) coverage (i.e., the number of persons or households served by the projects), and (iii) targeting (i.e., gauging the distribution of project benefits according to social, ethnic, or gender grouping).

As pointed out before, a major concern was the absence of a baseline survey which could be used as a reference for impact assessment. This required reconstructing the "before project" situation. By the same token, it was clear that the observed results could not simply be attributed to the evaluated interventions. In addition to exogenous factors such as weather changes, other important factors were at play, for example, changes in government strategies and policies (such as the increased support to grassroots associations by Moroccan public agencies) or operations supported by other development organizations in the same or in adjacent zones. This meant that the evaluated interventions would interplay with existing dynamics and interact with other interventions. Understanding synergies or conflicts between parallel dynamics could not be done simply through inferential statistical instruments but required interaction with a wider range of stakeholders.

The third step in the process was the fielding of a data collection survey (after pre-testing the instruments) that would help the evaluation cope with the dearth of impact data. The selected techniques for data collection included a quantitative survey with a range of 200–300 households (including both project and control groups) and a more reduced set of focus group discussion with groups of project users and "control groups" stratified based on the economic activities in which they had engaged and the area they were leaving.

In the quantitative survey standardized questionnaires were administered to final project users (mostly farmers or herders) as well as to non-project groups (control observations) on the situation before (recall methods) and after the project. Recall methods were adopted to make up for the absence of a baseline.

In the course of focus group interviews, open-ended discussion guidelines were adopted; results were mostly of a qualitative nature. Some of the focus group facilitators had also been involved in the quantitative survey and could refer the discussion to observations previously made. After the completion of data collection and analysis, a first cross-checking of results could be made between the results of quantitative and qualitative analysis.

As a fourth step, an interdisciplinary evaluation team would be fielded. Results from the preliminary data collection exercise were made available to the evaluation team. The data collection coordinator was a member of the evaluation team or in a position to advise its members. The evaluation would conduct field visits and conduct a further validation survey and collect focus group data through participant observations and interviews with key informants (and further focus group discussions if necessary). The team would also spend adequate time with project management units to gather a better insight of implementation and business processes.

The final impact assessment would be made by means of triangulation of evidence captured from the (scarce) existing documentation, the preliminary data collection exercise, and the main interdisciplinary mission (figure A11.1).

4. Constraints in data gathering and analysis

Threats to the validity of recall methods. According to the available literature sources[3] and our own experience, the reliability of recall methods may be questionable for monetary indicators (e.g., income) but higher for easier-to-remember facts (e.g., household appliances, approximate herd size). Focus group discussions helped identify possible sources of bias in the quantitative survey and ways to address them.

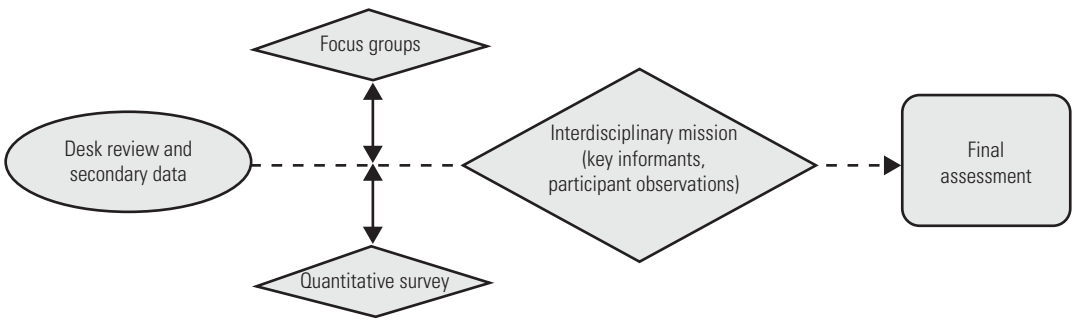
Finding "equivalent" samples for with and without-project observations. One of the challenges was to extract a control sample that would be "similar" in the salient characteristics to the project sample. In other words, problems of sampling bias and endogeneity should have been controlled for (e.g., more entrepreneurial people are more likely to participate in a rural finance intervention). In sampling control observations, serious attempts were made to match project and non-project households based on similarity of main economic activities, agro-ecological environment, household size, and resource endowment. In some instances, household that had just started to be served by the projects ("new entries") were considered control groups, on the grounds that they would broadly satisfy the same eligibility criteria at entry as "older" project clients. However, no statistical technique (e.g., instrumental variables, Heckman's procedure or propensity score) was adopted to test for sampling bias, due to limited time and resources.

Coping with linguistic gaps. Given the broad scope of the evaluations, a team of international sector specialists was required. However, international experts were not necessarily the best suited for data collection analysis, which calls for fluency in the local vernacular, knowledge of local practices, and skills to obtain the most possible information within a limited time frame. Staggering the process in several phases was a viable solution. The preliminary data collection exercise was conducted by a team of local specialists, with university students or local teachers or literate nurses serving as enumerators.

5. Main value added of mixed methods and opportunities for improvement

The choice of methods was made taking into account the objectives of the evaluations and the resource constraints (time, budget, and expertise) in conducting the exercise. The combination of multiple methods allowed us to cross-check the evidence and understand, for example, when survey questions were likely to be misinterpreted or generate over- or under-reporting. In contrast, quantitative evidence

Figure A11.1: Final impact assessment triangulation



allowed us to shed light on the prevalence of certain phenomena highlighted during the focus group discussion. Finally, the interactions with key informants and project managers and staff helped us better understand the reasons for under- or over-achievements and come up with more practical recommendations.

The findings, together with the main conclusions and recommendations in the report, were adopted to design new projects or a new country strategy. There was also interest from the concerned project implementation agencies in adopting the format of the survey to conduct future impact assessments on their own. Due to time constraints, only inferential analysis was conducted on the quantitative survey data. A full-fledged econometric analysis would have been desirable. By the same token, further analysis of focus group discussion outcomes would be desirable in principle.

6. A few highlights on the management

The overall process design, as well as the choice of methods and the design of the data collection instruments, was made by the lead evaluator in the Office of Evaluation of IFAD, in consultation with international sectoral specialists and the local survey coordinator. The pre-mission data collection exercise was coordinated by a local rural sociologist, with the help of a statistician for the design of the sampling framework and data analysis.

The time required for conducting the survey and focus groups was as follows:

- Develop draft questionnaire and sampling frame, identify enumerators: 3 weeks.
- Conduct a quick trip on the ground, contact project authorities and pre-test questionnaires: 3 days.
- Train enumerators' and coders' team: 3 days.
- Survey administering: depending on the length of the questionnaire, on average an enumerator will be able to fill no more than three to five questionnaires per day. In addition, time needs to be allowed for travel, rest. With a team of 6 enumerators, in 9–10 working days up to 200 questionnaires can be filled in, in the absence of major transportation problems.
- Data coding: it may vary depending on the length and complexity of the questionnaire. It is safe to assume 5–7 days.
- Time for conducting focus groups discussions: 7 days based on the hypothesis that around 10 FGD would be conducted by 2 teams.
- Data analysis. Depending on the analysis requirement, it will require one to two weeks only to generate the tables and summary of focus group discussions.
- Drafting survey report: 2 weeks.

Note: As some of the above tasks can be conducted simultaneously, the total time for conducting a preliminary data collection exercise may be lower than the sum of its parts.

Case 3: Combining qualitative and quantitative descriptive methods—Impact evaluation: Agricultural development projects in Guinea[4]

1. Summary

The evaluation focused on impact in terms of poverty alleviation; the distribution of benefits was of particular interest, not just the mean effect. All data gathering was conducted after the intervention had been completed; mixed methods were used, including attention to describing the different implementation contexts. Assessing causal contribution is the major focus of the case study. A counterfactual was created by creating a comparison group, taking into account the endogenous and exogenous factors affecting impacts. Modeling was used to develop an estimate of the impact. With regard to the management of the impact evaluation, it should be noted that the study was undertaken as part of doctoral dissertation work; the stakeholder engagement and subsequent use of it was limited.

This impact evaluation concerned two types of agricultural projects based in the Kpèlè region, in Guinea. The first one[5] was the Guinean Oil Palms and Rubber Company (SOGUIPAH). It was founded in 1987 by the Guinean govern-

ment to take charge of developing palm oil and rubber production at the national level. With the support of several donors, SOGUIPAH quickly set up a program of industrial plantations[6] by negotiating the ownership of 22,830 ha with villagers. In addition, several successive programs were implemented between 1989 and 1998 with SOGUIPAH to establish contractual plantations[7] on farmers' own land and at the request of the farmers (1,552 ha of palm trees and 1,396 ha of rubber trees) and to improve 1,093 ha of lowland areas for irrigated rice production.

The impact evaluation took place in a context of policy debates among different rural stakeholders at a regional level: two seminars had been held in 2002 and 2003 between the farmers' syndicates, the state administration, the private sector, and development partners (donors, NGOs) to discuss a regional strategy for agricultural development. These two seminars revealed that there was little evidence of what should be done to alleviate rural poverty, despite a long history of development projects. The impact of these projects on farmers' income seemed to be particularly relevant to assess, notably to compare the projects' efficiency.

This question was investigated through doctoral thesis work that was entirely managed by the AGROPARISTECH.[8] It was financed by AFD, one of the main donors in the rural sector in Guinea. This thesis proposed a new method, the systemic impact evaluation, aiming at quantifying impact using a qualitative approach. It enabled the understanding of the process through which impact materializes and rigorous quantification of the impact of agricultural development projects on the farmers' income, using a counterfactual. The analysis is notably based on the comprehension of the agrarian dynamics and the farmers' strategies, and permits the quantification of ex post impact but also to devise a model of ex ante evolution for the following years.

2. Gathering evidence of impact

The data collection was carried out entirely ex post. Several types of surveys and interviews were used to collect evidence of impact.

First, a contextual analysis realized all along the research work with key informants was necessary to describe the project implementation scheme, the contemporaneous events, and the existing agrarian dynamics. It was also used to assess qualitatively whether those dynamics were attributable to the project. A series of surveys and historical interviews (focused on the pre-project situation) were conducted to establish the most reliable baseline possible. An area considered "witness" to the agrarian dynamic that would have existed in the project's absence was identified.

Second, a preliminary structured survey (of about 240 households) was implemented, using recall to collect data on the farmers' situation in the pre-intervention period and during the project. It was the basis of a judgment sample to realize in-depth interviews (see below), which aimed at describing the farming systems and rigorously quantifying the farmers' income.

3. Assessing causal attribution

By conducting an early contextual analysis, the evaluator was able to identify a typology of farming systems that existed before the project. To set up a sound counterfactual, a judgment sample was realized among the 240 households surveyed, by choosing 100 production units that had belonged to the same initial types of farming system and that had evolved with (in the project area) or without the project (in the witness area).

In-depth understanding of the endogenous and exogenous factors influencing the evolution and possible trajectories of farming systems enabled the evaluator to rigorously identify the individuals whose evolution *with or without* the project was comparable. This phase of judgment sample was followed by in-depth interviews with the hundred farmers. The evaluator's direct involvement in data collection was then essential, hence the importance of a small sample. It would not have been possible to gather reliable data on yields, modifications to production structures over time, and producers' strategies from a large survey sample in a rural context.

Then, based on the understanding of the way the project proceeded and of the trajectories of these farmers, with or without the project, it was possible to build a quantitative model, based on Gittinger's method of economic analysis of development projects (Gittinger, 1982). As the initial diversity of production units was well identified before sampling, this model was constructed for each type of farming system existing before the project. Understanding the possible evolution of each farming system with and without the project allowed for the estimation of the differential created by the project on farmers' income, i.e., its impact.

4. Ensuring rigor and quality

Although the objective differences between each production unit studied appear to leave room for the researcher's subjectivity when constructing the typology and sample, the rationale behind the farming system concept made it possible to transcend this possible arbitrariness. What underlies this methodological jump from a small number of interviews to a model is the demonstration that a finite number of types of farming systems exists in reality.

Moreover, the use of a comparison group, the triangulation of most data collected by in-depth interviews through direct observation and contextual analysis, and the constant implication of the principal researcher were key factors to ensure rigor and quality.

5. Key findings

The large survey of 240 households identified 11 trajectories related to the implementation of the project. Once each trajectory and impact was characterized and quantified through in-depth interviews and modeling, this survey permitted as well quantifying a mean impact of the project, on the basis of the weight of each type in the population. The mean impact was only 24 €/year/household in one village poorly served by the project, due to its enclosed situation, whereas it was 200 €/year/household in a central village.

Despite a positive mean impact, highly differentiated impacts also existed, depending on the original farming system and the various trajectories with and without the project, which could not be ignored. Whereas former civil servants or traditional landlords beneficiated large contractual plantations, other villagers were deprived of their land for the needs of the project or received surfaces of plantations too limited to improve their economic situation.

Therefore, it seems important that the impact evaluation of a complex development project include an analysis of the diversity of cases created by the intervention, directly or indirectly.

The primary interest of this new method was to give the opportunity to build a credible impact assessment entirely ex post. Second, it gave an estimate of the impact on different types of farming systems, making explicit the existing inequalities in the distribution of the projects' benefits. Third, it permitted a subtle understanding of the reasons why the desired impacts materialized or not.

6. Influence

The results from this impact assessment were available after four years of field work and data treatment. They were presented to the Guinean authorities and to the local representatives of the main donors in the rural sector. In the field, the results were delivered to the local communities interviewed and to the farmers' syndicates. The Minister of Agriculture declared that he would try to foster more impact evaluations on agricultural development projects. Unfortunately, there is little hope that the conclusions of this research will change the national policy about these types of projects, in the absence of an institutionalized forum for discussing it among the different stakeholders.

Case 4: A theory-based approach with qualitative methods—Global Environment Facility impact evaluation 2007[9, 10]

Evaluation of three GEF-protected area projects in East Africa

1. Description of evaluation

The objectives of this evaluation included—

- To *test evaluation methodologies* that can assess the impact of GEF interventions. The key activity of the GEF is "providing new and additional grant and concessional funding to meet the agreed incremental costs of measures to achieve agreed global environmental benefits."[11] The emphasis of this evaluation was therefore on verifying the achievement of agreed global environmental benefits.
- Specifically, to test a *theory of change approach* to evaluation in GEF's biodiversity focal area, and assess its potential for broader application within GEF evaluations.
- To assess the *sustainability and replication of the benefits of GEF support* and extract lessons. It evaluated whether and how project benefits have continued, and will continue, after project closure.

Primary users

The primary users of the evaluation are GEF entities. They include the GEF Council, which requested the evaluation; the GEF Secretariat, which will approve future protected area projects; implementing agencies (such as the World Bank, UN agencies and regional development banks); and national stakeholders who will implement future protected area projects.

2. Evaluation design

Factors driving selection of evaluation design

The Approach Paper to the impact evaluation[12] considered the overall GEF portfolio to develop an entry-point which could provide a good opportunity to develop and refine effective and implementable impact evaluation methodologies. Themes and projects that are relatively straightforward to evaluate were emphasized. The Evaluation Office adopted the DAC definition of impact, which determined that closed projects would be evaluated to assess the sustainability of GEF interventions.

Biodiversity and protected areas

The biodiversity focal area has the largest number of projects within the GEF portfolio of currently active and completed projects. In addition, biodiversity has developed more environmental indicators and global data sets than other focal areas, both within the GEF and in the broader international arena. The Evaluation Office chose protected areas as the central theme for this phase of the Impact Evaluation because protected areas are one of the primary approaches supported by the GEF biodiversity focal area and its implementing agencies, and the GEF is the largest supporter of protected areas globally; previous evaluations have noted that an evaluation of the GEF support for protected areas has not been carried out and recommended that such a study be undertaken; protected areas are based on a set of explicit change theories, not just in the GEF, but in the broader conservation community; in many protected area projects, substantial field research has been undertaken, and some have usable baseline data on key factors to be changed by the intervention; a protected areas strategy can be addressed at both a thematic and regional cluster level (as in East Africa, the region chosen for the study); and the biodiversity focal area team has made considerable progress in identifying appropriate indicators for protected areas through its "managing for results" system.

The choice of projects

Lessons from a set of related interventions (or projects) are more compelling than those from an isolated study of an individual project. To test the potential for aggregation of project results, enable comparisons across projects and ease logistics, it was decided to adopt a sub-regional focus and select a set of projects that are geographically close to each other. East Africa is the sub-region with the largest number of complete and active projects in the GEF portfolio with a protected area component, utilizing large GEF and cofinancing expenditure.

The following three projects were selected for evaluation:

- Bwindi Impenetrable National Park and Mgahinga Gorilla National Park Conservation Project, Uganda (World Bank)
- Lewa Wildlife Conservancy, Kenya (World Bank)

- Reducing Biodiversity Loss at Cross-Border Sites in East Africa, Regional: Kenya, Tanzania, Uganda (UNDP).

These projects were implemented on behalf of the GEF by the World Bank and UNDP. They have a variety of biodiversity targets, some of which are relatively easy to monitor (gorillas, zebras, rhinos). Also, these projects were evaluated positively by terminal and other evaluations and the continuance of long-term results was predicted. The *Bwindi Impenetrable National Park and Mgahinga Gorilla National Park Conservation Project* is a $6.7 million full-size project and the first GEF-sponsored trust fund in Africa. The *Lewa Wildlife Conservancy* is a medium-sized project, within a private wildlife conservation company. The *Reducing Biodiversity Loss at Cross-Border Sites in East Africa* Cross project is a $12 million project, implemented at field level by government agencies, that aims to foster an enabling environment for the sustainable use of biodiversity.

The advantages of a theory of change approach

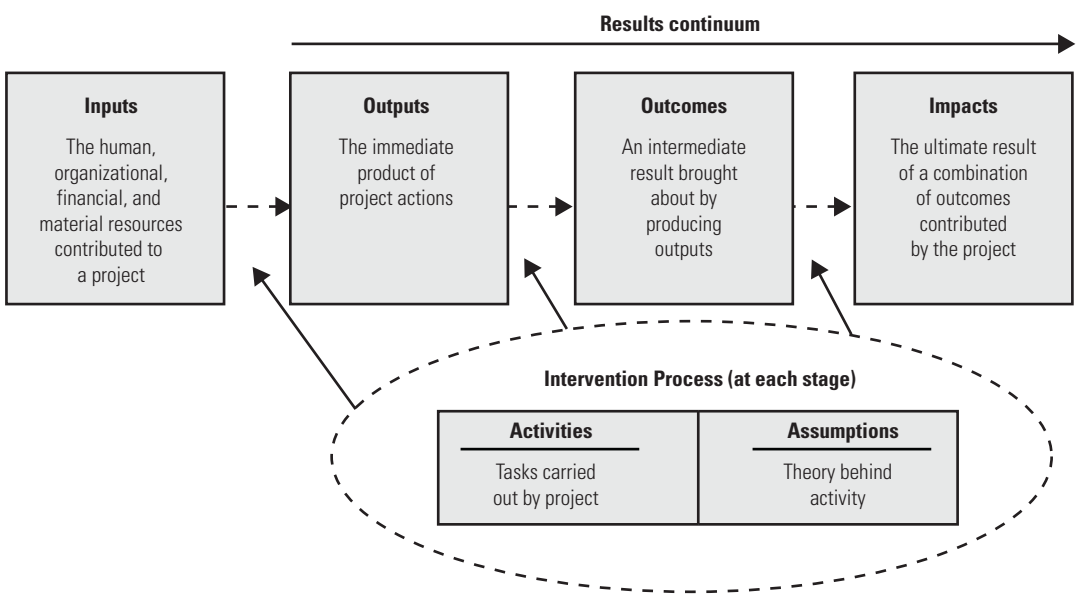
An intervention generally consists of several complementary activities that together produce intermediate outcomes, which are then expected to lead to impact (see figure A11.2). The process of these interventions, in a given context, is determined by the contribution of a variety of actions at multiple *levels*, some of which are *outside* the purview of the intervention (e.g., actions of exterior actors at the local, national, or global levels or change in political situations, regional conflicts, and natural disasters). Subsequently, an intervention may have different levels of achievement in its component parts, giving mixed results towards its objectives.

The use of a hybrid evaluation model

During field testing it was decided that, given the intensive data requirements of a theory of change approach and the intention to examine project impacts, *the evaluation would mainly focus on the later elements of each project's theory of change, when outcomes are expected to lead to impact.* Based on this approach, the evaluation developed a methodology composed of three components (see figure A11.3):

- *Assessing implementation success and failure:* To understand the contributions of the

Figure A11.2: Generic representation of a project's theory of change



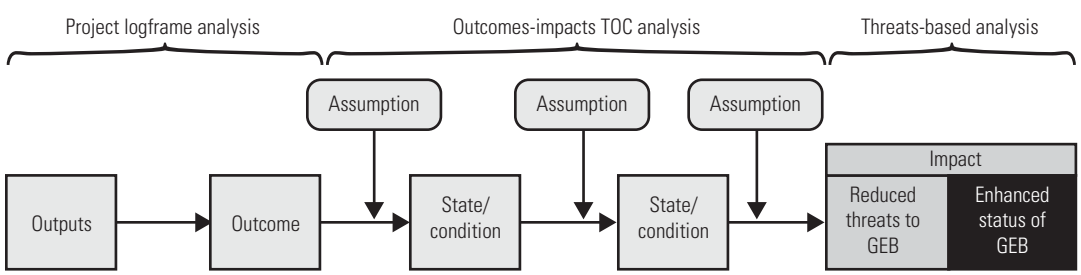
project at earlier stages of the results continuum, leading to project outputs and outcomes, a *logframe analysis* is used. Though the normally complex and iterative process of project implementation is not captured by this method, the logframe provides a means of tracing the realization of declared objectives. GEF interventions aim to "assist in the protection of the global environment and promote thereby environmentally sound and sustainable economic development."[13]

- *Assessing the level of contribution (i.e., impact):* To provide a direct measure of project impacts, a *targets-threats analysis (threats-based analysis)* is used to determine whether global environmental benefits have actually been produced and safeguarded.[14] The robustness of global environment benefits identified for each project (targets) is evaluated by collecting information on attributes relating to the targets' biological composition, environmental requirements, and ecological interactions. This analysis of targets is complemented by an assessment of the level of "threat" (e.g., predation, stakeholder attitude, and behavior) faced by the global environment benefits. For targets and significant threats, trends over time (at project start, at project close, and currently), and across project and non-project areas are sought, so that a comparison is available to assess levels of change.
- *Explanations for observed impact:* To unpack the processes by which the project addresses and contributes to impact, an *outcomes-impacts theory of change analysis* is used. This theory of change approach constructs and validates the project logic connecting outcomes and ultimate project impact. It involves a comprehensive assessment of the activities undertaken after project closure, along with their explicit and implicit assumptions. This component enables an assessment of the sustainability and/or catalytic nature of project interventions and provides a composite qualitative ranking for the achievements of the projects. Elements of the varied aspects of sustainability include behavior change and the effectiveness of capacity-building activities, financial mechanisms, legislative change, and institutional development.

The model incorporates three different elements that may be involved in the transformation of project outcomes into impacts. These are as follows, and were each scored for the level of achievement of the project in converting outcomes into impacts:

- *Intermediary states.* These are conditions that are expected to be produced on the way to delivering the intended impacts.
- *Impact drivers.* These are significant factors or conditions that are expected to contribute to the ultimate realization of project impacts. Existence of the impact driver in relation to the project being assessed suggests that there is a good likelihood that the intended project impact will have been achieved. Absence of these suggests that the intended impact may not have occurred or may be diminished.

Figure A11.3: Components of impact evaluation framework



- *External assumptions.* These are potential events or changes in the project environment that would negatively or positively affect the ability of a project outcome to lead to the intended impact, but that are largely beyond the power of the project to influence or address.

3. Data collection and constraints

Logical framework and theory of change model

The approach built on existing project logical frameworks, implying that a significant part of the framework could be relatively easily tested through an examination of existing project documentation, terminal evaluation reports and, where available, monitoring data. Where necessary, targeted consultations and additional studies were carried out.

Assessing conservation status and threats to global environment benefits

A data collection framework for assessing the status of the targets and associated threats was developed, identifying indicators for each, along with the potential sources of information. For the Bwindi and Lewa projects, the task of collecting and assessing this information was undertaken by scientists from the Institute of Tropical Forest Conservation, headquartered in Bwindi Impenetrable National Park, and the Lewa Research Department respectively. For the Cross-Borders project, this exercise was done by the Conservation Development Center, based on the existing project documentation, a field visit to the project site, and consultations with key informants. The objective of this exercise was to provide quantitative measures for each indicator from *before the project* (baseline), at the *project close*, and *present day*. Where quantitative data were not available, detailed qualitative data were collected.

Improving rigor

Internal validity: The evaluation used a participatory approach with substantial involvement of former project staff in drawing out theories of change and subsequently providing data for verification. These data were verified by local independent consultants, via a process of triangulating information from project documentation and external sources. Given that all three projects are now closed, the participation from former project staff enabled a candid and detailed exchange of information (during workshops in Uganda and Kenya). The participants in return found the process to be empowering, as it clarified and supported the rationale for their actions (by drawing out the logical connections between activities, goals and assumptions) and enabled them to plan for future interventions.

External validity: Given the small number of projects, their variety, and age (approved in varied past GEF replenishment phases), the evaluation did not expect to produce findings that could be directly aggregated. Nevertheless, given the very detailed analysis of the interventions a few years after project closure, it did provide a wealth of insights into the functioning of protected area projects, particularly elements of their sustainability after project closure. This allowed limited generalization on key factors associated with achievement of impact, on the basis of different levels of results related to a set of common linkages in the theoretical models. On this basis, the Evaluation Office recommended that the GEF Secretariat ensure specific monitoring of progress toward institutional continuity of protected areas throughout the life of a project.

Weaknesses

Impact evaluations are generally acknowledged to be highly challenging. The objective of this particular study, examining GEF's impact at a "global" level in biodiversity, makes the study particularly complex. A few concerns:

- The nature of changes in biodiversity is still under debate. Such changes are often nonlinear, with uncertain time scales even in the short run, interactions within and across species, and exogenous factors (e.g., climate change). Evidence regarding the achievement of global environment benefits and their sustainability must therefore be presented with numerous caveats.
- Numerous explanations and assumptions may be identified for each activity that is carried out.

- The approach may not always uncover unexpected outcomes or synergies, unless they are anticipated in the theories or assumptions of the evaluation team. However, fieldwork should be able to discern such outcomes, as was the case in the Bwindi case study, which produced evidence of a number of unexpected negative impacts on local indigenous people.
- The association between activities and outcomes in the Theory of Change approach depends on measuring the level of activities carried out, and then consciously (logically) linking them with impact through a chain of intermediate linkages and outcomes. Information on these intermediate outcomes may be difficult to obtain, unless former project implementers participate fully in the evaluation.

4. Concluding thoughts on the evaluation approach

For biodiversity, GEF's first strategic priority is *catalyzing sustainability of protected area systems*, which aims for an expected impact whereby "biodiversity [is] conserved and sustainably used in protected area systems."

The advantage of the hybrid evaluation model used was that by focusing toward the end of the results chain, it examined the combination of mechanisms in place that led to a project's impacts and ensure sustainability of results. It is during this later stage, after project closure, that the ecological, financial, political, socio-economic and institutional sustainability of the project are tested, along with its catalytic effects. During project conceptualization, given the discounting of time, links from outcome to impact are often weak. Once a project closes, the role of actors, activities, and resources is often unclear; this evaluation highlighted these links and assumptions.

Adopting a theory of change approach also had the potential to provide a mechanism that helped GEF understand what has worked and what has not worked and allows for predictions regarding the probability of success for similar projects. The Evaluation Office team concluded that the most effective combination for its next round of impact evaluation (phase-out of ozone-depleting substances in eastern Europe) should seek to combine Theory of Change approaches with opportunistic use of existing data sets, which might provide some level of quantifiable counterfactual information.

Application: Impact of Lewa Wildlife Conservancy (Kenya)[15]

Context

The Lewa GEF medium-sized project provided support for the further development of Lewa Wildlife Conservancy ("Lewa"), a not-for-profit private wildlife conservation company that operates on 62,000 acres of land in Meru District, Kenya. The GEF awarded Lewa a grant of $0.75 million for the period 2000 to the end of 2003, with co-financing amounting to $3.193 million.

Since the GEF grant, Lewa has been instrumental in initiating the formation of the Northern Rangelands Trust (NRT) in 2004. NRT is an umbrella local organization with a goal of collectively developing strong community-led institutions as a foundation for investment in community development and wildlife conservation in the Northern Rangelands of Kenya. The NRT membership comprises community conservation conservancies and trusts, local county councils, the Kenya Wildlife Service, the private sector, and NGOs established and working within the broader ecosystem. The establishment and functioning of the NRT has therefore been a very important aspect in understanding and assessing the ultimate achievement of impacts from the original GEF investment.

The Lewa case study implemented the three elements of the Impact Evaluation Framework, which are summarized below.

Assess implementation success and failure

Given that no project logical framework or outcomes were defined as such in the original GEF project brief, the GEF Evaluation Office team for the Study of Local Benefits in Lewa, with the participation of senior Lewa staff, identified project outcomes and associated outputs

that reflected the various intervention strategies employed by the project and identified missed opportunities in achieving the project goals. The assessment provided an understanding of the project logic used (figure A11.2) and a review of the fidelity with which the project activities were implemented (figure A11.3).

Assess the level of contribution (i.e., impact)

A *targets-threats analysis* of those ecological features identified as global environment benefits (black rhinos and Grevy's zebra) was undertaken with input from scientists from Lewa and the NRT research departments. Tables A11.2 and A11.3 provide an overview of the variables considered

Figure A11.4: Project outputs and outcomes

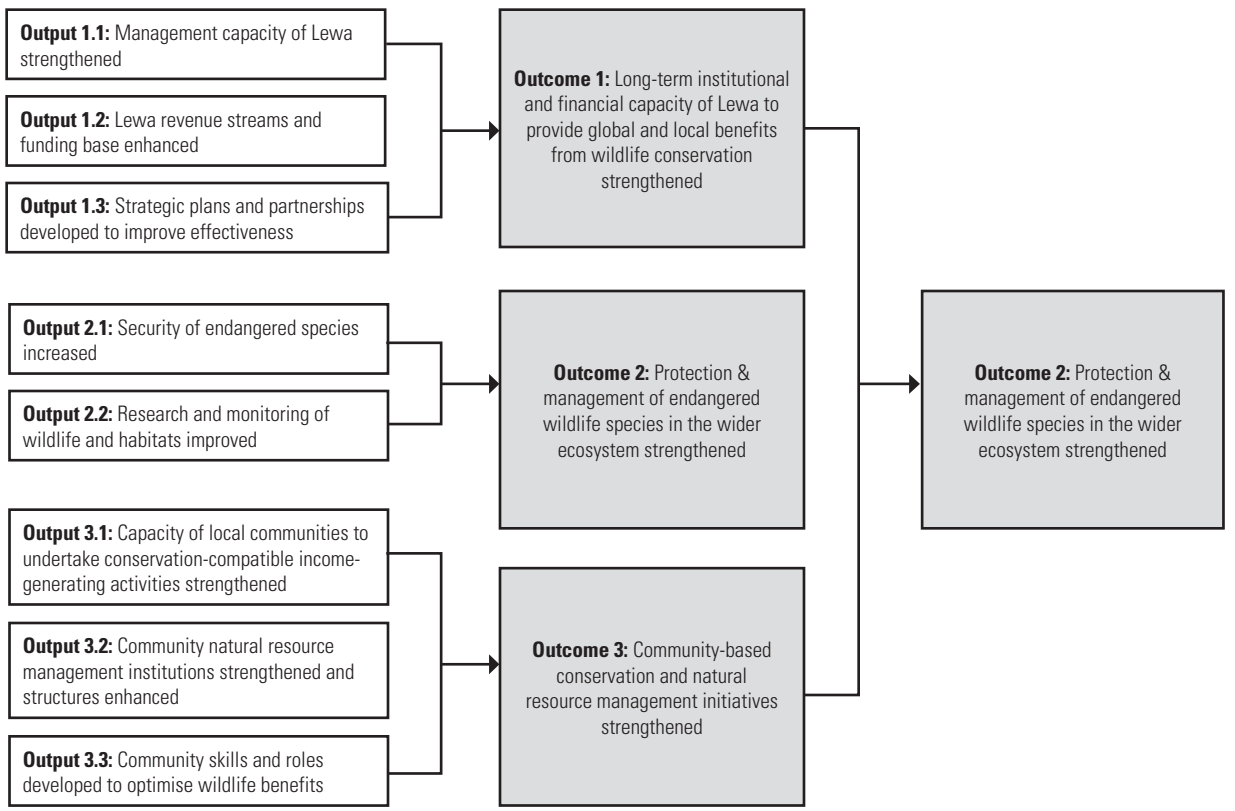


Table A11.1: Project Outcomes

Outcomes	Assessment
Outcome 1: Long-term institutional and financial capacity of Lewa to provide global and local benefits from wildlife conservation strengthened	Fully achieved (5)
Outcome 2: Protection and management of endangered wildlife species in the wider ecosystem strengthened	Well achieved (4)
Outcome 3: Community-based conservation and natural resource management initiatives strengthened	Well achieved (4)

Table A11.2: Change in key ecological attributes over time

Key ecological attribute	Indicator	Unit	Conservation Status			Trend
			Baseline	Project end	Now	
Black rhino						
Population size	Total population size of black rhino on Lewa	Number	29	40	54	⇧
Productivity	Annual growth rates at Lewa	Percent	12	13	15	⇧
Suitable secure habitat	Size of Lewa rhino sanctuary	Acres	55,000	55,000	62,000	⇧
Genetic diversity	Degree of genetic variation	—	No data available			
Grevy's zebra						
Population size	Total population size of Grevy's zebra on Lewa	Number	497	435	430	⇔
Productivity	Annual foaling rates on Lewa	Percent	11	11	12	⇔
Population distribution	Number of known sub-populations and connectivity		No data available			
Suitable habitat (grassland and secure water)	Community conservancies set aside for conservation under NRT	Number	3	4	15	⇧
Genetic diversity	Degree of genetic variation		No data available			

Table A11.3: Current threats to the global environment benefits

	Severity[a] score (1–4)	Scope[b] score (1–4)	Overall ranking
Black rhino			
Poaching and snaring	3	3	**3**
Insufficient secure areas	2	3	**2**
Habitat loss (due to elephant density)	1	1	**1**
Grevy's zebra			
Poaching	2	2	**2**
Disease	4	2	**3**
Predation	3	1	**2**
Habitat loss/ degradation	3	3	**3**
Insufficient secure areas	2	2	**2**
Hybridization with Burchell's zebra	1	1	**1**

[a] Severity (level of damage): Destroy or eliminate GEBs/Seriously degrade the GEBs/Moderately degrade the GEBs/Slightly impair the GEBs.

[b] Scope (geographic extent): Very widespread or pervasive/Widespread/Localized/Very localized.

to increase robustness of the understanding of ecological changes that have taken place since before the project started.

Provide explanations for observed impact

Theory of change models were developed for each project outcome to establish contribution; the framework reflected in figure A11.5 was used. This analysis enabled an examination of the links between observed project interventions and observed impact. As per GEF principles, factors that were examined as potentially influencing results included the *appropriateness* of intervention, the *sustainability* of the intervention and its *catalytic effect*—these are referred to as impact drivers. The next step involved the identification of *intermediary states*, examining whether the successful achievement of a specific project outcome would directly lead to the intended impacts and, if not, identifying additional conditions that would need to be met to deliver the impact. Taking cognizance of factors that are beyond project control, the final step identified those factors that are necessary for the realization and sustainability of the intermediary state(s) and ultimate impacts, but outside the project's influence.

An example is provided by a consideration of Outcome 3 that via *community-based conservation and natural resource management initiatives strengthened*, expected to achieve enhanced conservation of black rhinos and Grevy's zebras. The *theory of change* model linking Outcome 3 to the intended impacts is illustrated below, in figure A11.6. The overall logframe assessment of the project's implementation for community-based conservation and natural resource management was *well achieved*. All intermediate factors/impact drivers/external assumptions that were identified received a score of *partially to well achieved*, indicating that together with all its activities, this component was well-conceived and implemented.

In sum for Lewa

The analysis provided indication that the black rhino and Grevy's zebra populations on the Lewa Conservancy are very well managed and protected. Perhaps the most notable achievement has been the visionary, catalytic, and support role that Lewa has provided for the conservation of these endangered species in the broader ecosystem, beyond Lewa. Lewa has played a significant role in the protection and management of about 40% of Kenya's black rhino population and is providing leadership in finding innovative ways to increase the coverage of secure sanctuaries for black rhinos. Regarding the conservation of Grevy's zebra, Lewa's role in the establishment of community conservancies, which have added almost 1 million acres of land set aside for conservation, has been unprecedented in East Africa and is enabling the recovery of Grevy's zebra populations within their natural range. However, the costs and resources required to manage and protect this increasing conservation estate are substantial, and unless the continued and increasing financing streams are maintained, it is possible that the substantial gains in the conservation of this ecosystem and its global environmental benefits could eventually be reversed.

In conclusion

The assessment of project conceptualization and implementation of project activities in Lewa has been favorable, but, this is coupled with indications that threats from poaching, disease, and habitat loss in and around Lewa continue to be severe. Moreover, evaluation of the other case studies, Bwindi Impenetrable National Park and Mgahinga Gorilla National Park Conservation Project, Uganda and Reducing Biodiversity Loss at Cross-Border Sites in East Africa, Regional: Kenya, Tanzania, Uganda, confirmed that to achieve long-term results in the generation of global environment benefits the absence of a specific plan for institutionalized continuation would, in particular, reduce results over time—this was the major conclusion of the GEF's pilot impact evaluation.

Figure A11.5: Framework to establish contribution

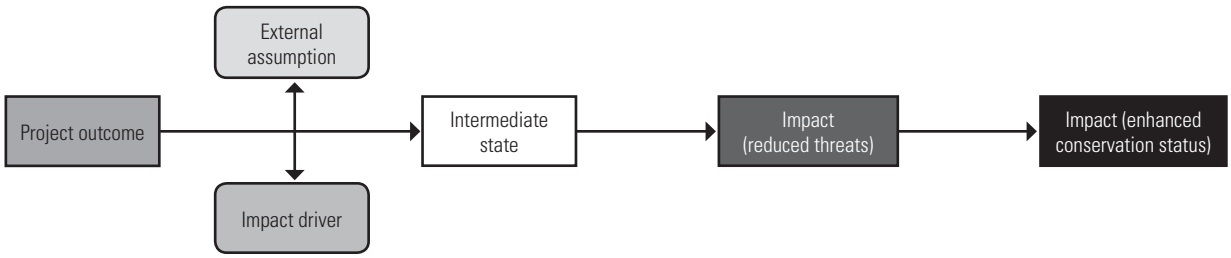
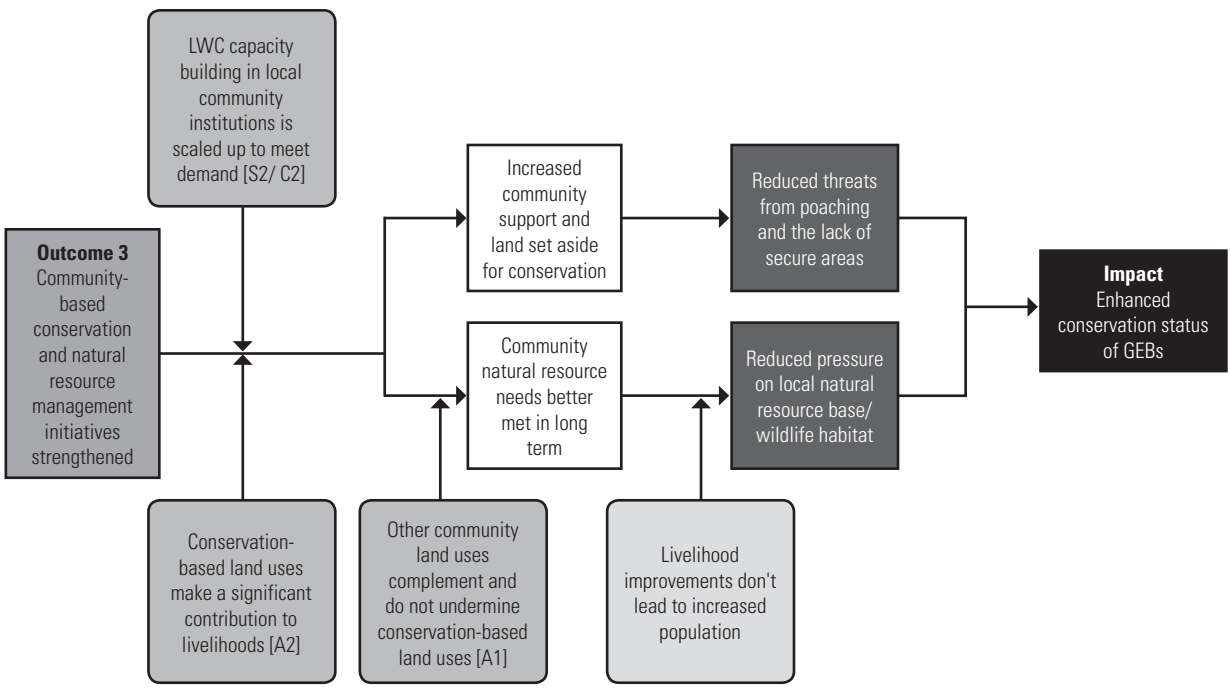


Figure A11.6: Model linking outcome to impact



Realist synthesis

This approach is different from the systematic research reviews. It conceptualizes *interventions, programs, and policies* as theories and collects earlier research findings by interpreting the specific policy instrument that is evaluated, as an example or specimen of *more generic instruments and tools (of governments).* Next it describes the intervention in terms of its context, mechanisms (what makes the program work), and outcomes (the deliverables).

Instead of synthesizing results from evaluations and other studies *per intervention* or *per program*, realist evaluators first open the black box of an intervention and synthesize knowledge about social and behavioral mechanisms. Examples are Pawson's study of incentives (Pawson, 2002), on naming and shaming, and Megan´s law (Pawson, 2006) and Kruisbergen's work (2005) on fear-arousal communication campaigns trying to reduce the smuggling of cocaine.

Contrary to producers of systematic research reviews, realist evaluators do *not* use a hierarchy of research designs. For realists an impact study using the RCT design is not necessarily better than a comparative case study design or a process evaluation. The problem (of an evaluation) that needs to be addressed is crucial in selecting the design or method, not vice versa.

Combining different meta approaches

In a study on the question which public policy programs designed to reduce and/or prevent violence in the public arena work best, Van der Knaap et al. (2008) have shown the relevance of *combining* the *systematic research review* and the *realist synthesis*. Both perspectives have something to offer. Opening the black box of an intervention under review will be helpful for experimental evaluators if they want to understand *why* interventions have (no) effects and/or side effects. Realists are confronted with the problem of the selection of studies to be taken into account, ranging from opinion surveys, oral history, and newspaper content analysis to results based on more sophisticated methodologies. As the methodological quality of evaluations can be and sometimes is a problem, particularly with regard to the measurement of the impact of a program, realists can benefit from a *stricter methodology and protocol,* like the one used by the Campbell Collaboration, when doing a synthesis. For example, knowledge that is to be generalized should be credible and valid.

To combine Campbell standards and the realist evaluation approach, Van der Knaap et al. (2008) *first* conducted a *systematic review* according to the Campbell standards. The research questions were formulated, and next the inclusion and exclusion criteria were determined. This included a number of questions. What types of interventions are included? At which participants should interventions be aimed? What kinds of outcome data should be reported? At this stage, criteria were also formulated for inclusion and exclusion of study designs and methodological quality. As a third step, the search for potential studies was explicitly described. Once potentially relevant studies had been identified, they were screened for eligibility according to the inclusion and exclusion criteria.

After selecting the relevant studies, the quality of these studies had to be determined. Van der

Knaap et al (2008) used the Maryland Scientific Methods Scale (MSMS) (Sherman et al., 1998; Welsh and Farrington, 2006). This is a five-point scale that enables researchers to draw conclusions on methodological quality of outcome evaluations in terms of the internal validity. Using a scale of 1–5, the MSMS is applied to rate the strength of scientific evidence, with 1 being the weakest and 5 the strongest scientific evidence needed for inferring cause and effect.

Based on the MSMS scores, the authors then classified each of the 36 interventions that were inventoried by analyzing some 450 English, German, French, and Dutch articles and papers into the following categories: effective, potentially effective, potentially ineffective, and ineffective.

However, not all studies could be grouped in one of the four categories: in 16 cases the quality of the study design was not good enough to decide on the effectiveness of a measure. The (remaining) *nine interventions were labeled effective and the (final) six were labeled potentially effective.* Four interventions were labeled potentially ineffective and one was labeled ineffective in preventing violence in the public and semi-public domain.

To combine Campbell standards and the realist evaluation approach, the realist approach was applied *after finishing the Campbell-style systematic review*. This means that only then the underlying mechanisms and contexts as described in the studies included in the review were on the agenda of the evaluator. This was done for the four types of interventions, whether they were measured as being effective, potentially effective, potentially ineffective, or ineffective. As a first step, information was collected concerning social and behavioral mechanisms that were assumed to be at work when the program or intervention was implemented. Pawson (2006: 24) refers to this process as "to look beneath the surface [of a program] in order to inspect how they work." One way of doing this is to search articles under review for statements that address the why question: why will this intervention be working or why has it not worked? Two researchers independently articulated these underlying mechanisms. The focus was on behavioral and social "cogs and wheels" of the intervention (Elster, 1989; 2007).

In a second step the studies under review were searched for information on *contexts* (schools, streets, banks, etc., but also types of offenders and victims and type of crime) and *outcomes*. This completed the C[ontext], M[echanism] and O[utcome] approach that characterizes realist evaluations. However, not every original evaluation study described which mechanisms are assumed to be at work when the program is implemented. The same goes for contexts and outcomes. This meant that in most cases missing links in or between different statements in the evaluation study had to be identified through *argumentational analysis*.

Based on the evaluations analyzed, Van der Knaap et al. (2008) traced the following three mechanisms to be at work in programs that had demonstrated their impact or the very-likely-to-come-impact:

- The first is of a *cognitive nature*, focusing on *learning, teaching, and training*.
- The second (overarching) mechanism concerns the way the *(social) environment is rewarding or punishing behavior* (through bonding, community development, and the targeting of police activities).
- The third mechanism is *risk reduction*, for instance, promoting protective factors.

Concluding remarks on review and synthesis approaches

Given the "fleets" (Weiss, 1998) and the streams of studies (Rist and Stame, 2006) in the world of evaluation, it is not recommended to start an impact evaluation of a specific program, intervention, or tool of government *without making use of the accumulated evidence to be found in systematic reviews and other types of meta-studies*. One reason concerns the efficiency of the investments: what has been sorted out does not need (always) to be sorted out again.

If over and over again it has been found that awareness-raising leads to behavior changes only under specific conditions, then it is wise to have that knowledge ready before designing a similar program or evaluation. A second reason is that by using results from synthesis studies the test of an intervention theory can be done with more rigor. The larger the discrepancy between what is known about mechanisms a policy or program believes to be at work and what the policy in fact tries to set into motion, the smaller the chances of an effective intervention.

Different approaches in the world of (impact) evaluation are a wise thing to have, but (continuous) paradigm wars ("randomistas versus relativistas"—realists versus experimentalists) run the risk of developing into intellectual ostracism. Wars also run the risk of vesting the image of evaluations as a "helter-skelter mishmash [and] a stew of hit-or-miss procedures" (Perloff, 2003), which is not the best perspective to live with. Combining perspectives and paradigms should therefore be stimulated.

Introduction

In 1986 the government of Ghana embarked on an ambitious program of educational reform, shortening the length of pre-university education from 17 to 12 years, reducing subsidies at the secondary and tertiary levels, increasing the length of the school day, and taking steps to eliminate unqualified teachers from schools. These reforms were supported by four World Bank credits—the Education Sector Adjustment Credits I and II, the Primary School Development Project, and the Basic Education Sector Improvement Project. An impact study by IEG looked at what had happened to basic education (grades 1–9, in primary and junior secondary school) over this period.

Data and methodology

In 1988–89 the Ghana Statistical Service (GSS) undertook the second round of the Ghana Living Standards Survey (GLSS 2). Half of the 170 areas surveyed around the country were chosen at random to have an additional education module, which administered math and English tests to all those aged 9–55 years with at least three years of schooling and surveyed schools in the enumeration areas. Working with bothGSS and the Ministry of Education, Youth and Sport (MOEYS), IEG resurveyed these same 85 communities and their schools in 2003, applying the same survey instruments. In the interests of comparability, the same questions were kept, although new ones were added pertaining to school management, as were two whole new questionnaires—a teacher questionnaire for five teachers at each school and a local language test in addition to the math and English tests. The study thus had a possibly unique data set—not only could children's test scores be linked to both household and school characteristics, but this could be done in a panel of communities over a 15-year period. The test scores are directly comparable because exactly the same tests were used in 2003 as had been applied 15 years earlier.

There was no clearly defined project for this study, rather support to the sub-sector through four large operations. The four projects had supported a range of activities, from rehabilitating school buildings to assisting in the formation of community-based school management committees. To identify the impact of these various activities a regression-based approach was adopted that analyzed the determinants of school attainment (years of schooling) and achievement (learning outcomes, i.e., test scores). For some of these determinants—notably books and buildings—the contribution of the World Bank to better learning outcomes could then be quantified.

The methodology adopted a theory-based approach to identify the channels through which a diverse range of interventions were having their impact. As discussed below, the qualitative context of the political economy of education reform in Ghana at the time proved to be a vital piece of the story.

Findings

The first major finding from the study was the factual. Contrary to official statistics, enrollments in basic education had been rising steadily over the period. This discrepancy was readily explained: in the official statistics, both the numerator and denominator were wrong. The numerator was wrong as it relied on the administrative data from the school census, which had incomplete coverage of the public sector and did not cover the rapidly growing private sector. A

constant mark-up was made to allow for private sector enrollments, but the IEG analysis showed that that had gone up fourfold (from 5% to 20% of total enrollments) over the 15 years. The denominator was based on the 1984 census, with an assumed rate of growth that turned out to be too high once the 2000 census became available, thus underestimating enrolment growth.

More strikingly still, learning outcomes have improved markedly: 15 years ago nearly two-thirds (63%) of those who had completed grades 3–6 were, using the English test as a guide, illiterate. By 2003 this figure had fallen to 19%. The finding of improved learning outcomes flies in the face of qualitative data from many, though not all, key informant interviews. But such key informants display a middle class bias that persists against the reforms that were essentially populist in nature.

Also striking are the improvements in school quality revealed by the school-level data:

- In 1988, fewer than half of schools could use all their classrooms when it was raining, but in 2003 over two-thirds could do so.
- Fifteen years ago over two-thirds of primary schools reported occasional shortages of chalk. Only one in 20 does so today, with 86% saying there is always enough.
- The percentage of primary schools having at least one English textbook per pupil has risen from 21% in 1988 to 72% today, and for math books in junior secondary school (JSS) these figures are 13% and 71%, respectively.

School quality has improved across the country, in poor and non-poor communities alike. But there is a growing disparity within the public school sector. Increased reliance on community and district financing has meant that schools in relatively prosperous areas continue to enjoy better facilities than do those in less-well-off communities.

The IEG study argues that Ghana has been a case of a quality-led quantity expansion in basic education. The education system was in crisis in the seventies; school quality was declining and absolute enrolments falling. But by 2000, more than 90% of Ghanaians 15 and older had attended school, compared to 75% 20 years earlier. In addition, drop-out rates have fallen, so completion rates have risen: by 2003, 92% of those entering grade 1 complete JSS (grade 9). Gender disparities have been virtually eliminated in basic enrolments. Primary enrolments have risen in both disadvantaged areas and amongst the lowest income groups. The differential between both the poorest areas and other parts of the country, and between enrollments of the poor and non-poor, have been narrowed but still exist.

Statistical analysis of the survey results showed the importance of building school infrastructure based on enrollments. Building a school, and so reducing children's travel time, has a major impact on enrollments. Although the majority of children live within 20 minutes of school, some 20% do not, and school building has increased enrollments among these groups. In one area surveyed, average travel time to the nearest school was cut from nearly an hour to less than 15 minutes, with enrollments increasing from 10% to 80%. In two other areas, average travel time was reduced by nearly 30 minutes and enrollments increased by more than 20%. Rehabilitating classrooms so that they could be used when it is raining also positively affects enrollments. Complete rehabilitation can increase enrollments by as much as one-third. Across the country as a whole, the changes in infrastructure quantity and quality have accounted for a 4% increase in enrolments between 1988 and 2003, about one-third of the increase over that period. The World Bank has been the main source of finance for these improvements. Before the first World Bank program, communities were responsible for building their own schools. These structures collapsed after a few years. The Bank has financed 8,000 school pavilions around the country, providing more permanent structures for the school that can better withstand the weather.

Learning outcomes depend significantly on school quality, including textbook supply. Bank-financed textbook provision accounts for around one-quarter of the observed improvement in test

scores. But other major school-level determinants of achievement, such as teaching methods and supervision of teachers by the head teacher and circuit supervisor, have not been affected by the Bank's interventions. The Bank has not been heavily involved in teacher training and plans to extend in-service training have not been realized. Support to "hardware" has been shown to have made a substantial positive contribution to both attainment and achievement. But when satisfactory levels of inputs are reached—which is still far from the case for the many relatively deprived schools—future improvements could come from focusing on what happens in the classroom. However, the Bank's one main effort to change incentives—providing head teacher housing under the Primary School Development Project in return for the head teacher signing a contract on school management practices—was not a great success. Others, notably DFID and USAID, have made better progress in this direction but with limited coverage.

The policy context, meaning government commitment, was an important factor in making the Bank's contributions work. The government was committed to improving the quality of life in rural areas, through the provision of roads, electricity, and schools, as a way of building a political base. Hence there was a desire to make it work. Party loyalists were placed in key positions to keep the reform on track, the army distributed textbooks in support of the new curriculum in the early 1990s to make sure they reached schools on time, and efforts were made to post teachers to new schools and make sure that they received their pay on time.

Teachers also benefited from the large civil service salary increase in the run up to the 1992 election. Better education leads to better welfare outcomes. Existing studies on Ghana show how education reduces fertility and mortality. Analysis of IEG's survey data shows that education improves nutritional outcomes, with this effect being particularly strong for children of women living in poorer households. Regression analysis shows there is no economic return to primary and JSS education (i.e., average earnings are not higher for children who have attended primary and JSS than for children who have not), but there is a return to cognitive achievement. Children who attain higher test scores as a result of attending school can expect to enjoy higher income; but children who learn little in school will not reap any economic benefit.

Some policy implications

The major policy finding from the study relates to the appropriate balance between hardware and software in support for education. The latter is now stressed. But the study highlights the importance of hardware: books and buildings. It was also of course important that teachers were in their classrooms; the government's own commitment (borne out of a desire to build political support in rural areas) helped ensure this happened.

In the many countries and regions in which educational facilities are inadequate, then hardware provision is a necessary step in increasing enrollments and improving learning outcomes. The USAID project in Ghana encourages teachers to arrange children's desks in groups rather than rows—but many of the poorer schools don't have desks. In the words of one teacher, "I'd like to hang posters on my walls but I don't have posters. In fact, as you can see, I don't have any walls."

These same concerns underlie a second policy implication. Central government finances teacher's salaries and little else in basic education. Other resources come from donors, districts, or the communities themselves. There is thus a real danger of poorer communities falling behind, as they lack both resources and the connections to access external resources. The reality of this finding was reinforced by both qualitative data—field trips to the best and worst performing schools in a single district in the same day—and the quantitative data, which show the poorer performance of children in these disadvantaged schools. Hence children of poorer communities are left behind and account for the remaining illiterate primary graduates, which should be a pressing policy concern.

The study highlighted other areas of concern: first, low teacher morale, manifested through increased absenteeism; and second, the growing importance of the private sector, which now accounts for 20% of primary enrolments compared to 5% 15 years earlier. This is a sector that has had limited government involvement and none from the Bank.

APPENDIX 14: HIERARCHY OF QUASI-EXPERIMENTAL DESIGNS

	Start of project (pre-test)	Project intervention (process not discrete event)	Mid-term evaluation	End of project (post-test)	The stage of the project cycle at which each evaluation design can be used
Quantitative Impact Evaluation Design	T_1		T_2	T_3	
Relatively robust quasi-experimental designs					
1. *Pre-test/post-test non-equivalent control group design with statistical matching of the two groups.* Participants are either self-selected or are selected by the project implementing agency. Statistical techniques (such as propensity score matching), drawing on high-quality secondary data used to match the two groups on a number of relevant variables.	P_1 C_1	X		P_2 C_2	Start
2. *Pre-test/post-test non-equivalent control group design with judgmental matching of the two groups.* Participants are either self-selected or are selected by the project implementing agency. Control areas usually selected judgmentally and subjects are randomly selected from within these areas.	P_1 C_1	X		P_2 C_2	Start
Less robust quasi-experimental designs					
3. *Pre-test/post-test comparison where the baseline study is not conducted until the project has been under way for some time* (most commonly this is around the mid-term review).		X	P_1 C_1	P_2 C_2	During project implementation (often at mid-term)
4. *Pipeline control group design.* When a project is implemented in phases, subjects in Phase 2 (i.e., who will not receive benefits until some later point in time) can be used as the control group for Phase 1 subjects.	P_1 C_1	X		P_2 C_2	Start
5. *Pre-test/post-test comparison of project group combined with post-test comparison of project and control group*	P_1	X		P_2 C_2	Start
6. *Post-test comparison of project and control groups*		X		P_1 C_1	End
Non-experimental designs (the least robust)					
7. *Pre-test/post-test comparison of project group*	P_1	X		P_2	Start
8. *Post-test analysis of project group*		X		P_1	End

Source: Bamberger et al. (2006).

Note: T = time; P = project participants; C = control group; P_1, P_2, C_1, C_2 = first and second observations; X = project intervention (a process rather than a discrete event).

APPENDIX 15: INTERNATIONAL EXPERTS WHO CONTRIBUTED TO THE SUBGROUP DOCUMENTS

Marie-Hélene Adrien: President and Senior Consultant, Universalia

Paul Balogun: Consultant, Author

Michael Bamberger: Consultant, Author

Fred Carden: Director of Evaluation Unit, IDRC Canada

Stewart Donaldson: Professor and Chair of Psychology, Director of the Institute of Organizational and Program Evaluation Research, and Dean of the School of Behavioural and Organizational Sciences, Claremont Graduate University

Osvaldo Feinstein: Consultant. Author, Editor

Ted Freeman: Consultant and Partner, Gross Gilroy, Inc.

Sulley Gariba: Consultant, Executive Director, Institute of Policy Alternatives

Jennifer Greene: Professor, Educational Psychology, University of Illinois at Urbana-Champaign

Ernie House: Emeritus Professor, School of Education, University of Colorado

Mel Mark: Professor of Psychology, Penn State University

John Mayne: Consultant, Author, Adviser on public sector performance

Masafumi Nagao: Research Professor, Center for the Study of International Cooperation in Education, Hiroshima University

Michael Quinn Patton: Consultant, Author, Former President of AEA

Ray Pawson: Professor of Social Research Methodology, School of Sociology and Social Policy, University of Leeds

Bob Picciotto: Visiting Professor, Kings College, London

Patricia Rogers: Professor in Public Sector Evaluation, CIRCLE (Collaboration for Interdisciplinary Research, Consulting and Learning in Evaluation), Royal Melbourne Institute of Technology

Thomas Schwandt: University Distinguished Teacher/Scholar and Professor of Education, University of Illinois at Urbana-Champaign

Nicoletta Stame: Professor, University of Rome "La Sapienza"

Bob Williams: Consultant, Author, member of the Editorial Boards of the American Journal of Evaluation and New Directions in Evaluation

Executive Summary

1. Available at www.worldbank.org/ieg/nonie.

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

Introduction

1. The history of impact evaluations in some countries goes back many decades (Oakley, 2000).

2. The Maryland Scientific Methods Scale (MSMS) is, for example, used in parts of criminology and in several countries (see Leeuw, 2005). RCTs are believed to be the top design (level 5).

Chapter 1

1. An interesting overview of public-private partnerships and their evaluation is given by Utce Ltd. and Japan Pfi Association (2003).

2. "We probably also under-invest in evaluative research on types of interventions that tend to have diffused, wide-spread benefits" (Ravallion, 2008: 6). See also Jones et al. (2008), who have identified geographical and sectoral biases in impact evaluation.

3. Complexity in terms of the nature of change processes induced by an intervention.

4. For example, Elbers et al. (2008) directly assess the impact of a set of policy variables (i.e., the equivalent of a multi-stranded program) by means of a regression-based evaluation approach (see chapter 4) on outcome variables.

5. Though not necessarily easy to measure.

6. Please note that the two levels should not be regarded as a dichotomy. In fact, a particular intervention might induce a "cascade" of processes of change at different institutional levels (e.g., national, provincial government, cooperatives) before finally affecting the welfare of individuals.

7. A *third and fourth level of impact*, more difficult to pinpoint, respectively refer to the replicatory impact and the wider systemic effects of interventions. Both replicatory and systemic effects can result from processes of change at institutional or beneficiary levels. With respect to the first, evaluations that cover replicatory effects are quite scarce. This is in dire contrast with the manifest presence of replication (and the related concept of scaling up) as explicit objectives in many policy interventions. For further discussion on replication, see, for example, GEF (2007). These dimensions can be addressed in a theory-based impact evaluation framework (see chapter 3).

8. This is the interpretation that has received the most attention in methodological guidelines of international organizations working on impact evaluation, such as the World Bank or the Asian Development Bank.

9. In this context one can distinguish between the effect of aid modalities on "the way business is being done" (additionality of funding, direction of funding, public sector performance, coherence of policy changes, quality of intervention design, etc.; see, e.g., Lawson et al., 2005), i.e., what we call institutional-level impact, and subsequently the impact of interventions funded (in part) by general budget support, sector budget support, or debt relief funds at the beneficiary level. In the latter case, we are talking about impact evaluation as it is understood in most of the literature.

Chapter 2

1. "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).

2. For a discussion on different dimensions of sustainability in development intervention, see Mog (2004).

Chapter 4

1. Economists employ several useful techniques for estimating the marginal impact of an extra dollar invested in a particular policy intervention. See, for

example appendix 1, second example. We consider these methods to be complementary to impact evaluation and beyond the scope of this guidance.

2. The larger the sample size, the more likely it is that groups are equivalent, on average.

3. We would like to thank Antonie de Kemp of IOB for insightful suggestions. See also SG1 (2008).

4. Alternative, more nuanced classifications distinguish between experimental, quasi-experimental, and passive observational (correlational) research designs. Features that distinguish one type of design from another are (i) control over exposure to the treatment; (ii) control over the nature of the treatment; and (iii) control over the timing and nature of measurement. In experiments one has control over i, ii, and iii; in quasi-experiments one usually controls ii and iii only; and in passive observational studies one does not have full control over any of these features (see, e.g., Posavac and Carey, 2002; personal communication, J. Scott Bayley).

5. We discuss only a selection of available methods. See Shadish et al. (2002) or Mohr (1995) for additional (quasi-experimental and regression-based) methods.

6. It is difficult to identify general guidelines for avoiding these problems. Evaluators have to be aware of the possibility of these effects affecting the validity of the design. For other problems, as well as solutions, see Shadish et al. (2002).

7. For further discussion on the approaches discussed below, see appendices 3–6.

8. For an explanation, see Wooldridge (2002), chapter 18.

9. This subsection comes largely from Bamberger (2006).

10. The approach is similar to a *fixed-effects regression* model that uses deviations from individual means to deal with (unobserved) selection effects.

11. Although in reality one will not find such a clear linear correlation as figure 4.2.

12. With instrumental variables one may try to get rid of an expected bias, but the technique cannot guarantee that endogeneity problems will be solved completely (the instrumental variable may also be endogenous). Moreover, with weak instruments the precision of the estimate may be low.

13. Alternatively, impact evaluation in the case of complex interventions or complex processes of change can rely on several statistical modeling approaches to capture the complexity of a phenomenon. For example, an extension of reduced form regression-based approaches to impact evaluation referred to earlier are structural equation models that can be used to model some of the more complex causal relationships that underlie interventions, using, for example, an intervention theory as a basis.

14. In general, regression-based techniques (and quasi-experimental techniques that rely on existing data) are primarily constrained by the availability of existing data (see chapter 8). In contrast, experimental and quasi-experimental techniques that rely on design-based group comparisons face more pressing constraints in terms of the need for ex ante involvement of evaluators in a policy intervention (see appendix 14). Consequently, there is probably more scope for extending the use of the former group of techniques.

15. This might need to be analyzed using other methods (see §4.4 and chapter 5).

16. See appendices 7 and 8 for brief discussions on additional approaches applicable to impact evaluation problems in multi-level settings.

17. However, as explained below, in some cases these methods can be articulated to quantitative methods of impact evaluation (see also chapter 5).

18. See also SG2 (2008).

19. One of the methods that relies on the reconstruction of stakeholder perspectives is called the *strategic assessment approach*, also known as *assumptional analysis.* It can be found in a series of studies (Jackson, 1989) but has as its core knowledge basis Mason and Mitroff's (1981) book *Challenging Strategic Planning Assumptions* (see also Leeuw, 2003; see also chapter 3).

20. Participatory Learning and Action as a generic approach with an associated set of methods has its origins in rapid rural appraisal and participatory rural appraisal. Participatory poverty assessment processes have built strongly on this tradition.

21. Although particular case studies of localized intervention activities within the sector program might be conducted in a participatory manner.

22. When addressing the attribution problem, the role of participatory approaches is also restricted because perceptions and experiences of participants collected through participatory methods run the risk of making an evaluation "partnerial." In such a situation, the distinction between evaluator and evaluated is blurred. As policies and programs often—implicitly or explicitly—deal with interests, incentives,

and disincentives, this complicates the process and the reliability of the evaluation outcomes. (See also §8.3 for a wider discussion of data quality issues.)

23. Throughout this document we have used the rather generic terms "quantitative" and "qualitative" methods of research/evaluation. Although we are aware of the limitations of these concepts, we have opted to use them because of their widespread accepted use. In practice, *often but not always,* a distinction can be made between methods of data collection and methods of data analysis. In addition, one should distinguish between the type of method and the scale of measurement (type of data). For example, quantitative data (that is, data measured on interval or ratio scales) can be collected using what are often called qualitative methods. Rather than spending a lot of effort on coherently separating these issues, we decided to keep things simple for the sake of argument (and space).

24. Please note that different methods rely on different types of sampling or selection of units of analysis. For example, quantitative descriptive analysis (preferably) relies on data based on random (simple, stratified, clustered) samples or on census data. In contrast, many qualitative methods rely on nonrandom sampling techniques such as purposive or snowball sampling or do not rely on sampling at all, as they might focus on a relatively small number of observations.

25. Appendix 9 presents a list of qualitative methodological frameworks that combine several qualitative (and occasionally quantitative) methods for the purposes of evaluating the effects of an intervention (see also chapter 5 on combining methods).

Chapter 5

1. This dimension is only addressed by quantitative impact evaluation techniques.

2. The most commonly used term is mixed methods (see for example Tashakkori and Teddlie, 2003). In the case of development research and evaluation, see Bamberger (2000) and Kanbur (2003).

3. This is true for the broad interpretation of the concept of triangulation as used by Mikkelsen (2005). Other authors use the concept in a more restrictive way (e.g., Bamberger [2000] uses triangulation in the more narrow sense of validating findings by looking at different data sources).

4. 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.

Chapter 6

1. This step may rely on statistical methods (meta-analysis) for analyzing and summarizing the results of included studies, if quantitative evidence at the level of single-intervention studies is available and if interventions are considered similar enough.

Chapter 8

1. In some cases, talking about the "end" of an intervention is not applicable or is less applicable, for example, in institutional reforms, new legislation, fiscal policy, etc.

2. For example, with secondary data sets, what do we know about the quality of the data collection (e.g., sampling errors, training and supervision of interviewers) or data processing (e.g., dealing with missing values, weighting issues)? We cannot simply take for granted that a data set is free from error and bias. Lack of information on the process of generating the database inevitably constrains any subsequent data analysis efforts.

Chapter 9

1. An example from Europe stresses this point. In some situations, educational evaluators of the Danish Evaluation Institute discussed their reports with up to 20-plus stakeholders before the report was cleared and published (Leeuw, 2003).

2. For a broader discussion on ethics in evaluation, see Simons (2006).

Appendix 2

1. The text is a literal citation of Scriven (2008: 21–22).

Appendix 4

1. In traditional usage, a variable is endogenous if it is determined within the context of a model. In econometrics, it is used to describe any situation in which an explanatory variable is correlated with the disturbance term. Endogeneity arises as a result of omitted variables, measurement error, or in situations where one of the explanatory variables is determined along with the dependent variable (Wooldridge, 2002: 50).

2. The approach is similar to a fixed-effects regression model, using deviations from individual means.

Appendix 5

1. For further examples see White (2006).

Appendix 9

1. Source: SG2 (2008).

Appendix 11

1. This case study is drawn from the 2002 report published by the Ministry of Foreign Affairs, Denmark (SG2, 2008).

2. Source: SG2 (2008).

3. Typical problems with recall methods are that of incorrect recalling and telescoping, i.e., projecting backward or forward onto an event: for example, the purchase of a durable good that took place seven years ago (before the project started) could be projected to four years ago, during project implementation (see, e.g., Bamberger et al., 2004).

4. Source: SG2 (2008).

5. The second project was inland valley development for irrigated rice cultivation and is not presented here.

6. Industrial plantations are the property of SOGUIPAH and are worked by salaried employees.

7. A contract between SOGUIPAH and the farmer binds the farmer to reimburse the cost of the plantation and deliver his production to SOGUIPAH.

8. AGROPARISTECH is a member of the Paris Institute of Technology, which is a consortium of 10 of the foremost French Graduate Institutes in Science and Engineering. AGROPARISTECH is a leader Institute in life sciences and engineering.

9. Source: SG2 (2008).

10. The GEF Evaluation Office section of the GEF website contains the 11 papers produced by the impact evaluation in 2007, under the heading of "ongoing evaluations."

11. Instrument for the Establishment of the Restructured Global Environment Facility.

12. GEF Evaluation Office, "Approach Paper to Impact Evaluation," February 2006.

13. See the Preamble, "Instrument for the Establishment of the Restructured Global Environment Facility."

14. This is based on Nature Conservancy's conservation action planning methodology.

15. Full case study at http://www.thegef.org/uploadedFiles/Evaluation_Office/Ongoing_Evaluations/Ongoing_Evals-Impact-8Case_Study_Lewa.pdf.

Appendix 13

1. White (2006).

ADB (2006) *Impact Evaluation—Methodological and Operational Issues*, Economics and Research Department, Asian Development Bank, Manila.

Agresti, A., and B. Finlay (1997) *Statistical Methods for the Social Sciences*, Prentice Hall, New Jersey.

Baker, J.L. (2000) *Evaluating the Impact of Development Projects on Poverty*, The 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. (2006) *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints*, World Bank, Washington, D.C.

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) *Real-World Evaluation Working Under Budget, Time, Data, and Political Constraints*, Sage Publications, Thousand Oaks, CA.

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

Bemelmans-Videc, M.L., and R.C. Rist (eds.) (1998) *Carrots, Sticks and Sermons: Policy Instruments and their Evaluation*, Transaction Publishers, New Brunswick.

Booth, D., and H. Lucas, H. (2002) "Good Practice in the Development of PRSP Indicators", *Working Paper* 172, Overseas Development Institute, London.

Bourguignon, F., and M. Sundberg (2007) "Aid effectiveness, opening the black box", *American Economic Review* 97(2), 316–321.

Brinkerhoff, R. (2003) *The Success Case Method*, Berrett Koehler, San Francisco.

Bryman, A. (2006) "Integrating quantitative and qualitative research: How is it done?" *Qualitative Research* 6(1), 97–113.

Bunge, M. (2004) "How Does It Work? The Search for Explanatory Mechanisms", *Philosophy of the Social Sciences* 34(2), 182–210.

Campbell, D.T. (1957) "Factors relevant to the validity of experiments in social settings", *Psychological Bulletin* 54, 297–312.

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.

Casley, D.J., and D.A. Lury (1987) *Data Collection in Developing Countries*, Oxford University Press, New York.

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. (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.

Clarke, A. (2006) "Evidence-Based Evaluation in Different Professional Domains: Similarities, Differences and Challenges", in: I.F. Shaw,

J.C. Greene and M.M. Mark (eds.) *The SAGE Handbook of Evaluation*, Sage Publications, London.

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

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.) (2000) *Program Theory in Evaluation: Challenges and Opportunities*, New Directions for Evaluation, 87, Jossey-Bass, San Francisco.

Cook, T.D., and D.T. Campbell (1979) *Quasi-Experimentation: Design and Analysis for Field Settings*, Rand McNally, Chicago.

Cooke, B. (2001) "The Social Psychological Limits of Participation?" in: B. Cooke and U. Kothari (eds.) *Participation: The New Tyranny?*, Zed Books, London.

Connell, J.P., A.C. Kubisch, L.B. Schorr and C.H. Weiss (eds.) (1995) *New Approaches to Evaluating Community Initiatives*, The Aspen Institute, Washington, D.C.

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.

Davies, R., and J. Dart (2005) *The 'Most Significant Change' Technique,* http://www.mande.co.uk/docs/MSCGuide.pdf (last consulted May 12, 2009).

Deaton, A. (2005) "Some remarks on randomization, econometrics and data", in: G.K. Pitman, O.N. Feinstein and G.K. Ingram (eds.) *Evaluating Development Effectiveness*, Transaction Publishers, New Brunswick, NJ.

Dehejia, R. (1999) "Evaluation in multi-site programs", Working paper, Columbia University and NBER, http://emlab.berkeley.edu/symposia/nsf99/papers/dehejia.pdf (last consulted January 12, 2009).

De Leeuw, E.D., J.J. Hox and D.A. Dillman (eds.) (2008) *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates, London.

Duflo, E. and M. Kremer (2005) "Use of randomization in the evaluation of development effectiveness", in: G.K. Pitman, O.N. Feinstein and G.K. Ingram (eds.) *Evaluating Development Effectiveness*, Transaction Publishers, New Brunswick.

Elbers, C., J.W. Gunning and K. De Hoop (2008) "Assessing sector-wide programs with statistical impact evaluation: a methodological proposal", *World Development* 37(2), 513–520.

Elster, J. (1989) *Nuts and Bolts for the Social Sciences*, Cambridge University Press, Cambridge.

Elster, J. (2007) *Explaining Social Behavior—More Nuts and Bolts for the Social Sciences*, Cambridge University Press, Cambridge.

Farnsworth, W. (2007) *The Legal Analyst—A Toolkit for Thinking about the Law*, University of Chicago Press, Chicago.

GEF (2007) "Evaluation of the Catalytic Role of the GEF", Approach Paper, GEF Evaluation Office, Washington, D.C.

Gittinger, J.P. (1982) *Economic Analysis of Agricultural Projects,* Johns Hopkins University Press, Baltimore.

Greene, J.C. (2006) "Evaluation, democracy and social change", in: I.F. Shaw, J.C. Greene and M.M. Mark (eds.) *The SAGE Handbook of Evaluation*, Sage Publications, London.

Greenhalgh, T., G. Robert, F. Macfarlane, P. Bate and O. Kyriakidou (2004) "Diffusion of Innovations in Service Organizations: Systematic Review and Recommendations", *The Milbank Quarterly* 82(1), 581–629.

Hair, J.F., B. Black, B. Babin, R.E. Anderson and R.L. Tatham (2005) *Multivariate Data Analysis*, Prentice Hall, New Jersey.

Hansen, H.F., and Rieper, O. (2009) "Institutionalization of second-order evidence producing organizations", in: O. Rieper, F.L. Leeuw and T. Ling (eds.) *The Evidence Book: Concepts, Generation and Use of Evidence*, Transaction Publishers, New Brunswick.

Hedström, P. (2005) *Dissecting the Social: On the Principles of Analytical Sociology*, Cambridge University Press, Cambridge.

Hedström, P., and R. Swedberg (1998) *Social Mechanisms: An Analytical Approach to Social Theory*, Cambridge University Press, Cambridge.

Henry, G.T. (2002) "Choosing Criteria to Judge Program Success—A Values Inquiry", *Evaluation* 8(2), 182–204.

House, E. (2008) "Blowback: Consequences of Evaluation for Evaluation", *American Journal of Evaluation* 29(4), 416–426.

IDRC (2001) *Outcome Mapping: Building Learning and Reflection into Development Programs,* International Development Research Centre (IDRC), Ottawa.

IEG (2005) "OED and Impact Evaluation: A Discussion Note," Operations Evaluation Department, World Bank, Washington, D.C.

IFAD (2002) *Managing for Impact in Rural Development: A Practical Guide for M&E*, IFAD, Rome.

Jackson, M. C. (1989) "Assumptional analysis", *Systems Practice* 14, 11–28.

Jerve, A.M., and E. Villanger (2008) *The challenge of Assessing Aid Impact: A Review of Norwegian Practice,* Study commissioned by NORAD, Chr. Michelsen Institute, Bergen.

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.

Kanbur, R. (ed.) (2003) *Q-Squared: Combining Qualitative and Quantitative Methods in Poverty Appraisal*, Permanent Black, Delhi.

Kellogg Foundation (1991) *Information on Cluster Evaluation*, Kellogg Foundation, Battle Creek.

Kraemer, H.C. (2000) "Pitfalls of Multisite Randomized Clinical Trials of Efficacy and Effectiveness." *Schizophrenia Bulletin* 26, 533–541.

Kruisbergen, E.W. (2005) Voorlichting: doen of laten? Theorie van afschrikwekkende voorlichtingscampagnes toegepast op de casus van bolletjesslikkers, *Beleidswetenschap* 19(3), 3–1.

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.

Lawson, A., D. Booth, M. Msuya, S. Wangwe and T. Williamson (2005) *Does General Budget Support Work? Evidence from Tanzania*, Overseas Development Institute, London.

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. (2005) "Trends and Developments in Program Evaluation in General and Criminal Justice Programs in Particular", *European Journal on Criminal Policy and Research* 11, 18–35.

Leeuw, F.L., and J.E. Furubo (2008) "Evaluation Systems – What Are They and Why Study Them?", *Evaluation* 14(2), 157–169.

Levinsohn, J., S. Berry, and J. Friedman (1999) "Impacts of the Indonesian Economic Crisis: Price Changes and the Poor", Working Paper 7194, National Bureau of Economic Research, Cambridge.

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.

Lister, S., and R. Carter (2006) *Evaluation of General Budget Support: Synthesis Report*, Joint evaluation of general budget support 1994–2004, Department for International Development, University of Birmingham.

Maluccio, J. A., and R. Flores (2005) "Impact evaluation of conditional cash transfer program: The Nicaraguan Red de Protección Social", International Food Policy Research Institute, Washington, D.C.

Mansuri, G., and V. Rao (2004) Community-Based and -Driven Development: A Critical Review, *The World Bank Research Observer* 19(1), 1–39.

Mark, M.M., G.T. Henry and G. Julnes (1999) "Toward an Integrative Framework for Evaluation Practice", *American Journal of Evaluation* 20, 177–198.

Mason, I., and I. Mitroff (1981) *Challenging Strategic Planning Assumptions*, Wiley, New York.

Mayne, J. (2001) "Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly", *Canadian Journal of Program Evaluation* 16(1), 1–24.

Mayntz, R. (2004) "Mechanisms in the Analysis of Social Macro-phenomena", *Philosophy of the Social Sciences* 34(2), 237–259.

McClintock, C. (1990) "Administrators as applied theorists", in: L. Bickman (ed.) *Advances in Program Theory*, *New Directions for Evaluation*, Jossey-Bass, San Francisco.

Mikkelsen, B. (2005) *Methods for Development Work and Research*, Sage Publications, Thousand Oaks, CA.

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

Mohr, L.B. (1995) *Impact Analysis for Program Evaluation,* Sage Publications, Newbury Park, CA.

Morgan, S.L., and C. Winship (2007) *Counterfactuals and Causal Inference—Methods and Principles for Social Research*, Cambridge University Press, Cambridge.

Mukherjee, C., H. White and M. Wuyts (1998) *Econometrics and Data Analysis for Developing Countries*, Routledge, London.

North, D.C. (1990) *Institutions, Institutional Change and Economic Performance*, Cambridge University Press, New York.

Oakley, A. (2000) *Experiments in Knowing: Gender and Method in the Social Sciences*, Polity Press, Cambridge.

OECD-DAC (2000) *Effective Practices in Conducting a Multi-donor Evaluation*, OECD-DAC, Paris.

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

Oliver, S., A. Harden, R. Rees, J. Shepherd, G. Brunton, J. Garcia and A. Oakley (2005) "An Emerging Framework for Including Different Types of Evidence in Systematic Reviews for Public Policy", *Evaluation* 11(4), 428–446.

Patton, M.Q. (2002) *Qualitative Research and Evaluation Methods*, Sage Publications, Thousand Oaks, CA.

Pawson, R. (2002) "Evidence-based Policy: The Promise of 'Realist Synthesis'", *Evaluation* 8(3), 340–358.

Pawson, R. (2005) "Simple Principles for The Evaluation of Complex Programmes", in: A. Killoran, M. Kelly, C. Swann, L. Taylor, L. Milward and S. Ellis (eds.) *Evidence-Based Public Health*, Oxford University Press, Oxford.

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*, Sage Publications, London.

Pawson, R., and N. Tilley (1997) *Realistic Evaluation,* Sage Publications, Thousand Oaks, CA.

Perloff, R. (2003) "A potpourri of cursory thoughts on evaluation", *Industrial-Organizational Psychologist 40*(3), 52–54.

Picciotto, R. (2004) "The value of evaluation standards: A comparative assessment" Paper presented at the European Evaluation Society's 6th biennial Conference on Democracy and Evaluation, Berlin.

Picciotto, R., and E. Wiesner (eds.) (1997) *Evaluation and Development: The Institutional Dimension*, World Bank Series on Evaluation and Development, Transaction Publishers, New Brunswick.

Pollitt, C. (1999) "Stunted by stakeholders? Limits to collaborative evaluation", *Public Policy and Administration* 14 (2), 77–90.

Posavac, E.J., and R.G. Carey (2002) *Program Evaluation: Methods and Case Studies,* Prentice Hall, Englewood Cliffs, NJ.

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.

Rieper, O., F.L. Leeuw and T. Ling (eds.) (2009) *The Evidence Book: Concepts, Generation and Use of Evidence*, Transaction Publishers, New Brunswick, NJ.

Rist, R., and N. Stame (eds.) (2006) *From Studies to Streams—Managing Evaluative Systems*, Transaction Publishers, New Brunswick.

Robilliard, A.S., F. Bourguignon and S. Robinson (2001) "Crisis and Income Distribution: A Micro-Macro Model for Indonesia", International Food Policy Research Institute, Washington, D.C.

Roche, C. (1999) *Impact Assessment for Development Agencies: Learning to Value Change*, Oxfam, Oxford.

Rogers, P. J. (2008) "Using programme theory for complex and complicated programs", *Evaluation* 14(1), 29–48.

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, Jossey-Bass, San Francisco.

Rosenbaum, P.R., and D.B. Rubin (1983) "The central role of the propensity score in observational studies for causal effects", *Biometrika* 70, 41–55.

Rossi, P.H., M.W. Lipsey, and H.E. Freeman (2004) *Evaluation: A Systematic Approach*, Sage Publications, Thousand Oaks, CA.

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

Salmen, L., and E. Kane (2006) *Bridging Diversity: Participatory Learning for Responsive Development,* World Bank, Washington, D.C.

Scriven, M. (1976) "Maximizing the Power of Causal Investigations: The Modus Operandi Method", in: G. V. Glass (ed.) *Evaluation Studies Review Annual,* Vol. 1, Sage Publications, Beverly Hills, CA.

Scriven, M. (1998) "Minimalist theory: The least theory that practice requires", *American Journal of Evaluation 19*(1), 57–70.

Scriven, M. (2008) "Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research", *Journal of Multidisciplinary Evaluation* 5(9), 11–24.

SG1 (2008) *NONIE: Impact Evaluation Guidance—Sections 1 and 2*, Subgroup 1, Network of Networks on Impact Evaluation.

SG2 (2008) *NONIE Impact Evaluation Guidance*, Subgroup 2, Network of Networks on Impact Evaluation.

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 (1998) "Preventing crime: What works, what doesn't, what's promising", *National Institute of Justice Research Brief*, July 1998, Washington, D.C.

Simons, H. (2006) "Ethics in evaluation", in: I.F. Shaw, J.C. Greene and M.M. Mark (eds.) *The SAGE Handbook of Evaluation,* Sage Publications London.

Snijders, T., and R. Bosker (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, London.

Späth, B. (2004) *Current State of the Art in Impact Assessment: With a Special View on Small Enterprise Development,* Report for SDC.

Straw, R.B., and J.M. Herrell (2002) "A Framework for understanding and improving Multisite Evaluations", in: J.M. Herrell and R.B. Straw (eds.), *Conducting Multiple Site Evaluations in Real-World Settings*, New Directions for Evaluation 94, Jossey-Bass, San Francisco.

Swedberg, R. (2005) *Principles of Economic Sociology*, Princeton University Press, Princeton, NJ.

Tashakkori, A., and C. Teddlie (eds.) (2003) *Handbook of Mixed Methods in Social and Behavioral Research*, Sage Publications, Thousand Oaks, CA.

Trochim, W.M.K. (1989) "An introduction to concept mapping for planning and evaluation", *Evaluation and Program Planning* 12, 1–16.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, PA.

Turpin, R.S., and J.M. Sinacore (eds.) (1991) *Multisite Evaluations*. New Directions for Evaluation 50, Jossey-Bass, San Francisco.

Utce Ltd & Japan Pfi Association (2003*) Impact Evaluation Study on Public-Private Partnerships: The Case of Angat Water Supply Optimization Project and the Metropolitan Waterworks and Sewerage System*, Republic of the Philippines.

Vaessen, J., and J. De Groot (2004) "Evaluating Training Projects on Low External Input Agriculture: Lessons from Guatemala", *Agricultural Research & Extension Network Papers* 139, Overseas Development Institute, London.

Vaessen, J., and D. Todd (2008) "Methodological challenges of evaluating the impact of the Global Environment Facility's biodiver-

sity program", *Evaluation and Program Planning* 31(3), 231–240.

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.

Van De Walle, D., and D. Cratty (2005) "Do Donors Get What They Paid For? Micro Evidence on the Fungibility of Development Project Aid", World Bank Policy Research Working Paper 3542, World Bank, Washington, D.C.

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*, Transaction Publishers, New Brunswick.

Webb, E.J., D.T. Campbell, R.D. Schwartz and L. Sechrest (2000) *Unobtrusive measures,* Sage Publications, Thousand Oaks, CA.

Weiss, C.H. (1998) *Evaluation—Methods for Studying Programs and Policies*, Prentice Hall, New Jersey.

Welsh, B., and D.P. Farrington (eds.) (2006) *Preventing crime: What Works for Children, Offenders, Victims and Places*, Springer, Berlin.

White, H. (2002) "Combining quantitative and qualitative approaches in poverty analysis", *World Development* 30(3), 511–522.

White, H. (2006) *Impact Evaluation Experience of the Independent Evaluation Group of the World Bank*, World Bank, Washington, D.C.

White, H. (2009) "Some Reflection on Current Debates in Impact Evaluation", Working Paper 1, International Initiative for Impact Evaluation, New Delhi.

White, H., and G. Dijkstra (2003) *Programme Aid and Development: Beyond Conditionality*, Routledge, London.

Whitmore. E (1991) "Evaluation and empowerment: it's the process that counts", *Empowerment and Family Support Networking Bulletin*, 2(2), 1–7.

Wholey, J.S. (1987) "Evaluability Assessment: Developing Program Theory", in: L. Bickman (ed.) *Using Program Theory in Evaluation, New Directions for Program Evaluation*, Jossey-Bass, San Francisco.

Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge.

World Bank (2003) *A User's Guide to Poverty and Social Impact Analysis*, Poverty Reduction Group and Social Development Department, World Bank, Washington, D.C.

Worrall, J. (2002) "What evidence in evidence-based medicine?" *Philosophy of Science*, 69, 316–330.

Worrall, J. (2007) "Why there's no cause to randomize", *The British Journal for the Philosophy of Science* 58(3), 451–488.

Worthen, B.R., and C.C. Schmitz (1997) "Conceptual Challenges Confronting Cluster Evaluation." *Evaluation* 3(3), 300–319.

Yang, H., J. Shen, H. Cao and C. Warfield (2004) "Multilevel Evaluation Alignment: An Explication of a Four-Step Model", *American Journal of Evaluation* 25(4), 493–507.

ISBN 978-1-60244-120-0