

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

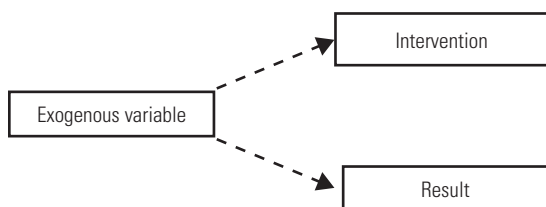
- i. The general premise is the deterministic principle: all macro events (or conditions, etc.) have a cause. This is only false at the micro-level, where the uncertainty principle applies, but the latter principle has essentially no detectable effect on the truth of macro determinism (though it is easy 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
S E L E C T I O N	O B S E R V E D	Propensity score	Regression analysis
	U N O B S E R V E D	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

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

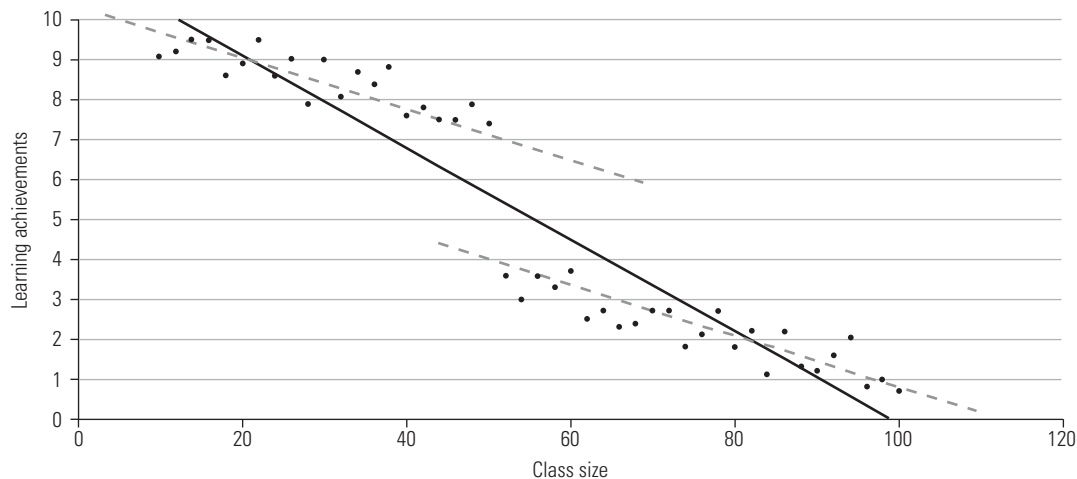
where e_i is a random error term and e_x is the effect of the unobserved variable. P and e_x are correlated and therefore P is *endogenous*. Ignoring this correlation results in a biased estimate of b . When the source of the selection bias (X) is known, inclusion of this variable (or these variables)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

$$\ln (p_i / (1-p_i)) = a + bX_i + cY_i + dZ_i + e_i,$$

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:

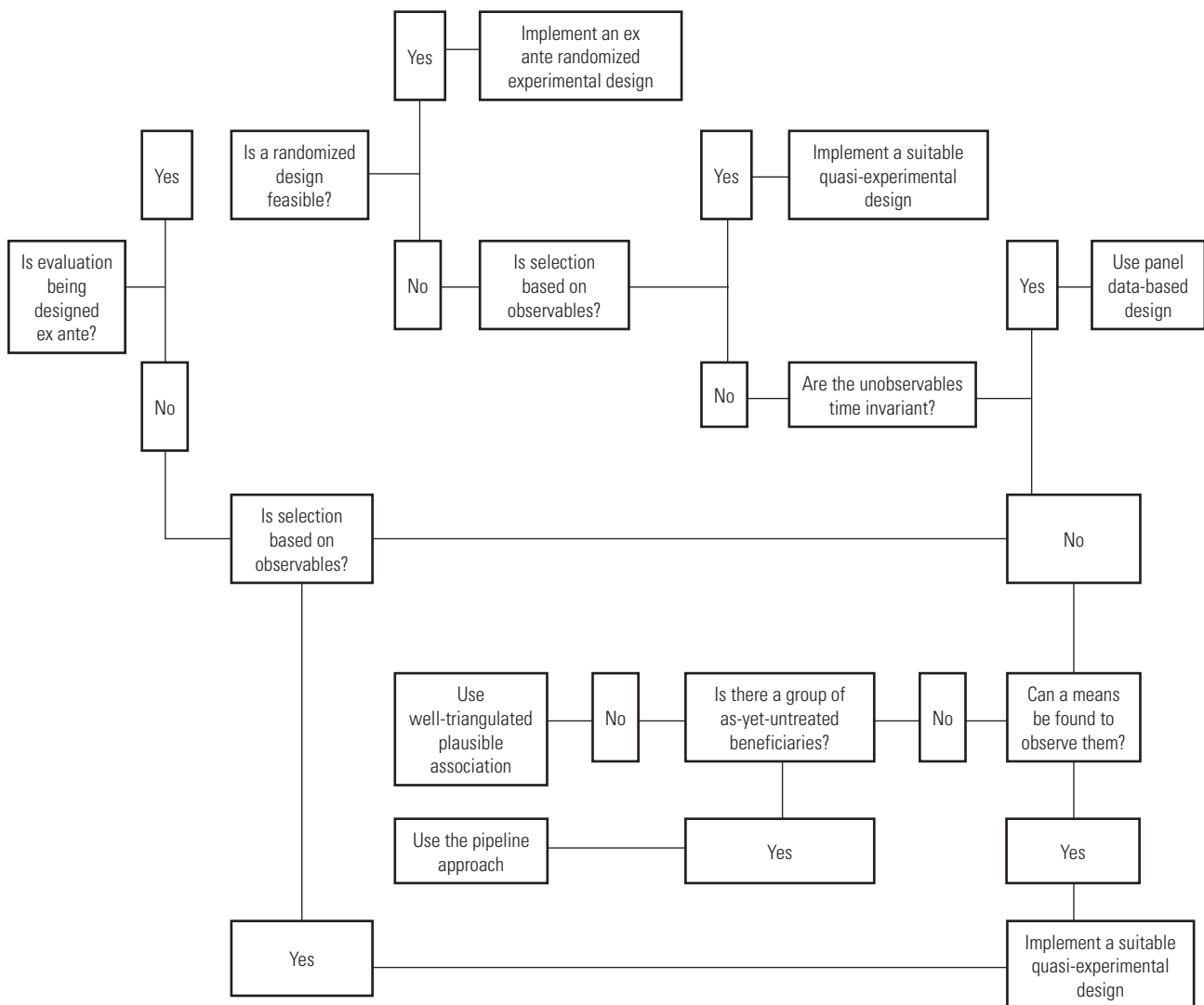
- Program participation: Among children participating in GMP, all service delivery indicators (age at enrolment, regular attendance of sessions, administration of vitamin A, and de-worming) show a substantial increase between 1982 and 1986, though subsequently they declined to around their initial levels. 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 contemporaneous 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.

APPENDIX 6: DECISION TREE FOR SELECTING QUANTITATIVE EVALUATION DESIGNS TO DEAL WITH SELECTION BIAS

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 METHODS¹

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.

APPENDIX 10: WHERE TO FIND REVIEWS AND SYNTHESIS STUDIES ON MECHANISMS UNDERLYING PROCESSES OF CHANGE

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 interna-

tional 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 Bangladesh¹

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 Morocco²

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 defini-

tion 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 sources³ 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 consid-

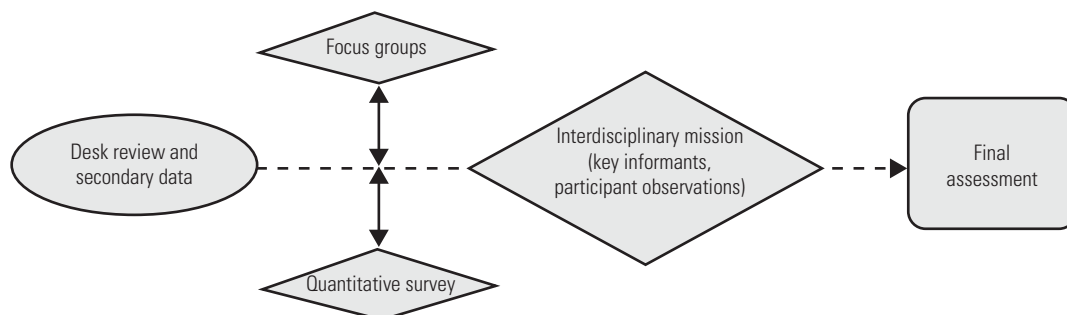
ered control groups, on the grounds that they would broadly satisfy the same eligibility criteria at entry as “older” project clients. However, no statistical technique (e.g., instrumental variables, Heckman’s procedure or propensity score) was adopted to test for sampling bias, due to limited time and 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 Guinea⁴

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 one⁵ 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 plantations⁶ 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 plantations⁷ 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 benefited 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 evaluation¹² 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 Mghinga 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 Mgabinga 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

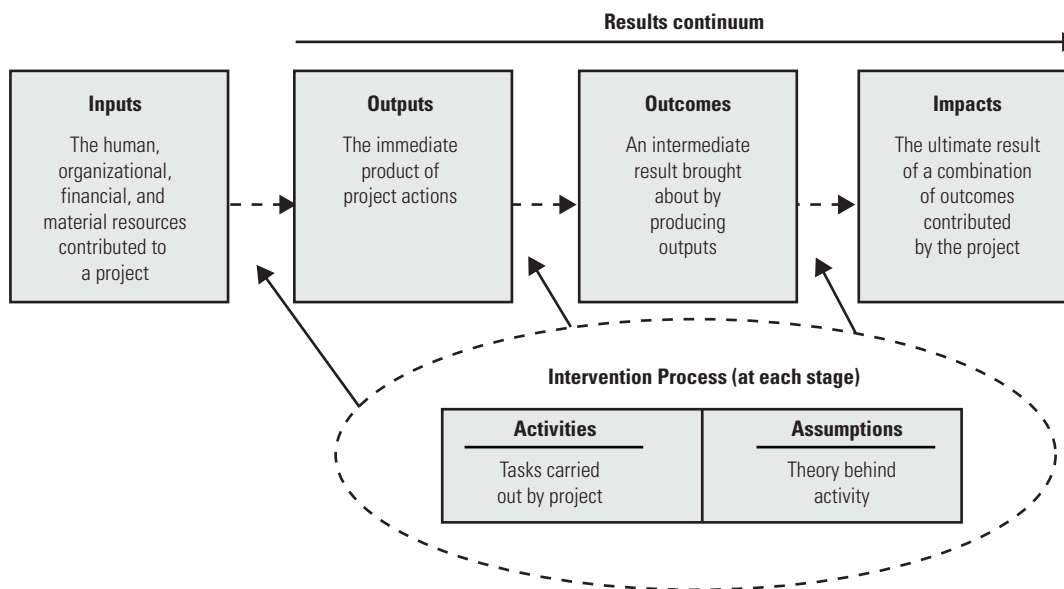
intermediate outcomes, which are then expected to lead to impact (see figure A11.2). The process of these interventions, in a given context, is determined by the contribution of a variety of actions at multiple *levels*, some of which are *outside* the purview of the intervention (e.g., actions of exterior 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

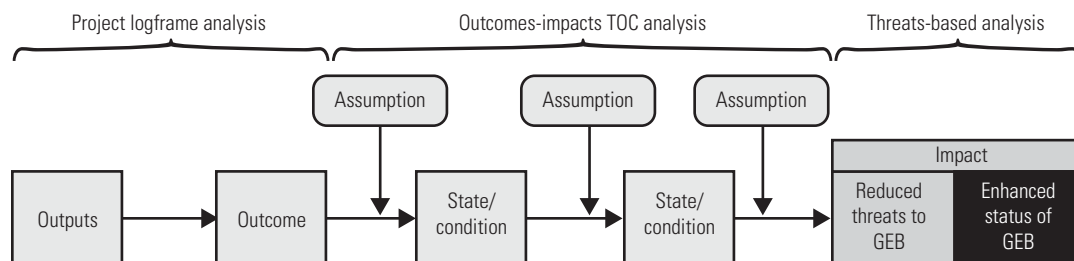
- *Assessing the level of contribution (i.e., impact):* To provide a direct measure of project impacts, a *targets-threats analysis (threats-based analysis)* is used to determine whether global environmental benefits have actually been produced and safeguarded.¹⁴ The robustness of global environment 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 non-linear, 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

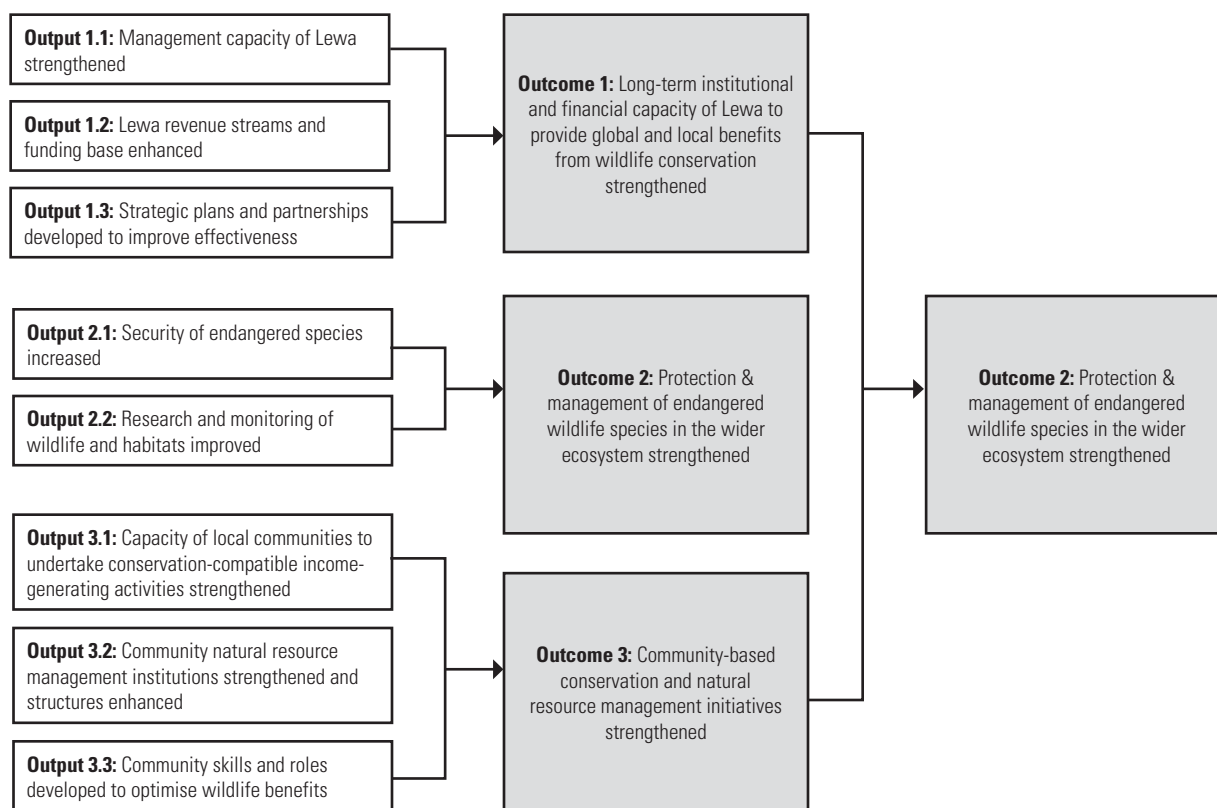


Table A11.1: Project Outcomes

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

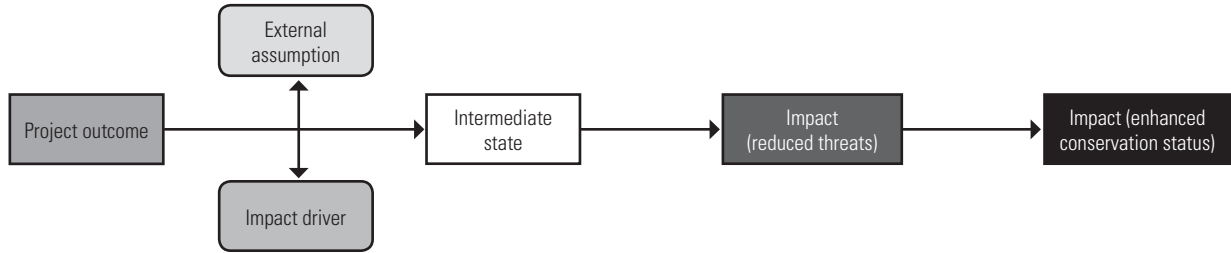
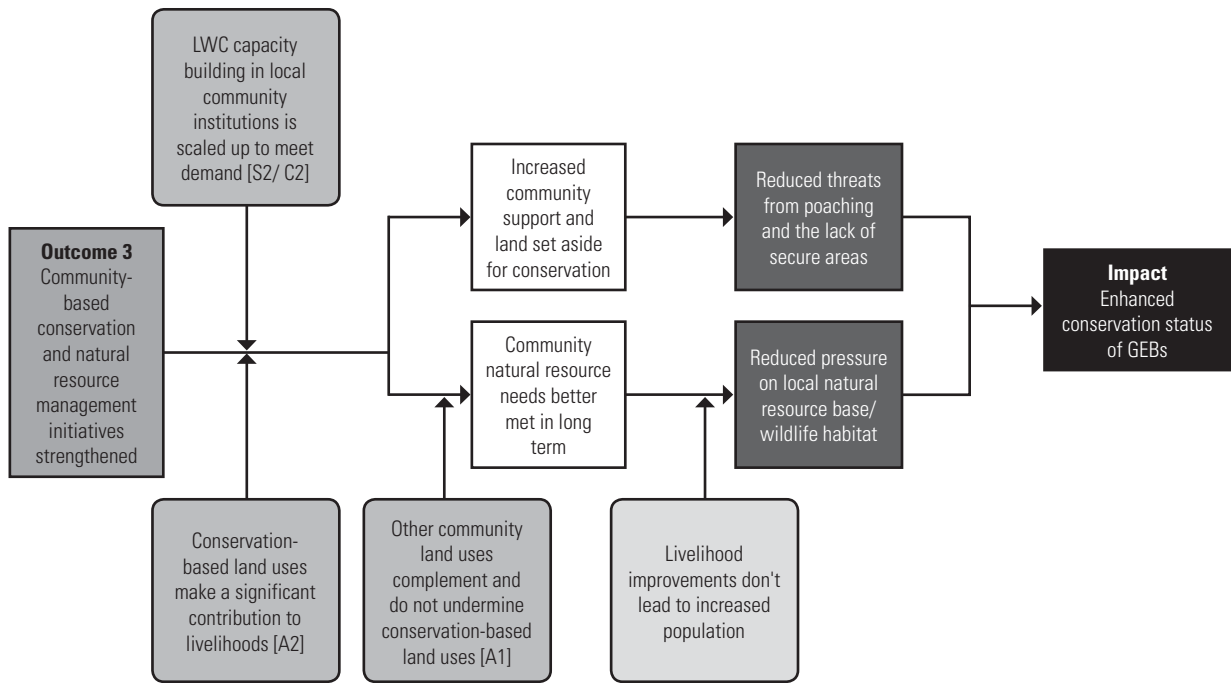


Figure A11.6: Model linking outcome to impact



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 both GSS 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; P₁, P₂, C₁, C₂ = 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-Hélène 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
- Oswaldo 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 interven-

tions. 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 direct 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).