Impact Evaluation of Development Assistance

Edition 2

A Practical Handbook for Designing Methodologically Sound Impact Evaluations under Budget, Time, and Data Constraints

FASID

Impact Evaluation of Development Assistance

Designing
Methodologically Sound Impact Evaluations
Under Budget, Time and Data Constraints

A Practical Handbook
Edition 2

by Michael Bamberger
& Nobuko Fujita

FASID
Foundation for Advanced Studies on International Development

Foundation for Advanced Studies on International Development
International Development Research Institute
1-16-17 Kudan-minami, Chiyoda-ku, Tokyo 102-0074, Japan
Telephone: 03-5226-0305
Fax: 03-5226-0023

Preface

This handbook is based on materials prepared for an ODA evaluators' training workshop organized in Tokyo in November 2002 by FASID (Foundation for Advanced Studies on International Development).

Simple logical framework-based evaluations are sometimes inadequate to assess the extent to which observed changes in target populations are attributable to program interventions and not to factors unrelated to the program under study. However, most ODA evaluations are conducted under budget and time constraints which make application of many rigorous and costly evaluation designs difficult. Consequently there is a growing demand for uncomplicated, rapid, economical designs which also can ensure acceptable standards of methodological rigor.

The "shoestring evaluation approach" (now known as "real-world evaluation") presented in this handbook draws on the experiences of ODA agencies, national governments and NGOs who have conducted evaluations in many Asian, African, and Latin American countries under budget, time, and data constraints. The approach has five steps for improving the quality of evaluations conducted with such limitations: (1) reduce data collection cost and time by reviewing and assessing different quasi-experimental designs applicable under these conditions; (2) reconstruct baseline data and control groups when information on conditions was unavailable before the project began; (3) compile a checklist of factors which can affect the validity of the evaluation's findings; (4) strengthen evaluation design and correct factors affecting validity; and (5) build evaluation data generation into the design of new programs. These

steps will assist ODA evaluators to produce persuasive and methodologically valid findings even when working under budget, time or data constraints.

After the workshop, requests for the course materials were received from people who were unable to attend. To meet these requests in a way that would facilitate independent study, the materials were edited, exercises involving case studies were added and a handbook was produced. In response to continued interest in the challenges of conducting evaluations under real-world constraints, this year we have updated the handbook, making some editorial changes with the assistance of Mrs. Suzanne Akiyama. We hope this will help evaluators in constructing simple, economical and robust evaluation designs.

For more detailed discussion, see Bamberger, Rugh, Marby (2006) *Real World Evaluation*, Sage. Also, a recent summary of the material discussed in this handbook has been compiled as a 35 pages color booklet, "*Conducting quality impact evaluations under budget, time and data constraints*"(2006), which can be downloaded free of charge from www.worldbank.org/ieg/ecd.

March 2008

Michael Bamberger
Nobuko Fujita

TABLE OF CONTENTS

Tables and Figures

Part I

Introduction to Impact Evaluation

Chapter 1

Overview: Opportunities and Challenges

1-1 Why impact evaluations?

In program theory-based evaluations, often applied to ODA projects, observed changes in performance or output indicators are compared with the pre-project situation ("before/after" comparison). This approach assumes that observed outcomes are results of the project, "more or less." Unfortunately, this assumption is almost never valid because almost all projects operate in dynamic environments where many social, economic, political, environmental and demographic changes are occurring, any of which might have influenced the observed outcomes. An assessment of impacts, changes actually attributable to a particular project, requires some "with/without" comparison.

Impact evaluation is a way of dealing with this issue. For example, in the evaluation of the impact on poor households of the EI Salvador housing improvement project discussed in Chapter 2, a 70.0 per cent increase was observed in the average household income of project families when the situation "before" the project began was compared with the situation "after" the project had been implemented and families had moved into their new houses. Using the theory-based model described above, the evaluator would assume that the project had had a significant impact on household income and might have reported that "evaluation of this project suggests that investment in low-cost housing is an effective way to increase the income of poor households."

However, the EI Salvador evaluation included a comparison with a control group of similar families who were interviewed at the same time as the project group, both before and after the project. The comparison revealed that over the

same time period, average household income for the control group had increased by 74.6 per cent, slightly more than for the project beneficiaries (although the difference between the two groups is not statistically significant). Using the control group comparison, the evaluation report might now draw exactly the opposite conclusion, that "evaluation of this project did not find any evidence that low-cost housing is an effective way to increase the income of poor households."

Impact evaluation permits the evaluator to assess whether the observed changes can be attributed to the project impact and the extent to which projects, programs, and policies have produced their intended impact and benefited the intended target populations.

"Impact" here refers to "net project impact," which means "total observed change" minus "change which should be attributed to other factors not related to the project" (see the following figure). Impact can be positive or negative and intended or unintended. Positive, intended net impact indicates that the program theory was valid and accurately explains how and to what extent the project contributed to the observed changes in the target population.

Identification of causal attribution, while one of the most difficult issues in evaluation, is necessary to be sure about the effectiveness of investment. Whether or not the result or outcome of an intervention was due to the project or to other separate factors has serious implications when deciding whether a project should be repeated elsewhere or implemented on a larger scale. An understanding of the true magnitude and direction of project impact could result in savings, more effective allocation of resources and avoidance of investments in projects which do not really produce significant benefits.

Impact evaluation is useful also for determining which components of the program theory worked the best. Suppose that a project with several different components (activities) was considered to have been very successful. By conducting impact evaluation, the relative contribution of each component to overall success can be assessed; in addition, an understanding can be gained of which elements of the theory proved to be correct in particular context. This information can be very useful when planning whether and how to replicate a

project. Impact evaluation can be useful also for assessing the validity of the program theory and for testing some of the critical assumptions and hypotheses on which it is based.

1-2 Challenges and constraints in designing impact evaluations

Unfortunately, ODA projects/programs, and even most evaluations of social and economic development programs in industrialized countries, face serious limitations in conducting rigorous social experiments to identify causality accurately. Evaluations are carried out with limited funds and time, and frequently without baseline data. Evaluators develop innovative ways to produce "acceptable" findings even while unable to follow all of the standard evaluation design principles. As a result, a series of "quasi-experimental designs"(QEDs) have been devised to provide the best possible estimates of project impacts under real-life situations which almost never approximate laboratory research conditions. Chapter 2 describes the strongest quasi-experimental design (Model 1) which development evaluators normally seek to follow. It also describes four methodologically weaker models which are used quite widely in situations where it is not possible to follow all of the conditions for Model 1.

Following are some of the limitations, constraints and challenges encountered in designing impact evaluation for development projects/programs:

a. *The evaluator is not consulted until the project is already being implemented.* The evaluator frequently is not called in until the project is already quite advanced. This creates numerous difficulties, including lack of baseline data (discussed below) and pressure to start the studies immediately without taking the time to fully understand the project and plan the evaluation. This sense of urgency also may create pressure to ignore some of the basic principles of good evaluation design. A much more favorable situation allows for initial consultations during the project design stage so that appropriate measures can be taken to collect and organize information for eventual evaluation.

b. *Limited baseline data*. Evaluation studies which do not start until late in the project cycle usually have little or no baseline data on the conditions of the beneficiary group prior to the beginning of the project. Even if records are available, they tend not to be organized in the form that evaluators require for before and after comparative analysis.
c. *No control group*. Even if some baseline data exists, most project managers collect data on the target groups or affected populations only. The possibility that the collection of information on non-beneficiaries might create in these people an expectation of financial or other compensation further discourages such data gathering. For these reasons, efforts to create a control (or comparison) group at mid-project or after project completion usually are hampered by insufficient data.
d. *Time Constraints*. Many evaluation studies must be conducted within a period of several months with only a few weeks of site visits. Nonetheless, the results of the analysis are expected to be presented in a timely-manner.
e. *Budgetary constraints*. Frequently funds for evaluation are not included in the project budget. As a result, funds to apply data collection instruments (tracer studies or sample surveys, for example) are often not available, making it difficult to use some of these methods to reconstruct baseline data or create control groups.

◆ ***Additional complications for sector and national level/policy evaluations***

While for the reasons given above it may be impossible to use the ideal impact evaluation design, a project evaluator still has a good idea of the evaluation design she or he would like to approximate. But as the scope of the evaluation broadens to sector or national level, it becomes more difficult even theoretically to define the appropriate design. Additional complications this might raise include the following:

f. *Defining the appropriate counterfactual*[1]. Sector and national interventions operate in a very complex environment where many economic, social, political and other changes are taking place simultaneously and many actors are involved. This increases the challenge to develop an appropriate

[1] A counterfactual is an estimate of what the situation would have been if the policy or program had not been implemented. Chapter 8 discusses some of the methods for defining counterfactuals.

counterfactual to define what would have been the situation in the absence of the particular program or policy intervention.

g. Data deficiency. Sector or country level evaluations frequently have to rely on secondary data collected for other purposes. Even when data is available, it may not include all of the needed information for the right time periods.

Chapter 3 describes ways to reduce the cost and time required to conduct impact evaluations. Chapter 4 deals with reconstructing baseline data and control groups.

1-3 Threats to Validity

Because of the constraints discussed above, there are a large number of factors which could lead the evaluator to make a wrong assessment about project impacts. In some cases it might be inferred incorrectly that the project has produced certain impacts; in other cases the evaluator might overlook some impacts which the project has actually produced. These factors are called "threats to validity." (See Chapter 5 for detailed discussion.)

Many of the approaches to reducing evaluation cost and time mean that principles of technically sound impact evaluation design cannot easily be applied. Less time and resources may be available for activities such as developing sampling frames, instrument development and testing, ensuring adequate sample size, and general issues of quality control. Often the demand for rapid and low cost delivery of evaluation findings takes precedence over methodological rigor.

An additional set of issues concerns the representativeness and generalizability of findings from qualitative/participatory studies. Frequently very little documentation is presented on how interviewees were selected, who actually attended a group discussion and the level of participation by people of different categories (for example, women or young adults) in the discussion. There may be a similar lack of documentation with respect to the selection of the communities or groups for whom case studies are prepared. These factors all can affect representativeness and generalizability. The fact is that methods do exist for addressing all of these issues in qualitative research (see, for example, Patton 2002) but the necessary controls are often not used in ODA

evaluations due to time and resource constraints.

Countering threats to validity requires particular care when assessing and interpreting the findings of rapid and low cost impact evaluations. The following chapters explain a framework which identifies and assesses four categories of threats to the validity of interpretations about program impact. Each category identifies factors which may yield misleading interpretations about the impact or effects of program interventions. Discussed also are actions which can be taken to address the different threats, thereby improving validity. The four categories of threats to validity are the following:

a. *Threats to statistical conclusion validity*. The statistical design and analysis may assume incorrectly that program interventions have contributed to the observed changes (impacts), or that some potential impacts may have been overlooked.
b. *Threats to internal validity*. Reasons inferred why the observed relationship between a project intervention and an output or impact is causal may be incorrect. Internal validity affirms the following questions "Does the evaluation design prove what it is supposed to prove about the effects of the treatment on the subjects actually studied?" "Does it prove that the treatment produced the claimed effect on the experimental subjects?" Common threats to internal validity include poor instrumentation (data collection methods), participant maturation (respondents learn over time how they should respond to the questions), spontaneous change or assignment bias (people in the project group are different from those in the control group).
c. *Threats to construct validity*. The underlying constructs (hypotheses/ concepts) on which the evaluation design (logic model) is based may not adequately describe the actual indicators of outputsÅCimpacts and settings used in the study. For example, observed changes in income may not adequately measure the construct "changes in household welfare," particularly in rural communities where there is a subsistence economy and most farmers use very little money.
d. *Threats to external validity*. Reasons inferred how study findings might be generalized to other settings (regions, social or economic groups, etc.) may not be valid.

1-4 Realistic approaches and useful ways for impact evaluation

Despite the problems and challenges described above, the increasing interest in rapid and low cost impact evaluation has produced a number of promising approaches, some of which will be discussed in the following chapters.

a. *Refinements to quasi-experimental designs*. Recognition of the need to adapt QEDs to the real-life circumstances under which evaluations are conducted has yielded many ways to simplify and streamline evaluation designs. These are discussed in the following chapters.

b. *Participatory methods*. A wide range of participatory methods have been developed to obtain the perspectives of project beneficiaries and other affected groups. Many of these methods can be used also to reduce the time and cost of data collection. For example, groups rather than individual households can be asked to estimate the impact of a project; i.e., a women's group can be asked to estimate the average time household members spend daily collecting water, or groups can assess improvements in the management of community services such as water supply or health clinics. Applicable participatory techniques include PRA (participatory rural appraisal), described in publications such as Voices of the Poor (Narayan and Petesch 2002), and The Poverty Reduction Sourcebook (World Bank 2002)[2]. Use of focus groups also can be considered a participatory method, but an experienced facilitator, using explicitly participatory methods, is required to ensure that discussions are not directed or dominated by a few people[3].

c. *Statistical methods to improve the use of cross-sectional studies*. Multivariate analysis is frequently used to statistically adjust for differences between project and control groups; hence improving the utility of cross-sectional studies is a tool for impact assessment. Statistical methods are used primarily to improve the analytical strength of sample survey designs, but these methods can be used also to improve the way in which participants in qualitative studies are selected and the results of the studies

---

[2] See also Chambers, Robert 1995. *Poverty and Livelihoods: Whose Reality Counts?*

[3] It should also be mentioned that focus groups are frequently used in market research in a non-participatory way with participants responding to a set of questions prepared by the researcher.

are analyzed[4]. Using various designs and approaches, rapid and low cost impact evaluations have become possible not only at the project level (Chapter 6), but also at the sector and program level (Chapter 7), and at the country and policy level (Chapter 8). It is now understood that the most efficient way to run impact evaluation is to build evaluation into project, program, and policy design, and Chapter 9 discusses some practical measures for ensuring that the collection of data needed for possible future impact evaluation is built into the project operating systems -- even in cases where the possibility of conducting a future impact evaluation has not yet been discussed with project management.

[4] A counterfactual is an estimate of what the situation would have been if the policy or program had not been implemented. Chapter 8 discusses some of the methods for defining counterfactuals.

Chapter 2

Introduction to the Theory and Practice of Impact Evaluation Design

2-1 True experimental design and quasi-experimental design

◆ ***True experimental design***

True experimental design is used in such fields as medicine, animal behavior and educational research studies conducted under carefully controlled laboratory conditions. In the simplest design, subjects are randomly assigned to the Experimental [E] Group, which will receive the treatment [X] (for example, a new drug, or rewards/punishments used in animal research or school programs), and the Control Group [C] which does not receive the treatment. A test is applied to both groups in Time Period 1 [T_1] before the experiment begins to measure the behavior, physiological reaction or other variables the treatment is intended to influence. The measurements are repeated in T_2 following the application of the experimental treatment. The measurements in T_1 and T_2 are defined as E_1 and E_2 for the experimental group and C_1 and C_2 for the control group. The research design is described below:

The simplest true experimental design

	T_1	Experimental Treatment	T_2
Experimental group	E_1	X	E_2
Randomized control group	C_1		C_2

Assuming that the assignment of subjects to the two groups was truly randomized and that the experiment was conducted under carefully controlled laboratory conditions, the impact [I] produced by the experimental treatment [X] can be estimated by comparing the measured change in the treatment group with that in the control group[5].

Normally a number of different indicators will be used to assess the effects (impacts) of the treatment. In assessing the effects of a drug on learning ability in first grade students, the indicators might include, for example, the number of new words learned, the number remembered after one hour, after one day, after one week etc. These different impact indicators are normally referred to as I_1, I_2, ... I_n. If the value of impact indicator I_1 differs significantly from zero (either positively or negatively) then there is some preliminary evidence that the treatment did have an impact. However, experiments have to be repeated many times under different conditions and usually with different groups before it is possible to speak with confidence of the efficacy of the treatment.

◆ ***Quasi-experimental designs***

When evaluating the impact of development projects (water supply, road construction, micro-credit, teacher training, provision of teaching materials, etc.), it is almost impossible to approximate the true experimental design level of experimental control. For example, it is rarely possible to randomly assign subjects to treatment groups and control groups, and treatments cannot be applied in such a precise way. Consequently a series of quasi-experimental designs (QEDs) have been developed to approximate as closely as possible the true experimental design, for the following purposes:

- To make the best possible estimate of the extent to which a project, program or policy has produced its intended impact.
- To identify the factors which positively or negatively influence the

[5] If the observations E_1, E_2 and C_1 and C_2 refer to the mean scores for each group (for example, average income, average educational test score or average anthropometric score), then a statistical test such as the t-Test for the difference of means is used to determine whether the observed difference is statistically significant. If, on the other hand, the values refer to proportions (for example the proportion of children attending school or the proportion correctly answering a test question), then the appropriate statistical test would be a measure such as the Z score for difference of proportions (Herbert Blalock Social Statistics Chapter 13. McGraw Hill 1972)

magnitude and direction of the impact.

In the real world, evaluators using QEDs typically face the following problems:

- It is almost never possible to randomly assign subjects to experimental groups and control groups. For logistical reasons, most projects are accessible to or affect everyone in a given community or area. For example, a school or water supply system will be accessible to all families and it is clearly not possible to tell some families they cannot use the water or send their children to the school.
- Some projects use a self selection process, whereby, for example, people decide if they wish to apply for micro-credits, enroll in a literacy class, or plant new varieties of seed. In these cases it is likely that the people who do decide to participate will be different in important ways from those who do not participate.
- Typically people who take the initiative to participate are economically better off, better educated, and have more self-confidence. Consequently, it is difficult to know whether observed changes in income, reading skills, health, etc. are due to the effects of the project or to the differences in the initial conditions of participants and non-participants.
- It is very difficult to find a control group which closely matches the experimental group on the key indicators. Project communities are often selected because of special characteristics. In some cases project planners choose the poorest communities, in other cases they choose communities which have the greatest likelihood of success. In either case it will be difficult to find a control group which closely matches the project population.
- In many cases, for political or ethical reason it is difficult to use any kind of control group at all. Frequently politicians and community leaders in control group areas will exert pressure for their community to be included in the project. It is often considered ethically inappropriate to ask families to spend time responding to surveys if they will not receive any benefit. The fact that families are being interviewed sometimes creates false expectations that they will be eligible to participate in a later phase of the project.

- It is also difficult to ensure that treatments (services) are administered in exactly the same way to all project sites and families. Sometimes the delivery of materials and equipment is delayed, in other cases there are major differences with respect to the organization of the project and delivery of services in different sites. In one micro-credit program the local administrator may speak the local language and may create a welcoming atmosphere, encouraging families to visit the project to discuss loans. In another site the administrator may not speak the local language, the project may be seen as a hostile to the community and fewer people visit the center. For these reasons it is difficult to determine whether differences in project performance are due to differences in the responsiveness of different communities, or whether the differences are due to the way the project was differently administered.
- Finally, each project operates within a unique economic and political context. Each must interact with a number of government or non-government organizations which have their own particular characteristics. Also, the social, economic and cultural characteristics of target population may vary significantly among project sites. All of these contextual factors can have an important influence on the project outcome. As a result of these contextual factors, even when a project is administered in exactly the same way at each site, there may be significant differences in the outcomes.

Several lessons can be drawn from these evaluation design difficulties. First, it is important to understand the problems facing a particular study and to select the methodologically strongest design possible under the particular circumstances. Second, the strengths and weaknesses of the evaluation design should be carefully analyzed and the implications for the interpretation of findings and recommendations assessed. In some cases the methodological weaknesses may not seriously affect the kinds of recommendations to be prepared, whereas in other cases they may be very serious. For example:

- The lack of a control group may not be very important if the purpose of the evaluation is to assess whether indigenous communities participating in pilot projects are able to manage and sustain community water supply

projects; or whether women will apply for small loans if a loan office staffed by local language speakers is established in the community.

- On the other hand, if the purpose of the evaluation is to estimate whether a pilot project could be replicated on a national scale; or whether it would offer a more cost-effective way to deliver a particular service, then the lack of a control group might be a serious problem.

Finally, if methodological problems are identified which seriously affect the purposes of the evaluation, then the evaluator should consider measures to rectify the problems (see later chapters for discussions of these measures). One example of a serious problem is the common situation in which no baseline study was conducted at the time the project was planned so there is no reliable quantitative information available on, for example, school enrolment, distances traveled in vehicles or on foot, or water consumption before the project began. Consequently, even if precise information on enrolment, travel or water consumption is collected after the project is implemented, it is difficult to assess the magnitude of changes which have occurred.

Under these circumstances, some of the possible tools which could be used to estimate the baseline conditions are as follows:

- focus groups in which community residents are asked to estimate the impact of the project
- a rapid sample survey in which families are asked to recall which children went to school, how much water was consumed, etc., before the project began.
- key informants such as community leaders, local health authorities, school teachers etc. could also be asked to assess the impact of the project

An important aspect of this approach is the use of triangulation (consistency checks) to compare information obtained from different sources. If the information from all of the sources is more or less consistent, then the evaluator can have confidence in the findings. If, on the other hand, the information from different sources is inconsistent or even contradictory, then further analysis is required to determine if the inconsistencies can be reconciled.

2-2 The evaluation framework for project impact evaluation

Figure 1 identifies seven stages in the project cycle which can be considered in impact evaluation design. Readers familiar with logical frameworks will see many similarities to the LogFrame format and this evaluation model can be coordinated with the LogFrame if that format is already being used. The evaluator will often find that there is no written documentation defining the project model and she or he will have to work with planners and project managers to agree on a definition of project objects and critical assumptions. The seven stages are as follows:

1. ***Project planning and design :*** This examines the following

- the project's approach to planning (for example, central planning or participatory consultations)
- the information sources on which the project is based and their adequacy . For example, how adequately do the surveys (and other data sources) cover all sectors of the target population and how well do they provide the information required for planning this project.
- whether the surveys (and other data) provide information on the different needs and constraints of adults and children, men and women, different ethnic groups and people engaged in different kinds of economic activities
- whether a systematic stakeholder analysis was conducted to ensure that all sectors of the target group were consulted
- the critical assumptions on which the project design was based. For example, for a micro-credit program intended to benefit both male and female farmers some of the critical assumptions might include the following:
  - Lack of credit is a major constraint to women's economic activities.
  - If women receive credit they will be able to start up or expand economic activities.
  - If women start economic activities they will be able to control how the profits are used.
  - Increased income in the hands of women will improve their economic and social welfare, giving them more influence in household and

community decision-making.

This information can be used in at least two ways in the evaluation design: first, to assess how well the project was planned and the quality of the information used; second, to develop indicators to monitor the validity of the key assumptions as the project evolves. For example, if women's social and economic welfare did not improve, was this due to the following situations:

- Even when women received loans they did not start up businesses.
- If they did start up businesses, a male household member often controlled use of the profits.
- Even if women controlled the profits, this might not affect conventional indicators of economic and social welfare (income, consumption, expenditures, etc.). For example, a number of studies in India have shown that women often save all of the profits from their business to provide a dowry for their daughters.

2. ***Project inputs:*** This identifies the materials, money, staff, equipment, extension workers, consultants, capacity building and other resources identified in the project plan. The use of these inputs should be monitored because one of the main reasons many projects do not achieve their intended impacts is that a high proportion of the resources never reach the schools, clinics or other service centers through which project is implemented on ground[6].

3. ***The project implementation process:*** Projects can be implemented in different ways. Some involve the community in planning and administration and others are managed directly by the implementing agency (Ministry of Transport, Agricultural Investment Bank, etc.). Projects also vary in terms of the ease of community access to the services. For example, if a micro-credit program is administered by the local branch of a large agricultural development bank, it may be difficult for poor families without transport, or women with small children, to reach the bank. Consequently this program may

[6] A number of recent public expenditure tracking studies conducted by the World Bank in countries such as Uganda, Tanzania and Ghana have found that in some cases as little as 10% of the project resources reached the frontline services. There was some corruption, but often most of the education funds transferred from the Ministry of Finance to the General Fund of a local government agency were simply diverted to other purposes.

not reach the poorest farmers.

4. *Project outputs or products:* Projects are intended to achieve a set of quantifiable outputs or products: i.e., the number of children attending school or continuing from primary on to secondary school, the number of families with access to good drinking water, the number of micro-credit loans approved and the number of small businesses started, the kilometers of roads or footpaths constructed or maintained. There may also be outputs which are assessed qualitatively, such as the quality of leadership training or the strength of community groups created.

5. *Outcomes or short-term impacts:* These are the impacts which are achieved within a relatively short period of time, perhaps 6 to 12 months after project completion. Poverty reduction programs frequently identify four types of impacts[7]:

- Opportunity: Access to economic resources and improved economic conditions
- Capability: Access to public services (health, education etc.) and the affect on human development indicators such as anthropometric measures, years of schooling, frequency of use of public transport
- Security: Economic, environmental and personal security
- Empowerment and voice: Participation in decisions affecting the social, economic and political life at the household, community and local government level. This may also include access to information and control of the means of communication.

6. *Medium and long-term impacts*: These are assessed on the same four dimensions as the short term impacts, but given the longer time frame, broader assessments are possible. For example, access to education can also include access to labor markets after school completion.

7. *Sustainability*: The overall objective of a project is not simply to produce impacts during the life of the project, but to ensure that the impacts are

[7] These categories are based on the 2000/2001 World Development Report: Attacking Poverty; and the World Bank Poverty Reduction Strategy Sourcebook (see especially the chapter on gender).

sustained over time. For example:

- Schools and clinics continue to function after donor funding has ended
- Communities are able to maintain minor irrigation works, rural roads and bridges, and the bus company is able to maintain its fleet

◆ ***Contextual factors affecting project outcomes***

An important feature of this evaluation framework is its recognition that each project is implemented in a particular economic, political, institutional and socio-cultural context. Consequently, due to contextual factors even projects always implemented according to the same design may vary significantly in outcomes, impacts and sustainability from one project site to another. The model identifies three sets of contextual factors to be taken into account in the evaluation:

- *Economic and political factors*: A job training program is likely to have different outcomes in areas where the economy is growing than in areas with high unemployment and economic decline. Similarly, families may be less inclined to invest in their children's education if the labor market is very tight. The local political context is also relevant. A project in a region where the local government is in the hands of an opposition party may find it more difficult to obtain support from national authorities, or an impending local or national election might affect project dynamics. There have been cases, for example, where a local political candidate told farmers not to apply for small business loans or pay service charges for water because if he were elected all of these services would become free to the poor.
- *Institutional and organizational factors*: Projects depend to a considerable extent on the efficiency and support of local government agencies. For example, if the local office of the Ministry of Health, Education or Transport is poorly managed, or if it has an acting director and is short of staff, this will likely affect the efficiency and impact of project implementation. Non-government agencies also play an important role in project implementation and an assessment of the efficiency of their operation also may be needed. Finally, among government agencies or between government agencies and NGOs there may be conflicts which affect project implementation.
- *Social, economic and cultural characteristics of participating*

communities: Often important differences in the social and cultural characteristics of participating communities can influence project implementation and impact. This factor is particularly important where there are ethnic differences but the analysis should be conducted for all projects.

Contextual analysis can be used at any stage of the project cycle. For the purposes of impact evaluation, it is particularly useful for explaining differences in project impacts at different sites which are not explained by how well or badly the project was implemented. Contextual analysis is usually based on qualitative methods: participant observation, meetings with community leaders, focus groups and interviews with key informants (journalists, academics, NGOs, religious organizations, local government agencies etc.). Secondary sources such as newspapers and university studies can also be useful.

2-3 The most commonly used quasi-experimental designs

All of the following models can be strengthened if used in combination with the evaluation framework described in section 2-2 and by using some of the shoestring evaluation methods of later chapters. For the models which include a control group (Models 1, 2 and 4) multivariate analysis may also be used to statistically control for differences in characteristics. See Table l for a summary of the strengths and weaknesses of each model.

Model 1: The strongest general purpose quasi-experimental design

There are a number of changes in the symbols used to describe the QED models reflecting differences from the true experimental design[8]. One of the

[8] The differences are: 1) [E] (experimental group) is replaced by [P] (project group) reflecting the fact that this is an impact evaluation rather than an experiment; 2) [X] represents the project intervention' rather than the experimental treatment; and 3) in most cases "non-equivalent control group" is used rather than "randomized control group" to reflect the fact that for most QEDs it is not possible to randomly assign subjects to control and project groups. Italics are used to reflect the non-randomization of the control group.

differences is that the post-project observation for the QEDs is defined as T_3 rather than T_2, as in the case of the true experimental design. This is because during the project implementation process some QEDs include an observation which is defined as T_2. The evaluation of the Eritrean Feeder Roads Project described in Annex 6 is an example of a longitudinal design which includes several observations during the period of project implementation.

For most purposes Model 1 is the strongest and preferred QED. This model is described as follows:

	T_1	Project Intervention	T_3
Project group	P_1	X	P_3
Non-equivalent control group	C_1		C_3

In this model a non-equivalent control group [C] is selected at the start of the project to approximate as closely as possible the project beneficiary group [P]. The term "non-equivalent" in italics reminds that it is rarely possible to assign subjects randomly to the project and control groups; thus, between the characteristics of the project and the control groups there may be differences which could distort the interpretation of the findings. The project and control groups are both interviewed in time period 1 [T1] before the project begins, and information is obtained on a set of indicators [I1, I2 . . . In] measuring the changes (impacts) the project is intended to produce (for example, increase in household income, reduced daily travel time, number of children attending school). Information is also collected on the social and economic characteristics of the individuals or families [x1, x2, . . xn], called intervening variables, which might affect project outcomes . Data collection is repeated in time period 3 [T3] after the project has been in operation long enough to have produced its intended impacts. Ideally the analysis will include the contextual factors discussed in the previous section.

The analysis can be considerably strengthened if multiple regression analysis is used to control statistically for differences in the social and

economic characteristics of the project and control groups[9]. Multiple regression analysis matches subjects statistically on such characteristics as age, income and education to ensure that observed differences in the impact indicator are not due to differences between the project and control groups on these intervening variables. The analysis determines whether after controlling for these household characteristics, differences remain between the two groups with respect to the impact indicator (income, years of schooling, water consumption etc.). The analysis does not guarantee that the differences are due necessarily to the project, but the more that other factors are eliminated as possible explanations, the more likely it is that the project contributed to the observed changes.

Many refinements can be introduced into the basic QED design to assess multiple treatments or to capture impacts which evolve gradually over time (Shadish, Cook and Campbell 2002; Valadez and Bamberger 1994).

Example of Model 1:
Evaluating the impact of improved housing on households in El Salvador

A four year evaluation was conducted in 1976-80 in EI Salvador to assess the impact of improved housing on poor households in San Salvador, the capital. In 1976 a randomly selected sample of households was interviewed shortly before they entered a self-help housing construction project. A control group was selected by combining samples of randomly selected families from the three types of inner-city housing from which the project participants had been selected. The samples were repeated in 1980.

It was found that between T_1 and T_3 the average household income for project participants had increased by 70.0 per cent compared to an increase of 74.6 per cent for the control group. This means that there was no evidence that improved housing had a positive impact on income; in fact, the income of the control group rose slightly faster. This illustrates the importance of a carefully selected control group. If only project participants had been studied one might have

[9] The simplest form of a multiple regression analysis of project impact is specified as follows:
$I_1 = f [D_1, x_1, x_2, \ldots x_n]$ where: I_1 = impact coefficient; D_1 = dummy variable' for project and control groups; and $x_1, x_2, \ldots x_n$ = attributes of the individual (age, sex, education), the household (income, family size, type of housing etc.), or the community which might affect the estimated impact. The analysis is testing whether there is a statistically significant difference in the impact coefficient for the project and control groups after adjusting for individual, household and community characteristics.

concluded that "improved housing has a significant impact on household income because the income of participants in the low-cost housing project increased by70 per cent in four years.'' (Source: Valadez and Bamberger 1994. pp. 237-8)

◆ ***Cheaper and faster -- but usually weaker-- QED***

There are many situations in which it is not possible to use Model 1. In some cases time and budget constraints do not permit the use of a control group. In other cases the evaluator is not called in until the project is being implemented so it is not possible to go back in time and collect baseline data. A number of simpler and more economical QEDs can be used in these situations. However, each successive model sacrifices one or more essential elements of a sound evaluation design, becoming vulnerable to a wider range of methodological problems.

Model 2: No pre-test control group

In Model 2 a baseline survey is conducted with the project's intended beneficiaries before the project begins, but no control group is used at this stage. A control group is selected once the project is operational and an ex-post survey is conducted in time 3 [T_3]with both project and control groups. The model is represented as follows:

	T_1	Project Intervention	T_3
Project group	P_1	X	P_3
Non-equivalent control group			C_3

This design works reasonably well for assessing how a project is being implemented and whether it is able to produce the intended outputs. It also allows comparison of the project and control group characteristics. For example, with a rural road construction project, surveys and participatory consultations with the community may have identified a number of factors affecting the willingness of the community to participate in the project and the benefits they obtain from it. These factors might include the following: whether local culture permits women to participate in road construction and to

travel to market, the community's distance from the local market and the agricultural surplus available to sell. A control group, if well selected, could rate other local communities on these variables and hence determine the likelihood that the project would be well received and might have an impact in other areas. The project and control groups could also be compared on indicators such as amount of produce sold in the local markets, average number of trips and distance traveled and kinds of consumer goods available in community shops.

However, this design has some weaknesses. Most importantly the lack of control group baseline data means that it is not possible to determine whether observed differences between the project and control groups in T_3 are due to the project or were pre-existing before the project began. Another weakness is that we cannot control for local history which might have affected outcomes. This is particularly significant for projects seeking to increase agricultural output or sales. Sales of maize or wheat may have increased because of good rains and not because of the project. The *ex-post* control group can provide some information on this but the analysis will obviously be much stronger if changes in the project and control areas can be compared over time.

Example of Model 2:
Comparing the effects of resettlement on project beneficiaries and non-beneficiaries in the second Maharashtra Irrigation Project, India

Sample surveys were conducted periodically between 1978-85 in areas from which families were to be resettled as a consequence of a large-scale irrigation project. The study only covered families who were eligible to receive land or housing plots in the relocation areas. The surveys were repeated in 1990 after the relocations had taken place. An *ex-post* control group survey was conducted in 1990 with a sample of families who had remained in the irrigation project command. This was not an ideal control as many of the sample households received families who had been forced to move as a result of the dam, so their situation did not really represent families not affected by the project. Recognizing that no information was available on the approximately 45 per cent of families who were forced to relocate but who were not eligible for compensation, a tracer study' was conducted in 1990 to try to identify them. The study found that the economic conditions of most

families receiving compensation had improved. The situation concerning the families who had not received compensation was more mixed, but in general, forced resettlement appeared to have had less negative consequences than had been expected. (Source: Valadez and Bamberger. 1994. pp. 264-266)

Model 3: No control group

In this model there is no control group and the analysis is based on a comparison of the project group before and after implementation. The model is described as follows:

	T_1	Project Intervention	T_3
Project group	P_1	X	P_3
Non-equivalent control group			

This model works reasonably well for projects having large and clearly defined impacts; for example, the construction of a village school or clinic where there had previously been no such facility within easy access. It can also work well when the purpose of the evaluation is to understand the project implementation process and where quantitative assessment of impacts is less important.

This model does not work well, however, when precise estimates of the magnitude of project impacts are required. It also does not control for the influence of local history. The lack of comparative data on the project group and other communities also means that it is difficult to assess the potential for replicating the project on a larger scale. For example, if the project was successful because communities with higher than average levels of education and income had been selected, the lack of control group would make it difficult to assess how successful a larger project would be if extended to more typical communities with lower education and income.

Example of Model 3:
Evaluating the impact of the feeder roads component of the Eritrea Social Fund

In 1999 local consultants were commissioned to conduct a rapid evaluation of each of the eight components of the Eritrea Social Fund. Due to budget and time constraints, it was decided that for six of the eight project components only an *ex-post* survey of beneficiary communities would be possible. However, for the feeder roads component (as well as one other component) it was decided to conduct a longitudinal impact study in which surveys and other forms of measurement (direct observation, key informants, counts of road traffic, etc.) would be taken at the start of the project, and again after 6 and 9 months. No control group was used. The study found that the construction of the road had a major impact on the production and marketing of agricultural produce, mobility on foot and by vehicle, and use of schools and health facilities. Secondary data (such as market records and vehicle registration records) were used to compare changes in surrounding regions (reconstruction of a control group) but it was difficult to get precise estimates due to the limited availability of records and the considerable fluctuations in market prices. (Source: Annex 6 case study)

Model 4: No baseline data

This model relies entirely on data collected after the project has been implemented and no baseline data is collected on either project or control areas. The model is represented as follows:

	T_1	Project Intervention	T_3
Project group	X		P_3
Non-equivalent control group			C_3

This model can be used to obtain an approximate estimate of project impacts. It works best in isolated communities where the project is the only major outside intervention. Under these circumstances there is no need to

isolate the effects of other interventions which might be taking place at the same time. It can also be used to compare the characteristics of project participants with people from other similar communities. If project households have characteristics similar to other communities, then it is more likely that the results of the pilot project can be generalized. If, on the other hand, there are significant differences between the groups it will be more difficult to generalize.

This model does not control for historical events which may have affected outcomes, and it has the same weaknesses as the earlier models which do not collect baseline data. It is also not possible to evaluate the project implementation process.

Example of Model 4:
Assessing the impact of micro-credit on the social and economic conditions of women and families in Bangladesh

In 1991-92 a random sample of households were interviewed from a sample of rural Bangladeshi villages in which village banks were operating. A control group was interviewed in villages where no village bank programs were operating. The surveys were conducted *ex-post*, when the village banks had been operating for several years, and no baseline information was collected on the condition of the families prior to the banks' operation. It was found that borrowing from a village bank had much greater impact on women than on men (although the latter also benefited). Per capita household expenditures increased almost twice as fast for women, housing conditions improved and personal savings increased. Interestingly, it was found that contraceptive usage declined for women borrowers and their fertility increased. The lack of baseline data made it difficult to determine the extent to which the observed differences between the project and control groups were due to the effects of the project; or, at least in part, to differences existing before the project began. (Sources: Khandker 1998; Pitt and Khandker 1998; Baker 2000 (Annex 1. 2); and World Bank. 2001).

Model 5: Eliminating baseline data and control groups

This is the weakest QED. Only the project population is studied and only

surveyed after the project has been implemented. The model is represented as follows:

	T_1	Project Intervention	T_3
Project group	X		P_3
Non-equivalent control group			

This model works reasonably well for exploratory studies, when the purpose is to get a general idea of whether the model works. It can also be used to get a very approximate estimate of impacts. This model works better for relatively isolated projects where the potential impact is expected to be quite large. Among its limitations are the following: It cannot be used to obtain reasonably precise estimates of impact. It cannot control for local history events which might affect outcomes. And, since it does not provide any comparative data on the characteristics of the project population, it cannot be generalized to a wider population.

Example of Model 5:
Assessing the education impacts of the Eritrea Social Fund.

In the evaluation of the Eritrea Social Fund referred to earlier, an *ex-post* survey was conducted in 48 communities representing the catchment areas for 10 newly constructed primary schools. No control group was used but baseline data on school attendance prior to the construction of the schools was estimated by asking families to recall the situation before the schools were built. Recall data seems to have been very reliable because it was easy for families to recall whether their children attended school before the village school was built and because they had no incentive to give wrong information. The analysis focused on the following topics:

- *Process evaluation*: More than 90% confirmed that the school was a high priority but only 37% had attended meetings to participate in planning the project.

- *Accessibility impact and gender*: The schools were successful in reaching the poorest sectors of the community but it was more difficult to involve recently returned refugees as they were still unsettled and not motivated to send their children to school. Families are equally motivated to send boys and girls to school, but if they were to choose for economic reasons, they would normally give priority to the boy.
- *Social impact*: School construction reduced travel time for students by one half to two thirds.
- *Sustainability*: Despite extreme poverty almost all households contributed the required 10% of the cost of the school in cash, labor or materials.

Part II (Chapters 3-5) will present the main elements of the shoestring evaluation approach, which is designed specifically for use in situations where evaluators are working under budget and time constraints and where they frequently do not have access to baseline data on the conditions of the project population before the project began. The approach can also be used more generally, to ensure that the highest possible standards of methodological rigor are employed in all evaluations conducted under budget, time and data constraints. The approach comprises the following main elements:

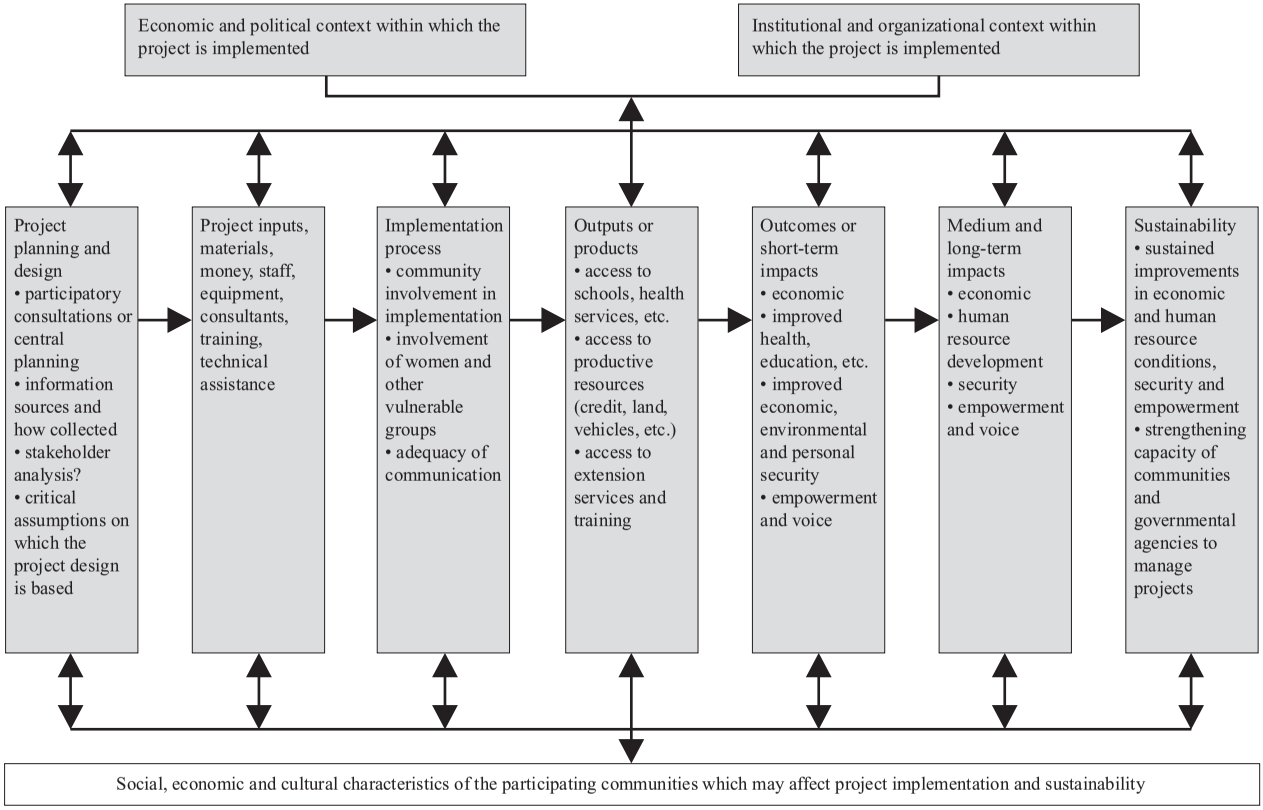
- Reducing costs and time required to conduct impact evaluations (Chapter 3)
- Reconstructing baseline data and control groups when these were not included in the original evaluation design (Chapter 4)
- Identifying factors which may affect the interpretation of whether and how the project contributed to producing the intended impacts, i.e., threats to validity (Chapter 5)
- Addressing potential problems affecting the validity of evaluation conclusions once the problems have been identified (Chapter 5)

Part III (Chapters 6-9) will then discuss how the shoestring evaluation approach can be applied at the project, sector and country levels.

Table 1: The Strengths and Weaknesses of the Five Most Frequently Used Quasi-Experimental Designs

Model	Works reasonably well to	Does not work well to
1.Pre- and post-test survey of project and control groups	This is the strongest QED. With a well-selected control group, it provides good estimates of project impacts.	
2.No pre-test control group	• Assess if the project model works and produces the intended outputs. • Assess similarities and differences between project and control areas. • Assess the extent to which the project could potentially be replicated	• Assess whether observed ex-post differences between the project and control groups are due to the project or to pre-existing differences between the two groups • Control for local history which might affect outcomes
3.No control group	• Evaluate projects which have large impacts or which operate in isolated areas where here is no interference from other outside interventions. • Understand the project implementation process	• Estimate the exact magnitude of project impacts. • Control for local history. • Assess potential for replication on a larger scale
4.No baseline data	• Obtain an approximate estimate of probable project impacts, particularly in small or isolated communities. • Compare project with other communities. • Control for the effect of intervening variables through the use of multivariate analysis.	• Estimate the exact magnitude of project impacts. • Control for local history.
5.No control groups or baseline data	• Conduct exploratory studies to get a general idea of how well the project model works. • Obtain a first, approximate estimate of impacts, particularly for small or isolated project.	• Obtain reasonably precise estimates of project impact. • Feel confident that the observed changes are due to the project and not to other factors or interventions. • Control for external events. • Obtain comparative data to estimate potential replicability.
Note that the strength of all of these models can be increased by combining them with the impact evaluation framework and analysis of contextual factors discussed in section 2-2; and with some of the shoestring evaluation techniques discussed in the following chapters. For Models 1, 2 and 4, which use control groups, the analysis can be greatly strengthened by using multiple regression analysis to statistically control for differences in the characteristics of the project and control groups.		

Figure 1: The Project Evaluation Framework



Part II

The Shoestring Evaluation Approach

Chapter 3

Reducing the Cost and Time Required to Conduct Impact Evaluations

3-1 Ways to reduce the cost and time required for data collection

◆ ***Simplifying evaluation design***

The strongest quasi-experimental design: It is rarely possible in development evaluations to use a true experimental design in which subjects are randomly assigned to treatment (project) and control groups. A large number of quasi-experimental designs (QED) have been developed which seek to approximate as closely as possible true experimental design. Model 1 (described in Chapter 2 Section 2-3) describes the QED considered to be the strongest for most purposes. As indicated in Chapter 2, the analytical power of this model can be strengthened if multiple regression analysis is used to statistically control for differences between the project and control groups. A number of additional refinements can be included to handle more complex evaluation designs, including, for example, multiple treatments.

Cheaper and faster, but usually weaker, QEDs: Where cost or time is a major concern, many evaluations eliminate one or more of the four observations. The most common options are:

- ***Model 2*** **:** the elimination of a pre-test control group
- ***Model 3*** **:** the elimination of a control group in both pre and post-test data collection
- ***Model 4*** **:** the elimination of baseline studies for both project and control groups
- ***Model 5*** **:** the elimination of both a control group and a pre-test

baseline study

Each successive model becomes progressively weaker as it fails to control for a greater number of the threats to validity. However, many situations exist in which the use of one of these less robust models is the only available option.

◆ ***Defining information needs to avoid collecting unnecessary information***

The timing, focus and level of detail of an evaluation should be determined by the information needs of key stakeholders and the types of decisions to which the evaluation must contribute. Typical questions that decision-makers must address do not require a high level of statistical precision, as the following:

- Is the project achieving its objectives? Which objectives are and are not being achieved?
- Are all sectors of the target population benefiting from the project? Are any groups being excluded or benefiting significantly less?
- Is the project sustainable and are benefits likely to continue?
- What are the contextual factors determining the degree of success or failure of a program or project?

Though many of the questions may not require a high level of statistical precision, some require reliable answers to such questions as the following:

- Are there measurable and significant changes in the target population with respect to the impacts the project is trying to produce?
- Is it reasonable to assume that the changes were due in a significant measure to the project and not to other unrelated factors?
- Is the project reaching all sectors of the target population, including the poorest and most vulnerable groups? Are both women and men benefiting? Are there any ethnic or religious groups who do not benefit?
- Why have these observed changes occurred? Are the conditions which facilitated these changes likely to continue and are the impacts likely to be sustainable?
- Were the target communities or groups reasonably typical of broader population groups (such as poor farmers or urban slum dwellers) and is

it likely that the same impacts could be achieved if the project were replicated on a larger scale?

The key design questions concern such issues as: (a) careful measurement of the key impact indicators; (b) ensuring that reliable information is obtained on participation and access to benefits by vulnerable groups such as women and ethnic minorities; (c) understanding the economic, social and political context within which the project is being implemented; and (d) ensuring that the observed changes are due to the project and not to unrelated factors. In many cases good estimates on most or all of these questions can be obtained with relatively simple evaluation designs. For the evaluation of larger and more complex projects, more rigorous and costly designs may be required.

◆ ***Reducing sample size and simplifying the sample design***

Sample size can often be reduced by accepting a lower level of precision for the estimates or by reducing the types of disaggregated data analysis; for example, accepting global estimates of project impacts and not comparing impacts in different regions. The use of cluster sampling can often significantly reduce interviewer costs by reducing distances and travel time between interviews. Of course, it will be necessary to assess the trade-off in each case between reduced cost and lower precision or less detailed analysis.

◆ ***Rapid and low cost methods for data collection and analysis***

A wide range of rapid and low cost data collection methods are available (Kumar 1993, Valadez and Bamberger 1994 Chapter 7), including: direct observation, automatic counters, focus groups and community fora, key informants, survey instruments that respondents can complete by themselves, use of secondary sources rather than interviews, etc. (see Table 2). PRA and related participatory methods are a potentially effective way to reduce data collection cost. Through these, estimates of the degree and direction of community level impacts can be obtained from a few carefully designed and executed group sessions rather than through large numbers of individual interviews.

◆ ***Integrating quantitative and qualitative approaches***

The integration of quantitative and qualitative data collection and analysis

methods is a requirement for any good evaluation design. Integration is particularly important to cost-effectiveness, as the use of a number of independent estimators can help validate methods which reduce sample size or the cost of data collection. This is an important application of the triangulation principle. Integrated approaches are particularly valuable also for understanding the contextual factors discussed above. Bamberger 2000 (Chapter 1) argues that an integrated evaluation approach is more than simply combining different data collection methods; it affects the way in which research hypotheses are generated, how the research team is constituted, how the research budget is allocated, and how time is allocated among different phases of the research process.

Participatory (largely qualitative) methods also can increase the validity and utility of information. Discussions with intended beneficiaries or groups who may have been negatively affected by projects can often identify unanticipated consequences of projects which may not be captured in surveys. Participatory methods are very useful for understanding contextual factors which may influence the level and distribution of project impacts and for assessing sustainability and replicability (Hentschel 1999).

Table 2: Rapid and Low Cost Methods for Impact Evaluations

Simplify evaluation design	Note: There is an important trade-off between reducing the costs and time of data collection. on the one hand, and the quality and validity of the evaluation findings, on the other. Each successive model described below sacrifices methodological rigor and is subject to an increasing range of threats to validity (see Chapter 5). It is important to compensate for some of these threats by reconstructing baseline and control group data through the use of secondary sources and the other measures discussed later in this handbook. Commonly used approaches for simplifying evaluation designs include: • eliminate data collection for pre-test control group (Model 2) • eliminate data collections for pre-test and post-rest control group (Model 3) • eliminate pre-test measurement for both project and control groups (Model 4) • eliminate all baseline measurements and also post-test control group (model 5)
Reduce sample size and data collection costs	• lower the level of required precision • reduce the types of disaggregation required • stratify sample designs • use cluster sampling • use university students, student nurses and community residents to reduce data collection costs
Reconstruct baseline data and control groups	• use secondary data • redesign project records to incorporate impact indicators • use recall • use key informants • use PRA and other participatory methods
Reduce the cost of quantitative data collection	• use self-administered questionnaires • reduce length and complexity of survey instruments
Qualitative data collection methods	• direct observation • automatic counters and other non-obtrusive methods • focus groups and community fora • key informants • PRA and other participatory methods
Integrated, multi-method data collection	• use triangulation (multi-method approaches) so that through independent estimates of key variables it may be possible to reduce sample size while at the same time increase reliability and validity.

Chapter 4

Reconstructing Baseline Data and Control Groups

Reliable information on the condition of project participants or control groups is often not readily available at the time the project is launched (the "baseline'' period referred to as T1). There are a number of sources which can be drawn on to reconstruct baseline conditions, including the following:

- *Secondary data* on factors such as morbidity, access to health services, school attendance, farm prices, and travel time and mode of transport can often be obtained from surveys conducted by sectoral agencies or from household surveys conducted by central statistical agencies. NGOs also may have conducted studies in some of the project areas. Many of these secondary sources are less than ideal for baseline references, lacking the exactly desired coverage, not conducted at the right time or not including all of the required information. Consequently, while these sources can provide a useful approximation of baseline conditions, their strengths and weaknesses as proxy baselines must always be assessed. Factors to be assessed include the following: differences in time periods and their significance (for example did economic conditions change significantly between the survey date and the project launch); differences in the populations covered (for example did the surveys include employment in the informal as well as in the formal sector and were both women and men interviewed); was information collected on all key project variables and potential impacts.
- *Project records* from micro-credit agencies, health centers, schools and water projects often contain information on conditions prevailing before

the project began. Again,the reliability of each source must be assessed[10].

- *Recall:* Asking individuals, key informants or focus groups to recall the social and economic conditions at an earlier point in time can be sometimes be used to estimate household or community conditions prior to the launch of the project. While recall is generally not reliable for obtaining precise numerical data on income, numbers of incidences of diarrhea or farm prices etc., it may provide useable information on major changes in the welfare conditions of the household. For example, families can usually recall which children traveled outside the community to attend school before the community village school was opened, how the children traveled to school, and the travel time and cost. Families may also be able to provide reasonably reliable information on use of health facilities prior to the project, or where they previously obtained water, how much they used and how much it cost. As relatively few studies have been conducted to assess the reliability of these kinds of recall estimates in developing countries, it is particularly important to identify and assess potential sources of bias in the estimates. For example, families might be reluctant to admit that their children had not been attending school, or that they had been using certain kinds of traditional medicine. They might also wish to underestimate how much they had spent on water if they are trying to convince the project that they are too poor to pay the proposed water charges.

The U.S. research literature has identified two common sources of recall bias. First, there is substantial evidence that the underestimation of small and routine expenditures increases as the recall period increases. Second, there is a telescoping of recall concerning major expenditures, so that major expenditures made outside of the recall period will often be reported as having

[10] For example, in the Eritrea impact evaluation referred to in Chapter 2, it was found that the project health centers kept records only of individual patient visits, it did not have records of the number of different patients or families visiting the center over a given period of time. These records would have been extremely time-consuming and expensive to sort by individual patient or to arrange by family. For this reason, it was not possible to use the records to estimate changes in the proportion of the population using health services.

occurred within the period. Most of the systematic research on recall bias has been carried out by U.S. government survey agencies (such as the U.S Expenditure Survey) or by universities; but some of their findings are relevant to developing country research. The LSMS (Living Standards Measurement Study) program also has conducted some assessments on the use of recall for estimating consumption in developing countries[11].

- *Key informants* such as community leaders, doctors, teachers, local government agencies, NGOs and religious organizations may be able to provide useful reference data on baseline condition . However, many of these sources have potential biases (such as health officials or NGOs wishing to exaggerate health or social problems, or community leaders downplaying community problems in the past by romanticizing conditions in the "good old days").
- *Participatory methods* such as PRA can be used to help the community to reconstruct information on past conditions and to identify critical incidents in the history of the community or region.

◆ ***Reconstructing control groups***

Many of the above methods can also be used to reconstruct control groups. However, this poses additional difficulties; it becomes necessary to identify appropriate control areas as well as to assess the conditions in these areas. With few exceptions project areas are selected purposively (for example to target the poorest areas or those with the greatest development potential) rather than randomly[12], so it can be a challenge to identify reasonably similar control areas.

Many ex-post quantitative impact assessments use statistical techniques to control for differences in individual and household characteristics, and hence

[11] LSMS is a World Bank program extending over more than 10 years to develop standard survey instruments and guidelines for the measurement of living standards in developing countries. For a review of the literature see Angus Deaton and Margaret Grosh "Consumption" Chapter 5 Margaret Grosh and Paul Glewwe (editors) *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*. World Bank 2000.

[12] One of the situations in which randomization is used in the selection of project areas occurs when demand significantly exceeds supply and some kind of lottery or other random process is used. This sometimes occurs with social funds (see Baker 2000 for a discussion of the Bolivia Social Fund) or with community supported schools (see for example Kim, Alderman and Orazem 1999 for a discussion of the Pakistan Community School project).

to approximate a control group, by identifying households or individuals who did not receive particular project services or who received less of the services[13]. While this kind of multivariate analysis offers a useful statistical control for individual characteristics, it cannot usually control for historical events or for differences in non-household attributes (such as different employment opportunities).

[13] For example subjects may be categorized according to their distance from a project-constructed road or water source, by whether any family attended literacy classes, or by the amount of food aid they received.

Chapter 5

Identifying and Addressing Threats to Validity

5-1 Threats to validity

In their efforts to reduce time and costs, evaluators frequently ignore some of the basic principles of evaluation design such as: random sampling, specification of the evaluation model, instrument development and full documentation of the data collection and analysis process. As a consequence, many rapid evaluations suffer from serious methodological weaknesses which threaten the validity or generalizability of their findings.

It is often assumed that evaluators working under budget and time constraints cannot be held to the same high methodological standards as would be the case normally when designing an impact evaluation. Two central premises of the shoestring evaluation approach address this point:

- The evaluator must make every effort to achieve the maximum possible methodological rigor in a given research context.
- The evaluator must explicitly recognize and make every effort to control for methodological weaknesses in the evaluation design.

To guide the researcher in identifying potential weaknesses in the evaluation design, a checklist of "threats to validity in the interpretation of evaluation findings" has been developed (Annex 1). The checklist is based on the recently published update and expansion of the seminal work of Cook and Campbell (1979) (Shadish, Cook and Campbell 2002). The table in Annex 1 is based on the four categories of threats to validity in Shadish, Cook and Campbell (2002). Three of the four categories have been expanded here to

include additional factors of particular importance for assessing the validity of shoestring evaluation designs. The discussion in this chapter also applies to many other impact evaluation designs which have relaxed or eliminated some of the requirements of a fully randomized experimental design (Valadez and Bamberger 1994 pp. 244-45). The four categories of threats to validity are discussed below.

◆ ***Threats to statistical conclusion validity*** (Annex 1 Section 1)

This deals with why the statistical design and analysis may incorrectly assume that program interventions have contributed to the observed changes (impacts). It also considers situations in which the statistical analysis may have overlooked some potential impacts. Of the 10 reasons given in Annex 1 Section 1, some of the most common issues for shoestring evaluations are the following:

- [1.1] *Low statistical power*: If the sample is too small, the conclusion may be incorrectly drawn that the relationship between treatment (project input) and outcome (project impact) is not significant. For example, many community water supply and health projects include reductions in infant mortality as one of the intended impacts. Infant mortality rates tend to change quite slowly because they are influenced by many factors and because only a small proportion of households have children born in a given year; for these reasons, even a successful initiative will probably produce only a small reduction in the short-run. It will often be necessary to have a sample size of several thousand households in order to identify a significant change but many evaluations of community projects only use samples of several hundred households. Thus, it will usually be impossible to identify a significant change, particularly over a 2-3 period year which is the reference period for many evaluations.
- [1.4] *Unreliability of measures*: Measurement error weakens the relationship between two variables. This is a common problem in project evaluations where administrative records on who received which benefits can be unreliable. For example, local health centers may provide treatment for common local diseases and infections and they may also provide advice to mothers on child-care and nutrition. Due to heavy work pressure on health center staff and unreliable supplies of

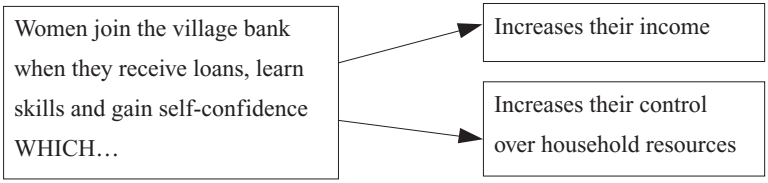
medicines, it frequently happens that one mother receives all required medicines as well as guidance on child-care and nutrition while another mother does not receive all of the medicines and other mothers do not receive guidance. Health center records normally will not record these details so in the analysis it is difficult to determine whether differences in impacts are due to contextual factors or simply to the fact that different people received different services.

- [1.5] *Restriction of range*: When a sample is truncated and covers only part of the population range, usually the relationship is weakened between the impact indicators and another variables, such as household income or family size. For example, if the sample is targeted to cover only low-income families, detecting a statistically significant relationship may be more difficult. This can be an important statistical issue when assessing the impact of poverty reduction programs which target the poorest segments of the population.
- [1.10] *Extrapolation from a truncated or incomplete data base*: Statistical analysis is often based on data sets which exclude part of the target population. Linear extrapolation or weighting methods used to estimate values for the excluded groups can provide biased population estimates. For example, enterprise surveys often cover only firms with more than a certain number of employees (often 10 or 25), excluding the large number of small informal enterprises where most of the poor work. Similarly, censuses and household surveys often miss the very poorest households, such as pavement dwellers, illegal immigrants or slum dwellers living outside the money economy.

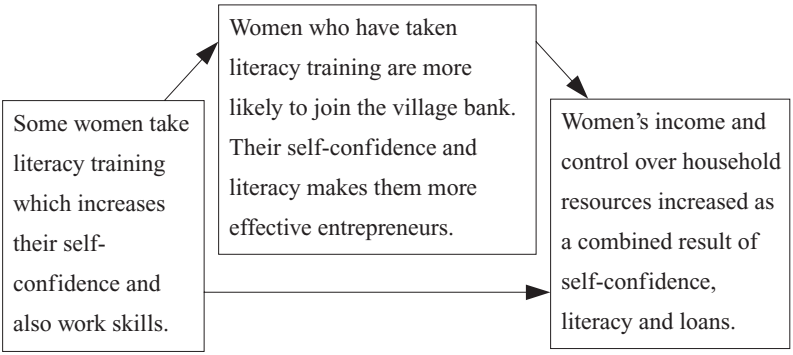
◆ ***Threats to internal validity*** (Annex 1 Section 2)

This deals with reasons why inferences of causality in the observed relationship between a project intervention and an output/impact may be incorrect. For example, the analysis may find a positive association between membership in a village bank and women's income and control over household resources and decision-making. It might be inferred that membership in the village bank caused women's income to increase and caused their enhanced control over household income. However, for a number of reasons, the assumed causal relationship may not be correct. Possibly women's decision-making power in the household had increased even before they joined the village bank

(perhaps as the result of their participation in a literacy or leadership training program). Or there may be a selection bias, so that women with higher incomes are more likely to be accepted by the village bank. The following figure presents the evaluation design hypothesis about the impact of village bank membership on women's income and control of household resources:



The figure below presents an alternative explanation of the observed changes in which the temporal sequence of events is changed.



The Annex 1 table lists a range of other factors which could result in incorrect inferences about causal relationships. Of the 12 reasons given in Section 2 of the table, some of the most common issues for shoestring evaluations include the following:

- [2.1] *Ambiguous temporal precedence* (confusion as to whether "impacts" occurred before or after the start of the project). Lack of clarity about precisely when the assumed impacts occurred may yield confusion about which is cause and which is effect.

- [2.10] *Inappropriate proxy indicators*. Proxy indicators are frequently used when it is not possible to directly measure the desired change or impact. For example, ownership of consumer durables, quality of housing construction or ownership of land may each be used as a proxy to measure changes in household welfare even though it is understood that none will fully capture all of the dimensions of welfare. Proxies frequently rely on project documents or other secondary data collected for purposes other than the evaluation, thus these indicators may not adequately measure access or impact. A typical example is the use of income as a proxy for welfare or economic condition. In many rural areas the majority of families practice subsistence agriculture and barter; they use very little money. Clearly in this context monetary income is not a good proxy for welfare or economic condition.
- [2.11] *Reliance on qualitative indicators*. The evaluation may rely, for ideological[14], methodological or logistical reasons, on qualitative indicators which may not permit generalization nor control for other explanations of the hypothesized causal relations. For example, participants in a community discussion group may be asked to agree on whether the quality of life of the community has improved or worsened over the past years. Normally it would be impossible to include this kind of very general qualitative indicator in a statistical analysis of factors determining changes in the social and economic conditions of a community.
- [2.12] *Unreliable respondent memory or deliberate distortion*. Recall, which is frequently used to recreate baseline conditions (see Chapter 4), is subject to biases due to memory failure and to deliberate distortion.

◆ ***Threats to external validity*** (Annex 1 Section 4)

This deals with reasons why inferences about how to generalize study findings to other settings (regions, social or economic groups, etc.) may not be valid. For example, the evaluation report on a community-managed village water supply project may say that the approach used in the projects was successful. The report may also recommend that the same approach be applied at the

[14] For example, some NGOs, women's organizations and academic researchers question the appropriateness of using quantitative surveys to understand living conditions in a village or local community. They argue that research must be based on empathy with the community and personal experience with their way of life.

national level. However, the pilot areas may have been selected because of characteristics likely to make the project a success (for example, traditions of community management of social infrastructure projects) and these characteristics might not exist in other regions of the country. Consequently the large scale replication of this project may be much less successful than expected.

Of the 8 reasons given in this section, some of the most common issues for shoestring evaluations include the following:

- [4.6] *Policymaker indifference*. Policymakers may impede or fail to implement a program they perceive to be irrelevant or threatening to their own priorities. Due to this kind of hindrance from local officials, a potentially effective project model may have disappointing results. Such a situation may result in under-estimation of the project's potential impact, which might have been achieved without the interference.
- [4.7] *Pro-active political interference*. The opposite may also occur. Policymakers and politicians may be anxious for the project to succeed and through their direct or indirect support may change the program model in ways that managers cannot control. For example, additional resources (equipment, materials, staff etc.) may be provided without charge, or they may remove normal administrative bottlenecks. This may result in an over-estimate of how successful the project would be if replicated in other areas that might lack this special level of support[15].
- [4.8] *Seasonal cycles*. Results may be attributable to irregular seasonal variations rather than to the program impact. This is particularly critical in assessing the impact and replicability of agricultural and rural development projects.

5-2 Strategies for addressing common threats to validity

Table 3 gives examples of approaches which can be used to address some of the common threats to validity once they have been identified. The fourth

[15] For example, internationally funded low-cost housing programs in the 1970s and 1980s sometimes found that governments (often in the hope of ensuring continued funding) wanting to demonstrate these programs' success would transfer to the project areas schools, health centers and other facilities originally intended for other areas of the city. As a consequence, the evaluation of the housing projects might find unexpectedly large improvements in health or education indicators which should have been attributed, at least in part, to these non-project interventions.

column indicates the stage of the evaluation at which different correction measures can be used. Most of the measures should be used during the design of the evaluation, though some can be used while the evaluation is being implemented (to conduct quality control on the validity of the information being collected or to consult with key informants and others to understand or correct some of the data collection issues). Still other measures are used during the analysis phase (for example to find secondary data sources to compensate for a lack of control groups). Time and resources should be allocated to allow for rapid follow-up field visits to check on inconsistencies or questions arising from the analysis or to explore in depth some of the interesting issues arising from the analysis.

The following are examples of the corrective measures which can be taken:

Threat 1.4: *Unreliability of measures*: This is an issue to be addressed in the design phase of the evaluation. Three possible approaches are recommended:

- Ensure that sufficient time and resources are allocated for developing and testing data collection instruments.
- Incorporate multi-method data collection approaches so that at least two independent measures are used for all key variables.
- Use triangulation to check the reliability of the information.

Threat 2.2: *Selection biases*: This refers to the possible differentiation of project participants from non-participants with respect to unique characteristics such as age, sex, economic status, ethnicity or motivation. Four possible approaches are recommended, some of which can be used during the evaluation design, others while the evaluation is being implemented, and still others during the analysis phase:

- Compare participant characteristics with non-participants, either informally or, if possible, through the construction of a control group. Ideally this should be done during evaluation design; but if not done at this stage, it is important to build some kind of comparison into the analysis and interpretation of the findings.
- During analysis, statistically control for participant characteristics to assess how they affect impact indicators.

- If there is no formal control group, use key informants and other available sources to compare participant characteristics with those of other people/families in the project and similar areas. This can be done during evaluation design, implementation, or analysis.

Use direct observation in focus groups and other settings to assess the unique psychological characteristics of participants, such as self-confidence or motivation, which might help explain project outcomes.

Table 3: Strategies for Addressing Common Threats to Validity in Shoestring Evaluations

No	Threat	How to address the threat	Applicable evaluation stage
1. Threats to Statistical Conclusion Validity			
1.1	Low statistical power	• Increase sample size • Use stratified sample to increase number of observations in critical cells	• Design • Design
1.4	Unreliability of measures	• Spend sufficient time and resources to develop and test data collection instruments • Multi-method approaches to provide at least two independent measures of key variables • Use triangulation to check on reliability of measures	• Design • Design • Design
1.5	Restriction of range	• Broaden sample to cover wider segment of the population • Use secondary data for comparative purposes • Conduct small, rapid sample of excluded population	• Design • Design and analysis • Design and analysis
1.10	Extrapolation from truncated or incomplete data base	• Same as for 1.5	• Design and analysis
2. Threats to Internal Validity			
2.1	Ambiguous temporal precedence	• Conduct rapid studies to clarify temporal sequence	• Design

No	Threat	How to address the threat	Applicable evaluation stage
2.2	Selection Bias	• Compare participant characteristics (age, sex, income, etc.) with control group • Statistically control for particular characteristics to assess how this affects impact indicators • If no control group, use key informants to compare participant characteristics with those of the total project area • Use direct observation in focus groups and other settings to assess psychological characteristics of participants such as self-confidence and motivation	• Design and analysis • Analysis • Design implementation and analysis • Design and implementation
2.3	Local or national history	• Collect secondary data on market prices, educational enrolment, morbidity patterns, wage rates, etc. • Consult key informants and experts on national trends.	• Design and analysis • Design and analysis
2.10	Inappropriate indicators	• Same as for 1.4.	• Design
2.12	Unreliable respondent memory or deliberate distortion	• Find and use independent sources (key informants, secondary sources, etc.) to check reliability for information. • Combine individual recall questions with group interviews to rest for consistency of information	• Design and implementation • Design and implementation
3. Threats to Construct Validity			
3.1	Inadequate explanation of constructs	• Meet with key stakeholders to understand more fully the implicit project model. • Develop a project theory model and meet individually or in groups with stakeholders to check and elaborate the model.	• Design • Design
3.8	Reactivity to the experimental situation	• Use exploratory studies, observation, etc., to understand respondent expectations and to identify potential response bias.	• Design
4. Threats to External Validity			
4.6	Policymaker indifference	• If the project is implemented in different locations, identify differences in the attitudes of policy makers in each location (through interviews, secondary sources or key informants) and assess how these differences appear to affect the project.	• Design

No	Threat	How to address the threat	Applicable evaluation stage
4.7	Pro-active political interference	• Same as for 4.6	• Design and analysis
4.8	Seasonal cycles	• Try to construct a control group. • Consult key informants • Use secondary data to identify seasonal patterns.	• Design • Design and implementation • Design and implementation

Part III

Using the Shoestring Evaluation Approach at the Project, Sector, and National/Policy Levels

Chapter 6

Shoestring Evaluation at the Project Level

6-1 Applications

There is a growing demand for rapid and low cost impact evaluations. In all areas evaluators face time and budget constraints as well as limited access to data; in this regard, shoestring approaches can make a major contribution. The following are some of its applications at the project level:

- *During project preparation*: rapid assessment of the outputs and impacts of earlier projects
- *During the mid-term review*: rapid assessment of actual and potential project impacts and of factors potentially limiting impacts
- *Project completion report*: better assessment of actual and potential impacts and identification of factors affecting future impacts and sustainability
- *Ex-post evaluations*: conducted by internal evaluation and audit units or by external consultants
- *PRSP*: periodic evaluations of the outputs and impacts of PRSP components.
- *Gender mainstreaming strategy*: evaluating the impacts of integrating gender into projects, programs and policies
- *Pilot projects*: rapid assessment of the actual and potential impacts of pilot projects

6-2 Challenges

Evaluators face all of the methodological and organizational challenges discussed in Chapter 1.

6-3 Useful approaches

◆ ***Simplifying the evaluation design***

There are normally two questions to be addressed with respect to the types of QED discussed in Chapter 2:

- Which design to use?
- What are the implications for the purposes of a given study of using one of the simpler but less robust evaluation designs?

The range of options is greatest when the evaluator is called in at the start of the project. Ideally the evaluation would use Model 1 with pre- and post-test measurement of project and control groups and with documentation on the project implementation process. The designs can be strengthened if they are combined with the development of a project theory model (see Chapter 2) so that the process of project implementation can be studied and so that the influence of contextual factors can be taken into account. However if time or budgets are constrained or if finding a control group will be difficult, the evaluator can consult with project management to clarify the precise purpose of the impact study and on this basis decide how important it is to have a control group:

- If the project is a relatively small pilot project whose purpose is to assess whether a particular new service or delivery system works, then it may be less important to have a control group.
- If, on the other hand, the purpose of the evaluation is to assess the cost-effectiveness of the project compared to alternative approaches, and to recommend whether the project could be replicated on a larger scale; then a control group may be more important.

When the evaluator is not called in until the project is completed or implementation is well advanced, the options for the evaluation design are more 1imited. However, it is still important to consult with the key stakeholders to answer the following questions:

- How will the evaluation findings be used?
- What are the key questions to be answered?
- What level of detail and precision is required?

The evaluator should then use the evaluation worksheet (see following section) to assess how well the current evaluation design can address the priority information needs of the key stakeholders. If the evaluation design does not provide the required analysis and information, the evaluator should consider using one or more of the following approaches:

- Reconstruct project baseline data using the methods described in Chapter 4
- Define a control group and conduct a rapid ex-post survey
- Reconstruct baseline data for the control group using the methods described in Chapter 4
- Conduct rapid assessment studies or focus groups to provide more information on contextual factors and how they affect project performance
- Complement ex-post survey information with other independent data collection methods to test the reliability and validity of the information being collected

◆ ***Check sample design and statistical power (threats to statistical conclusion validity***

If the sample size is too small to for statistically significant associations between project inputs and intended outputs/impacts, there are various possible options:

- Assuming the interviews have not yet been conducted, increase the sample size or improve the efficiency of the sample through stratification

- Reduce the planned level of disaggregation (for example by presenting estimates for the whole program only, and not for each project or region) or reduce the required level of precision
- Complement the sample survey with qualitative group interviews or consultations to obtain independent estimates of key variables or impacts

◆ ***Threats to internal validity***

Check the potential causes of threats to internal validity in Annex 1 Section 2. Depending on the stage at which the evaluation is conducted, the analysis can either be used to improve project performance or to better understand the strengths and weaknesses of the program model. For example:

- [2.1] Conduct rapid assessment studies to clarify temporal precedence of project inputs and changes in dependent variables
- [2.2] Remove biases from participant selection or recruitment procedures
- [2.6] Try to reduce attrition or note the characteristics of those who drop out so as to take account of this in the analysis
- [2.10] Review carefully the key indicators and assess their appropriateness for defining and measuring the key constructs

◆ ***Threats to construct validity***

A key recommendation is to work closely with all key stakeholders to develop a detailed and logical program theory model that specifies a set of logically consistent hypotheses about the processes through which the project is intended to produce its impacts. The model will also permit clear specification of the indicators required to measure all of the constructs. Once the model has been developed and agreement reached with stakeholders on the project design model and the key hypotheses, the evaluation design can be checked to assess which threats to construct validity exist (see Annex 1 Section 3) and how important they are for the objectives of the evaluation.

◆ ***Threats to external validity***

The evaluator should consult with stakeholders to define what kinds of generalizations and projections are to be made on the basis of the evaluation

findings. For example, will the evaluation results be used for the following purposes:

- To recommend whether the project could be replicated on a larger scale?
- To define for which population groups or in which regions the project could be replicated?
- To define the contextual factors which positively and negatively affect the success of the project and which should be taken into consideration when recommending how and where it could be replicated?

The threats to external validity checklist (see Annex 1 Section 4) should be used to identify any factors affecting the reliability of these projections. Depending on the stage of the project, appropriate measures should then be taken to correct factors affecting the reliability of projections. For example, if the attitude (positive or negative) of local politicians and policy makers appears to be a key determinant of project success, an analysis should be made of the implementation experience and outcomes in areas with different types of political participation. This will provide a firmer basis for anticipating the extent to which project outcomes are likely to be affected by the policy and political environment.

6-4 Using the Shoestring Project Evaluation Worksheet

Annex 2 presents a Shoestring Project Evaluation Worksheet which can be used as a checklist to assess the strengths and weaknesses of evaluation designs (particularly rapid and low cost designs) to identify methodological weaknesses which might affect the validity of findings and reduce their practical utility. Annex 5 gives an example of how to apply the worksheet to the evaluation of the impact of low-cost housing on income and employment discussed in Chapter 2 (as an example of QED Model 1). The worksheet contains the following sections:

◆ ***Coversheet***

1. *Stage of the evaluation at which the worksheet is being used*. This is important because the types of changes and interventions which can be effected will vary depending on how advanced the evaluation is.

2. *Evaluation design*. The design should be identified according to the classification system presented in Chapter 2. Potential problems should be identified concerning the following:
   - baseline data
   - control group
   - data collection methods and quality of data
   - analysis of contextual factors
   - other potential problems

The problems can be listed on the front page of the worksheet or on a separate page of notes are to be added.

3. *Objectives of the evaluation*. It is essential for the reviewer to fully understand the objectives of the evaluation. If the objectives are not explained in the evaluation design, the evaluator should meet with project stakeholders to discuss their expectations and to understand the decisions or actions to which the evaluation findings will contribute. It is particularly important to understand whether the evaluation is intended to be any of the following:
   - An initial exploratory study
   - An evaluation of a small pilot project
   - A rigorous multivariate statistical analysis of a large and complex project

Determining which of these options best describes the purpose of the evaluation is critical to understanding the information needs and the required level of methodological rigor.

4. *Time and resource constraints*. The reviewer must understand the primary concern of project management:
   - To comply with a tight time and budget deadline
   - To comply with a tight deadline, but where additional resources might be available
   - To produce a high quality product which can withstand scrutiny from

evaluation professionals and the project's critics

The recommendations which can be made on possible revisions to the design will be different under each of these scenarios.

◆ ***Analysis of individual threats to validity***

A separate page should be completed for each potentially important threat to validity. Importance is a judgment call which will be based on the reviewer's understanding of the objectives of the evaluation. For example, a more rigorous assessment of the sample design and the standardization of data collection procedures will be required for a large-scale statistical evaluation. The points to be covered for each threat to validity include the following:

A. *How is the problem manifested in the evaluation design?* For example:
- Very few women attended the focus group session or community meeting in which feedback on the project was obtained.
- The village banks were launched in communities with previous entrepreneurial experience so that the project is expected to be more successful than in typical communities.

B. *What are the potential effects on the study findings and generalizations?* For example:
- The finding that the community was generally satisfied with the choice of projects may not reflect the views of most women.
- The findings may over-estimate the potential for replication of the pilot project.

C. *How big a problem is this for the evaluation?* Reasons should be given for the judgment. For each threat the reviewer must decide whether it is critical, important, of minor importance or unimportant for the evaluation.

D. *What are the proposed actions?* The proposed actions will vary according to the seriousness of the problem and the resource constraints. In some cases the reviewer may be in a position to initiate the recommended actions, in other cases recommendations are made to the client concerning actions which he could take. Examples of possible actions include the following:

- Prepare a report for the client indicating areas in which the results and conclusions of the evaluation might be misleading or where they should be interpreted with reservations.
- Secondary data could be analyzed to assess whether the observed changes in the project population are similar to or different from those occurring in the wider population.
- Rapid assessment studies could be conducted to validate findings or to provide a stronger framework for the interpretation of the findings.

E. *How adequate are the proposed actions?* A note should indicate whether the actions are likely to correct, reduce, or have little impact on the problem. The note should also indicate why the problem cannot be resolved at this stage of the analysis. Is it a question of time and resource constraints, is it due to resistance or lack of interest from the client, or are there external political factors involved?

Chapter 7

Shoestring Evaluation at the Sector and Program Level

7-1 Applications

There is also a growing demand for rapid and low cost impact evaluations at the program and sector level. The following are some of the present and potential applications:

- Assessments by development agencies of their sector work (for example, social funds, irrigation, transport)
- Assessments of the agencys' social policies such as gender mainstreaming, indigenous peoples, etc., at the country, regional or international levels
- Assessment of the impacts of rural development programs on the poor
- Evaluating the impacts of the initiatives to combat HIV/AIDS

7-2 Challenges

Evaluators face all of the methodological and organizational challenges discussed in Chapter 6, with the following additional complications:

- As sector programs usually comprise many different interventions, all of the methodological questions discussed in the previous chapters (sampling, constructing baseline and control data, etc.) become even more complicated.
- The issue of identifying a control group is particularly difficult as different projects and programs in the sector are organized differently

and address different groups.

- Many different agencies are involved, in most sectors – each with different objectives and ways of organizing projects.
- Several additional organizational complications must be addressed. First, different agencies organize projects and programs differently. Second, each funding agency has a slightly different set of questions of concern to its national clients (Parliament, Ministry of Finance, NGOs, etc.) or international stakeholders. Consequently, developing countries are often required to divide their 1imited within-country evaluation capacity resources among many different projects. It has proved extremely difficult to achieve the cooperation required for any kind of coordinated evaluation effort.

7-3 Useful approaches

◆ ***Simplifying the evaluation design***

Sector and program evaluations frequently combine global information on the sector with data compiled from a number of different project level evaluations. The guidelines for conducting and reviewing individual project level evaluations will be similar to those discussed in Chapter 6. However, there are added complications when the findings of different project evaluations are combined.

New challenges must be faced when compiling sector-wide information. The analysis is easiest when the findings of similar projects are combined (for example when different donors fund similar education or water projects in different regions). The analysis becomes much more difficult when synthesizing projects with different objectives, scope and implementation methods. Rapid assessment methods offer a useful way to obtain a quick overview of project level experiences. Some of the possible evaluation approaches include:

- Focus groups and workshops with representatives from different projects and agencies
- Using recall at the project (rather than household) level to obtain overall assessments of implementation experience and impact
- Using project records and secondary data to reconstruct baseline

conditions
- Control group information (for example, on education, health, water-supply, etc.) is sometimes easier to obtain at the sector level than at the project level

◆ ***Assess sample design and statistical power (threats to statistical conclusion validity***

In many cases the challenge is to compare data from different projects to determine whether observed differences in outputs and impacts are attributable to project design or to other factors. This often is very difficult because the number of observations is quite small. Due to the small number of observations, frequently an interpretation of differences must rely more on qualitative data than on rigorous statistical analysis.

◆ ***Threats to internal validity***

The issues and approaches are similar to Chapter 4 with the added complication that the assessment often must be made on a number of different projects. The following are some of the special issues which may have to be addressed:

- [2.1] *Temporal Precedence*: The same set of checks on temporal precedence (see Chapter 5) often must be applied to different projects. There is the additional complication that the implementation schedule of different projects may interfere with each other. For example, if one agency is organizing women's leadership training programs in areas where another is starting a micro-credit program; it becomes difficult to assess whether women's enhanced empowerment is due to the impacts of the second project, or whether the women participating in the second project had already been empowered by the leadership training of the first project.
- [2.2] *Selection bias*: Selection procedures and biases are likely to be different for different projects. Biases must be identified and the implications of different selection procedures compared.
- [2.6] *Attrition*: Again experiences with different projects must be compared.
- [2.10] *Inappropriate indicators*: Frequently different projects use

slightly different indicators, something must be controlled for. While it is desirable to use standard indicators, care must be taken not to obscure differences in approach between different projects[16].

◆ ***Threats to construct validity***

Where projects using different models are being compared, the importance of well defined theory models becomes even more critical. These should define the program theory models used by each project, but also must identify critical differences between models.

◆ ***Threats to external validity***

In addition to the project level analysis discussed in Chapter 6, market or demand saturation is a factor which must be addressed at the sector level. When projects are analyzed separately each one may appear to have a capacity for replication, but when an aggregate analysis is conducted, the issue of saturation, or competition between projects, must also be assessed.

7-4 Using the Shoestring Program Evaluation Worksheet

Annex 3 presents a Shoestring Program Evaluation Worksheet. As with the project worksheet in Annex 2, this can be used either to plan an evaluation or to review and modify an ongoing evaluation. The program worksheet is similar to the project worksheet except for the following:

◆ ***Coversheet***

Type of evaluation study: Clarify whether the study evaluates a single program implemented in a standard way in many sites, or whether there are a number of different projects each using a different design.

◆ ***Analysis of individual threats to validity***

Level of generality: Does the discussion refer to a threat to one or more individual projects or interventions, or to the overall program?

---

[16] For example, one micro credit program may provide only training on financial management, while another might include also entrepreneurial development or empowerment training. It is obviously important to make sure that while indicators are comparable, they also capture these differences.

Chapter 8

Using Shoestring Evaluation at the Country and Policy Level

8-1 Applications

There is also a growing demand by development agencies for rapid and low cost evaluations of the effectiveness of their country programs and policy work. Although there is less experience in the use of shoestring evaluation at the national level, the shoestring approach has potential to contribute to the unique methodological challenges facing effectiveness analysis at this level. Some potential applications include the following:

- Evaluating the effectiveness of country assistance strategies.
- Contributing to periodic reviews of agency performance; for example, the World Bank Annual Review of Development Effectiveness prepared by the Bank's Independent Evaluation Group (IEG)
- Evaluating the effectiveness of special financing mechanisms, such as targeted development grants, and cooperative financing mechanisms; for example, the World Bank recently evaluated the special PHRD project preparation facility financed by the Government of Japan
- Evaluating the effectiveness of the contributions of an agency to the international HIV/AIDS initiative

8-2 Challenges

Country and policy evaluation poses an additional set of challenges on top of the methodological and organizational questions discussed in earlier

chapters. These include the following:

- Finding the right counterfactual
- How to make before-and-after comparisons for complex programs which include many components and operate on many different levels. Frequently this involves aggregating the impacts of different projects, often using different methodologies, which form part of a national program.
- How to distinguish the impact of development agency interventions from:
  - Changes attributable to country performance
  - Changes due to other development partners
  - Changes due to exogenous factors
- Identifying and measuring often subtle changes in policies

8-3 Useful approaches

◆ ***Distinguishing levels of impact***

It is important to define which of the following levels of effectiveness are to be examined:

- *Outcomes and impacts within a development agency (corporate responsiveness)*. For example, the World Bank evaluation of the implementation of its gender policies includes, among other things, the increasing use of gender analysis by the Bank and the allocation of funds for Bank gender initiatives.
- *Mainstreaming policies and approaches within client county agencies*
- *Evaluating impacts of policies on target populations within client countries*

◆ ***Finding the right counterfactual***

A counterfactual is an estimation of what would have been the situation if a particular policy or country program had not been implemented. There are two main approaches:

- Comparing before-and-after conditions

- Analysis of client country performance relative to:
  - Similar countries
  - Development indicators defined in country assistance strategies, United Nations Human Development Indicators, country gender reviews, PRSPs, country sector work and policy documents, etc.
  - Endogenous factors (economic conditions, demographic changes, terms of trade, etc.)

◆ ***Approach 1: Before-and-after studies***

Country program and policy evaluations can be assessed in either or both of the following ways:

- Aggregating the outcomes and impacts of a number of different project and sectoral interventions, all of which form part of a national program; for example, programs which: provide grants to communities and local level organizations in different regions to construct or upgrade community infrastructure, or which providing loans and tax breaks to small entrepreneurs to offer transport services to rural communities. In both cases the assessment of the effectiveness of the country program (or policy) requires evaluating and aggregating the impacts of large numbers of different projects.
- Directly assessing a program or policy which is implemented in a uniform manner throughout the country

There are three main types of evaluation studies which can be used to evaluate either of these kinds of national programs/policies:

- Exploratory evaluations
- Evaluations of pilot interventions, at either the project or program levels
- Large-scale quantitative analysis of (large-scale) program/policy interventions

Each of these evaluation types must address the following three methodological challenges:

- Defining indicators for each dimension

- Selecting cost-effective and reliable data collection methods
- Assessing the extent to which observed changes are due to the program strategy

◆ ***Exploratory studies***

Exploratory studies provide initial feedback on the ability of individual projects or national programs/policies to achieve their objectives. These studies normally use rapid participatory assessment methods with relatively small samples and are completed at a modest cost in a relatively short period of time. The sample size and cost will normally increase for larger and more complex programs. Rapid assessment methods typically include the following:

- Focus groups and other community consultations
- Participant observation
- Unstructured interviews with individuals, households and community groups
- Key informant interviews
- Stakeholder analysis
- Rapid surveys to obtain basic quantitative data

◆ ***Evaluations of pilot interventions at the project or program levels***

The assessment of (usually small) pilot interventions is intended to assess the potential of the model for achieving its objectives and impacts. The assessments are also intended to help understand factors contributing to the success, accessibility, and potential replicability of the project model. Again, rapid assessment studies are used, but usually in combination with rapid sample surveys to provide a stronger base for generalization.

◆ ***Large-scale quantitative analysis of (large-scale) interventions***

Large-scale quantitative evaluations are normally justified only for large projects or programs which have been operating for several years and which are already producing clearly defined, uniform, and quantifiable results. They are also particularly useful where comparative analysis is required for projects or programs operating in several different contexts. These expensive studies are justified only if the findings can contribute to decisions on future investments.

Large-scale evaluations will normally have the following stages:

- Exploratory studies to identify key variables and issues
- Analysis of secondary data
- Definition of evaluation design model
- Definition of control group
- Selection of a representative sample of project beneficiaries
- Application of the data instruments
- Use of a multimethod approach that combines surveys with qualitative methods

Shoestring methods can be introduced at most stages to reduce cost or time, for purposes of triangulation, or to help fill gaps in the available data (for example, reconstructing baseline studies or creating control groups).

◆ ***Approach 2: Comparing client performance to benchmarks***

In order to assess the potential impacts of ODA interventions, client performance in countries where the agency is active can be assessed relative to the following:

- Similar countries
- Development indicators included in country assistance strategies, country gender reviews, United Nations Human Development Indicators, etc.
- Client performance relative to endogenous factors such as national and international economic environment, demographic changes, etc.

◆ ***Strengthening the logical robustness of attributions***

Several approaches can be used to strengthen the logical robustness of attributions:

- Self-assessment by agencies
- Independent assessment by other agencies
- Independent assessment by internal evaluation and audit departments and other development partners
- Comparison of impacts in different policy and program areas

- Rapid assessment studies to obtain feedback from different stakeholders (intended beneficiaries, implementing agencies, civil society, etc.)

The assessments can include short-, medium-, and long-term indicators. It is important to define clearly the time-horizon over which different types of impacts are expected to occur.

Example of short-term indicators (within the first 12 months):

- Are required inputs being mobilized and used effectively?
  - Were funds approved and released?
  - Were staff and consultants hired?
  - Were training/capacity building activities launched?

Examples of medium-term indicators for projects or programs (within 2-3 years)

- Were intended products created (schools built, health workers trained, community environmental action plans prepared, etc.)?
- Do all sectors of the target group have access to services and benefits?

Examples of long-term indicators (beyond 2-3 years)

- Welfare indicators for the target populations
- Indicators of sustainability:
  - There is accessibility and use of services.
  - There is regular maintenance by community and responsible agencies.
  - The community makes the required financial contribution.
  - Finance and other resources are contributed by responsible agencies.
  - There is capacity building at the community and government levels.

◆ ***Evaluating the effectiveness of programs in achieving five broad development criteria***

The evaluation should also assess program performance in terms of broad development criteria. The following five criteria are used by the World Bank Independent Evaluation Group in all of its program evaluations, and similar criteria are used by many other development agencies:

- Efficiency

- Efficacy
- Development Impact
- Relevance
- Sustainability

◆ ***Threats to validity***

The issues to be addressed here are similar to those for program and sector evaluations. Some points to be kept in mind are the following:

Threats to statistical conclusion validity:

The need to combine estimates from a number of different studies requires caution in the interpretation of statistical conclusions about outcomes and impacts, and particularly their attribution to projects.

Threats to internal validity:

Issues of internal validity are particularly difficult to assess because the administration of programs and policies is often not done in a sufficiently systematic and controlled way. This may be due to the often weak administrative structures and the need to compromise with different groups who oppose the programs.

Threats to construct validity:

It is particularly important to have a clearly defined program model with explicit definition of the intended links between inputs, processes, outputs and expected impacts. This will help define key indicators and assess their appropriateness for measuring the underlying constructs.

Threats to external validity:

As most policies are intended for application at the national level, it is usually not possible to generalize them to a broader population group in the same way that projects and sector programs are generalized. However, some of the generalizability issues which must be addressed include the following:

- Policies may be implemented first in certain sectors or regions with the intention of subsequent replication in other sectors or regions. Because the characteristics of different sectors often vary in important ways, it

can be difficult to assess how easily this replication can be achieved. For example, a policy to reduce subsidies and increase user charges for the provision of public services is likely to have a very different implementation experience in primary health services, education, and public transport. If this is the case, experience in the health sector may not provide good guidelines for assessing what will happen in education.

- Are the policies sustainable?

8-3 Using the Shoestring Policy and Country Evaluation Worksheet (Annex 4)

The policy and country program worksheet is very similar to the sector worksheet (Annex 3) except for the following:

◆ ***Covershee***

1. *Type of evaluation*: Is this an evaluation of a policy or of a country program?
2. *Evaluation design:* Does the evaluation use a before-and-after design, or does it assess effectiveness by comparing client performance with similar countries, development indicators or exogenous factors?

◆ ***Analysis of individual threats to validity***

Level of generality: Does the discussion refer to the overall policy or to certain components of it?

Chapter 9

Building Evaluation into Project and Program Design

9-1 Building evaluation elements into project and program design

Many of the problems facing evaluators who are called in late in the project cycle occur because no information was collected on the characteristics of the future project beneficiaries and non-beneficiaries before the project began. Consequently, one of the main challenges facing the evaluator is to reconstruct information on pre-project conditions to form the basis for comparison with post-project data on beneficiaries and (hopefully) a control group. If a way could be found to convince policy makers and project managers to collect baseline data, the task of the future evaluator would be much easier and the quality of the evaluation analysis would be much better. Unfortunately, at the time when new projects and programs are being planned and launched, concerns about future impact evaluations are very low on the project manager's list of priorities. When a project is launched there is no guarantee it will survive or that there will be any results to measure. In most cases the evaluation staff are not even recruited and trained until the project has been underway for some time.

Much of the basic baseline data could be collected with almost no additional effort or cost. The question addressed in this chapter is the following: Given the real-life conditions under which projects are launched – keeping in mind that the evaluation staff have probably not yet been recruited and the manager has no time to worry about future evaluations –are there are any simple measures that could be taken to ensure that the data is generated?

9-1 Strengthening the treatment of evaluation at four phases of the project cycle

Focus in this regard is on four critical moments in the project cycle described below. All of the following recommendations are considered to be minimal guidelines, to be followed by project management regardless of whether or not there is a plan to conduct an impact evaluation and whether or not an evaluation unit has been established. In other words, the proposed measures provide useful information for strengthening program management. If management does choose to plan for a future evaluation, then these guidelines can be used to prepare it.

◆ ***During project planning***

Despite the fact that extensive data collection and analysis is frequently undertaken during project appraisal and planning, very little attention is normally given to how the information could be used for future evaluations. The concern is to produce the information required for the project approval and launch. Even though many development agencies require that project planners produce a LogFrame and a set of performance indicators, little serious thought is normally given to laying the basis for a future evaluation. The following are some simple measures which could be taken during project planning:

- Require that the project planning process include a discussion on how to collect a minimum set of evaluation indicators. The plan should include:
  - Definition of a preliminary theory model
  - A simple evaluation design
  - A discussion of sampling (how to select a representative sample of future project beneficiaries and a control group of similar families who will not benefit from the project)
  - A preliminary set of measurable performance indicators
  - Definition of organizational responsibilities for both monitoring and evaluation

- Ensure that all surveys and other forms of data collection and analysis conducted during appraisal are archived and documented so that they

will be accessible to and usable by future evaluators.

- Require that the project planning documents include at least a brief discussion of the characteristics of the target population, and their similarities and differences from the rest of the low-income population. Ideally there should be some minimal collection of information on the control population and a discussion of how a control group could be created.
- Recommendations on how routinely collected administrative information on the project could be organized to make it more useable for a possible future evaluation. For example:
  - Records of local health facilities could be organized so that there is a file for households and not just for individuals. In this way it would be possible to quickly estimate what proportion of households use the health facilities.
  - Consideration should also be given to collecting brief information on the basic living conditions of patients using the health facilities (quality of housing, type of water supply and sanitation, and possibly a simple indicator of economic status), although it will often not be feasible to collect all of the information.
  - Schools might be asked to collect some basic social and economic information on students (similar to the above).
  - Village banks and credit programs, many of which already collect a lot of information, might be asked to prepare a brief socio-economic profile of borrowers

◆ ***During project implementation***

- Ensure that the proposed revisions to the administrative records (discussed above) are put into place.

◆ ***In conjunction with the mid-term review*** (MTR)

In many cases rapid assessment studies are conducted in preparation for the MTR in order to provide an initial assessment of potential impacts and to identify potential constraints to the achievement of project objectives. A number of simple measures could be taken to ensure the information can contribute to a possible future impact evaluation:

- When rapid assessment studies are commissioned, attention should be paid to the sample procedures (however simple they may be) to ensure the representativeness of the findings. These studies can often provide an approximation of a baseline study.
- Data should be archived and documented so that it can be easily located and used for future evaluations.
- The MTR should produce estimates on the status of all of the critical performance indicators identified during project planning.
- The Logic Model should be reviewed and the validity of its key assumptions assessed and, if necessary, updated.

◆ ***In conjunction with the project Completion Report***

Similar guidelines should be followed as for the MTR.

References

Alderman, Harold, Jooseop Kim and Peter Orozem. 2000. *Design, Evaluation and Sustainability of Private School for the Poor: The Pakistan Urban and Rural Fellowship School Experiments*. HDNED.

Baker, Judy. 2000. *Evaluating the impact of Development Projects on Poverty: a Handbook for Practitioners*. Directions in Development. World Bank.

Bamberger, Michael. 2002. *Longitudinal Impact Evaluation of the Community Development Fund Feeder Road Component*. Unpublished.

Bamberger, Michael. 2000. *Integrating Quantitative and Qualitative Research in Development Projects*. Directions in Development. World Bank.

Chambers, Robert. 1995. *Poverty and Livelihoods: Whose Reality Counts*? Institute for Development Studies. Discussion Paper No. 347. University of Sussex.

Deaton, Angus and Margaret Grosh. 2000. "Consumption" Chapter 5 Margeret Grosh and Paul Glewwe (editors) *Designing Household Survey Questionnaires for Developing Countries: Lessons from Fifteen Years of the Living Standards Measurement Study*. Oxford University Press.

FASID 2000. *Project Cycle ManagementÅDMonitoring and Evaluation based on the PCM Method*.

Grosh, Margaret and Paul Glewwe. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from Fifteen Years of Living Standard Measurement Study*. Oxford University Press.

Khandker,Shahidur 1998 *Fighting Poverty with Microcredit: Experience in Bangladesh*. New York. Oxford University Press for the World Bank.

Kumar, Krishna (editor). 1993. *Rapid Appraisal Methods*. World Bank.

Reitbergen-McCracken and Deepa Narayan. 1997. *Participatory Tools and Techniques: A Resource Kit for Participation and Social Assessment*. Environment Department. The World Bank. Washington D.C.

Miles, Matthew and Michael Huberman. 1994. *Qualitative Data Analysis*. Sage Publications.

Narayan, Deepa and Patti Petesch. 2002. *Voices of the Poor: From Many Lands*. World Bank. Washington D.C.

Newman, Constance. 2001. *Gender Time-Use, and Change: Impacts of Agricultural Export Employment in Ecuador*. Policy Research Report

on Gender and Development. Working Paper Series No. 18. URL: www. worldbank. org/gender/prr

Nizam, Ahmed. 2000. *Study on Gender Dimensions of the Second Bangladesh Rural Roads and Markets Improvement and Maintenance Project*. URL: www. worldbank. org/gender/transport/grantsandpilots

Patton, Michael. 2002. *Qualitative Research and Evaluation Methods* (3rd edition). Sage Publications.

Picciotto, Robert. 2002. *Development Cooperation and Performance Evaluation: the Monterrey Challenge*. OED, World Bank.

Pitt, Mark; Shahidur Khandker, Signe-Mary McKernan, M. Abdul Latif. 1999. "Credit Programs for the Poor and Reproductive Behaviour in Low-Income Countries: Are the Reported Causal Relationships the Result of Heterogeneity Bias?" *Demography* 36 (1) 1-2 1

Pitt, Mark and Shahidur Khandker. 1998. "The Impact of Group-based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy*. 106: 958-96.

Shadish, William, Thomas Cook and Donald Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin. New York.

Valadez, Joseph and Michael Bamberger. 1994. *Monitoring and Evaluating Social Programs in Developing Countries*. World Bank.

World Bank. 2000. *Attacking Poverty: World Development Report 2000/2001*. Oxford University Press.

World Bank. 2001. *Engendering Development through Gender Equality in Rights, Resources and Voice*. pp. 160-162. A World Bank Policy Research Report.

World Bank. 2002. *Poverty Reduction Strategies Sourcebook*. World Bank. Washington D.C.

Annexes

Annex 1

Threats to Validity in the Interpretation of Evaluation Findings

Threats to Validity in the Interpretation of Evaluation Findings
1. Threats To Statistical Conclusion Validity[17]: Reasons why inferences about covariation between two variables may be incorrect
1.1 Low Statistical Power 1.2 Violated Assumptions of Statistical Tests 1.3 Fishing and the Error-Rate Problem 1.4 Unreliability of Measures 1.5 Restriction of Range 1.6 Unreliability of Treatment Implementation 1.7 Extraneous Variance in the Experimental Setting 1.8 Heterogeneity of Units 1.9 Inaccurate Effect Size Estimation 1.10 *Extrapolation from a Truncated or Incomplete Data Base*
2. Threats to Internal Validity[18]: Reasons why inferences that the relationships between two variables is causal may be incorrect
2.1 Ambiguous Temporal Precedence

[17] Source: Adapted from Shadish, Cook and Campbell 2002. Table 2.2 page 45. The threats in italics, representing additional factors of particular importance for Shoestring Evaluations, have been added by the present authors.

[18] Source: Adapted from Shadish, Cook and Campbell 2002. Table 2.4 page 55. The final 3 threats (in italics) have been added by the present authors.

2.2 Selection 2.3 History 2.4 Maturation 2.5 Regression 2.6 Attrition 2.7 Testing 2.8 Instrumentation 2.9 Additive and Interactive Effects Of Threats To Internal Validity. 2.10 *Inappropriate proxy indicators* 2.11 *Reliance on qualitative indicators* 2.12 *Unreliable respondent memory or deliberate distortion*
3. Threats To Construct Validity[19]: Reasons why inferences about the constructs that characterize study operations may be incorrect
3.1 Inadequate explanation of constructs 3.2 Construct confounding 3.3 Mono-operation bias 3.4 Mono-method bias 3.5 Confounding Constructs with Levels of Constructs 3.6 Treatment sensitive factorial structure 3.7 Reactive self-report changes 3.8 Reactivity to the experimental situation 3.9 Experimental expectancies 3.10 Novelty and disruption effects
4. Threats To External Validity[20]: Reasons why inferences about how study results would hold over variations in persons, settings, treatments and outcomes may be incorrect
4.1 Interaction of the causal relationship with units

[19] Source: Adapted from Shadish, Cook and Campbell 2002. Table 3.1 page 73.

[20] Source: Adapted from Shadish, Cook and Campbell Table 3.2 Page 87. The three final threats (in italics) have been added by the present authors.

4.2 Interaction of the causal relationship over treatment variations
4.3 Interaction of the causal relationship with outcomes
4.4 Interactions of the causal relationships with settings
4.5 Context-dependent mediation
4.6 *Policy maker indifference*
4.7 *Political interference*
4.8 *Seasonal cycles*

Annex 2

Shoestring *Project* Evaluation Worksheet

Shoestring *Project* Evaluation Worksheet

Part I: Description of the Evaluation

Name of Evaluation: _____

1. ***Stage of the study at which the worksheet prepared:***
   - Evaluation design stage
   - Pilot testing of instruments
   - During data collection
   - Data analysis
   - Report writing

2. ***Evaluation design***
   - Which evaluation design was used? [See Chapter 2]
   - Explain any modifications to the designs in Chapter 2.
   - Potential problems (threats to validity) concerning: [see Chapter 5]
     - Baseline data
     - Control group
     - Data collection methods and quality of data
     - Analysis of contextual factors
     - Other: _____

[*Pages may be added to describe any additional problems or issues not covered above.*]

3. ***Objectives of the evaluation***
   - Why was the evaluation commissioned?
   - What are the specific decisions or actions which will be taken on the basis of the findings ?
   - Purpose of the evaluation:
     - Exploratory study to provide initial indications on whether the project model "works"
     - Assessing the efficiency and potential impacts of a small pilot project to recommend whether it is worth replicating on a larger scale
     - Rigorous multivariate statistical analysis of a large-scale, multi-component project to compare costs and benefits with alternative investment options

4. ***Time and resource constraints. Which of the following describes the current situation:***
   - There is a very tight deadline and no possibility of additional resources.
   - There is a tight deadline but additional resources could be obtained for use within this deadline.
   - The priority is to produce a high-quality product with a solid methodology which can withstand scrutiny from the critics of the project.

Note: Complete a Part II sheet for each important threat to validity.

Part II: Analysis and discussion of each threat to validity

Threat [Number and name]

A. How are the threats manifested in the evaluation ?

B. What are the potential affects on the study findings and generalizations ?

C. How big a problem is this for the evaluation ?

D. Proposed actions

E. How adequate are the proposed actions ?

Annex 3

Shoestring *Program* Evaluation Worksheet

Shoestring *Program* Evaluation Worksheet

Part I: Description of the Evaluation

Name of Evaluation: _____

1. ***Type of program being evaluated:***
   - Single program with standard activities in all sites
   - Different projects or intervention strategies

2. ***Stage of the evaluation at which the worksheet prepared:***
   - Evaluation design stage
   - Pilot testing of instruments
   - During data collection
   - Data analysis
   - Report writing

3. ***Evaluation design***
   - Which evaluation design was used ? [See Chapter 2]
   - Explain any modifications to the designs in Chapter 2.
   - Potential problems (threats to validity) concerning: [see Chapter 5]
     - Baseline data
     - Control group

- Data collection methods and quality of data
- Analysis of contextual factors
- Other _____

[*Pages may be added to describe any additional problems or issues not covered above.*]

4. Objectives of the evaluation

- Why was the evaluation commissioned?
- What are the specific decisions or actions which will be taken on the basis of the findings?
- Purpose of the evaluation:
  - Exploratory study to provide initial indications on whether the program model "works"
  - Exploratory study to assess the relative efficacy of different projects or interventions
  - Rigorous multivariate statistical analysis of a large-scale, multi-component project to compare costs and benefits with alternative investment options

5. Time and resource constraints. Which of the following describes the current situation:

- There is a very tight deadline and no possibility of additional resources.
- There is a tight deadline but additional resources could be obtained for use within this deadline.
- The priority is to produce a high-quality product with a solid methodology which can withstand scrutiny from the critics of the project.

Note: Complete a Part II sheet for each important threat to validity.

Part II: Analysis and discussion of each threat to validity

Threat [Number and name]

A. ***Level of generality:*** *Does the discussion refer to one or more individual projects or interventions or to the overall program?*

B. How are the threats manifested in the evaluation ?

C. What are the potential affects on the study findings and generalizations ?

D. How big a problem is this for the evaluation ?

E. Proposed actions

F. How adequate are the proposed actions ?

Annex 4

Shoestring *Policy and Country* Evaluation Worksheet

Shoestring Policy and Country Evaluation Worksheet

Part I: Description of the Evaluation

Name of the Evaluation

1. ***Type of evaluation:***
   - Policy evaluation
   - Country program evaluation [describe the country program]

   _____

2. ***Stage of the evaluation at which the worksheet prepared:***
   - Evaluation design stage
   - Pilot testing of instruments
   - During data collection
   - Data analysis
   - Report writing

3. ***Evaluation design***
   - Before and after design
   - Assessing effectiveness in terms of client performance compared to:
     - Similar countries
     - Development indicators
     - Exogenous factors

4. ***Potential problems [see chapter 5 and chapter 8]***
   - Baseline data
   - Control group
   - Data collection methods and quality of data
   - Analysis of contextual factors
   - Other _____

[*Pages may be added to describe any additional problems or issues not covered above.*]

5. ***Objectives of the evaluation***
   - Why was the evaluation commissioned?
   - What are the specific decisions or actions which will be taken on the basis of the findings ?
   - Purpose of the evaluation:
     - Exploratory study to provide initial indications on whether the policy or country program model "works"
     - Exploratory study to assess the relative efficacy of different policy components or application in different areas
     - Rigorous multivariate statistical analysis of a large-scale, multi-component project to compare costs and benefits with alternative investment options

6. ***Time and resource constraints. Which of the following describes the current situation:***
   - There is a very tight deadline and no possibility of additional resources.
   - There is a tight deadline but additional resources could be obtained for use within this deadline.
   - The priority is to produce a high-quality product with a solid methodology which can withstand scrutiny from the critics of the project.

Note: All important threats to validity should be listed and a worksheet prepared for each one.

Part II: Analysis and discussion of each threat to validity

Threat [Number and name]

A. ***Level of generality:*** *Does the discussion refer to the overall policy or to certain components?*

B. ***How are the threats manifested in the evaluation ?***

C. ***What are the potential affects on the study findings and generalizations ?***

D. ***How big a problem is this for the evaluation ?***

E. ***Proposed actions***

F. ***How adequate are the proposed actions ?***

Annex 5

Example of a Completed Shoestring *Project* Evaluation Worksheet

Using the worksheet to assess an already completed evaluation of a low-cost housing project

Shoestring *Project* Evaluation Worksheet

Part I: Description of the Evaluation

Name of Evaluation: **Evaluating the Impacts of a low-cost housing project in a Central American city.**

1. ***Stage of the study at which the worksheet prepared:***
   - Evaluation design stage
   - Pilot testing of instruments
   - During data collection
   - Data analysis
   - Report writing XXX

2. ***Evaluation design***
   - Which evaluation design was used? *Model 2 No pre-test control group*
   - Explain any modifications to the designs in Chapter 2
   - Potential problems concerning: [see Chapter 5]
     - Baseline data
     - Control group: XXX *No control group*
     - Data collection methods and quality of data:
       + *Only household head interviewed*

+ *Potential biases in recalling/ reporting income*
+ *Only labor market earnings reported. Ignores transfer income and underestimates earning from informal sector.*

- Analysis of contextual factors: *Not analyzed.*
- Other _____

[*Pages may be added to describe any additional problems or issues not covered above*]

3. Objectives of the evaluation

- Why was the evaluation commissioned? *To assess impacts of housing on poverty reduction and household income.*
- What are the specific decisions or actions which will be taken on the basis of the findings? *Decide if project should be replicated on a larger scale.*
- Purpose of the evaluation:
  - Exploratory study to provide initial indications on whether the project model "works"
  - Assessing the efficiency and potential impacts of a small pilot project to recommend whether it is worth replicating on a larger scale
  - Rigorous multivariate statistical analysis of a large-scale, multi-component project to compare costs and benefits with alternative investment options

4. Time and resource constraints Which of the following describes the current situation:

- There is a very tight deadline and no possibility of additional resources.
- There is a tight deadline but additional resources could be obtained for use within this deadline. XXX *Some [limited] additional resources available but the report must be published within 3-4 months.*
- The priority is to produce a high-quality product with a solid methodology which can withstand scrutiny from the critics of the project.

Note: Complete a Part II sheet for each important threat to validity

Part II: Analysis and discussion of each threat to validity

Threat: ***2.3 History***

A. How manifested in the evaluation

The evaluation estimates the project produced a 70% increase in household income. But no control group used to define the counterfactual and estimate how income might have changed if there had been no project.

B. Potential affects on the study findings and generalizations.

The study findings may be completely wrong as the impact of the project on income may be much less than estimated due to external factors such as improved economic conditions in the city.

C. How big a problem is this for the evaluation

Potentially a very serious problem.

D. Proposed actions

Reconstruct control group through:

- *Secondary data*
- *Key informants [Chamber of Commerce, Ministry of Planning, local industries, community leaders, NGOs]*

Select control group comparison areas:

- *Rapid survey of current economic conditions and recall of past conditions*
- *Focus groups*
- *Observation*

E. How adequate are the proposed actions

The procedures can identify potentially important external factors and can provide a rough estimate of their importance. This will significantly reduce, but not eliminate the error.

Threat: 2.10 ***Inappropriate indicators and measurements***

A. How manifested in the evaluation

- *Earning of other household members may be under-estimated or ignored.*
- *Informal earnings may be under-estimated.*
- *Transfer income is ignored.*

B. Potential affects on the study findings and generalizations.

- *Household earning and total income likely to be significantly under-estimated.*

C. How big a problem is this for the evaluation

- *Potentially very serious.*

D. Proposed actions

- *Rapid resurvey of sub-sample of project households to:*
  - *Interview other household members about their income.*
  - *Estimate transfer income*
  - *Estimate informal earnings*
- *Focus groups and PRA to obtain independent estimates of income.*
- *Direct observation of economic conditions.*

E. How adequate are the proposed actions

- *Improve estimates of household earnings and total income*
- *Harder to obtain reliable estimates of income before the project but can improve estimates.*
- *The proposed methods will significantly improve the estimates but will only partially correct potential errors.*

Annex 6

Case Study

Longitudinal Impact Evaluation of the Eritrean Community Development Fund Feeder Road Component

FEEDER ROAD IMPACT EVALUATION[21]

<Summary>

Feeder roads were one of the two ECDF (Eritrean Community Development Fund) project components where a longitudinal impact evaluation was conducted. 3 of the 5 feeder roads were selected for the evaluation. As construction had already started on all the roads, the evaluation design involved a rapid baseline survey which was conducted when the projects were already underway, and was repeated between 3-6 months later when most of the roads were nearing completion. Between these two surveys, the road sites were visited several times and observation methods were used to note changes in traffic patterns and pedestrian travel and the opening of new businesses along the road route. The surveys were complemented by focus groups, key informant case studies, traffic pattern surveys, and case studies on a number of families and villages. Baseline information was reconstructed by asking families about their travel patterns, economic conditions, and agricultural production and consumption before road construction began. While no control group was identified, the counterfactual condition (what would have happened without the road) was estimated by obtaining information on food prices and agricultural sales in other local villages which did not have access to the new road.

The evaluation found that the road had significantly improved transport and mobility which had positive effects on access to hospitals, schools and other public services; marketing of agricultural produce; prices of consumer goods; and increased visits to friends and relatives.

[21] This case was prepared by Michael Bamberger from an unpublished Eritrean consultant report.

1. OBJECTIVES AND METHODOLOGY OF THE STUDY

Objective of the study

The objective of this study was to evaluate the social, economic and service impacts of the feeder road projects implemented by the Eritrean Community Development Fund (ECDF) during the period 1996 – 2000 with reference to:

(i) the social, economic and service impacts of the feeder road projects to beneficiary communities,
(ii) the benefits of ECDF's project implementation process on community participation,
(iii) accessibility of project benefits to the target villages,
(iv) the impact of the project on community capacity building, and
(v) the projects' sustainability.

Methodology

The evaluation covers three sample feeder road projects implemented by ECDF during the 1995-2000 period. In order to assess the benefits and active involvement of beneficiaries and stakeholders, and to also gain insights on the beneficiaries' perception of the feeder road projects, participatory rural appraisal, rapid assessment, beneficiary assessment and participatory evaluation methods were used. Stakeholders analysis and SARAR[22] techniques were to a limited extent also applied.

The impact evaluation can be described as follows:

	T_1	X Project	T_2	T_3
Project group	$[P_1]$	X	P_2	P_3
Control group				$[C_3]$

Where:

T_1 = time period before start of project.

[22] Self-esteem, Associative strength, Resourcefulness, Action planning and Responsibility (A participatory methodology for empowering stakeholders at different levels to assess, prioritize, plan, create, and evaluate initiatives. See McCracken and Narayan 1997 Module IV).

T_2 = time period during project implementation
T_3 = time period after completion of the project
P = project group
C = control group
[ ] = these observations [baseline data on project group and ex-post control group] had to be recreated as interviews were not conducted with these groups.

Four kinds of impact were assessed:

- *Process impact*: impacts resulting from how the project was implemented.
- *Accessibility impact*: who had access to project services and resources and who did not.
- *Social and economic impact*: impacts on income and employment, increased access to schools, health facilities, etc.
- *Sustainability impact*: whether or not benefits continued after project was completed.

(i) Sample design

Three feeder road projects were sampled from the five under construction to reflect regional variations and to permit generalizations to be made on the overall ECDF's feeder road program. The feeder road program was studied more intensively than the other ECDF components because feeder road projects have relatively high investment costs, long completion period, and have wider impact that cut across many communities.

(ii) Household survey

A socio-economic and impact assessment survey of 90 households was undertaken with focus on socio-economic conditions, community participation in project implementation, accessibility and use of project services, management of road repair and maintenance, impacts and sustainability. For this purpose a questionnaire was developed and 30 households were interviewed in each sample project. During the selection of households, due consideration were taken to ensure fair representation of settlement areas along the road and the catchments areas who use the road services. Once the users distribution was established (through the assistance of the local administration and consultant's

observations), a sampling fraction (n) was determined for each village to select the 30 households and the final household selection was then made on every nth household.. Extra care was taken to interview male and female respondents in order to ensure fair representation of gender views. In almost all cases, the household heads or their spouse were interviewed.

(iii) Focus group discussion

Focus group discussions were conducted at each project area. The discussion centered on assessing the impacts of the projects and evaluating community participation at the various stages of the project cycle, sustainability, problems encountered and how they were solved, and accessibility of the feeder road projects to different community groups. The discussion was designed to provide additional input from key stakeholders, interpret the results of household surveys, and to generate information that could not be obtained from the survey.

The local administrations and the consultants jointly made the selection of focus group participants. Typically the focus group comprised the Director of Nus-Zoba, Village Administrator, Project Coordinator, road maintenance committee members, users of the feeder road (at least 2), community elders (at least 2), representative of the National Union of Eritrean Women and representatives of the Ministries of Health and Agriculture.

(iv) Key informant interviews

The evaluation process heavily depended on the views, opinions and assessment of various groups who not only interacted during the project implementation but also are direct beneficiaries and are directly or indirectly involved or concerned on the project.

The consultants interviewed 10 key informants in each area, namely village leaders and administrators, road maintenance committee members, local health officers, community elders, bus drivers, school directors, shop keepers, farmers and female and male road users.

(v) In-depth follow-up interviews

This assessment technique was primarily used to capture information that has not been properly covered in the questionnaire. In each project 5 households were interviewed to understand their knowledge and opinions of the project impacts. The interviewees were selected randomly, but with due consideration

to age, sex, occupational and social groupings.

(vi) Case study

Community and household case studies were conducted along one road. The case study focused on the process of project implementation, impacts produced, sustainability and community capacity building.

The beneficiary assessment involved conversational interviews, in-depth interviews, group discussions and consultants' observation. Discussion groups, each consisting of 7 people were organized in three villages along the road to discuss issues mentioned above. For conducting the case studies three visits were made to the villages.

(vii) Longitudinal studies

Longitudinal studies were conducted on three villages along one road and cluster of villages along another. A total of 60 households were interviewed and visited three times during the longitudinal study period. (May, August and October).

(viii) Limitations of the study

The study was conducted during the Ethiopian invasion, and this was a major limiting factor on the evaluation because transportation is one of the most affected sectors during a war. However, the consultants concluded that the war and the other technical difficulties described below did not substantially change the general conclusions and judgments made on the feeder road projects.

(a) The border conflict with neighboring Ethiopia created huge socio-economic imbalances on village communities mainly because most of the work force was mobilized for the defense of the country and thereby reducing the economic power of every household. Moreover, meeting people with first hand information and who participated from the early stage of the project was quite difficult and at times impossible due to the war mobilization. This was true not only with community dwellers but also with local administrators.

(b) Household's tendency to hide income was quite prevalent. As a result responses regarding income should be interpreted with caution. Likewise getting information about surplus production was difficult because

farmers do not clearly remember the quantity they sold to the market. But as the excess production capacity of sample projects was quite limited, it does not have a substantial effect on the overall study.

(c) Although price fluctuations for both consumer goods and agricultural products were quite sporadic, attempts were made to give changes in both absolute and real terms.

(d) Capturing baseline information and successive impact information for the longitudinal study was not easy. For example, at times all transport vehicles (trucks and buses) were mobilized for the defense effort, which significantly affected mobility of people and commodities. In such instances going back to the project areas again was necessitated.

(e) Getting full cooperation of drivers for estimating volume of freight on trucks and buses was sometimes difficult.

2. QUANTITATIVE IMPACTS

For the feeder road projects there are different levels of impacts to be considered: on the beneficiary communities who use the roads; on trade and agricultural produce, which move to and from village communities; on government service and civil governance; and on social cohesion and the economy at large.

Table 1: Synopses of impact and outcome indicators

Indicators	Outcome
1. Process and implementation indicator	
% of population who knew about the project	89%
% of population consulted about site selection	64%
% of population involved in project implementation	74%
% of women involved in project implementation	61%
% of population who contributed in labor and money	81%
Organization of work as rated by respondents	Not good (96.7%)

Indicators	Outcome
2. Indicators of accessibility and use of project	
% of population served	87%
Reliable access to motorized transport	Yes
Volume of traffic generated	Average weekly motorized traffic 95
Diverted traffic from pedestrian to motorized transport	83.3%
People mobility increased	56%
Access to social services increased	83.3%
Users of public transport	Housewives 33% Farmers 31% Students 26% Traders 2% Others 8%
Major motorized vehicle users	Religious & NGO's 61% Commercial vehicles 23% Government vehicles 14% Private vehicles 2%
3. Economic impact of the feeder road projects	
Cost of transport decreased	No
Travel time decreased	118min/week (50.6%)
Transporting goods by trucks increased	35%
Access to wider range of consumer goods increased	500%
Prices of consumer goods decreased	No
Availability of agricultural inputs increased	300% (for horticulture farms)
New business established	21
Employment during construction increased	38%
Income generated during construction	1379 Nakfa/ household
Average employment/km	58.4
Increased business activities and turnover	No
Village surplus to other market increased;	Yes
Number of road construction firms increased	7
4. Indicators on social impact	
% of population who knew about road maintenance committee	33%
Health and education services increased	Yes
Number of health visits increased	Yes
Number of schools increased	100%
Learning environment for students improved	Yes
5. Indicators on sustainability	
% of population who contributed for maintenance	Nearly 100%
Quality of maintenance	good
% of people involved in road maintenance	70%
Community share in maintenance increased	10%
Communities' attitude on shared responsibility raised	7%

A. Impact of the process of implementation

Communities of the three sample feeder road projects had repeatedly requested the government for the construction of feeder roads. According to survey findings 88.9% of the community attended public planning and information meetings indicating knowledge about the project prior to its implementation. 64.4% were involved in site selection of the road, and 74.4% were involved in project implementation.

81.1% of the population contributed to the construction of the feeder road, either in labor or money, while most of the others did not participate due to old age or health problems. The majority of the respondents (61.1%) believe women participation was the same as men.Survey reports also confirm, that 96.7% of the respondents believe that the community labor contribution was not well organized due to weakness of the organizers. The main reason for community dissatisfaction was that some villagers had to travel relatively long distances for labor contribution while other were able to work close to their villages.

In general, from focus group discussions and key informants, it was realized that community participation in the project implementation process was a learning process that enhances communities' action planning, responsibility, organization and problem solving capacity, which resulted in an observable improvements during subsequent road maintenance works.

B. Social impacts

(i) ***Beneficiaries:*** The total population of the project area[23], which is estimated at 14,188, will directly or indirectly benefit through increased access to social services, markets, travel, income, and employment created or facilitated by the construction of the road. In other words, all social groups, poor and rich, male and female, farmer and trader, young and old, etc. are direct or indirect beneficiaries of the road.

The major users of the public transport are housewives (33%), farmers (31%), students (26%), traders (2%) and others (8%). Motorized vehicles users or beneficiaries of the road are religious institutions and NGO's (61%), commercial vehicles[24] (23%), government vehicles (14%) and private vehicles (2%).

[23] Unless specified, benefits expressed hereafter are for the completed feeder roads (i.e. MTR and GR).

[24] Public transport, freight transport, hired vehicles and business vehicles, etc.

(ii) Increased access: One of the major impacts of the construction of feeder roads is increased access and ability of people to move to and from their villages. Comparison of baseline data and survey results as well as discussion with community members confirm that the feeder road have increased people's mobility for social visits, health, school, and markets. In Maeraba Tekelabi Road (MTR) average weekly[25] traffic generated from motorized transport is about 95. According to survey findings, after the project average travel per person per year has increased by 56% for the Gullie Road (GR) and MTR (table 2).

Social visits are an important part of Eritrean social life. The feeder road project has greatly enhanced social intercourse and profoundly affected the social welfare of the communities. Due to the road, communities who were far apart and seldom meeting are now visiting each other frequently. Social visits account for 26% of the public transport passengers. The communities report that even when transport vehicles are not available or unaffordable, the roads have decreased travel time for pedestrians and thereby increasing social visits and students ability to continue their schooling.

Table 2: Change in mobility

Description	Before the project			After the project		
	GR	MTR	Average	GR	MTR	Average
Average travel per person/ year	14.55	16.78	16	25.2	25.5	25

(iii) Impact on women: Survey findings reveal that 53% of the respondents believe that the road has benefited women through creating easy access to health facilities. This view is supported by the fact that over 70% of the health visits in Gullie and Hadida health facilities are women and children. However, the benefit is not only limited to health services but to other sectors of the social services and benefits. For example 33% of the public transport passengers are housewives who travel for various reasons, such as shopping, social visits, trade etc. Hence, with regard to access, the impact of the feeder road on women is much wider than reported by the survey.

[25] As public transports are only available during weekends and the majority of the people travel during this period, it would be more appropriate to talk about weekly-generated traffic.

C. Economic impact

The potential impacts of the feeder roads were probably reduced due to the war with Ethiopia which was ongoing at the time of the study.

(i) ***Local transport and travel:*** After the construction of the road the average travel per person/year has increased by 56%. Before the project only contracted vehicles used to come once or twice in three months for funerals. Vehicle traffic counts conducted during the longitudinal study showed the average weekly traffic of 95, of which 55% were on weekends. In other words 95 vehicles moved to and from the villages per week giving wide access of transport use for communities along the line and other people who went to travel to this area.

Table 3: Time saved on transporting goods to the market (week)

Type of transport	Time spent (hrs) MTR	
	Before project	After project
Carrying on the back	4	12.75
Using animals	88.75	-
On the back & using animals	-	2.25
Human back, animals & trucks	-	2
Trucks	24	40.5
Total time spent per sample population	116.75	57.5
Average time spent/household	3.89 hrs/week	1.92 hrs/week
Average time saved in minutes		118
Time saved in %		50.6

(ii) ***Small business and commerce:*** Immediately after the construction of the roads and prior to the border conflict many small enterprises started to sprout along GR and MTR. But due to the war mobilization, some of the new enterprises were closed. The table below shows the new enterprises established after construction of the new roads.

Table 4: Private small business and services (before and after the project)

Description	Maereba Tekelabi Road (MTR)		
	Before project	After project	% change
Shops	7	14	100%
Tea house	-	1	100%
Flour mills	2	4	100%
Agriculture implement distribution	-	1	100%
Water supply system	-	1	100%
Horticulture farms	6	14	133%
Hand loom weaving	-	1	100%

(iii) Access to wider range of consumer goods: The majority of the households (70%) surveyed rank availability of consumer goods in shops after the project as good, with an increase in the variety of consumer goods by 500%. Likewise the shopkeepers said their shop supply has increased in variety and quantity due to the availability of motorized transport and easy access, but complain that people tend to buy from the major towns of Asmara or Dekemhare rather than the local market. The reason for their complaints could either be because of easy access of transportation, people might prefer to do their shopping when they go to the major towns for a marginal price difference; or the local shopkeepers' prices are not very competitive. But in general communities' access to wide range of commodities has increased. According to the survey the following items were reported to have appeared for the first time in the market since the opening of the new roads: *soft drinks, beer, lentils, hot spice, pasta, rice, onion, tomato paste, sorghum, wheat, sugar, coffee, bread, shoe, cement, metal sheet, nail, biscuit, and laundry soap.*

(iv) Prices of consumer goods: Generally, prices of goods in the project areas have increased between 30% -118% in absolute terms, while in real terms it ranges from -9.3% to 34.1%. For example along the MTR prices of the basics have increased in absolute terms. These increases, however, are due to the general economic conditions. It appears that had the roads not been built, the level of price increase would have been much higher than what is now recorded. Table 5 shows the real percentage changes of prices for selected commodities. The difference between the absolute and real price changes is attributed to the feeder road indicating the impact of the road in reducing or suppressing the

inflationary price hikes. As seen on the table below the price for fuel, which is government regulated, has declined in real terms indicating the impact of the road, whereas the price of other consumer good has shown increase but relatively lower as compared to the absolute increases.

Table 5: Consumer price changes in MTR (1996 – 2000)

Items	Baseline Year		Impact study year		% change absolute	% change real
	Range	Average	Range	Average		
Sugar (kg)	2.00 ~ 5.50	3.48	4.00 ~ 5.50	4.51	30	+ 23.5
Tea (100gm)	1.50 ~ 2.50	1.86	2.50 ~ 5.00	3.40	82.8	+ 32
Salt (kg)	0.25 ~ 1.50	0.66	0.50 ~ 1.50	0.98	48.5	+ 7.14
Coffee (kg)	12.00 ~ 35.00	21.41	24.00 ~ 40.00	29.69	38.7	+ 18.15
Edible oil (liter)	4.00 ~ 15.00	10.04	12.00 ~ 20.00	13.36	33.1	+ 23.8
Kerosene (liter)	0.46 ~ 2.00	1.29	0.96 ~ 3.75	2.81	117.8	– 9.3
Laundry soap (pcs)	0.25 ~ 2.00	1.01	1.25 ~ 3.00	1.67	65.3	+ 34.1

(v) ***Impact on agriculture:*** Case studies on horticulture farms clearly indicate that fruits and vegetables, which perished frequently due lack of transportation, have easy access and thereby induced increase in production by about 300% to 500%. This had significant benefits not only for farm owners but also for the temporary employees in the farm and transport vehicles that work in this area. With regard to horticulture farm two major positive impacts are noted. First in contrast to earlier times where farmers had to carry their produces on pack animals to the market and spoilage accounted for about 20% to 30% of what reached the market, now trucks come and load fruits and vegetables from the farm. Secondly, farmers have, due to the road, a better bargaining edge over middlemen who come to buy farm produces. In general, despite price increases in absolute terms, farmers clearly see the benefits brought by the road.

(vi) ***Generated employment and income:*** Direct employment was generated during the construction and maintenance works of the feeder roads. Indirect

employment, though very difficult to capture, were also generated. Obtaining secondary data on actual employment of community members during the construction and maintenance period was difficult. Survey reports, however, indicate that 38% of the households were employed during the road construction for an average of 71 days with an average payment of 19 Nakfa/day. Accordingly, the average household income from employment in construction work is about 1379 Nakfa/household. Table 6 shows the direct income and employment generated by the feeder road. The total employment for the sample population is 1633 person/days.

Table 6: Employment and income of sample population during road construction

Description	MTR	GR	Total/average
Number of households employed	10	13	23 (38%)
Average days worked	110	41	71
Total human/days employed	1100	533	1633
Average payment (Nfa/day)	22.76	12.68	19.47
Total payment	25,036	6,758	31,794
Average income/household	2,504	520	1,382
Average number employed/km	40.7	76.1	49.5

Changes have also occurred in the establishment of new business (mostly household run businesses) and in increased motorized transport and increased surplus of horticulture products. Despite the fact these changes have generated some short-term and seasonal employment, significant permanent employment increases were not, however, reported or observed mainly because of the conflict situation.

3. QUALITATIVE IMPACTS

The qualitative impacts of the feeder road projects, as perceived by the beneficiaries, were obtained from focus group discussions, key informants, in-depth interview, observations and survey results.

A. Impact of implementation process

Although the feeder road committees evolved out of the traditional

committees, the intensity, responsibility, and procedure of performing their tasks changed significantly. The feeder road maintenance committees are required to periodically report the number of people involved, type and magnitude of work performed, payment made, etc. As a result committee work has become not only time consuming but also more transparent than traditional committees which were normally a source of disputes among communities.

Another major impact of the project implementation process was that community participation from the initial stage of the project implementation has instilled a sense of responsibility and influenced frequent discussions on problems or issues about the road with their "baitos[26]". Cases in point are the communities in Geleb frequently asked and demanded the road construction should be expedited and in Maereba they have requested that something has to be done with the steep gradient that has caused the infrequent bus traffic, etc. Hence one of the major impacts of ECDF's project implementation process is the opportunity it created for communities to participate in the decision making of community based projects.

B. Impact on provision of public services.

Communities report that government services have increased both in quality and quantity wise and visits by different level of government officials has increased. The construction of the roads has not only enabled the government to deliver social services but also has reduced the social cost of delivering it. Along the completed Gullie and Maereba Tekelabi roads, the government has built a health center in Hadida and an elementary school and a health station in Gullie.

The new health center and health station are changing the health condition of the local population. As a result, sick people who were being carried by stretchers or horse back on steep terrain are now easily transported by car or road with considerable improvement in travel speed and convenience. It was reported that many people, particularly pregnant women, used to die on this journey. Access to health care and improvement in family health have increased tremendously.

Daily heath attendance has on the average increased from almost nil to 215 for Hadida and 85 for Gullie. However, baseline year frequency of health visits could not be established mainly because respondents' responses were so

[26] baito: traditional community leader

unreliable. But what is clearly established is that before the road they used to only visit clinics for emergencies while now they visit health centers for minor symptoms of sickness and even for check-ups (such as pregnancy check ups).

Likewise, the impact of the road on education was quite significant. Building a new primary school in Gullie was possible due to the road. 8% of MTR's public transport passengers are high school students who shuttle for the weekends from Dekemhare. Although increased school attendance in the schools could not be clearly established, key informant in Maereba and Gullie report that the road has improved the learning environment in several different ways.

C. Environmental impacts

The construction of feeder roads has contributed for the degradation of land and the clearing of shrubs and trees. Particularly during construction, the scale of destruction requires environmental regulations mitigating destruction of trees.

In all the visited sample feeder road projects, it has been observed that a new, previously unknown plant is being widely spread. The plant is observed only few meters from both sides of the roads and is neither eaten by animals nor can be used for any domestic activities or otherwise. The communities along the new roads are also dumbfounded about the spread of this plant. It is, however, too early to say whether this is an environmental problem, but one thing is sure that the spread of this plant is caused by the road and there is no economic benefit for the communities. The consultants highly advise that the respective ministry undertake a detail scientific study before this gets out of hand.

Annex 7

Case Studies

Three Approaches to Evaluating the Gender Impacts of Micro-Credit Programs in Bangladesh: Different Evaluation Paradigms Produce Different Findings[27]

<Summary>

Micro-credit is generally considered to be one of the most effective ways to promote women's economic, social, and political empowerment. Bangladesh is usually seen as one of the countries where micro-credit programs have had the greatest impact on the status of women. This paper discusses three different studies which have all assessed the impacts on women and their families of the same micro-credit programs in Bangladesh. Each study addresses a different set of issues, and uses a different evaluation design to test a different set of hypotheses. The comparison of the three studies provides an almost unique opportunity to understand how evaluation theory determines the definition of hypotheses, the selection of indicators and the choice of data collection and analysis methods. The first study, comparing household survey data for communities where the micro-credit programs were operating and for control communities, found that women's access to micro-credit has a significant positive impact on a range of indicators of household welfare. The second study drew much more pessimistic conclusions on the impacts of credit on women's empowerment. Using reconstructed household histories, the study found that women exercised significant control over the loans in less than 40 per cent of the cases. The third study, using an 8 point empowerment index and taking a longer time perspective, came to a more optimistic conclusion concerning the impacts of credit on women's empowerment.

[27] Based on a paper presented by Michael Bamberger at the American Evaluation Association Annual Meeting in Washington D.C. November 6, 2002

The potential contribution of micro-credit to women's economic, social, and political empowerment

Reducing poverty around the world will take multiple interventions and strategies. One strategy that has shown great promise is micro-finance. For women who may have limited skills and access to employment, self-employment is a way to increase the income of their families. However, obtaining small amounts of money to start a small business or develop an existing small enterprise is difficult. Banks and other lending institutions typically focus on large scale loans and require collateral; people must be credit worthy. For low-income people, obtaining small loans are out of the question from traditional banks. Government programs that subsidized loans to the poor resulted in other problems and have not been effective. However, the concept of micro-credit has become one strategy that appears to work. Not only does it offer small loans to poor people, it enables them to become aware of how to run small businesses.

The micro-credit lending is targeted to landless/assetless borrowers. Small groups are formed which meet regularly. Each participant in the group takes joint responsibility for repayment and each contributes to the common account through savings. Each could borrow from that and it is expected that each would pay back their loans with interest. Loans are collateral-free and usually have a maturity of 50 weeks. Small loans are given initially and larger loans are given to repeat borrowers if their repayment performance is satisfactory. People will have access to credit for 8-10 years in order to accumulate enough assets to escape poverty. Borrowers choose the activity to be financed; these include small-scale gardens; small-scale production of milk, cheese, eggs; and handicrafts. For example, money that is borrowed can be used to buy a cow and a goat that result in income by selling milk, cheese and butter. Or it could be used to buy needed equipment, seeds, or hiring additional employees in order to create new businesses or increase the capacity of their current small business or enterprise.

The goals of micro-credit programs are to raise individual incomes of people and the gross national product per capita. With increased income, consumption increases; this provides income to others in the community who are able to sell their goods and services. It is also intended to improve the status of women within their households and the quality of lives of their children as well.

Case 1: Evaluating the impact of micro-credit on women's economic status and capital formation

<Summary of the Case 1>

The first evaluation uses a cross-sectional design to compare households where women and their families were targeted for micro-credit programs with households not targeted. Program effectiveness was defined in terms of the following household-level outcomes: per capita spending, net worth, boys' and girls' school enrollment, boys' and girls' height for weight, contraceptive use and recent fertility. These outcomes were compared for female borrowing and male borrowing, and it was found that many of the variables relating to household welfare were affected more by female borrowing than male borrowing. Women obtaining loans were also associated with female capital accumulation.

The study

The purpose of the studies was to examine the gender-differentiated impacts of female and male borrowing from three micro-credit programs in Bangladesh on a range of household welfare indicators including income and assets, nutrition, school enrollment, fertility behavior and contraceptive usage, and empowerment. The micro-credit programs studied were the Grameen Bank, the Bangladesh Rural Advancement Committee (BRAC) and the Rural Development 12 (RD-12) project of the Bangladesh Rural Development Bank.

Methodology

The evaluation is based on a 1991-92 Household Survey conducted by the Bangladesh Institute of Development Studies. The sample covered 29 randomly selected *thanas*[28] from the 391 *thanas* in Bangladesh (with *thanas* affected by the 1991 cyclone being excluded). A total of 24 of the *thanas* had at least one of the three micro-credit programs operating while five had none. Several thanas had more than one micro-credit program operating but no household was a member of more than one. A total of 1798 households were selected using stratified random sampling. 1538 were *target* households (in communities with one of the micro-credit program operating, of whom 905 were participating in one of these programs). The remaining 260 were *non-target* households. The evaluation design can be represented as follows:

[28] A "thana" is the administrative center for a number of villages

Cross-Sectoral Comparison of Household Surveys of Borrowers and Non-Borrowers Using Statistical Controls to Adjust for Sample Selection Bias.

	T_1	Intervention (X)	T_2
Project group (P)		X ———	P_2
Control group (C)			C_2

Where:
T_2 = time period after families had received loans
P_2 and C_2 = observation of project and control groups after the project intervention (loans approved and used)

A detailed household questionnaire covering income, employment, education, consumption, borrowing, asset ownership, savings, children's schooling, fertility behavior and contraceptive use was administered to all households. For the 315 household included in the nutrition survey, anthropometric data was also collected. A village survey questionnaire was also administered to collect information on crop prices, fertilizers, wages for men, women and children, access to credit markets and access to roads and public services.

Impact assessments were based on cross-sectional analysis comparing households that did and did not use micro-credit programs with respect to the impact indicators (see table 1). Econometric methods were used to correct for differences between target villages and non-target villages and between borrowers and non-borrowers with respect to attributes such as wealth, land-holding, etc., likely to be correlated with the impact indicators. The analysis found that target villages were on average wealthier than non-target villages, and adjustment for these differences reduced in many cases the magnitude of the estimated program impacts, although in most cases they remained significant.

Findings: micro-credit programs have different impacts on female and male borrowers.[29]

Two related studies examine the impact of female and male borrowing –

[29] The findings section of this case study is taken directly from Engendering Development (cited below) pp. 160-162 with additional material from Shahidur Khandker, 1998 (cited below) p.12.

from Grameen Bank, the Bangladesh Rural Advancements Committee, and government program RD-12 on such outcomes as per capita household expenditure (income and girls' and boys' schooling and nutritional status (Khandker 1998; Pitt and Khandker 1998). The impacts often differ substantially based on whether the borrower is a woman or a man – and often the marginal impacts of borrowing are greater for women than for men.

For all three micro finance programs the impact of female borrowing on per capita household expenditure (income) is about twice as large as the impact of male borrowing (Table 1). A 10 percent increase in female borrowing is associated with a roughly .4 percent increase in per capita expenditure – an effect that is strongly significant statistically. Compare this with a roughly .2 percent increase in per capita expenditure associated with the same percentage increase in male borrowing. Female borrowing also has a greater impact than male borrowing on households' ability to "smooth" consumption over time (Khandker 1998).

Women also benefit from program participation through the cash income generated by self-employment and the assets they acquire in the process. Estimates indicate that micro-finance reduces poverty among program participants and reduces aggregate poverty in program villages (even after controlling for observable village characteristics that partially determine the extent of village poverty).

As with other forms of resource control, female borrowing also appears to have a greater impact on children's welfare than male borrowing does. For example, except for BRAC, female borrowing has a greater positive impact on children's school enrollments than male borrowing does. Moreover, in contrast to male borrowing, female borrowing has a large and statistically significant impact on children's nutritional well-being.

At the same time, male borrowing has a greater impact on household net worth than female borrowing. This suggests that while at the margin women seem to invest relatively more than men in the human capital of their children, men appear to invest more than women in physical capital.

Female and male borrowing also have different impacts on household reproductive behavior, suggesting that women and men do not share the same preferences relating to contraception or fertility. For example, female borrowing decreases contraceptive use and, except for Grameen Bank borrowing, increases fertility, whereas male borrowing increases contraceptive use and, except for

BRAC borrowing, decreases fertility. At first glance the findings on the impact of female borrowing on contraceptive use may seem counterintuitive, since a body of empirical literature suggests that factors increasing the opportunity cost of women's time – tend to reduce fertility. But low-income women in Bangladesh may see additional children as assets capable of assisting them with what are often home-based, self-employment activities.

Table 1: Impacts of female and male borrowing on selected household outcomes in Bangladesh

(Percentage change for a 10 percent increase in borrowing)

Household outcome	Grameen Bank		BRAC		RD-12	
	Male borrowing	Female borrowing	Male borrowing	Female borrowing	Male borrowing	Female borrowing
Per capita spending	0.18	0.43	0.19	0.39*	0.23*	0.40*
Net worth	0.15*	0.14*	0.20*	0.09*	0.22*	0.02
Boys school enrollment	0.07*	0.61*	-0.08	-.03	0.29	0.79
Girls' school enrollment	0.30	0.47*	0.24	0.12	0.07	0.23
Boys' height for age	-2.98	14.19*	-2.98	14.19	-2.98	14.19
Girls' height for age	-4.92	11.63*	-4.92	11.63*	-.492	11.63*
Contraceptive use	4.25*	-0.91*	0.40	-0.74*	0.84	-1.16
Recent fertility	-0.74*	-0.35	0.54	0.79*	-0.74*	0.50

* indicates coefficient estimate which is statistically significant at the 10 percent level or better.
Source: Khandker 1998 cited in World Bank 2001.

Increasing women's access to credit also empowers them in other dimensions. For example, female borrowing increases female control of non-land assets (Pitt and Khandker 1998; Khandker 1998).

Lessons for the evaluation of gender impacts

The study provides a good example of how one can plan survey design and data collection to study gender differences in program impacts. It also emphasizes the importance of assessing potential sample selection biases, and shows how this can be done through using econometric methods to control for differences in household characteristics such as income, labor force

participation, education and household size which may be correlated with the outcome (impact) indicators. It should, however, be pointed out this kind of cross-sectional analysis does not address many of the "Threats to Validity" of Quasi-Experimental Designs (Valadez and Bamberger, 1994, Box 8.1) such as local history, political interference, interaction between project and local context. The design also does not address the specific problems of gender impact assessment discussed in this chapter such as potential biases or omissions concerning information collected from, or about women.

Case 2: Who takes the credit? Assessing the extent to which women in Bangladesh control the use of the micro-credit loans they obtain

<Summary of the Case 2>

The second study, which used a purpose sample designed to include different types of female borrowers, addressed the question of the degree to which women actually controlled the resources from the loans they had obtained. Using an historical analysis to obtain information on the degree of control exercised by women at each stage of the loan approval and use, it was found that women completely controlled the use of the loan in less than 20% of cases, and exercised substantial control in less than 40% of cases. The authors emphasized that in the social context of rural Bangladesh it would be extremely difficult for a woman to create and manage a small business completely independently (it would be particularly difficult for her to directly market produce), and it is reasonable to assume that most married women would try to use the loan to improve their status within the household rather than trying to achieve economic independence. Consequently the lessons concerning the contributions of credit to women's economic and social empowerment must be drawn with care.

The study

The purpose of the study was to challenge the frequently stated assumption that women's obtaining and repaying loans is a good indicator of the role of micro-credit in promoting women's empowerment. The study sought to estimate the degree of control which women actually exercised over the loans they obtained and the implications this has for a fuller understanding of empowerment.

Methodology

A purposive sample was selected of women who had obtained loans from four micro-credit programs in Bangladesh [N=253]. The sample was selected to include a variety of group and loan characteristics, such as years of membership in the credit program and size of loan, and the marital status of the women. Loan histories were obtained on all of the women with a range of questions about women's control over the productive process. For example women were asked what activity they invested in, where the inputs and productive assets came from and who procured them, what they cost, how they were put to use, where outputs were marketed, for what price, what were the problems involved in the productive process, who the main user of the loan was in terms of labor input, and in terms of controlling accounts and general management. The evaluation design can be represented as follows:

A purposive sample of women borrowers from 4 credit programs using recall to assess women's participation in decision-making and control over use of loans.

	T_1	Intervention (X)	T_2
Project group (P)	$[P_1]$ ———	X ———	P_2
Control group (C)			

Where:
T_1 and T_2 = time periods before the projects began and after women had received loans respectively.
$[P_1]$ = baseline information recreated through recall.

No control group was used because the study was not testing a hypothesis but rather focusing exclusively on women who had received loans.

On the basis of these questions an index of loan control was developed:

- FULL = full control over the entire productive process, including marketing
- SIGNIFICANT = control over every aspect of the productive process with the sole exception of marketing
- PARTIAL = loss of managerial control over the productive process,

but the provision of substantial inputs of labor
- VERY LIMITED = minimal input to the production process
- NO INVOLVEMENT = cases where women provided no labor for activities which are culturally ascribed as masculine

The study relied heavily on recall to obtain information on how the loans were managed and the authors stress the reliability issues inherent in this method.

Findings of the study

The study found the following percentages of control by the women borrowers:

- Full control 17.8%
- Significant control 19.4%
- Partial control 24.1%
- Very limited control 17%
- No control 21.7%

The initial conclusion is that women retain full control of less than 20% of the loans [17.8%] and at least significant control in less than 40% [37.2%] of the loans. These figures clearly indicated that borrowing and loan repayment cannot be automatically equated with women's empowerment without a fuller understanding of the dynamics of loan control within the household.

The authors emphasize that the figures must be interpreted within the social context of rural Bangladesh where it is virtually impossible for a women to retain complete control over all stages of the productive process as social controls limit her geographical mobility and her ability to directly market goods that she has produced. This is evidenced by the fact that almost all of the women who retained full control of the loans were either divorced or widowed. They also argue that in a context such as rural Bangladesh, where a woman's economic and social welfare and physical security is almost exclusively defined by her ability to maintain a satisfactory marriage; women must be expected to use a tool such as credit to strengthen their position in the household rather than to seek economic independence.

Lessons for the Evaluation of Gender Impacts

The findings clearly demonstrate the need to broaden the range of indicators used in the evaluation of the impacts of micro-credit on the welfare of women.

The technique of historical analysis, in which subjects provide detailed information on how the loan was obtained and managed is shown to be a useful tool for studying the degree of women's participation at each stage of the loan process.

One potential weakness of the methodology is that the research relies exclusively on information provided by women. Within the cultural context of rural Bangladesh it may be difficult for women to speak freely, particularly with respect to issues such as the control of a loan which could be perceived as a criticism of her husband. Consequently there is some danger of bias in the information provided. The findings could have been strengthened through the use of triangulation whereby other independent sources would be consulted (such as other female household members, neighbors or members of the credit banks) to provide a consistency check on the information.

Case 3: Evaluating the impact of micro-credit on women's empowerment

<Summary of the Case 3>

This study, which also adopts an empowerment approach, assesses the impacts of micro-credit programs in Bangladesh on a number of dimensions of empowerment. The study combined an ex-post sample survey with comparison group and the preparation of longitudinal case studies (over a four year period) on six villages. Based on observation and informal interviews, eight indicators of empowerment were identified and transformed into ordinal scales. It was found that participation in the Grameen and BRAC credit programs had a statistically significant impact on women's: mobility, ability to make purchases and major household decisions, ownership of productive assets, participation in political activities and protests, and legal and political awareness. It also reduced women's exposure to domestic violence.

1. The Study

The study compared two programs providing micro-credit to women in Bangladesh in terms of the impact of the programs on women's empowerment.

2. Methodology

The study combined the following data-collection methods:

- A six-village ethnographic study conducted between 1991 and 1994. The study combined participant observation with informal interviews. The researchers also observed and documented the two credit programs and interviewed a sample of program participants. The study focused on a quasi-random sample of 120 families – 20 from each village of which half participated in one of the credit programs and half did not. A structured "household survival matrix" was administered to all households once a month over a period of a year to collect information on economic activities and earnings, as well as household responses to crises and economic stress. Information was also collected on the performance of the women's micro-enterprises and women's control over household enterprises and income.
- A sample survey of approximately 1,300 women under the age of 50 was conducted. This included both credit program participants and comparison groups of women living in the same areas but not receiving credit. The sample of communities was selected purposively to include communities where BRAC and Grameen Bank had been operating for at least 6 years and also communities where they had only recently started to operate. Communities where both programs were operating were avoided. The issue of selection bias was addressed through statistical comparisons of household characteristics of project participants and control groups. In general the differences were not found to be too large although Grameen Bank members were likely to have more education.

Eight indicators of empowerment were identified on the basis of observation, personal interviews during the ethnographic studies and information from the baseline study. The operational definition of the indicators tried to take into account specific characteristics of each community such as distance from the market, configuration of roads and tracks [affecting women's mobility] and the types of goods which were sold door to door [and which women could purchase without breaking purdah[32] rules affecting their mobility]. For each

[32] "Purdah" is defined as the Hindu or Muslim system of sex-segregation, practiced especially by keeping women in segregation.

indicator the definition of empowerment was based on the top 25% - 30% of responses. Each of these was developed into an ordinal scale:

1. Mobility
2. Economic security
3. Ability to make small purchases
4. Ability to make larger purchases
5. Involvement in major decisions
6. Relative freedom from domination by the family
7. Political and legal awareness
8. Participation in public protests and political campaigning

The eight indicators were combined into a composite empowerment indicator. A woman was defined as empowered if she had a positive score on at least five indicators.

The evaluation design can be described as follows:

Combining a sample survey with a comparison group and longitudinal village case studies

	T_1	Intervention (X)	T_2
Project group (P)	P_1	X	P_2
Control group (C)			C_2

Where:
T_2 = time period after families had received loans
P_2 and C_2 = observation of project and control groups after the project intervention (loans approved and used)

Findings

Using logistical regression analysis, it was found that participation in the Grameen Bank and BRAC credit programs had statistically significant impacts on women's:

- Mobility

- Ability to make purchases
- Ability to participate in major decisions
- Ownership of productive assets
- Legal and political awareness
- Participation in public campaigns and protests
- Reduced vulnerability to domestic violence

4. Lessons for the design of impact evaluations

The study shows the value of ethnographic and similar in-depth qualitative methods for developing complex indicators such as those used to measure empowerment. The study also emphasizes the need to combine standard measures with the need to reflect the unique characteristics of each community [such as the amount of goods which women can buy without having to break purdah]. The study also demonstrates the value of multi-method approaches and the ways in which quantitative and qualitative methods can complement each other in the study of complex and culturally sensitive questions such as empowerment.

Conclusions from a comparison of the three studies

Each of the above studies comes to very different, but all methodologically valid, conclusions about the impacts of access to micro-credit on women and their families in Bangladesh. The comparison of the three studies demonstrates the importance of a clear definition of the explicit or implicit theory model [Weiss 2001] on which the evaluation design is based. Even when a study is not explicitly using a Theory Based Evaluation approach, the implicit theory model defines the key assumptions on which the evaluation design is based, and this in turn identifies the hypotheses to be tested and the indicators to be developed.

Each of the three studies is based on different implicit or explicit assumptions. The first study, implicitly assumes that women automatically benefit from access to credit, and that the benefits can measured in terms of a set of social and economic indicators of household welfare. The second study is based upon the assumption that in a society such as Bangladesh it is likely that women will not be able to fully control the loans authorized in their names, and that this reduced control will significantly reduce the empowerment impacts of the program. The concern with control of the loan meant that the study did

not seek to assess the welfare benefits to women and their households of the loan. The third study is based on the assumption that the process of women's empowerment will be more subtle and gradual, and that even when women do not fully control the loan, the negotiation process of obtaining it will start a gradual process of empowerment which can be manifested both by women's increasing (albeit modest) control over household resources, and also by their level of participation in community and local political activities.

The very interesting findings produced by each of the three studies, suggest the great potential of a multi-method approach combining the theory models and research methods used in each of these studies.

References cited in the text (Annex 7)

Baker, Judy. 2000 *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Annex 1.2 "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh" Directions in Development. The World Bank.

Goetz, Anne Marie and Rita Sen Gupta. 1996. "Who Takes the Credit? Gender, Power, and Control Over Loan Use in Rural Credit Programs in Bangladesh." World Development. Vol. 24. No. 1 pp 45-63.

Hashemi, Syed Sydney Ruth Schuler and Ann P. Riley. "Rural Credit Programs and Women's Empowerment in Bangladesh." World Development Vol 24 No. 4 pp 635-653 1996.

Khandker, Shahidur. 1998. *Fighting Poverty with Micro-credit: Experience in Bangladesh.* New York. Oxford University Press for the World Bank.

Pitt, Mark; Shahidur Khandker, Signe-Mary McKernan, and M. Abdul Latif. 1999, "Credit Programs for the Poor and Reproductive Behaviour in Low-Income Countries: Are the Reported Causal Relationships the Result of Heterogeneity Bias?" *Demography* 36(1) 1-21

Pitt, Mark and Shahidur Khandker. 1998. "The Impact of Group-based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy*. 106: 958-96.

Weiss, Carol. 2001. "Theory Based Evaluation: Theories of Change for Poverty Reduction Programs" in Osvaldo Feinstein and Robert Picciotto (editors) *Evaluation and Poverty Reduction.* World Bank Series on Evaluation and Development. World Bank. Washington D.C.

World Bank. 2001. *Engendering Development through Gender Equality in Rights, Resources, and Voice*. pp. 160-162. A World Bank Policy Research Report.

FASiD

Foundation for Advanced Studies on International Development