

AEA Professional Development Session:

Impact Evaluations When Time and Money are Limited

Lessons from International Development on the Design of rapid and economical, but methodologically sound, impact evaluations.

November 5, 2002

Handouts 1, and 3--8

- 1. The Growing Demand for Rapid and Economical Impact Evaluations**
- 3. Approaches to the design of low-cost impact evaluations**
- 4. Rapid and Economical Methods for Impact Evaluations**
- 5. Introduction to the theory and practice of impact evaluation design**
- 6. Threats to the validity of interpretations about program impacts.**
- 7. Shoestring Project Evaluation worksheet**
- 8. Case Studies for Group Exercises: Three Approaches to Evaluating the Gender Impacts of Micro-Credit Programs in Bangladesh**

Handout 1

The Growing Demand for Rapid and Economical Impact Evaluations

Michael Bamberger

Evaluation textbooks are frequently based on the premise that policy makers and program managers have at least a minimal interest in having their programs evaluated and that some provision will be made for the evaluation when the project or program is being planned. While it is recognized that the evaluator will frequently encounter opposition, lack of data or limited access to key actors, and that budgets are always limited; it tends to be assumed that the evaluator will be called in relatively early in the project and that most key actors accept, perhaps reluctantly, the benefits from (or need for) program evaluation.

Evaluators brought up on these text books are likely to have quite a shock when they are asked to help design or implement an impact evaluation in many developing countries. They will often not be called-in until the project is well underway, and even nearing completion, and they will frequently find that the project did not include any plans for generating baseline data or identifying a control group. On top of this many of the key stakeholders may not accept the utility of evaluations and may perceive them as of little practical utility or even threatening; and the monitoring and evaluation system, if one exists, is often considered as placing too heavy a time and resource burden on the organization. Finally many groups may consider that the M/E system has been imposed by the external donor and is designed to respond to the donor's information needs rather than to those of the national agency [Bamberger 2001, Horton and Mackay 1999].

Even in the increasing number of cases where local agencies are interested in evaluation formidable challenges can still be encountered. These include: (a) no attention was paid to evaluation until the program was well advanced with the consequence that there is no systematically collected base-line data, no definition of a control group and the objectives of the program may not have been clearly defined; (b) only a very modest budget has been allocated for evaluation studies; (c) the project is nearing completion so that there is considerable time-pressure to complete the evaluation making it difficult to collect longitudinal data; (d) little secondary data has been collected; and (e) there is only a limited pool of local evaluation expertise.

It should be added that this situation should not be attributed exclusively to a lack of understanding or motivation by national agencies, as a similar situation will often be faced when conducting evaluations for international development agencies. A major part of the problem is the low priority given to impact evaluation by most of the major actors involved in international development including national governments, international development agencies, and to quite a large extent NGOs [Bamberger 1995, 2000].

Despite these constraints there is an increasing demand to assess the effectiveness of projects in achieving their objectives and to estimate the impacts they have produced on the intended beneficiaries. Interest in impact assessments has also received a major boost from the international debt reduction initiative for highly indebted poor countries (HIPC), where the approval of future debt reduction is based in part on an assessment of the effectiveness of initial HIPC supported poverty reduction strategies. At the same time budgets continue to be

constrained and the demand for evaluations often only becomes effective once the projects are already well advanced. Consequently there is an increasing demand for rapid, cost-effective methodologies for assessing project impacts. It is also probably true to say that a significant portion, if not the majority, of these impact evaluations must be designed once the project has been underway for sometime. For all of these reasons the evaluator is asked to produce a methodologically robust impact evaluation under circumstances in which it is impossible to fully comply with many of the conventional standards for good evaluations.

The microfinance sector illustrates the growing demand for rapid, cost-effective evaluations. Monique Cohen [2001] shows that the earlier focus on rigorous and expensive impact evaluations of micro-finance programs is being increasingly replaced by demand for middle-range, client-level impact evaluations. Practitioners are mainly concerned to use impact evaluation to answer basic questions such as: (a) which clients are receiving most and least benefits and why (b) the growth, decline and saturation of different market sectors (c) how to improve institutional understanding of what products and services clients prefer, what barriers they face, and what they value in a program. A series of virtual conferences conducted by the Impact Assessment Methodologies Working Group of the Consultative Group to Assist the Poorest (CGAP) proposed that lower-cost or middle-range impact assessment of micro-credit programs should include:

- A small set of key hypotheses grounded in a solid conceptual framework
- A series of well-defined, contextually meaningful variables
- A set of relative measures that reflect the degree of precision needed
- A comprehensive assessment of the program and of local contextual factors
- Clients at all stages of the impact assessment
- A longitudinal design
- Control groups
- A mixture of quantitative and qualitative research methods
- A systematic analysis of the data.

The present paper shows that similar concerns are evidenced in many other sectors. For example, the evaluations of the Reflect Adult Literacy Programs in Section E (Case Study No. 3), and the rapid assessments of the impacts of the different components of the Eritrea Social Fund raise very similar issues. The purpose of the present paper is to review recent experiences and approaches in conducting rapid, cost-effective impact evaluations - frequently without the benefit of systematically collected baseline data or identified control groups.

Handout 3

Approaches to the design of low-cost impact evaluations

Michael Bamberger

This section reviews some of the approaches used in reducing the cost, time and complexity of impact evaluations.

Simplifying evaluation designs to reduce costs and time

Model 1 describes the simplest form of the quasi-experimental design [QED] in which efforts are made to approximate a random assignment of subjects to the experimental and control groups in Time Period 1 (T1) before the project begins. Similar information is collected again in Time Period 2 (T2) after the project is expected to have produced its impacts. One of the main differences between true experimental designs in medicine, chemistry and physics and the kinds of quasi-experimental designs which evaluators of social and economic development programs normal use, is that in most field evaluations it is not possible to achieve random assignment of subjects to experimental and control groups. Many refinements can be introduced into the basic QED design to assess multiple treatments, or to capture impacts that evolve gradually over time (Shadish, Cook, and Campbell 2002; Valadez and Bamberger 1994).

Model 1 The Basic Quasi-Experimental Design

	T1		Project Treatment		T2
Project Group	P1	-----	X	-----	P2
Control Group	C1	-----		-----	C2

Where cost is a major concern, many evaluations will eliminate one or more of the four observations. The most common options are:

- The elimination of a pre-test control group (Model 2)
- The complete elimination of a control group in both the pre and post-test data collection (Model 3)
- The complete elimination of baseline studies for both the treatment and control groups (Model 4)
- The complete elimination of both a control group and also the pre-test baseline study (Model 5).

Each of these models becomes progressively weaker as successive model fail to control for a greater number of the “threats to validity” (discussed in Section E). However, many situations exist in which the use of one of these “less robust” models is the only available option (Valadez and Bamberger 1994). One of the objectives of this article is to assess the practical utility and pitfalls of the use of each of these “less robust” models.

Defining information needs

The timing, focus and level of detail of the evaluation should be determined by the information needs of key stakeholders and the types of decisions to which the evaluation must contribute. Typical decisions or questions that decision-makers must address include:

- Is there evidence that the project is 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?

Many of these questions do not require a high level of statistical precision, but what they do require are reliable answers to questions such as:

- Are there measurable and significant changes in the characteristics of the target population with respect to the impacts the project was trying to produce?
- Is it reasonable to assume that the changes were due in a significant measure to the project and not to external (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 that 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 can be achieved if the project is replicated on a larger scale?

The key design questions concern issues such as: (a) careful measurement to ensure that correct responses are obtained on 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 factors unrelated to the project. In many cases good estimates on most or all of these questions can be obtained with relatively simple evaluation designs. Obviously the larger and more complex the project, the longer the time period being studied and the more diverse the areas in which it is operating, the more important it becomes to use more rigorous evaluation designs.

Reconstructing baseline conditions

The evaluator will frequently find that no baseline surveys were conducted at the start of the project so that no reliable information is available on the conditions of project participants or control groups before the project. There are a number of approaches that can be used to reconstruct the baseline conditions. These include:

- *Secondary data* on factors such as morbidity, access to health services, school attendance, farm prices and travel time and mode can often be obtained from surveys

conducted by health, education and agriculture agencies or from household surveys conducted by central statistical agencies. NGOs may also have conducted surveys in some of the project areas. Frequently the secondary sources do not have exactly the desired coverage and have not been conducted at the desired moment in time and may not have asked exactly the desired questions. However these sources can often provide a useful approximation to baseline conditions. Whenever, these sources are used it is essential to include an assessment of their strengths and weaknesses for use as a baseline. Factors which must be assessed include: differences in time periods and their potential significance (for example had economic conditions changed significantly between the survey date and the project launch?); differences in the population covered (for example did the surveys include employment in the informal as well as the formal sectors and were both women and men interviewed?); was information collected on key project variables and potential impacts.

- *Project records* from micro-credit agencies, health centers, schools and water projects can often provide information on conditions before the project began. For example, surveys are often conducted to estimate the numbers of children not attending school, sources of water supply or availability of credit. An assessment must be made of the reliability of these sources and the utility of the data for the purpose of evaluation¹.
- *Recall* can be used to estimate conditions prior to the project. While recall is generally agreed not to be a reliable way to obtain precise numerical data such as income, numbers of incidences of diarrhea or farm prices, it may be a valid way to obtain information on major changes in the welfare conditions of the household. For example, families do not have a problem in recalling which, if any, of their children attended school before the community school opened. They can also usually recall how children traveled to school, travel time and cost. Families can also usually provide 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 this is an area in which few studies have been conducted to assess the reliability of these kinds of recall estimates, it is particularly important to identify and assess potential sources of bias. 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.
- *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 conditions. 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”).

¹ For example, in the Eritrea impact evaluation referred to in section E it was found that the project health centers only kept records of each individual patient visit but did not have records of the number of different patients visiting the center over a given period of time. Consequently it would have been extremely time consuming to use these records to estimate changes in the proportion of the population using health services.

- *Participatory methods* such as PRA can be used to help the community to reconstruct 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, there are additional difficulties as it is 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*², so it can be a challenge to identify areas that are reasonably similar to the project areas.

Many ex-post quantitative impact assessments use statistical techniques to control for differences in individual and household characteristics and hence to approximate a control group by identifying households or individuals who did not receive particular project services or who received less of the services³. While this kind of multivariate analysis offers a useful statistical control for a number of characteristics. This kind of ex-post analysis cannot usually control for historical events or for differences in non-household attributes (such as different employment opportunities).

Rapid and economic methods for data collection and analysis

Often one of the most effective ways to reduce data collection costs is to reduce the sample size or simplify the sample design. Sample size can be often be reduced by accepting a lower level of precision of the estimates, or by reducing the types of disaggregation which are required. The use of cluster sampling can often significantly reduce interviewer costs by reducing distances and travel time between interviews. It is of course necessary to assess the trade-off in each case between reduced cost and lower precision or less detailed analysis.

A wide range of rapid and economical 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, designing survey instruments so that respondents can complete the information themselves, using secondary sources rather than interviews etc.

Participatory approaches

In addition to helping reduce the costs of data collection, participatory (largely qualitative) methods can also increase the validity and utility of the information. The perspective of intended beneficiaries, and of groups who may have been negatively affected by projects can often identify unanticipated consequences of projects that will frequently not be captured in surveys. Participatory methods are also very useful for understanding the contextual factors

² One of the cases in which randomization is used in the selection of projects occurs when demand significantly exceeds supply and some kind of lottery or random selection 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 to a discussion of the Pakistan Community School project).

³ 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

that may influence the level and distribution of project impacts and for assessing sustainability and replicability (Hentschel 1999).

Integrating quantitative and qualitative approaches

The integration of quantitative and qualitative data collection and analysis methods is a requirement for a good evaluation design. Integration is particularly important as part of a cost-effective evaluation design as the use of a number of independent estimators can help validate methods which reduce sample size or the costs of data collection. This is a particularly important application of the triangulation principle.

Table 1: Rapid and Economical Methods for Impact Evaluations	
Simplify evaluation designs	<p>Note: each successive model is sacrificing methodological rigor and is subject to a wider range of “threats to validity” (see Section C). 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 paper.</p> <p>Commonly used approaches for simplifying sample designs include:</p> <ul style="list-style-type: none"> • Eliminate data collection for pre-test control group (Model 2) • Eliminate data collection for pre-test and post-test 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	<ul style="list-style-type: none"> • Lower the level of required precision • Reduce the types of disaggregation required • Stratified sample designs • Use of cluster sampling • Use university students, student nurses and community residents to reduce data collection costs
Reconstruct baseline data and control groups	<ul style="list-style-type: none"> • Using secondary data • Redesigning project records to incorporate impact indicators • Using recall • Key informants • PRA and other participatory methods
Reducing the costs of quantitative data collection	<ul style="list-style-type: none"> • Self-administered questionnaires • Reduce length and complexity of survey instruments
Qualitative data collection methods	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Using triangulation (multi-method approaches) so that independent estimates of key variables may make it possible to reduce sample size, while at the same time increasing reliability and validity.

Integrated approaches are also particularly valuable 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, and that 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.

Handout No. 4

Rapid and Economical Methods for Impact Evaluations

Michael Bamberger

Simplify evaluation designs	<p>Note: each successive model is sacrificing methodological rigor and is subject to a wider range of “threats to validity” (see Section C). 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 paper. Commonly used approaches for simplifying sample designs include:</p> <ul style="list-style-type: none"> • Eliminate data collection for pre-test control group (Model 2) • Eliminate data collection for pre-test and post-test 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	<ul style="list-style-type: none"> • Lower the level of required precision • Reduce the types of disaggregation required • Stratified sample designs • Use of cluster sampling • Use university students, student nurses and community residents to reduce data collection costs
Reconstruct baseline data and control groups	<ul style="list-style-type: none"> • Using secondary data • Redesigning project records to incorporate impact indicators • Using recall • Key informants • PRA and other participatory methods
Reducing the costs of quantitative data collection	<ul style="list-style-type: none"> • Self-administered questionnaires • Reduce length and complexity of survey instruments
Qualitative data collection methods	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Using triangulation (multi-method approaches) so that independent estimates of key variables may make it possible to reduce sample size, while at the same time increasing reliability and validity.

Handout 5

Introduction to the theory and practice of impact evaluation design

Michael Bamberger

1. True experimental designs and quasi-experimental designs

The true experimental design is used in fields like medicine, animal behavior and some kinds of educational research where studies can be 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] a new drug, rewards and punishments [used in animal research] or learning programs and the Control Group [C] which does not receive the treatment [but may receive a placebo so that neither subjects or researchers are aware who has received the treatment]. A test is applied to both groups in Time Period 1 [T1] before the experiment begins to measure the behaviour, physiological reactions or other variables the treatment is intended to influence. The measurements are repeated in T2 following the application of the experimental treatment. The measurements in T1 and T2 are defined as E1 and E2 for the experimental group and C1 and C2 for the control group. The research design is described below:

The simplest true experimental design

	T1	Experimental Treatment	T2
Experimental group	E1	----- X -----	E2
Control Group	C1	-----	C2

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 change produced by the experimental treatment XI can be estimated by comparing the change for the treatment group with the change for the control group.⁴

If the value of XI 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.

When evaluating the impacts of development projects [water supply projects, road construction, micro-credit, teacher training and provision of teaching materials etc], it is almost never possible to approximate this level of experimental control. For example it is

⁴ This can be expressed in the following equation:
$$XI = \frac{E[2] - E[1]}{C[2] - C[1]}$$

rarely possible to randomly assign subjects to treatment and control groups, and treatments cannot be applied in as precise a way. Consequently a series of Quasi-Experimental Designs [QED] have been developed which seek to approximate as closely as possible the true experimental design in order to:

- make the best possible estimate of the extent to which a project , program or policy has produced its intended impacts
- identify the factors that positively or negatively influence the magnitude and direction of the impacts.

In the real world QED's typically face the following problems:

- It is almost never possible to randomly assign subjects to experimental 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, when, 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 initial conditions of participants and non-participants.
- It is very difficult to find a control group that matches the experimental group on the key indicators. Project communities are often selected because of special characteristics. In some cases project planners chose the poorest communities, in other cases they chose communities that have the greatest likelihood of success. In either case it will be difficult to find a control group that closely matches the project population.
- In many cases it is difficult to use any kind of control group at all for political or ethical reasons. Frequently politicians and leaders in control group areas will pressure for their community to be included in the project. From the ethical perspective it is often considered inappropriate to ask families to spend a long time responding to surveys if they will not receive any benefit. In some cases, the fact that families are being interviewed 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 for several months, 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 that encourages families to visit the project to discuss loans. In another site the administrator may not speak the local language and the project may be seen as a hostile to the community so that less people visit the center. For all of 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 administered in different sites.

- Finally, each project operates within a unique economic and political context, and interacts with a number of government and possibly non-government organizations, each of which has its 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 outcome of the project. Consequently, even when a project is administered in exactly the same way in each site, there may be significant differences in the outcomes as a result of these contextual factors.

Several lessons can be drawn from these evaluation design difficulties. First, it is important to understand the problems facing a particular study and to try to produce 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 and if 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 what measures can be taken to address the problems [see later chapters for a discussion of these measures]. Let us take as an example the common situation in which no baseline study was conducted making it difficult to assess the magnitude of the changes in school enrolment, travel patterns, water usage etc. which have occurred since the project began. Some of the possible measures that could be considered include:

- 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, schoolteachers 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 more confidence in the findings. If, on the other hand, the information from different sources is inconsistent, or even contradictory then further analysis will be required to determine whether the inconsistencies can be reconciled.

2. The evaluation framework for project impact evaluations

Figure 1 identifies the seven stages of the project cycle that can be considered in the design of impact evaluations. Readers familiar with Logical Frameworks will see many similarities to the LOGFRAME format, and this evaluation model can be coordinated with the LOGFRAME if this 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:

Project planning and design. This examines:

- the project's approach to planning [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 the information required for planning this project.
- Do 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.
- Was a systematic stakeholder analysis conducted to ensure that all sectors of the target group were consulted?
- What were the critical assumptions on which 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:
 - 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 several 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:

- even when women received loans they did not start up businesses,
- if they did start up businesses a male household member often controls the use of the profits, or
- even if women controlled the profits it might not affect their economic and social welfare. 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.

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 why 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 the project is implemented on the ground⁵.

The project implementation process: Projects can be implemented in different ways, some of which involve the community in planning and administration and others which are directly managed by the implementing agency [Ministry of Transport, Agricultural Investment Bank etc]. Projects also vary in terms of the ease of access of the community to their 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 the program may not reach the poorest farmers.

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

Outcomes or short-term impacts: These are the impacts that are achieved within a relatively short period of time (for example 6 or 12 months after project completion. Poverty reduction programs frequently identify four types of impacts⁶:

- *Opportunity:* Access to economic resources and improved economic conditions.
- *Capability:* Access to public services (health, education etc) and their impacts 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.

Medium and long-term impacts: These are assessed on the same four dimensions but given the longer time frame it is possible to assess broader impacts. For example, access to education can also include access to labor markets after completing school.

Sustainability: The overall objective of a project is not simply to produce impacts over a certain period of time but to ensure the impacts are sustained over time. For example:

- Do schools and clinics continued to function after donor funding has ended?
- Are communities able to maintain minor irrigation works, rural roads and bridges etc, is the bus company able to maintain its transport fleet?

⁵ A number of recent Public Expenditure Tracking studies conducted by the World Bank in countries such as Uganda, Tanzania and Ghana found that less than 10% of the project resources never reached the frontline service agencies. It should be stressed that although there was some corruption, most of the resources were not stolen but rather diverted to other uses. For example, funds were often transferred from the Ministry of Finance to the General Fund of a Local Government agency. The funds were then used for a different purpose – which might have produced other benefits but obviously undermined the project for which they were intended.

⁶ These categories on the 2000/2001 World Development Report *Attacking Poverty, and the World Bank Poverty Reduction Strategy Sourcebook* (see especially the *Gender* chapter)

Contextual factors affecting project outcomes

An important feature of this framework is that it recognizes that each project is implemented in a particular economic, political, institutional and socio-cultural context. Consequently, even when a project is always implemented according to the same design; the outcomes, impacts and sustainability may vary significantly from one project site to another due to these contextual factors. The model identifies three sets of contextual factors that should 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 compared to 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 must also be examined. For example, a project in a region where local government is in the hands of an opposition party may find it more difficult to obtain support from national authorities. Also if a local or national election will take place soon this might affect the dynamics of the project. There have been cases, where, for example a local candidate told farmers not to apply for small business loans or to pay service charges for water, because if he/she were elected all of these services would be provided free to the poor.
- *Institutional and organizational factors:* the success of projects depends to a considerable extent on the efficiency and support of local government agencies. Consequently, if the local ministry of health, education or transport is poorly managed, or if it only has an acting director and is short of staff; this is likely to affect the efficiency with which projects are implemented and their impacts. Non-government agencies are also coming to play an important role in project implementation, and an assessment of the efficiency of their operation may also be needed. Finally, there are often conflicts among government agencies or between government agencies and NGOs that may affect project implementation.
- *Social, economic and cultural characteristics of the participating communities:* there are often important differences in the characteristics of participating communities that can influence project implementation and impacts. This analysis is particularly important where there are ethnic differences but the analysis should ideally be conducted for all projects.

The contextual analysis can be used at any stage of the project cycle, but for the purposes of impact evaluations it will be particularly useful for helping explain differences in project impacts in different sites which cannot be explained by how well or badly the project was implemented in each site.

Contextual analysis is usually based on qualitative methods such as participant observation, meetings with community leaders and other groups, focus groups, and interviews with key informants [journalists, academics, NGOs, religious organizations, local government agencies etc.]. Secondary sources such as newspapers, university studies can also be useful.

3. The most commonly used Quasi-Experimental Designs

[see Table 1 for a summary of the strengths and weaknesses of each model]

All of the following models can be strengthened if used in combination with the evaluation framework described in section 2, and by using some of the shoestring evaluation methods described in later chapters. For the models that include a control group [models 1,2 and 4] multivariate analysis may also be used to statistically control for differences in the characteristics of the two groups.

Model 1: The strongest general-purpose quasi-experimental design

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

	T1		Project Intervention		T2
Project Group	P1	-----	X	-----	P2
Control Group	C1	-----			C2

In this model a control group [C] is selected at the start of the project to approximate as closely as possible the project beneficiary group [P]. 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 [$I_1, I_2 \dots I_n$] measuring the changes the project is intended to produce. Information is also collected on the social and economic characteristics of the individuals or families [intervening variables] that might affect project outcomes. Data collection is repeated in Time Period 2 (T2) after the project has been in operation long enough to have produced its intended results. Ideally the analysis should also include the contextual factors discussed in the previous section.

In the simplest form of the analysis the potential project impact X for variable I_1 is defined as:

$$XI_1 = \frac{I_1P(2) - I_1P(1)}{I_1C(2) - I_1C(1)}$$

The analysis can be considerably strengthened if multiple regression analysis is used to statistically control for differences in the social and economic characteristics of the project and control groups. Multiple regression statistically matches subjects on characteristics such as age, income, education and other relevant variables in order 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 there is still a difference 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 in the impact indicator are necessarily due to the project, but the more other factors which are excluded [as possible explanations of the

changes] the more likely it is that the project at least partially contributed to the observed changes.

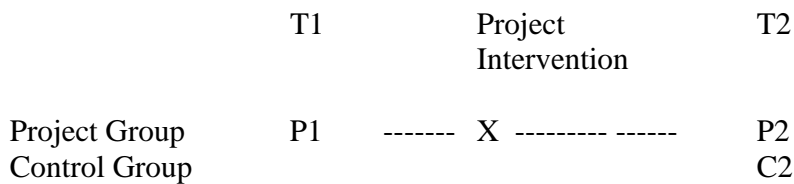
Many refinements can be introduced into the basic QED design to assess multiple treatments, or to capture impacts that evolve gradually over time (Shadish, Cook, and Campbell 2002; Valadez and Bamberger 1994).

Cheaper and faster but usually weaker QED

There are many situations in which it is not possible to use Model 1. In some cases this is due to time and budget constraints that do not permit the use of a control group. In many other cases the evaluator is not called in until the project is already 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 and consequently becomes 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 intended project 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 2 (T2) with both project and control groups. The model is represented as follows:



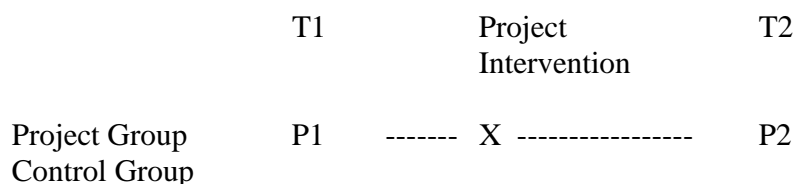
This design works reasonably well for assessing how well a project is being implemented and whether it is able to produce the intended outputs. It is also able to compare the characteristics of the project and control groups. For example, with a rural road construction project surveys and participatory 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 the road. These factors might include: whether local culture permits women to participate in road construction and to travel to the market, the distance from the local market, and the agricultural surplus available to sell. A control group, if it is well selected, could rate other local communities on these variables and hence determine the likelihood that the project would be well received and would 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, kinds of consumer goods available in community shops.

However, this design has some important 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 after the project is implemented are due to the project or to differences that already existed between the groups before the project began. Another weakness is that we cannot control for *local history* that might have affected outcomes. This

is particularly important for projects seeking to increase agricultural output or sales. Sales of maize or wheat may have increased because of the good rains and not because of the project. The ex-post control group can provide some information on this but the analysis is obviously much stronger if changes in the project and control areas can be compared over time.

Design 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:

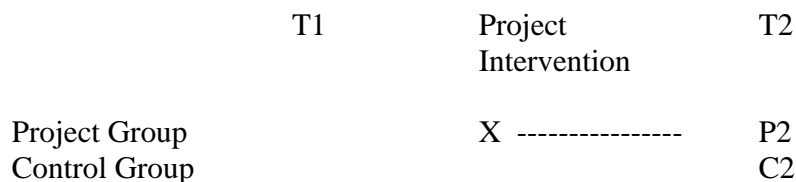


This model works reasonably well for projects having large and clearly defined impacts. For example the construction of a village school or clinic where there was previously no such facility within easy access. It can also work well when the purpose of the evaluation is to understand the project implementation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

However, this model does not work well when precise estimates of the magnitude of project impacts are required. It also does not control for the influence of local history, and the lack of comparative data on the project and other communities 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 it had selected communities with a higher than average levels of education and income, it would be difficult to assess how successful a larger project would be if it was extended to more typical communities with lower education and income.

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 the project or control areas. The model is represented as follows:



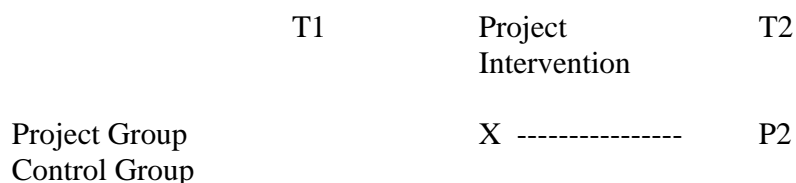
This model can be used to obtain an approximate estimate of project impacts. It works better in isolated communities there the project is the only major outside intervention, so that it is not necessary to isolate the effects of a number of other interventions which are taking place at the same time. It can also be used to compare the characteristics of project participants with people from other similar communities. If project households have similar characteristics 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 then it will be more difficult to generalize.

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

Model 5: Eliminating baseline data and control groups

This is the weakest QED. Only the project population is studied and they are only surveyed after the project has been implemented. The model is represented as follows:

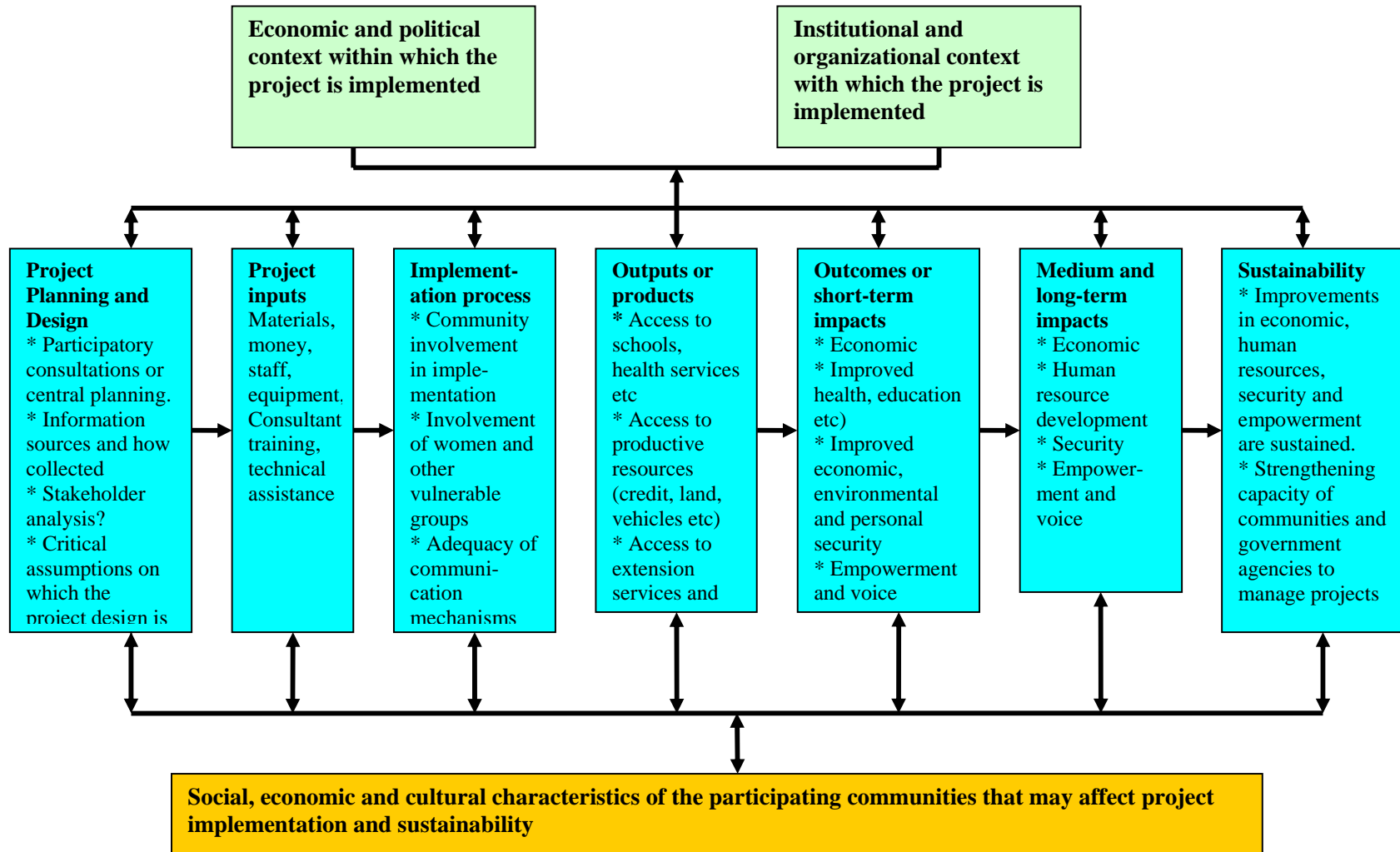


This model works reasonable well for exploratory studies where the purpose is to get a general idea of whether the model works. It can also be used to get a very approximate estimate of impacts, and works better for relatively isolated projects where the potential impact is expected to be quite large.

The model cannot be used to obtain reasonably precise estimates of impact. It also cannot control for local history events that might affect outcomes. It also does not provide any comparative data on the characteristics of the project population so it is not possible to generalize to wider population..

Table 1: The Strengths and Weaknesses of the 5 Most Frequently Used Quasi-Experimental Designs		
Model	Works reasonably well to	Does not work well to
1. Pre- and post-test surveys 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	<ul style="list-style-type: none"> ▪ 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 more widely. 	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ Evaluate projects which have large impacts or which operate in isolated areas where there is no interference from other outside interventions. ▪ Understand the project implementation process 	<ul style="list-style-type: none"> ▪ Estimate the exact magnitude of project impacts ▪ Control for local history ▪ Assess potential for replication on a larger scale
4. No baseline data	<ul style="list-style-type: none"> ▪ 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. 	<ul style="list-style-type: none"> ▪ Estimate the exact magnitude of project impacts ▪ Control for local history
5. No control groups or baseline data	<ul style="list-style-type: none"> ▪ 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 projects. 	<ul style="list-style-type: none"> ▪ 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
<p>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; and with some of the shoestring evaluation techniques discussed in the following chapters. For Models 1,2 and 4 that 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.</p>		

Figure 1: The Project Evaluation Framework



Handout 6

Threats to the validity of interpretations about program impacts.

Michael Bamberger

1. Threats to statistical conclusion validity.
2. Threats to internal validity
3. Threats to construct validity.
4. Threats to external validity.

Based on Shadish, Cook and Campbell 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. *Comments and additional threats of particular relevance for shoestring evaluation designs have been added by the author.*

Table 1 THREATS TO STATISTICAL CONCLUSION VALIDITY:
Reasons why the analysis may incorrectly assume the program intervention has contributed to the observed changes (impacts), or why some potential impacts have been overlooked.

<i>Threat to Validity</i>	<i>Measurement and inference problem</i>
1.1 Low Statistical Power.	○ <i>If sample is too small</i> (insufficiently powered) it may be incorrectly concluded that the relationship between treatment and outcome is not significant.
1.2 Violated Assumptions of Statistical Tests.	○ Violations of statistical test assumptions can lead to either overestimating or underestimating the size of an effect.
1.3 Fishing and the Error-Rate Problem.	○ Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
1.4 Unreliability of Measures.	○ Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
1.5 Restriction of Range.	○ Reduced range on a variable usually weakens the relationship between it and another variable. <i>For example, only targeting the poorest families makes it harder to identify causal relationships.</i>
1.6 Unreliability of Treatment Implementation	○ If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation. <i>A problem when not all beneficiaries receive all services.</i>
1.7 Extraneous Variance in the Experimental Setting.	○ Some features of an experimental setting may inflate error, making detection of an effect more difficult.
1.8 Heterogeneity of Units.	○ Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
1.9. Inaccurate Effect Size Estimation	○ Some statistics systematically overestimate or underestimate the size of an effect.

Source: Based on Shadish, Cook and Campbell 2002. Table 2.2 page 45. *Comments in italics added by present author.*

Table 2 THREATS TO INTERNAL VALIDITY <i>Reasons Why Inferences That The Relationships Between Two Variables Is Causal May Be Incorrect.</i>	
<i>Threat to Validity</i>	<i>Measurement and inference problem</i>
2.1 Ambiguous Temporal Precedence	○ Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2.2 Selection.	○ Systematic differences over conditions in respondent characteristics that could also cause the observed effect. <i>This is a problem when participants are self-selected [e.g. applicants for loans]</i>
2.3 History	○ Events occurring concurrently with treatment could cause the observed effect.
2.4 Maturation.	○ Naturally occurring changes over time could be confused with a treatment effect.
2.5 Regression.	○ When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
2.6 Attrition.	○ Loss of respondents to treatment or measurement can produce artifactual effects if that loss is systematically correlated with conditions.
2.7 Testing	○ Exposure to a test can affect scores on subsequent exposures, an occurrence that can be confused with a treatment effect.
2.8 Instrumentation.	○ The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
2.9 Additive And Interactive Effects Of Threats To Internal Validity.	○ The impact of a threat; can be added to that of another threat or may depend on the level of another threat.
2.10 <i>Inappropriate indicators</i>	○ <i>Selected indicators do not adequately capture critical information. This is a particular problem when using proxy indicators to measure poverty, welfare, empowerment etc.</i>
2.11 <i>Reliance on qualitative indicators</i>	○ <i>The evaluation, for ideological, methodological or practical reasons relies mainly on qualitative indicators which may not permit generalization or control for other explanations of the observed effects</i>
2.12 <i>Unreliable respondent memory or deliberate distortion</i>	○ <i>Recall is subject to potential biases due to memory failure or to deliberate distortion by respondents.</i>

Source: Shadish, Cook and Campbell 2002. Table 2.4 page 55. *Comments in italics have been added by the present author.*

Table 3 THREATS TO CONSTRUCT VALIDITY:	
<i>Reasons Why Inferences About The Constructs That Characterize Study Operations May Be Incorrect.</i>	
<i>The underlying constructs [hypotheses/concepts] on which the evaluation design [logic model] is based may not adequately describe the actual indicators of inputs, outputs, impacts and settings [operations] used in the study. For example changes in income may not adequately measure the construct “changes in household welfare”.</i>	
<i>Threat to validity</i>	<i>Measurement and inference problems</i>
3.1 Inadequate explanation of constructs	○ Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and construct.
3.2 Construct confounding	○ Operations usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inferences. <i>For example in a health program it is necessary to define who [sex, ethnicity, language spoken] provides the services as well as the services provided.</i>
3.3 Mono-operation bias	○ Any one operationalization of a construct both under represents the construct of interest and measures irrelevant constructs, complicating inference.
3.4 Mono-method bias.	○ When all operationalizations use the same method, that method is part of the construct actually studied.
3.5 Confounding Constructs with Levels of Constructs	○ Inferences about the constructs that best represent study operations may fail to describe the limited levels of the construct that were actually studied.
3.6 Treatment sensitive factorial structure	○ The structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used.
3.7 Reactive self-report changes	○ Self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made.
3.8 Reactivity to the experimental situation.	○ Participant responses reflect not just treatments and measures but also participant’s perceptions of the experimental situation, and those perceptions are part of the treatment construct actually tested.
3.9 Experimental expectancies.	○ The experimenter can influence participant responses by conveying expectations about desirable responses, and those expectations are part of the treatment construct as actually tested.
3.10 Novelty and disruption effects.	○ Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as part of the treatment construct description.

Table 3 (continued)	
3.11 Compensatory equalization.	<ul style="list-style-type: none"> ○ When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment, and this action must then be included as part of the treatment construct description.
3.12 Compensatory rivalry.	<ul style="list-style-type: none"> ○ Participants not receiving treatment may be motivated to show they can do as well as those receiving g treatment, and this compensatory rivalry must then be included as part of the treatment construct description .
3.13 Resentful demoralization	<ul style="list-style-type: none"> ○ Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must then be included as part of the treatment construct description.
3.14 Treatment diffusion	<ul style="list-style-type: none"> ○ Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.

Source: Shadish, Cook and Campbell 2002. Table 3.1 page 73. *Comments in italics have been added by the present author.*

Table 4 THREATS TO EXTERNAL VALIDITY: <i>Reasons why inferences about how study results would hold over variations in persons, settings, treatments and outcomes may be incorrect.</i>	
<p><i>Programs are usually implemented in communities which have special characteristics (e.g. the poorest or most remote, the most motivated) and in particular economic and political settings (e.g. the economy is in decline or is doing well) and political settings (e.g. strong government support or opposition from local politicians). The outcomes from these unique settings may be difficult to generalize to other settings in which a follow-up program would operate.</i></p>	
<p>4.1 Interaction of the causal relationship with units</p>	<ul style="list-style-type: none"> ○ An effect found with certain kinds of units might not hold if other kinds of units had been studied.
<p>4.2 Interaction of the causal relationship over treatment variations.</p>	<ul style="list-style-type: none"> ○ An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of the treatment is used.
<p>4.3 Interaction of the causal relationship with outcomes.</p>	<ul style="list-style-type: none"> ○ An effect found on one kind of outcome observation may not hold if other outcome observations were used.
<p>4.4 Interactions of the causal relationships with settings.</p>	<ul style="list-style-type: none"> ○ An effect found in one kind of setting may not hold if other kinds of settings were to be used.
<p>4.5 Context-dependent mediation.</p>	<ul style="list-style-type: none"> ○ An explanatory mediator of a causal relationship in one context may not mediate in another context.
<p>4.6 <i>Policy maker indifference</i></p>	<ul style="list-style-type: none"> ○ <i>Policy makers impede or do not implement a program because it is perceived as irrelevant or deleterious to their own priorities. This may result in under-estimation of potential impacts.</i>
<p>4.7 <i>Political interference</i></p>	<ul style="list-style-type: none"> ○ <i>Actions of political actors impede program implementation and thereby change the program model in ways that managers cannot control. This may result in inaccurate estimates of potential replicability.</i>
<p>4.8 <i>Seasonal cycles</i></p>	<ul style="list-style-type: none"> ○ <i>Results attributable to seasonal variations rather than to the program intervention. Particularly critical in assessing impacts and replicability many rural development programs</i>

Source: Shadish, Cook and Campbell Table 3.2 page 87. *Comments in italics have been added by the present author.*

Handout 7
Shoestring Project Evaluation Worksheet

Name of Evaluation Study

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]
 - Potential problems concerning: [see Chapter 3]
 - Baseline data
 - Control group
 - Data collection methods and quality of data
 - Analysis of contextual factors
 - Other

A note with additional details attached: Yes ___ No ___

Note any problems not covered in the following sheets:

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 that can withstand scrutiny from the critics of the project.

Note: Important threats to validity should be listed and discussed on the following pages.

Analysis and discussion of each threat to validity

Threat [Number and name]

A. How manifested in the evaluation

B. 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

Threat [Number and name]

A. How manifested in the evaluation

B. 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

Handout 8



Case Studies for Group Exercises

Three Approaches to Evaluating the Gender Impacts of Micro-Credit Programs in Bangladesh

Michael Bamberger

Overview

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 results 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.



The 3 case studies: different theory models – different findings.

The three case studies present approaches to the evaluation of the same credit programs. The first study illustrates a *technical/economic and mainly quantitative approach* to impact evaluation, while the second and third studies illustrate two different applications of the *social/empowerment and largely qualitative approaches*. When taken together the three studies show how the specification of the program theory model affects the issues that are studied and the kinds of conclusions that are drawn.

The **first study**, which adopts a *technical/economic approach*, 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 was also associated with female capital accumulation.

The **second study**, which adopts an *empowerment approach*, used a purposive sample designed to include different types of female borrowers. The study 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 **third study**, which also adopts an *empowerment approach*, assesses the impacts of the credit programs on a number of dimensions of empowerment. Using a combination of surveys and case study data it was found that the Grameen Bank and BRAC had a significant positive impact on eight dimensions of empowerment.



Group Exercise

For your case study answer the following questions:

- 1. What were the main strengths and weaknesses of the design?**
- 2. Which threats to validity were most critical to this evaluation?**
- 3. How well were they addressed in the design and in the presentation of findings?**
- 4. How could the evaluation design have been improved?**
- 5. What lessons did you learn concerning the use of shoestring evaluation methods?**

If you have time

- 6. How useful was the Shoestring Evaluation Worksheet for helping you assess this evaluation?**

Reporting back to the plenary session

In the plenary session groups will be asked in turn to respond briefly to one of the first 5 questions.

At the end we will ask for general comments on the utility of the Shoestring Evaluation Worksheet. Is it helpful? How could it be improved?



Case 1

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

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

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 female borrowing affected many of the variables relating to household welfare more than male borrowing. Women obtaining loans was also associated with female capital accumulation.

1. 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.

2. Methodology

The evaluation design can be described as follows:

	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)

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* 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.

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, and 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.

Two related studies examine the impact of female and male borrowing – from Grameen Bank, the Bangladesh Rural Advancements Committee, and government program RD-12 - on such outcomes as per capital 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	-0.03	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*	-4.92	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 that 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).

3. 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.

Acknowledgement: *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.*

Sources:

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



Case 2

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

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.

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, 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 that women actually exercised over the loans they obtained and the implications this has for a fuller understanding of empowerment.

Methodology

The evaluation design can be described as follows:

	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.



A purposive sample was selected of women who had obtained loans from four micro-credit programs in Bangladesh [N=253]. The sample included 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.

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.

Source: Anne Marie Goetz 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.



Case 3

Evaluating the impact of micro-credit on women’s empowerment

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

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 evaluation design can be described as follows:

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)



The study combined the following data-collection methods:

- A six-village ethnographic study conducted between 1991-94. The study combined participant observation with informal interviews.
- A sample survey of credit program participants and comparison groups of women living in the same areas but not receiving credit.

Eight indicators of empowerment were identified on the basis of observation and informal interviews. 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.

Findings

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

Source: Syed Hashemi, 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.