

Shoestring Evaluation: Designing Impact Evaluations Under Budget, Time, and Data Constraints

By
Michael Bamberger
World Bank

Jim Rugh
CARE

Dr. Mary Church, Program Evaluation Specialist
PREL's PRELSTAR Program

Lucia Fort
World Bank

This paper discusses the shoestring evaluation approach, which was developed to assist evaluators in conducting evaluations that are as methodologically sound as possible when operating with budget and time constraints and with limitations on the types of data to which they have access. The approach is used in two main scenarios. In the first scenario, the evaluator is not called in until the project or program has been operating for some time and typically no baseline data had been collected on the project population or a control group. As managers, policy makers, and funding agencies often only start to focus on assessing impacts when the time to make decisions on future funding is approaching, the evaluator is frequently required to work without an adequate timeline and often with a limited budget. In the second scenario, the evaluator is called in at the start of the project but for budget, political, logistical, or methodological reasons it was not possible to collect baseline data on a control group, or in some cases even on the project population itself, using methodologies that will be comparable to later evaluation.

In response to the growing demand for evaluation with these budget and time constraints, a number of rapid and economical assessment methods have been developed. Unfortunately, in an effort to deliver evaluation results on time and within budget, many of the basic principles of sound evalua-

tion design, such as random sampling, specification of the program theory, instrument development, control for researcher bias, and general quality control may be compromised. The shoestring evaluation approach provides tools for working within limitations of budget, time, and data, while at the same time providing a framework for the identification of threats to the validity or adequacy of the evaluation findings, and guidelines for addressing the different threats once they have been identified.

The approach was originally developed to assist evaluators working in developing countries, where the budget, time, and data constraints are often most severe; but feedback from colleagues working in the U.S. and other industrial nations suggests that the approach may be more widely applicable. However, all of the case studies in the paper are taken from the authors' experience in developing countries.

Most of the tools and methods used in the approach will be familiar to experienced evaluators. What is new is the way the tools are combined into a six-step strategy to ensure the best quality evaluation under the particular budget, time, and data constraints affecting an evaluation. Consequently most of the data collection and analysis methods are only referenced briefly. We do, however, discuss some of the less familiar methods, such as the use of recall and other methods for reconstructing baseline data and control groups, the strengths and weaknesses of different quasi-experimental designs for addressing the three sets of constraints, the development of an integrated framework for assessing the validity and adequacy of multi-method evaluation designs, and the strategies for

addressing the different threats to validity and adequacy. The goal is to conduct evaluations that are credible and adequately meet the needs of key stakeholders, given the conditions under which such evaluations need to be undertaken.

SHOESTRING EVALUATION SCENARIOS: TYPICAL TIME, DATA, AND BUDGET CONSTRAINTS FACING THE SHOESTRING EVALUATOR

Table 1 describes typical evaluation scenarios in which the evaluator is faced with constraints related to budget, time, and data. Sometimes the evaluator is faced with a single constraint. For example, in some cases the budget is limited but the evaluator does not face excessive time constraints, while in other cases the main constraint is time. Or sometimes the evaluation can be planned before the project begins and there is an adequate budget but the evaluator is told that for political or ethical reasons it will not be possible to collect data on a control group. Many unlucky evaluators find themselves simultaneously contending with two or all three constraints! The following paragraphs discuss some of the most common problems encountered under each of these constraints.

Time Constraints. The most common time constraint is when the evaluator is not called in until the project is already

well underway and the evaluation has to be conducted within a much shorter period of time than the evaluator considers necessary, either in terms of a longitudinal perspective over the life of the project, or in terms of the time allotted for conducting the end-of-project evaluation. Under this scenario it is not possible to conduct a baseline study using methodology that will be comparable with the preplanned final evaluation. The time available for planning stakeholder consultations, site visits and fieldwork, and data analysis may also have to be drastically reduced in order to meet the report deadline. These time pressures are particularly problematic for an evaluator who is not familiar with the area, or even the country, and who does not have time for familiarization and for building confidence with the communities and the agencies involved with the study. The combination of time and budget constraints frequently means that foreign evaluators can only be in the country for a short period of time—often requiring them to use shortcuts that they recognize as methodologically questionable.

Budget Constraints. Frequently, funds for the evaluation are not included in the original project budget, so the evaluation must be conducted with a much smaller budget than would normally be allocated. As a result, it may not be possible to apply the desirable data collection instruments (tracer studies or sample surveys, for example), or to apply the meth-

TABLE 1
Shoestring Evaluation Scenarios – Conducting Impact Evaluations with Time, Budget, or Data Constraints

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

ods for reconstructing baseline data or creating control groups. Lack of funds is also the cause of several of the time constraints discussed earlier.

Data Constraints. When the evaluation does not start until late in the project cycle, there is usually little or no comparable baseline data available on the conditions of the target group before the start of the project. Even if project records are available, they are often not organized in the form required for comparative before and after analysis. Project records and other secondary data often suffer from systematic reporting biases or poor record keeping standards. Even when secondary data is available for a period close to the project starting date it usually does not fully match the project populations. For example, employment data may only cover larger companies, whereas many project families work in smaller firms in the informal sector, or school records may only cover public schools, etc. Another problem is that survey data is often aggregated at the household level, so that information is not available on individual household members. This is a particular problem for gender analysis.

Most agencies are only interested in collecting data on the groups with which they are working. They may also be concerned that the collection of information on non-beneficiaries might create expectations of financial or other compensation for these groups, which further discourages the collection of data on a control group. It is also often difficult to identify a control group even if funds are available. Many project areas have unique characteristics that make it difficult to find comparable control areas. For example, the project may cover all of the poorest communities, or it may have selected all of the most dynamic communities, or it may only be organized in districts where there is strong political support and a commitment of local government funds.

In other cases, the project impacts concern sensitive topics such as women's empowerment, contraceptive usage, domestic or community violence, or corruption, on which information is difficult to collect even when funds are available. Similar data problems can arise when the project is working with difficult to reach groups such as drug addicts, criminals, ethnic minorities, migrants, illegal residents, or in some cases women.

THE SHOESTRING EVALUATION APPROACH: SIX STEPS TO BETTER IMPACT EVALUATION WHEN WORKING WITH BUDGET, TIME, AND DATA CONSTRAINTS

The shoestring evaluation approach proposes six steps for ensuring maximum possible methodological rigor in impact evaluations conducted under time, budget, or data constraints (see Figure 1). The approach can be used by evaluation practitioners, managers, or funding agencies and can be applied at

the start, mid-term, or end of the project (see Table 2). Managers can use the approach to identify ways to reduce the cost and time required for the evaluation (steps 2 and 3 of the approach). If the evaluation is being subcontracted to outside consultants, managers may also use the checklist in Table 7 to assess the strengths and weaknesses of the proposed evaluation design (step 5). Funding agencies may also find the approach useful when assessing the validity of the conclusions and recommendations produced by the evaluation, and in some cases to suggest measures to address and correct some of the weaknesses identified in the evaluation (step 6). Evaluation practitioners, on the other hand, will often be asked by managers and/or funding agencies to propose the minimum costs and time required to conduct the evaluation (steps 2 and 3). In some cases they will use the threats checklists (step 5) to negotiate with managers on the need to relax one or more of the budget or time constraints in order to avoid some of the major threats to validity and adequacy (for example they may make the case for conducting a household survey, or for the inclusion of a control group). Once all key stakeholders have agreed on these issues, the evaluators will then use steps 2, 3, and 4 to develop the best and most robust evaluation design within these constraints.

STEP 1: PLANNING AND SCOPING THE EVALUATION

Understanding Client Information Needs

A clear understanding of the client's priorities and information needs is an essential first step in the design of any evaluation and also an effective way for the shoestring evaluator to eliminate unnecessary data collection and analysis, hence reducing the cost and time of the evaluation. The timing, focus, and level of detail of the evaluation should be determined by the client information needs and the types of decisions to which the evaluation must contribute (Patton 1997). While it is usually a simple matter to define the client (the agency commissioning the evaluation), a more difficult issue is to define the range of stakeholders whose concerns should be taken into account in the evaluation design, implementation, and dissemination. This question is not unique to shoestring evaluations and consequently is not discussed here, other than to point out that time and budget constraints will often create pressures to limit the range of stakeholders who can be consulted and involved. The evaluator should assess early on whether these constraints may eliminate some important groups, particularly vulnerable groups that are often more difficult and expensive to reach.

The shoestring evaluator should meet as early as possible with clients and key stakeholders to ensure that the reasons for

FIGURE 1
The Shoestring Evaluation Approach

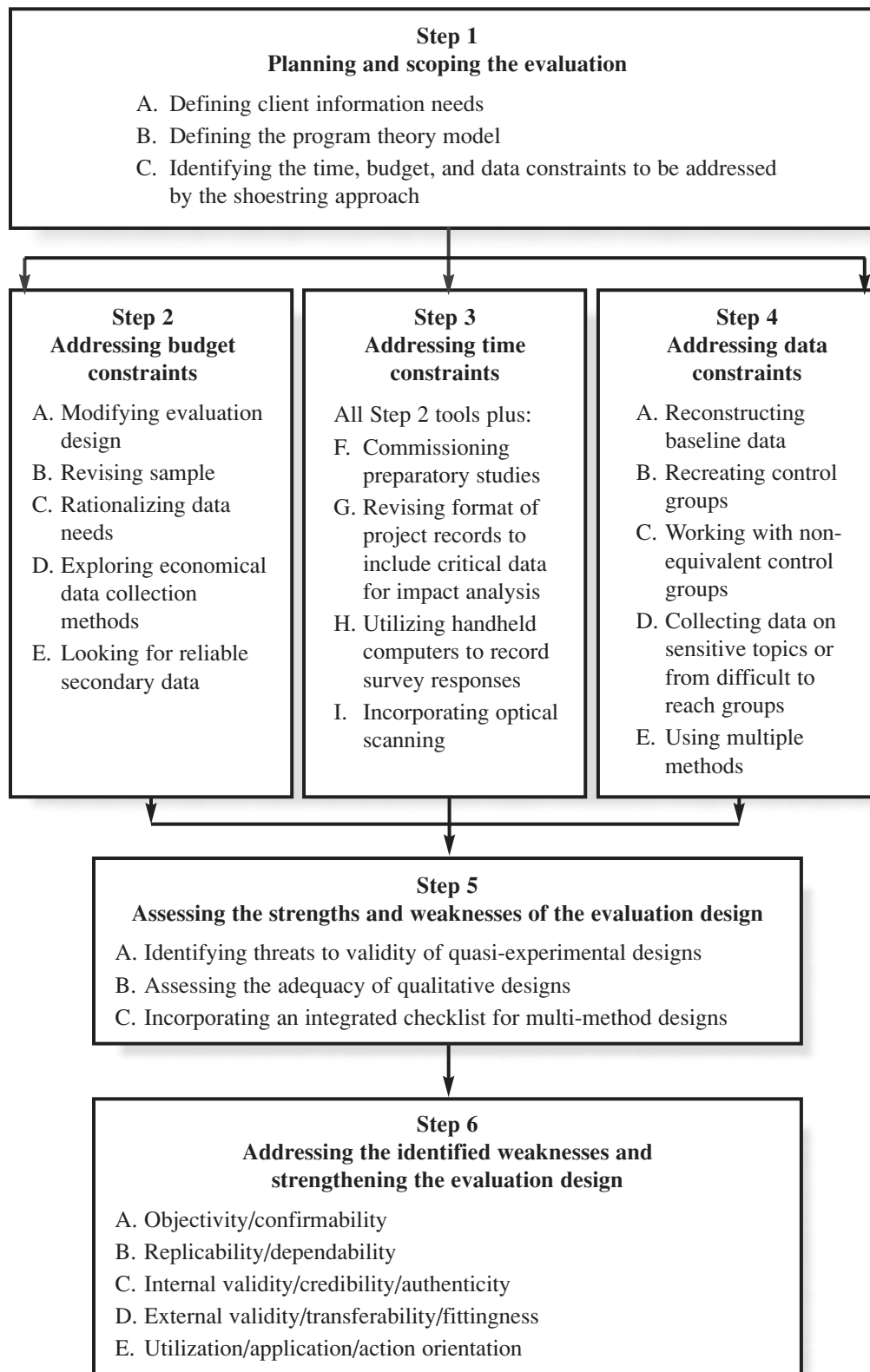


TABLE 2
Who Uses Shoestring Evaluation, For What Purpose and When?

When Does the Evaluation Start?	Evaluation Practitioners Who Design and Implement the Evaluation	Managers and Funding Agencies
At the beginning of a project (baseline)	<ul style="list-style-type: none"> • Advise management how to reduce costs and time while achieving evaluation objectives. • Negotiate with managers to relax some of the constraints in order to reduce some of the threats to validity and adequacy. • Identify ways to produce the best evaluation under budget, time, and data constraints. 	<ul style="list-style-type: none"> • Seek ways to reduce the costs and time of the evaluation. • Assess the quality of the proposed evaluation design.
During project implementation	<ul style="list-style-type: none"> • Identify ways to produce the best evaluation under budget, time, and data constraints. • Reconstruct baseline data. • Ensure maximum quality under existing constraints. 	<ul style="list-style-type: none"> • Identify ways to strengthen the ongoing evaluation (these measures may be directly implemented by management or funding agencies or recommended to the agency conducting the evaluation).
At the end of the project	<ul style="list-style-type: none"> • Identify ways to reduce costs and time. • Reconstruct baseline data. • Ensure maximum quality under existing constraints. 	<ul style="list-style-type: none"> • Identify ways to correct weaknesses in the evaluation within the budget and time constraints.

commissioning the evaluation are fully understood. The discussion of the program theory model with clients can help focus on these critical information needs. It is particularly important to understand the policy and operational decisions to which the evaluation will contribute and to agree on the level of precision required in making these decisions. Typical questions that decision makers must address include the following:

- Is there evidence that the project is achieving its objectives? Which objectives are and are not being achieved? Why?
- Are all sectors of the target population benefiting from the project? Are any groups being excluded?
- 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 they do require reliable answers to questions, including the following:

- 1) Are there measurable changes in the characteristics of the target population with respect to the impacts the project was trying to produce?
- 2) What impact has the project had on different sectors of the target population, including the poorest and most vulnerable groups? Are there different impacts on men and women? Are there any ethnic or religious groups who do not benefit or who are impacted negatively?

- 3) Were the target communities or groups reasonably typical of broader populations, such as poor farmers or urban slum dwellers, and is it likely that the same impacts could be achieved if the project were replicated on a larger scale?
- 4) Why have these observed changes occurred? Are the conditions that facilitated these changes likely to continue and are the impacts sustainable?
- 5) Is it reasonable to assume that the changes were due in a significant measure to the project and not to external factors not controlled by the project implementers?

The shoestring evaluator must understand which are the critical issues that must be explored in depth and which are less critical and can be studied less intensively. It is also essential to understand when the client needs rigorous (and expensive) statistical analysis to legitimize the evaluation findings to members of congress or parliament or funding agencies critical of the program, and when more general analysis and findings will be acceptable. The answer to these questions can have a major impact on the evaluation budget, particularly on the required sample design, size, and level of rigor.

Defining the Program Theory Model on Which the Project Is Based

Once the priorities and information needs of the clients and stakeholders have been defined, the evaluator should ascertain the program theory model or hypotheses on which the project

is based. While program theory models can be used in all evaluations, they are particularly useful for shoestring evaluations to identify the critical areas and issues on which the limited evaluation resources or time should focus, and help define the ways in which triangulation methods can be used most effectively. A program theory “consists of an explicit theory or model of how the program causes the intended or observed outcomes” (Rogers, Petrosino, Huebner, & Hacsí, 2000, p. 5). All projects and programs are based on an implicit theory about the most effective way to achieve the intended program outputs and impacts, and the factors constraining or facilitating their achievement. In some cases the program theory is spelled out in project documents, and may be summarized in the form of a logic model (e.g., log frame), while in other cases it can only be elicited by the evaluator through consultations with program staff, participants, and partner agencies. This will often be an iterative process in which an initial theory model is constructed by the evaluator on the basis of preliminary consultations and is then discussed and modified through further consultations.

Leeuw (2003) identifies three ways to reconstruct the underlying program theory: the policy-scientific approach, the strategic assessment approach, and the elicitation methodology. The second and third approaches are probably most useful for most shoestring evaluations. In the strategic approach, the theory model is identified in group discussions with key stakeholders. In the elicitation methodology, strategic documents are reviewed, managers are consulted, and decision making processes are observed. One of these two approaches can be applied in most shoestring evaluations.

The program theory model also helps assess whether failure to achieve program objectives is due to an inadequate program theory or to ineffective implementation procedures (Lipsey, 1993; Weiss, 1997). Many theory models define four stages of the project cycle: inputs, implementation, outputs, and outcomes or impacts. However, we consider it useful to add two additional stages to evaluate how the project was designed (for example: Was the project designed top-down or were participatory planning methods used? Was the project designed around a set of interventions that are expected to produce certain outcomes, or by identifying desired impact and then determining what the appropriate interventions should be?), and how effectively the stream of outcomes was sustained. The model should also identify contextual factors (the setting) that can affect implementation and outcomes. Contextual factors should include the economic, political, and organizational context as well as the socioeconomic characteristics of the affected population groups. Patton (2002) and Hentschel (1999) describe qualitative approaches for the analysis of many of these contextual factors.

A key element of theory models is the identification and monitoring of the critical assumptions on which the choice of inputs, the selection of implementation processes, and the

expected linkages between the different stages of the program cycle are based. Logical Framework Analysis (sometimes more generically referred to as Logic Models) is a widely used program theory approach that requires the critical assumptions to be identified and their validity assessed at each stage of project implementation.

Identifying the Constraints to Be Addressed by the Shoestring Approach

Step 1 also identifies the constraints facing the evaluation and determines which of steps 2, 3, and 4 (tools for addressing budget, time, and data constraints) will be required.

STEPS 2 AND 3: ADDRESSING BUDGET AND TIME CONSTRAINTS

This section describes five strategies for addressing typical budget and time constraints that evaluators face (see Table 3). While most of these strategies apply to both time and budget constraints, the following section discusses some additional strategies that can be used when adequate resources are available but where the evaluator is working under time constraints.

Simplifying the Evaluation Design

In their review of meta-analyses in the fields of psychological, educational, and behavioral treatments, Lipsey and Wilson (1993) report on 74 meta-analyses where the same topic was addressed with both randomized experiments and other types of study designs that are usually simpler and less costly. The average effect sizes did not differ between the two types of study, suggesting that the randomized experiments and simpler non-experiments tend to produce similar causal conclusions in the fields examined. This provides qualified support to the careful use of shoestring evaluation designs as one way to reduce impact evaluation costs, while recognizing that this increases the risk of threats to the validity of the conclusions.

When budget and time are not constraints, many impact evaluations would use one of the two robust designs described in Table 4. However, the shoestring evaluator must frequently select among the following five less robust designs in Table 4, which are less demanding in terms of time or budget. All of these less robust designs eliminate one or more of the pretest or posttest observations on the project or control group, and consequently are vulnerable to more of the threats to the validity of the evaluation conclusions (described in step 5 of the model). Step 4 discusses ways to strengthen some of these designs.

TABLE 3
Reducing Costs and/or Time of Data Collection and Analysis

A. Simplifying the evaluation design	<p>As Table 4 details, the following 5 “less robust” designs can be used to reduce costs or time:</p> <ul style="list-style-type: none"> • Begin evaluation at project midterm (model 3). • Evaluation combines pretest-posttest comparison of project group with posttest comparison of project and control groups (model 4). • Base evaluation on pretest-posttest comparison of project group (model 5). • Base evaluation on posttest comparison of project and control group (model 6). • Base evaluation on posttest data from project group (model 7).
B. Clarifying client information needs	<ul style="list-style-type: none"> • Often a discussion with the client can eliminate the collection of data not actually required for the evaluation objectives.
C. Reducing sample size	<ul style="list-style-type: none"> • Lower the level of required precision. • Reduce the types of disaggregation required. • Use stratified sample designs. • Use cluster sampling.
D. Reducing costs of data collection	<ul style="list-style-type: none"> • Use self-administered questionnaires (with literate populations). • Reduce length and complexity of survey instruments. • Use university students, student nurses, and community residents to collect data. • Utilize direct observation. • Use automatic counters and other non-obtrusive methods. • Convene focus groups and community forums. • Work with key informants. • Use Participatory Rapid Appraisal (PRA) and other participatory methods. • Use integrated 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. • Reorganize project data collection forms so that information on the target population and their use of project services can be more easily collected and analyzed.
E. Simplifying and speeding up data input and analysis	<ul style="list-style-type: none"> • Input survey data directly through notebook computers or other handheld devices. • Scan survey forms optically.

Table 4 begins by describing the two most methodologically robust quasi-experimental designs (QED), which can be used when evaluators have adequate time and resources, and when there are no major problems of access to data. Table 4 also describes five alternative models that can be used when one or more of the shoestring constraints are factors. It should be noted, however, that even the two most robust designs are subject to a number of threats to validity (see Shadish, Cook, & Campbell, 2002, for an extended discussion).

In model 2, which is probably the most widely used evaluation design when budget and time are not major constraints, observations are conducted on a randomly selected sample of project beneficiaries (P1) and a matched control group (C1) before the project (X) begins. The observations are repeated on both groups (P2 and C2) at the completion of the project. The impact of the project intervention X is estimated as the difference of means or proportions between the observed change in the project and control groups. This can either be

measured by a test for difference of differences of means or proportions or by using multivariate analysis to control for attributes such as income, age, education, and family size. It is normally not possible to randomly assign subjects to the project and control groups, so some element of subjectivity will be involved in selecting the most comparable control group.

Each of the five less robust models involves eliminating one or more of the pretest or posttest observations on the control or project groups:

- *Truncated longitudinal design* (model 3). The evaluation does not begin until the project has been underway for some time (no baseline data) but several observations are taken during project implementation.
- *No pretest control group* (model 4). A control group is only introduced in the posttest survey.
- *Pretest posttest comparison of project group* (model 5). There is no control group.

TABLE 4
Seven Evaluation Designs

Evaluation Design	Start of Project [Pretest]	Project Intervention (Continues on to End of Project)	Midterm Evaluation or Several Observations During Implementation	End of Project [Posttest]	Follow-Up After Project Operating for Some Time [Ex-post]	The Stage of the Project Cycle at Which Each Evaluation Design Can Begin to be Used
Two Strongest Evaluation Designs						
1. <i>Longitudinal design including posttest and ex-post observations</i> (the most comprehensive evaluation design but use limited by cost and time requirements). Observation at beginning, middle, and end of project, as well as after the project has ended. Permits assessment of project implementation as well as observing processes of change. Random assignment to project and control groups is rarely possible so this and all other designs normally use non-equivalent control groups.	P ¹ C ¹	X	P ² C ²	P ³ C ³	P ⁴ C ⁴	Start
2. <i>Pretest-posttest comparison of project and control groups</i> (the best general purpose impact design). For most purposes this is the best practical design when the evaluation can begin at the start of the project, there is a reasonable budget, and there are no particular constraints on use of control group or access to data.	P ¹ C ¹	X		P ² C ²		Start
Less Robust Evaluation Designs						
3. <i>Truncated longitudinal design.</i> Project and control groups observed at various points during project implementation but evaluation does not begin until project is underway so there are no baseline observations.		X	P ¹ C ¹	P ² C ²		Midterm
4. <i>Pretest-posttest comparison of project group combined with posttest comparison of project and control group.</i>	P ¹	X		P ² C		Start
5. <i>Pretest-posttest project group comparison of project group.</i>	P ¹	X		P ²		Start
6. <i>Posttest comparison of project and control groups.</i>		X		P C		End
7. <i>Posttest analysis of project group.</i>		X		P		End

Note. P = project participants; C = control group; P¹, P², C¹, C², etc. = first, second (third and fourth) observations of the project or control groups in a particular evaluation design; X = project intervention (this is normally a process rather than a discrete event).

- *Posttest comparison of project and control group* (model 6). The posttest control group is assumed to approximate the original conditions of the project population (no baseline on either group).
- *Only posttest project group* (model 7). There is no control group and no baseline data for the project group.

Each of models 3 through 7 can produce significant time and cost savings. However, there is a price to pay, as all of these designs are less able than the two more robust designs to address threats to validity and are thus more likely to lead to wrong conclusions concerning the contribution of the project intervention to the observed outcomes. However, when their strengths and weaknesses are fully understood and addressed, these designs can provide an acceptable level of precision for many, if not most, management needs at a much reduced cost.

Clarifying Client Information Needs

The costs and time required for data collection can sometimes be significantly reduced through a clearer definition of the information required by the client and the kinds of decisions to which the evaluation will contribute (see step 1A). For example, if the evaluand is a pilot microcredit project targeting women farmers in a region with strong social controls on women's ability to control productive resources, the client's main interest may be to assess whether it is possible for women to apply for loans and to control how money is used. In this context it may be possible to convince the client that it is not necessary to invest time and resources in a control group in order to compare outcomes until the basic question, "Can the program be implemented as planned?" has been answered. As another example, clients will often say, "It would be interesting to compare the outcomes of the project for (different religious groups, recent urban migrants vs. those who have been living in the city for a long time, etc.)." Often, discussions with the client will reveal there is no particular reason to think all of these factors will be important for the project. Once it is understood that each additional level of stratification of the sample will imply a significant increase in sample size and cost, the client will often agree that some or all of these factors could be eliminated from the sample design, resulting in significant reductions in sample size, cost, and time.

Reducing Sample Size

In many cases, a clearer understanding of the kinds of decisions to be made by clients and the level of precision required for these decisions can result in a significant reduction in sample size. As an illustration let us assume that an evaluator is asked to estimate whether there has been a significant change in the proportion of the target population attending school after an experimental school meals program has been introduced. The change would be assessed by a comparison of two groups: either a before and after comparison of the project population or a comparison of the ex-post attendance rates for the project and a control group. If the project is only expected

to produce a small change, the client may require that the sample be large enough to determine whether a change as small as, say, 5% is statistically significant. Using certain simplifying assumptions, it would be necessary, in this case, to interview a sample of 1,536 respondents. However, if a minimum difference of 10% is acceptable then the corresponding sample size could be reduced to 384 families, and for a 20% change only 96 interviews would be required.

To simplify the computations, the above examples are based on the estimation of differences between proportions. The same general principles apply when estimating differences between means, but in this case the required sample size is harder to estimate in advance as it depends on knowing the standard deviations of the two populations (for a more detailed discussion of the determinants of sample size see Lipsey, 1990).

It is important to avoid arbitrary reductions in sample size simply to save money and time; and the estimation of sample size must always be based on a full understanding of the client's information needs and the required level of precision. It is also important to explain to the client that each additional level of disaggregation of the estimates (e.g., by different regions, sex of household head, the type of benefits received) will require a corresponding increase in the sample size.

Reducing Costs of Data Collection and Analysis

Considerable savings in costs and time can often be achieved through reducing the length and complexity of the survey instrument. A ruthless pruning of survey instruments to eliminate nonessential information can often significantly reduce the length of the survey. Areas in which the amount of information to be collected can often be reduced include: demographic information on each household member; sources of household income and expenditures; and behavior such as travel patterns, time use, and agricultural activities. It is again important to define information requirements with the client and not to arbitrarily eliminate information simply to produce a shorter survey instrument. Be clear on what indicators are likely to significantly contribute to the results being evaluated. Be ruthless in leaving out what would simply be interesting to know for a researcher, but unaffordable for a shoestring evaluation.

Some additional ways to reduce the costs of data collection include the following:

- Collect information on community attitudes, time use, access to and use of services, etc., through focus groups rather than household surveys.
- Replace surveys with direct observation, for example, to study time use, travel patterns, and use of community facilities.
- Use key informants to obtain information on community behavior and use of services.

- Use self-administered instruments such as diaries to collect data on income and expenditure, travel patterns, or time use.
- Make maximum use of secondary data including project records.
- Use photography and videotaping, which can sometimes provide useful and economical documentary evidence on the changing quality of houses and roads, use of public transport services, etc. (Kumar, 1993; Valadez & Bamberger, 1994).

Box 1 presents three case studies illustrating ways to cut the cost and time of data collection.

Integrating Quantitative and Qualitative Approaches

While mixed method approaches are recommended for all evaluation designs, the integration of quantitative and qualitative data collection and analysis methods is particularly important for shoestring evaluators faced with budget and time constraints. The triangulation of several independent estimators can help validate information collected from smaller samples or when using the cost saving methods described above (Bamberger, 2000a, 2000b).

Specific Ways to Reduce the Time Required to Collect and Analyze Data

While most of the above methods can save both money and time, there are a number of additional ways to economize on the time required to collect and analyze data. Some of these methods may increase costs, so it is important to clarify with the client the relative importance of the budget and time constraints:

- When working with expensive and time constrained (usually foreign) consultants, it can be beneficial to commission exploratory studies to collect background information on the characteristics of the target population and the project context. These studies should be commissioned well ahead of the planned arrival of the consultants so that the information is available to the consultants when the evaluation officially begins. In this respect it should be noted that time constraints frequently only affect foreign consultants and there is much more flexibility when using local resources.
- By planning ahead, it is often possible to reorganize project monitoring records and data collection forms so that information on the target population and their use of project services can be more easily and rapidly analyzed.

Box 1. Rapid and Economical Methods of Data Collection

1. In Bulgaria a rapid midterm assessment was conducted of a project to reduce the environmental contamination produced by a major metallurgical factory. Key informant interviews, review of project records, and direct observation were combined to provide economical ways to assess compliance with safety and environmental regulations and to assess reductions in the level of environmental contamination. A survey of key stakeholders was conducted to obtain independent assessments of the findings reported in the evaluation. The evaluation cost less than \$ 5,000 and was completed in less than two months (Dimitrov, in press).
2. An evaluation of the impacts of a slum upgrading project in Manila, the Philippines, assessed the impact of the housing investments made by poor families on their consumption of basic necessities. A randomly selected sample of 100 households was asked to keep a daily record of every item of income and expenditure over a period of a year. Households recorded this information themselves in a diary and the evaluation team of the National Housing Authority made weekly visits to a sample of households to ensure quality control. The only direct cost, other than a small proportion of staff salaries, was the purchase of small gifts for the families each month. As the study only covered project participants, most of whom were very favorable towards the project, the response rate was maintained at almost 100% throughout the year. This proved to be a very economical way to collect high quality income and expenditure data and permitted the use of an interrupted time series design with 365 (daily) observation points (Valadez & Bamberger, 1994).
3. An assessment of the impacts of community management on the quality and maintenance of village water supply in Indonesia combined direct observation of the quality and use of water with participatory group assessments of water supply and interviews with key informants. The use of group interviews and direct observation proved a much more economical way to assess project impacts than conventional household sample surveys (Dayal, van Wijk, & Mukherjee, 2000).

- The direct inputting of survey data to notebook computers and other handheld devices can greatly reduce the time required for data processing and analysis.
- Optical scanning of survey instruments is another time saving device.

STEP 4: ADDRESSING DATA CONSTRAINTS

The shoestring evaluator may be faced with at least four sets of problems resulting from a lack of critical evaluation data:

- Lack of baseline data on the project population
- No control group
- Statistical problems in working with ex-post surveys using nonequivalent control groups
- Problems in collecting data on sensitive topics or from groups who are difficult to locate or to interview

Reconstructing Baseline Data on the Project or Control Groups

When the evaluation does not begin until midway through the project or even until the end of the project, the evaluator will frequently find that no reliable information is available on the conditions of project participants or control groups before the project interventions began. The following approaches can be used to reconstruct the baseline conditions. Though not as accurate as would be obtained from a good baseline study, the data may perhaps be sufficiently reliable for the purposes of a shoestring evaluation (see also examples in Box 2).

Using secondary data. Secondary data on previous years is often available on factors such as morbidity, access to health services, school attendance, farm prices, and travel time and mode from government agencies, central statistical bureaus, nongovernmental organizations (NGOs), and university researchers. While these sources can provide a useful (and often the only available) approximation to baseline conditions, it is essential to assess their strengths and weaknesses with respect to: differences in time periods (which are particularly important when economic conditions may have changed between the survey date and the project launch), differences in the population covered (e.g., did the surveys include employment in the informal as well as the formal sectors and were both women and men interviewed?), whether information was collected on key project variables and potential impacts, and whether or not the secondary data is statistically valid for the particular target population addressed by the project being evaluated.

Records from other projects in the same area can often provide information on conditions before the current project began. For example, surveys are often conducted to estimate the number of children not attending school, sources and costs of water supply, or availability of microcredit. An assessment must be made of the reliability and utility of these data for the purpose of the evaluation.

Using recall. Recall is a potentially valuable, although somewhat treacherous, way to estimate conditions prior to the start of the project and hence to reconstruct or strengthen baseline data. The limited available evidence suggests that while estimates from recall are frequently biased, the direction, and sometimes the magnitude, of the bias are often predictable so that usable estimates can often be obtained. Schwarz and Oyserman (2001) provide a useful review of cognitive and behavioral factors affecting recall and of ways to design data collection instruments to reduce some of the potential biases. Recall is therefore a potentially useful tool, particularly in the many situations where no other systematic baseline data is available. The utility of recall can often be enhanced if two or more independent estimates can be triangulated.

Box 2. Reconstructing Baseline Data

1. In Bangalore, India in 1999 a sample of households were asked to respond to a Citizen Report Card in which they assessed the changes in the quality of delivery of public services (water, sanitation, public hospitals, public transport, electricity, phones, etc.) since a first Report Card survey in 1993-1994. Families reported that although the quality of services was low, on average there had been an improvement in most services with respect to helpfulness of staff and proportion of problems resolved. The use of recall was an economical substitute for a baseline study (Paul, 2002).
2. In an evaluation of the impact of social funds in Eritrea, the program had been underway for several years before the evaluation began. Baseline conditions for access to health services were estimated by asking families how frequently they used health services before the village clinic was built, how long it took to reach these facilities, the costs of travel, and the consequences of not having better access. The information provided by the households was compared with information from health clinic records and key informants (nurses, community leaders, etc.) so as to strengthen the estimates through triangulation. While secondary data was useful, it was often found that the records were not organized in the way required to assess changes and impacts. For example, the village clinics kept records on each patient visit but did not keep files on each actual patient or each family so it was difficult to determine how many different people used the clinic each month/year and also the proportion of village families who used the clinic. Similar methods were used to reconstruct baseline data on village water supply and rural roads and transport for the evaluation of the water supply and road construction components ("The Impact of Social," n.d.).
3. The Operations Evaluation Department (OED) of the World Bank conducted an ex-post evaluation of the social and economic impacts of a resettlement program in Maharashtra State, India. Baseline data had been collected by project administrators on all families eligible to receive financial compensation or new land, but information was not collected on the approximately 45% of families who had been forced to move but were not entitled to compensation. A tracer study was conducted by OED in which families forced to relocate without compensation were identified through neighbors and relatives. A significant proportion of these families were traced in this way and were found, on average, to be no worse off as a result of resettlement, but it was not possible to assess how representative they were of families relocated without compensation (World Bank, 1993).

While recall is generally unreliable for collecting precise numerical data such as income, incidences of diarrhea, or farm prices, it can be used to obtain information on major changes in the welfare conditions of households. For example, families can usually recall which children attended a school outside the community before the village school opened, how children traveled to school, and travel time and cost. Also families can often provide reliable information on access to health facilities, where they previously obtained water, how much they used, and how much it cost. On the other hand, 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 deliberately underestimate how

much they had spent on water if they are trying to convince planners they are too poor to pay the water charges proposed in a new project.

Two common sources of recall bias have been identified. First, the underestimation of small and routine expenditures increases as the recall period increases. Second, there is a telescoping of recall concerning major expenditures, such as the purchase of a cow, bicycle, or item of furniture, so that expenditures made outside of the recall period (e.g., the past 12 months) will often be reported as having been made within the reference period. While most of the research on recall bias has been carried out by U.S. studies such as the Expenditure Surveys, the general results are potentially relevant to developing countries. The Living Standards Measurement Survey (LSMS) program has conducted some assessments on the use of recall for estimating consumption in developing countries. The LSMS program was launched in the 1980s by the World Bank to develop standard survey methodologies and questionnaires for comparative analysis of poverty and welfare in developing countries (Grosh & Glewwe, 2002). For a review of the recall bias literature, see Deaton & Grosh (2000).

The most systematic assessments of the reliability of recall data in developing countries probably come from demographic studies on the reliability of reported contraceptive usage and fertility. The existence of a number of large-scale comparative studies such as the World Fertility Survey means national surveys using comparable data collection methods are available for different points in time. For example, similar surveys were conducted in the Republic of Korea in 1971, 1974, and 1976, each of which obtained detailed information on current contraceptive usage and fertility, as well as detailed historical information based on recall for a number of specific points in the past. This permitted a comparison of recall in 1976 for contraceptive usage and fertility in 1974 and 1971 with exactly the same information collected from surveys in those two earlier years. It was found that recall produced a systematic underreporting, but that the underestimation could be significantly reduced through the careful design and administration of the surveys (Pebley, Goldman, & Choe, 1986). Similar findings are available from demographic analysis in other countries. The conclusion from these studies is that recall can be a useful estimating tool with predictable and to some extent controllable errors. Unfortunately, it is only possible to estimate the errors where large-scale comparative survey data is available, and there are few, if any, other fields with a similar wealth of comparative data.

Interestingly, there are a number of studies suggesting that recall can provide better estimates of behavioral changes in areas such as primary prevention programs for child abuse, vocational guidance, and programs for delinquents than conventional pretest and posttest comparisons based on self-assessment (Pratt, McGuigan, & Katzev, 2000). This is due to the fact that before entering a program, subjects often overes-

timate their behavioral skills or knowledge through a lack of understanding of the nature of the tasks being studied and the required skills. After completing the program they may have a better understanding of these behaviors and may be able to provide a better assessment of their previous level of competency or knowledge and how much these have changed. The present authors are not aware of any "response shift" studies examining things like self-assessment of poverty, empowerment, or community organizational capacity in developing countries, but these are all areas where the response shift concept could potentially be applied to reconstruction of baseline data for shoestrings evaluations.

Working with key informants. 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"). Caldwell (1985), reviewing lessons from the World Fertility Survey, uses some of these considerations to express reservations about the use of key informants for retrospective analysis in fertility surveys.

Using participatory methods. Participatory methods such as many of the Participatory Rural Appraisal (PRA) tools can be used to help the community reconstruct past conditions and identify critical incidents in the history of the community or region (Rietberger-McCracken & Narayan, 1997).

Reconstructing Control Groups

There are additional difficulties in constructing control groups, as this entails identifying appropriately comparable control areas as well as measuring the conditions in these areas. With few exceptions, project areas are selected purposively to target the poorest areas or those with the greatest development potential rather than randomly, so it can be a challenge to identify control locations that are reasonably similar to the project areas. One of the cases in which randomization is used in the selection process occurs when demand significantly exceeds supply and some kind of lottery or random selection is used. This sometimes occurs with social funds (Baker, 2000) or with community supported schools (Kim, Alderman, & Orazem, 1999). It will frequently be necessary to complement the limited quantitative data with judgment when deciding what is a good or acceptable control group. When the statistical data is available, cluster analysis can provide a powerful tool for selecting a comparison group that can be matched on the variables of most interest to the project (Weitzman, Silver, & Dillman, 2002).

It cannot be assumed the control group is “pure.” Rarely, if ever, in society are all factors equal between a project group and a control group (or comparison community), other than the project intervention itself. It is important to look for and document interventions by other organizations in the control community. The analysis at the time of the evaluation should then try to determine the relative influence of changes brought about by the project’s interventions compared to different internal and external influences in the comparison group.

It is sometimes possible to construct an internal control group within the project area. Households or individuals who did not participate in the project or who did not receive a particular service or benefit can be treated as the control for the project in general or for a particular service (for example, subjects may be categorized according to such factors as their distance from a road or water source, whether any family member attended literacy classes, or the amount of food aid they received). When projects are implemented in phases, it is also possible to use households selected for the second or subsequent phases as the control group for the analysis of the impacts of the previous phase. For example, the economic status of a new cohort of women about to receive their microfinance loans might serve as a control group to compare with those who received loans during the past year.

Selection bias. Throughout the discussion of control groups it is important to constantly check for potential selection bias. With respect to internal control groups, the families in a project area who did not participate in the project are likely to be different in potentially important ways from those who did participate. In some cases nonparticipants may have been excluded or discouraged on the basis of their political affiliation (or lack thereof), sex, ethnicity, or religion. In other cases they may not have had the motivation or self-confidence to apply or get selected. Similar factors may explain why some communities were not selected. The following section discusses some of the statistical procedures that can be used to at least partially address the selection bias issue and to improve the comparability of the control group. How effective the statistical controls are will depend on the adequacy of the control model and the reliability of the measurement of the control variables (Shadish et al., 2002).

Problems in Working with Nonequivalent Control Group Data from Surveys

Evaluations frequently compare the project population with nonequivalent control areas selected to match the project population as closely as possible. When subjects were not randomly assigned to the project and control groups, it is possible to strengthen the analytical value of available control groups by statistically matching subjects from the project and control areas on a number of relevant characteristics such as

education, income, and family size. The evaluations of Ecuador’s cut flower export industry and the Bangladesh microcredit programs are examples of this approach (see Box 3). If differences in the dependent variables (the number of hours men and women spend on household tasks, men’s and women’s savings and expenditure on household consumption goods, etc.) are still statistically significant after controlling for these household characteristics, this provides preliminary indications that the differences in the dependent variables may be due, at least in part, to the interventions of the project.

While this type of multivariate analysis is a powerful analytical tool, one important weakness is that the evaluation design does not provide any information on the initial conditions or attributes of the two groups prior to the project intervention. For example, the higher savings rates of women in the communities receiving microcredit in Bangladesh might

Box 3. Working With Nonequivalent Control Groups in Ex-Post Evaluation Designs

1. An evaluation was conducted in Guayaquil, Ecuador, to assess the impact of the cut-flower export industry (which employs a high proportion of women and pays well above average wages) on women’s income and employment and on the division of domestic tasks between husband and wife. Families living in another valley about 100 miles away and without access to the cut-flower industry were selected as a control group. This was a nonequivalent control group as families were not randomly assigned to the project and control groups. The project and control groups were interviewed after the flower industry had been operating for some time and no baseline data was available. Multivariate analysis was used to determine whether there were differences in the dependent variables (women’s employment and earnings and the number of hours spent by husband and wife on domestic chores) in the project and control areas after controlling for household attributes such as educational level of both spouses, family size, etc. Significant differences were found between the two groups on each of these dependent variables and it was concluded there was evidence that access to higher paid employment in the flower industry did affect the dependent variables (e.g., the distribution of domestic chores between men and women). While multivariate analysis created a stronger control group, it was not able to examine differences in the initial conditions of the two groups before the project began. For example, it is possible that the flower industry decided to locate in this particular valley because it was known that a high proportion of women worked and that husbands were prepared to assume more household chores, thus allowing their wives to work longer hours. This analytical model is not able to examine this alternative explanation (Newman, 2001).
2. An ex-post evaluation was conducted of the impact of microcredit on women’s savings, household consumption and investment, and fertility behavior in Bangladesh. The evaluation used household survey data from communities that did not have access to credit programs as a nonequivalent control group. Multivariate analysis was used to control for household attributes and it was found that women’s access to microcredit programs was significantly associated with most of the dependent variables. As in the case of the Ecuador study, this design did not control for preexisting differences between the project and control groups with respect to important explanatory variables such as women’s participation in small business training programs or prior experience with microcredit (Khandker, 1998).

be due to the fact that they had previously received training in financial management or that they already had small business experience. These nonequivalent control group designs can be strengthened if they incorporate some of the methods discussed above for reconstructing baseline data.

Collecting Data on Sensitive Topics or from Groups Who Are Difficult to Reach

A third set of problems, not unique to shoestring evaluations, concern the collection of data on sensitive topics such as domestic violence, contraceptive usage, or teenage violence; or from difficult to reach groups such as sex workers, drug users, ethnic minorities, the homeless, or in some cultures, women (Bamberger, Blackden, Fort, & Manoukian, 2001). These situations require the use of appropriate qualitative methods such as participant observation, focus groups, and key informants. These issues are particularly important for the shoestring evaluator as budget and time constraints may create pressures to ignore these sensitive topics or difficult to reach groups. Box 4 presents three case studies on the collection of data on sensitive topics.

STEP 5: IDENTIFYING THREATS TO THE VALIDITY AND ADEQUACY OF THE EVALUATION DESIGN AND CONCLUSIONS

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

The analysis of threats to validity of conclusions from quasi-experimental designs is familiar to quantitative evaluators at least since Cook and Campbell's 1978 publication. However, there is a continuing debate concerning the extent to which similar criteria can, and even should, be applied to qualitative evaluations. The challenge for the shoestring approach is to develop guidelines for assessing the validity and adequacy of multi-method evaluation designs.

Shadish et al. (2002) have updated Cook and Campbell's (1978) four categories of threats to conclusion validity, namely:

- *Statistical conclusion validity*, why inferences about covariation between two variables may be incorrect.
- *Internal validity*, why inferences that the relationship between the two variables is causal may be incorrect.

- *Construct validity*, why inferences about the constructs that characterize study operations may be incorrect.
- *External validity*, why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may be incorrect.

Box 4. Approaches to Difficult to Collect Data

1. In a study in Bangladesh to assess the impact of microcredit on women's empowerment, experience showed that conventional household survey methods would not allow women to speak freely about sensitive issues concerning control of household resources and male authority. Participant observation was used to observe women and family relations over a period of years in order to study changes in household power relations before and after women had obtained loans from a village bank. Observation was combined with the administration of an empowerment scale based on items identified by the women themselves in group discussions (Hashemi, Schuler, & Riley, 1996).
2. In Lima, Peru, it was believed that one of the reasons women did not use public transport was because of the fear of sexual harassment. However, women were unwilling to mention this in conventional transport surveys. Participant observation, in which researchers spent days traveling on public transport, was able to document the high incidence of harassment. This was confirmed and quantified in focus groups with women, men, and mixed groups stratified by age, conducted by a market research firm in their office in the center of town (i.e., away from the community) (Gomez, 2000).
3. Visits by representatives of donor agencies to rural health clinics in Nepal found that all of the health diagnosis and prescription of medicines was done by the resident doctors, most of whom had been transferred from the main cities to the villages. The untrained "peon," recruited from the local community, kept the clinic clean, made tea, etc. However, an anthropologist observed the clinics during normal periods when there were no outside visitors. She found that the doctors were absent for long periods of time and that the peon who, unlike the doctor, spoke the local language, regularly advised patients and even prescribed medicines during the long absences of the doctor. The donor agency was reluctant to accept this finding because during their visits only the doctor treated patients and the humble peon was very much in the background (Justice, 1986).

Other writers such as Miles and Huberman (1994) and Guba and Lincoln (1989) have proposed additional criteria of reliability and objectivity. Table 5 presents a checklist based on Shadish, Cook, and Campbell's four categories of threats to conclusion validity that can be used to assess potential weaknesses in all of the seven shoestring designs presented in Table 4. Additional subcategories pertinent to shoestring evaluations have been added to the checklist by the present authors.

TABLE 5
Threats to Validity of Quasi-Experimental Designs

<p>1. Threats to Statistical Conclusion Validity. Reasons Why Inferences About Covariation Between Two Variables May Be Incorrect</p>
<p>1.1. Low statistical power 1.2. Violated assumptions of statistical tests 1.3. Fishing and the error-rate problem 1.4. Unreliability of measures 1.5. Restriction of range 1.6. Unreliability of treatment implementation 1.7. Extraneous variance in the experimental setting 1.8. Heterogeneity of units 1.9. Inaccurate effect size estimation 1.10. <i>Extrapolation from a truncated or incomplete data base</i></p>
<p>2. Threats to Internal Validity. Reasons Why Inferences That the Relationships Between Two Variables Is Causal May Be Incorrect</p>
<p>2.1. Ambiguous temporal precedence 2.2. Selection 2.3. History 2.4. Maturation 2.5. Regression 2.6. Attrition 2.7. Testing 2.8. Instrumentation 2.9. Additive and interactive effects of threats to internal validity 2.10. <i>Deliberate respondent distortion when using recall</i> 2.11. <i>Use of less rigorous designs due to budget and time constraints</i></p>
<p>3. Threats to Construct Validity. Reasons Why Inferences About the Constructs That Characterize Study Operations May Be Incorrect</p>
<p>3.1. Inadequate explanation of constructs 3.2. Construct confounding 3.3. Mono-operation bias 3.4. Mono-method bias 3.5. Confounding constructs with levels of constructs 3.6. Treatment sensitive factorial structure 3.7. Reactive self-report changes 3.8. Reactivity to the experimental situation 3.9. Experimental expectancies 3.10. Novelty and disruption effects 3.11. <i>Inappropriate indicators</i> 3.12. <i>Unreliable respondent memory</i> 3.13. <i>Using indicators and constructs developed in other countries without pretesting in the local context</i></p>
<p>4. Threats to External Validity. Reasons Why Inferences About How Study Results Would Hold Over Variations in Persons, Settings, Treatments, and Outcomes May Be Incorrect</p>
<p>4.1. Interaction of the causal relationship with units 4.2. Interaction of the causal relationship over treatment variations 4.3. Interaction of the causal relationship with outcomes 4.4. Interactions of the causal relationships with settings 4.5. Context-dependent mediation 4.6. <i>Policy maker indifference</i> 4.7. <i>Political interference</i> 4.8. <i>Seasonal cycles</i> 4.9. <i>Reliance on qualitative indicators without assessing their representativity</i></p>

Note. Items in italics were added by present authors.

While this kind of checklist is often used for assessing the validity of quantitative evaluations (Cook & Campbell, 1978; Shadish et al., 2002), there is a continuing debate on the appropriate criteria for judging the adequacy or quality of conclusions drawn from qualitative evaluations. Schwandt (1990) argues that it is not possible to specify criteria for assessing qualitative research, while Patton (2002) proposes five different sets of criteria for judging the quality and credibility of different types of qualitative enquiry. However, other writers believe it is possible to establish uniform criteria for assessing qualitative evaluations. Guba and Lincoln (1989) proposed the use of four sets of “parallel” or “foundational” criteria for judging goodness or quality of qualitative evaluations, which parallel the post-positivist criteria (see Table 6):

TABLE 6
Guba and Lincoln’s Parallel or Foundational Criteria

Shadish, Cook, and Campbell’s Post-Positivist Criteria	Parallel or Foundational Qualitative Criteria
Internal validity	Credibility
External validity	Transferability
Reliability	Dependability
Objectivity	Confirmability

Although Guba and Lincoln were not completely comfortable with the use of their parallel criteria, considering them primarily methodological criteria and preferring to use “authenticity criteria” like fairness, ontological authenticity, educative authenticity, catalytic authenticity, and tactical authenticity, other authors such as Miles and Huberman (1994) and Yin (2003) have proposed the use of these parallel criteria as a way to move towards comparable criteria for assessing the validity and adequacy of quantitative and qualitative evaluation designs.

Table 7 presents a first attempt to develop an integrated checklist for assessing the validity and adequacy of multi-method shoestring evaluation designs. It uses the four sets of parallel criteria proposed by Guba and Lincoln plus a fifth category, utilization, proposed by Miles and Huberman. For evaluations that include a quasi-experimental design component, there is a cross-reference to the threats to internal, statistical, construct, and external conclusion validity given in Table 6. The shoestring evaluator can also find additional guidance from sources such as the American Evaluation Association’s “Guiding Principles for Evaluators,” which includes 23 points for assessing the quality of evaluations in terms of (1) systematic enquiry, (2) competence, (3) integrity/honesty, (4) respect for people, and (5) responsibilities for general and public welfare (Shadish, Newman, Scheirer, & Wye, 1995).

TABLE 7
Integrated Checklist for Assessing the Validity and Adequacy of Multi-Method Shoestring Evaluation Designs

A. Objectivity/Confirmability
<i>Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?</i>
1. Are the study's methods and procedures adequately described? Are study data retained and available for reanalysis?
2. Is data presented to support the conclusions?
3. Has the researcher been as explicit and self-aware as possible about personal assumptions, values, and biases? Were methods used to control for bias?
4. Were competing hypotheses or rival conclusions considered?
B. Reliability/Dependability
<i>Is the process of the study consistent and reasonably stable over time and across researchers and methods?</i>
1. Are the research questions clear, and is the study design congruent with them?
2. Are findings consistent or congruent across data sources?
3. Are basic paradigms and analytic concepts clearly specified?
4. Were data collected across the full range of appropriate settings, times, respondents, etc.?
5. Did all fieldworkers have comparable data collection protocols?
6. Were coding and quality checks made, and did they show adequate agreement?
7. Do the accounts of different observers converge?
8. Were peer or colleague reviews used?
9. Are the conclusions subject to threats to construct validity? If so were these addressed?
C. Internal Validity/Credibility/Authenticity
<i>Are the findings credible to the people studied and to readers? Do we have an authentic portrait of what we are studying?</i>
1. How context-rich and meaningful ("thick") are the descriptions?
2. Does the account ring true, make sense, seem convincing? Does it reflect the local context?
3. Did triangulation among complementary methods and data sources produce generally converging conclusions?
4. Are the presented data well linked to the categories of prior or emerging theory? Are the findings internally coherent, and are the concepts systematically related?
5. Were the rules used for confirmation of propositions, hypotheses, etc. made explicit?
6. Are areas of uncertainty identified? Was negative evidence sought, found? How was it used? Have rival explanations been actively considered?
7. Were conclusions considered accurate by the original observers?
8. Were any predictions made in the study and, if so, how accurate were they?
9. Are the findings subject to threats to internal validity? If so, were these addressed?
10. Are the findings subject to threats to statistical validity? If so, were these addressed?
D. External Validity/Transferability/Fittingness
<i>Do the conclusions fit other contexts and how widely can they be generalized?</i>
1. Are the characteristics of the sample of persons, settings, processes, etc. described in enough detail to permit comparisons with other samples?
2. Does the sample design theoretically permit generalization to other populations?
3. Does the researcher define the scope and boundaries of reasonable generalization from the study?
4. Do the findings include enough "thick description" for readers to assess the potential transferability?
5. Does a range of readers report the findings to be consistent with their own experience?
6. Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?
7. Are the processes and findings generic enough to be applicable in other settings?
8. Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?
9. Does the report suggest settings where the findings could fruitfully be tested further?
10. Have the findings been replicated in other studies to assess their robustness? If not, could replication efforts be mounted easily?
11. Are the findings subject to threats to external validity? If so, were these addressed?
E. Utilization/Application/Action Orientation
<i>How useful were the findings to clients, researchers, and the communities studied?</i>
1. Are the findings intellectually and physically accessible to potential users?
2. Do the findings provide guidance for future action?
3. Do the findings have a catalyzing effect leading to specific actions?
4. Do the actions taken actually help solve local problems?
5. Have users of the findings experienced any sense of empowerment or increased control over their lives? Have they developed new capacities?
6. Are value-based or ethical concerns raised explicitly in the report? If not, do some exist that the researcher is not attending to?

STEP 6: ADDRESSING AND REDUCING THREATS TO VALIDITY AND ADEQUACY OF THE EVALUATION DESIGN

A key element of the shoestring approach is that it recommends practical measures to correct or reduce threats to validity and adequacy once they have been identified. The following are examples of approaches for addressing problems identified in each of the four sets of threats to validity of quantitative evaluation designs presented in Table 5.

- *The unreliability of measures (Threat 1.4).* Three possible approaches can be considered to address this threat. First, ensure sufficient time and resources are allocated to develop and field test the data collection instruments. Second, incorporate multi-method data collection approaches so that at least two independent measures are used for all key variables. Third, use triangulation to check on the reliability and consistency of the different estimates.
- *Selection bias (Threat 2.2).* Four possible measures are proposed, including the following: compare characteristics of participants and control groups, statistically control for differences in participant characteristics in the two groups, use key informants (if no control group is used) to compare participants with the total population, and use direct observation of focus groups and other group settings to assess psychological characteristics such as self-confidence and motivation.
- *Reactivity to the experimental situation (Threat 3.8).* Use exploratory studies, observations, etc. to understand respondent expectations and to identify potential response bias.
- *Policymaker indifference or proactive political interference (Threats 4.6 and 4.7).* If a project is implemented in different locations, identify differences in the attitudes of policymakers in each location (through interviews, secondary sources, or key informants) and assess how these differences appear to affect the project.

The following are examples of ways to address threats to validity and adequacy of evaluation designs in each of the five categories of the integrated checklist given in Table 7:

- *Inadequate documentation of methods and procedures (threat A-1).* Request that the researchers document their methodology more fully or provide missing documents.
- *Data were not collected across the full range of appropriate settings, times, respondents, etc. (threat B-4).* If the study has not yet been conducted, discuss ways to revise the sample design or to use qualitative methods to cover the missing settings, times, or respondents. If data collection has already been com-

pleted, consider the possibility of using rapid assessment methods such as focus groups, interviews with key informants, participant observation, etc. to fill in some of the gaps.

- *The account does not ring true and does not reflect the local context (threat C-2).* Organize workshops or solicit the views of individual key informants to determine whether the problems concern missing information (for example, only men were interviewed), whether there are factual issues, or whether the problem concerns how the material was interpreted by the evaluator. Based on the types of problems identified, either return to the field to fill in the gaps or include the impressions of key informants, focus group participants, new participant observers, etc. to provide different perspectives.
- *The sampling is not theoretically diverse enough to permit generalization to other populations (threat D-3).* If the fieldwork has not yet been conducted, consider ways to broaden the sample to make it more representative of wider populations. If fieldwork has already been conducted, consider the possibility of interviews with small samples of the missing population groups. If this is not possible, try to meet with key informants to obtain information on the missing groups or use direct observation to obtain information on these groups (the elderly, women, the unemployed, ethnic minorities, etc.).
- *The findings do not provide guidance for future action (threat E-2).* If the researchers have the necessary information, ask them to make their recommendations more explicit. If they do not have the information, consider organizing brainstorming sessions with key groups in the community or the implementing agencies to develop more specific recommendations for action.

SUMMARY: SHOESTRING EVALUATION IN A NUTSHELL

It is unfortunate but probably true to say that planning for most international development impact evaluations does not begin until a project or program is well underway, and most of the evaluations must be conducted under budget and time constraints, often with limited access to baseline data and control groups. Despite these constraints, there is a growing demand for systematic assessments of the impacts of development projects and their potential replicability. Fortunately, most policy makers, managers, and funding agencies need answers to relatively straightforward questions, which do not require great methodological sophistication. Typical questions include: (1) Is the project achieving its basic objectives? (2)

Who has benefited and who has not? (3) Is the project sustainable? and (4) Is this approach replicable? The straightforward nature of these questions makes it possible to provide reasonably robust and useful answers, even within the typical constraints under which evaluators operate.

The pressures of working under budget and time constraints have often resulted in a lack of attention to sound research design, with limited attention given to identifying and addressing factors affecting the validity of the findings. The shoestring evaluation approach is being developed to respond to the demand for ways to work within budget, time, and data constraints while at the same time ensuring maximum possible methodological rigor within the given evaluation context. The shoestring evaluation guidelines discussed in this paper are summarized below. While many of these principles can be applied to all impact evaluations, each of them has a specific application to scenarios where the evaluator is subject to budget, time, or data constraints. For example, while all evaluations must understand the client's information needs (step 1A of the guidelines), a careful prioritization of these needs can help define areas in which time and costs can be reduced, and other areas where this is not possible without compromising the goals of the evaluation. Similarly, the checklists of threats to validity and adequacy (Step 5) should be applied to all evaluations, but they have particular relevance when assessing the implications of using designs that lack critical baseline or control group data.

Step 1. Planning and Scoping the Evaluation

1. Identify client information needs, the key issues to be addressed, and the required levels of precision.
2. Identify, or when necessary construct, the underlying program theory model to guide the evaluation design.
3. Identify the budget, time, and data constraints that the shoestring approach must address.

Step 2. Addressing Budget Constraints

4. Review the alternative evaluation designs and decide which is the strongest available model in a given context. Evaluators are encouraged to use the threats to validity and adequacy checklists (Tables 5 and 7) to assess the potential strengths and weaknesses of the selected design and to consider what measures can be used to strengthen it.
5. Consider possible ways to reduce sample size while ensuring the required level of statistical precision.

6. Rationalize data collection needs and ruthlessly eliminate questions not required for the purposes of the evaluation.
7. Review all possible options for data collection to identify the most economical methods consistent with quality concerns. Use multi-method approaches to strengthen the validity of cost saving methods through triangulation.
8. Look for reliable secondary data to reduce data collection costs.

Step 3. Addressing Time Constraints (in Addition to the Methods Discussed in Step 2)

9. Consider commissioning preparatory/exploratory studies prior to the arrival of time-constrained external consultants.
10. Revise the format of project records to include and facilitate the analysis of critical data for impact analysis.
11. Use notebook computers, handheld devices, and optical scanning to reduce data inputting and analysis time.

Step 4. Addressing Data Constraints

12. Identify and assess possible secondary data that could potentially provide baseline data on project or control groups.
13. Identify and assess the extent to which project records can be used to provide baseline data.
14. Consider using recall to reconstruct baseline data. Be aware of the potential weaknesses and biases of the information obtained and try to use triangulation to obtain at least two independent estimates of key indicators.
15. When using nonequivalent control groups, use some of the above methods to reconstruct baseline data to determine whether some of the variance attributed to project impacts is in fact due to preexisting differences between the project and control groups.
16. Consider the need for special data collection methods to obtain sensitive data or to include difficult to reach groups.

Step 5. Identifying Threats to the Validity and Adequacy of the Evaluation Design and Conclusions

17. [For quasi-experimental designs and other quantitative models] review Table 5 to identify potential threats to: (1) statistical conclusion validity, (2) internal validity, (3) construct validity, and (4) external validity.
18. Review the checklist in Table 7 to identify potential threats with respect to: (A) objectivity/confirmability, (B) reliability/dependability, (C) internal validity/credibility/authenticity, (D) external validity/transferability/fittingness, and (E) utilization/application/action orientation.
19. Assess the importance of each threat for the purposes of the evaluation and for the validity of the conclusions and recommendations.

Step 6. Addressing Identified Weaknesses and Strengthening the Evaluation Design and Analysis

20. For each of the important threats identified in Step 5, consider possible measures that could be taken to eliminate or reduce each threat, and apply the most appropriate measures within the given scenario.
21. In cases where it is not possible to correct the threat, ensure that the evaluation report clearly recognizes the problems and discusses the implications for the conclusions and recommendations of the evaluation.

In conclusion, despite the fact that major challenges arise when evaluators must work under serious budget, time, and data constraints, evaluators are constantly asked to address important operational and policy questions under these constraints. While the required methodological adjustments inevitably increase the range and seriousness of threats to the validity of the evaluation conclusions, it is hoped that the shoestring approach described in this paper can help support efforts to produce adequately robust and useful evaluation findings when working with real world constraints.

REFERENCES

- Baker, J. (2000). *Evaluating the impacts of development projects on poverty: A handbook for practitioners*. Washington, DC: World Bank.
- Bamberger, M. (Ed.). (2000a). *Integrating quantitative and qualitative research in development projects*. Washington, DC: World Bank.
- Bamberger, M. (2000b). The evaluation of international development programs: A view from the front. *American Journal of Evaluation*, 21, 95–102.
- Bamberger, M., Blackden, M., Fort, L., & Manoukian, V. (2001). Gender. In J. Klugman (Ed.), *A sourcebook for poverty reduction strategies* (Vol. 1, pp. 333–376). Washington, DC: World Bank. Retrieved October 22, 2004, from www.worldbank.org/poverty
- Caldwell, J. (1985). Strengths and limitations of the survey method approach for measuring and understanding fertility change: Alternative possibilities. In J. Cleland & J. Hobcraft (Eds.), *Reproductive change in developing countries: Insights from the World Fertility Survey* (pp. 45–63). Oxford, United Kingdom: Oxford University Press.
- Cook, T., & Campbell, D. (1978). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Dayal, R., van Wijk, C., & Mukherjee, N. (2000). *Methodology for participatory assessments with communities, institutions and policy makers: Linking sustainability with demand, gender and poverty*. Washington, DC: World Bank.
- Deaton, A., & Grosh, M. (2000). Consumption. In M. Grosh & P. Glewwe (Eds.), *Designing household survey questionnaires for developing countries: Lessons from 15 years of the Living Standards Measurement Study* (pp. 91–134). Washington, DC: World Bank.
- Dimitrov, T. (in press). Enhancing the performance of a major environmental project through a focused mid-term evaluation: The Kombinat za Cvetni Metali environmental improvement project in Bulgaria. In *Influential evaluations: Detailed case studies*. Washington, DC: World Bank.
- Gomez, L. (2000). Gender Analysis of Two Components of the World Bank Transport Projects in Lima, Peru: Bikepaths and Busways. World Bank internal report.
- Grosh, M., & Glewwe, P. (Eds.). (2002). *Designing household survey questionnaires for developing countries: Lessons from 15 years of the Living Standards Measurement Study*. Washington, DC: World Bank.
- Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage Publications.
- Hashemi, S., Schuler, S. R., & Riley, A. P. (1996). Rural credit programs and women's empowerment in Bangladesh. *World Development*, 24, 635–653.
- Hentschel, J. (1999). Contextuality and data collection methods: A framework and application to health service utilization. *The Journal of Development Studies*, 35(4), 64–94.
- The impact of social funds in Eritrea*. (n.d.). Unpublished report.
- Justice, J. (1986). *Policies, plans and people: Culture and health development in Nepal*. Berkeley: University of California Press.
- Kim, J., Alderman, H., & Orazem, P. (1999). Can private school subsidies increase schooling for the poor? The Quetta Urban Fellowship Program. *World Bank Economic Review*, 13, 443–466.
- Khandker, S. (1998). *Fighting poverty with microcredit: Experience in Bangladesh*. Oxford, United Kingdom: Oxford University Press.
- Kumar, K. (Ed.). (1993). *Rapid appraisal methods*. Washington, DC: World Bank.
- Leeuw, F. (2003). Reconstructing program theories: methods available and problems to be solved. *American Journal of Evaluation*, 24, 5–20.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.