
Chapter 2

Key Steps in Designing and Implementing Impact Evaluations*

Undertaking an impact evaluation study can be quite challenging and costly, with implementation issues arising at every step of the way. These challenges highlight the importance of a well-designed study, a committed and highly qualified team, and good communication between the evaluation team members. By incorporating the evaluation early into the design of a project, it will be possible to obtain results in a timely way so that the findings can be used for midproject adjustments of specific components.

Regardless of the size, program type, or methodology used for the evaluation, there are several key steps to be carried out as outlined below (Box 2.1). This chapter will provide a discussion of these steps as well as a discussion of the many issues that may arise in implementation. The sequencing of these steps is critical, particularly in ensuring the collection of necessary data before the project begins implementation. Early planning provides the opportunity to randomize, to construct *ex ante* matched comparisons, to collect baseline data, and to identify upcoming surveys that could be used in a propensity score matching approach.

All of the design work and initial data collection should be done during project identification and preparation. Ideally, some results will be available during the course of project implementation so they can feed into improving the project design if necessary. A good example of how a project incorporated evaluation plans from the earliest stages is illustrated in the Uganda Nutrition and Early Childhood Development Project (see chapter 4).

Determining Whether or Not to Carry Out an Evaluation

A first determination is whether or not an impact evaluation is required. As discussed above, impact evaluations differ from other evaluations in that they are focused on assessing causality. Given the complexity and cost in carrying out impact evaluation, the costs and benefits should be assessed, and consideration should be given to whether another approach would be more appropriate, such as monitoring of key performance indicators or a process evaluation. (These approaches should not

* This chapter draws heavily on a paper prepared by Laura Rawlings, *Implementation Issues in Impact Evaluation*, Processed, 1999.

Box 2.1 Main Steps in Designing and Implementing Impact Evaluations

During Project Identification and Preparation

1. Determining whether or not to carry out an evaluation
2. Clarifying objectives of the evaluation
3. Exploring data availability
4. Designing the evaluation
5. Forming the evaluation team
6. If data will be collected:
 - (a) Sample design and selection
 - (b) Data collection instrument development
 - (c) Staffing and training fieldwork personnel
 - (d) Pilot testing
 - (e) Data collection
 - (f) Data management and access

During Project Implementation

7. Ongoing data collection
8. Analyzing the data
9. Writing up the findings and discussing them with policymakers and other stakeholders
10. Incorporating the findings in project design

be seen as substitutes for impact evaluations; indeed they often form critical complements to impact evaluations.) And perhaps the most important inputs to the decision of whether or not to carry out an evaluation are strong political and financial support.

The additional effort and resources required for conducting impact evaluations are best mobilized when the project is innovative, is replicable, involves substantial resource allocations, and has well-defined interventions. For example, the impact evaluation of the Bolivian Social Investment Fund met each of these criteria. First, the new social fund model introduced in Bolivia was considered innovative and replicable; second, the social fund has been responsible for roughly 25 percent of all public investments in Bolivia since the beginning of the evaluation; and third, the interventions were well-defined by the social fund menu of subprojects.

Impact evaluations should also be prioritized if the project in question is launching a new approach such as a pilot program that will later be under consideration for expansion based on the results of the evaluation, or the new World Bank Learning and Innovation Loans. This rationale made the Nicaraguan school autonomy reform a good candidate for an impact evaluation. The evaluation study accompanied the government's testing of a new decentralized school management model from its pilot stage in the mid-1990s through its expansion to almost all secondary schools and about half of all primary schools today. The evaluation was managed by a closely coordinated international team including local staff from the Ministry of Education's research and evaluation unit and the World Bank's Primary Education Project coordination office in Managua. Their involvement ensured that the evaluation informed key policy decisions regarding the modification and expansion of the pilot.

Another important consideration is to ensure that the program that is to be evaluated is sufficiently developed to be subject to an impact evaluation. Pilot projects and nascent reforms are often prone to revisions regarding their content as well as how, when, and by whom they will be implemented. These changes can undermine the coherence of the evaluation effort, particularly experimental designs and other types of prospective evaluations that rely on baseline and follow-up data of clearly established treatment and control groups. Where the policies to be evaluated are still being defined, it may be advisable to avoid using an impact evaluation in order to allow for flexibility in the project.

Gaining support from policymakers and financiers for an impact evaluation can be challenging but is a prerequisite for proceeding. They must be convinced that the evaluation is a useful exercise addressing questions that will be relevant to decisions concerning the evaluated program's refinement, expansion, or curtailment. They must also be convinced of the legitimacy of the evaluation design and therefore the results, particularly when the results are not as positive as anticipated.

Financing for an impact evaluation remains a difficult issue for program managers and client counterparts alike. The financing issue is compounded by the fact that data on evaluation costs are usually difficult to obtain. And perhaps the stickiest issue arises from the public good value of the evaluation: if the results of the evaluation are going to be used to inform policies applied outside of the national boundaries within which the evaluation is conducted, as is often the case, why should an individual country bear the cost of the evaluation? Among the case studies that had information on sources of funding, the information shows that countries often assume the majority, but not the entirety, of the evaluation costs. As is discussed more fully in chapter 4, many of the cases reviewed suggest that successfully implementing an impact evaluation requires

not only a substantial resource commitment from the client countries but also the involvement of World Bank staff, or external researchers and consultants, necessitating resources beyond those provided by the country.

Clarifying Evaluation Objectives

Once it has been determined that an impact evaluation is appropriate and justified, establishing clear objectives and agreement on the core issues that will be the focus of the evaluation up front will contribute greatly to its success. Clear objectives are essential to identifying information needs, setting output and impact indicators, and constructing a solid evaluation strategy to provide answers to the questions posed. The use of a logical (log) framework approach provides a good and commonly used tool for identifying the goals of the project and the information needs around which the evaluation can be constructed.

The log frame, increasingly used at the World Bank, is based on a simple four-by-four matrix that matches information on project objectives with how performance will be tracked using milestones and work schedules, what impact project outputs will have on a beneficiary institution or system and how that will be measured, and how inputs are used to deliver outputs (see Annex 5 for examples). In other words, it is assumed that the project's intended impact is a function of the project's outputs as well as a series of other factors. The outputs, in turn, are a function of the project's inputs and factors outside the project. Quantifiable measures should then be identified for each link in the project cycle. This approach does not preclude the evaluator from also looking at the unintended impacts of a project but serves to keep the objectives of the evaluation clear and focused. Qualitative techniques are also useful in eliciting participation in clarifying the objectives of the evaluation and resulting impact indicators.

Although a statement of the objective would seem on the face of it to be one of the easiest parts of the evaluation process, it can be extremely difficult. For example, statements that are too broad do not lend themselves to evaluation. The objective statement in the Mexico PROBECAT evaluation (Annex 1.9) that the evaluation is about "the effect of the PROBECAT training program on labor market outcomes" would be more precise if it were narrowed down to the effect of PROBECAT on hours worked, hourly earnings, monthly salary, and time to first job placement for different types of workers. The Mexico PROGRESA evaluation provides a good example of creating a clear outline and delineating multiple objectives from the start with a separate discussion of each component—with objectives detailed in subcategories (Annex 1.10). This was particularly important because the

intervention was quite complex, with the evaluation having to address not only the program impact but also aspects of program operations targeting and timing.

Reviewing other evaluation components such as cost-effectiveness or process evaluations may also be important objectives of a study and can complement the impact evaluation. Cost-effectiveness may be of particular concern for policymakers whose decision it will be to curtail, expand, or reform the intervention being evaluated. On issues related to service delivery, a process evaluation may be relevant to assess the procedures, dynamics, norms, and constraints under which a particular program is carried out.

Exploring Data Availability

Many types of data can be used to carry out impact evaluation studies. These can include a range from cross-sectional or panel surveys to qualitative open-ended interviews. Ideally this information is available at the individual level to ensure that true impact can be assessed. Household-level information can conceal intrahousehold resource allocation, which affects women and children because they often have more limited access to household productive resources. In many cases, the impact evaluation will take advantage of some kind of existing data or piggyback on an ongoing survey, which can save considerably on costs. With this approach, however, problems may arise in the timing of the data collection effort and with the flexibility of the questionnaire design. Box 2.2 highlights some key points to remember in exploring the use of existing data resources for the impact evaluation.

With some creativity, it may be possible to maximize existing information resources. A good example is the evaluation of the Honduran Social Investment Fund (see chapter 4). This study used a module from the national income and expenditure survey in the social fund questionnaire, thereby allowing social fund beneficiaries' income to be compared with national measures to assess poverty targeting (Walker and others 1999).

At the most basic level, data on the universe of the population of interest will be required as a basis from which to determine sample sizes, construct the sampling frame, and select the sample. Other types of data that may be available in a given country and can be used for different impact evaluations include (see Valadez and Bamberger 1994): household income and expenditure surveys; Living Standards Measurement Studies (LSMSs); labor market surveys; records of cooperatives, credit unions, and other financial institutions; school records on attendance, repetition, and examination performance; public health records on infant mortality,

Box 2.2 Key Points for Identifying Data Resources for Impact Evaluation

- Know the program well. It is risky to embark on an evaluation without knowing a lot about the administrative and institutional details of the program; that information typically comes from the program administration.
- Collect information on the relevant “stylized facts” about the setting. The relevant facts might include the poverty map, the way the labor market works, the major ethnic divisions, and other relevant public programs.
- Be eclectic about data. Sources can embrace both informal, unstructured interviews with participants in the program and quantitative data from representative samples. However, it is extremely difficult to ask counterfactual questions in interviews or focus groups; try asking someone who is currently participating in a public program: “What would you be doing now if this program did not exist?” Talking to program participants can be valuable, but it is unlikely to provide a credible evaluation on its own.
- Ensure that there is data on the outcome indicators and relevant explanatory variables. The latter need to deal with heterogeneity in outcomes conditional on program participation. Outcomes can differ depending, for example, on whether one is educated. It may not be possible to see the impact of the program unless one controls for that heterogeneity.
- Depending on the methods used, data might also be needed on variables that influence participation but do not influence outcomes given participation. These instrumental variables can be valuable in sorting out the likely causal effects of nonrandom programs (box 1.2).
- The data on outcomes and other relevant explanatory variables can be either quantitative or qualitative. But it has to be possible to organize the information in some sort of systematic data structure. A simple and common example is that one has values of various variables including one or more outcome indicators for various observation units (individuals, households, firms, communities).
- The variables one has data on and the observation units one uses are often chosen as part of the evaluation method. These choices

(Box continues on the following page.)

Box 2.2 (*continued*)

should be anchored to the prior knowledge about the program (its objectives, of course, but also how it is run) and the setting in which it is introduced.

- The specific source of the data on outcomes and their determinants, including program participation, typically comes from survey data of some sort. The observation unit could be the household, firm, or geographic area, depending on the type of program one is studying.
- Survey data can often be supplemented with useful other data on the program (such as from the project monitoring database) or setting (such as from geographic databases).

incidence of different infectious diseases, number of women seeking advice on contraception, or condom consumption; specialized surveys conducted by universities, nongovernmental organizations (NGOs), and consulting groups; monitoring data from program administrators; and project case studies.

Using Existing Survey Data. Many surveys may also be in the planning stages or are ongoing. If a survey measuring the required indicators is planned, the evaluation may be able to oversample the population of interest during the course of the general survey (for example, to use for the propensity score matching approach) as was done for the Nicaraguan Social Investment Fund evaluation and the Argentine TRABAJAR workfare program evaluation (Jalan and Ravallion 1998). Conversely, if a survey is planned that will cover the population of interest, the evaluation may be able to introduce a question or series of questions as part of the survey or add a qualitative survey to supplement the quantitative information. For example, the Credit with Education program in Ghana included a set of qualitative interviews with key stakeholders as well as with nonparticipant and participant focus groups that provided qualitative confirmation of the quantitative results (Annex 1.6). The evaluation assessed the impact of the program on the nutritional status and food security of poor households. Quantitative data included specific questions on household income and expenditure and skills level, whereas qualitative data focused on women's empowerment—status and decisionmaking in the household, social networks, self-confidence, and so forth.

Designing the Evaluation

Once the objectives and data resources are clear, it is possible to begin the design phase of the impact evaluation study. The choice of methodologies will depend on the evaluation question, timing, budget constraints, and implementation capacity. The pros and cons of the different design types discussed in chapter 1 should be balanced to determine which methodologies are most appropriate and how quantitative and qualitative techniques can be integrated to complement each other.

Even after the evaluation design has been determined and built into the project, evaluators should be prepared to be flexible and make modifications to the design as the project is implemented. In addition, provisions should be made for tracking the project interventions if the evaluation includes baseline and follow-up data so that the evaluation effort is parallel with the actual pace of the project.

In defining the design, it is also important to determine how the impact evaluation will fit into the broader monitoring and evaluation strategy applied to a project. All projects must be monitored so that administrators, lenders, and policymakers can keep track of the project as it unfolds. The evaluation effort, as argued above, must be tailored to the information requirements of the project.

Evaluation Question. The evaluation questions being asked are very much linked to the design of the evaluation in terms of the type of data collected, unit of analysis, methodologies used, and timing of the various stages. For example, in assessing the impact of textbooks on learning outcomes, it would be necessary to tailor the evaluation to measuring impact on students, classrooms, and teachers during a given school year. This would be very different than measuring the impact of services provided through social fund investments, which would require data on community facilities and households. The case studies in Annex I provide the other examples of how the evaluation question can affect the evaluation design.

In clarifying the evaluation questions, it is also important to consider the gender implications of project impact. At the outset this may not always be obvious, however; in project implementation there may be secondary effects on the household, which would not necessarily be captured without specific data collection and analysis efforts.

Timing and Budget Concerns. The most critical timing issue is whether it is possible to begin the evaluation design before the project is implemented and when the results will be needed. It is also useful to identify up front at which points during the project cycle information

from the evaluation effort will be needed so that data collection and analysis activities can be linked. Having results in a timely manner can be crucial to policy decisions—for example, during a project review, around an election period, or when decisions regarding project continuation are being made.

Some methods require more time to implement than others. Random assignment and before-and-after methods (for example, reflexive comparisons) take longer to implement than ex-post matched-comparison approaches. When using before-and-after approaches that utilize baseline and follow-up assessments, time must be allowed for the last member of the treatment group to receive the intervention, and then usually more time is allowed for postprogram effects to materialize and be observed. Grossman (1994) suggests that 12 to 18 months after sample enrollment in the intervention is a typical period to allow before examining impacts. In World Bank projects with baselines, waiting for both the intervention to take place and the outcomes to materialize can take years. For example, in the evaluation of the Bolivian Social Investment Fund, which relied on baseline data collected in 1993, follow-up data was not collected until 1998 because of the time needed for the interventions (water and sanitation projects, health clinics, and schools) to be carried out and for effects on the beneficiary population's health and education outcomes to take place. A similar period of time has been required for the evaluation of a primary education project in Pakistan that used an experimental design with baseline and follow-up surveys to assess the impact of community schools on student outcomes, including academic achievement.

The timing requirements of the evaluation cannot drive the project being evaluated. By their very nature, evaluations are subject to the time frame established by the rest of the project. Evaluations must wait on projects that are slow to disburse and generate interventions. And even if projects move forward at the established pace, some interventions take longer to carry out, such as infrastructure projects. The time frame for the evaluation is also sensitive to the indicators selected because many, such as changes in fertility rates or educational achievement, take longer to manifest themselves in the beneficiary population.

Implementation Capacity. A final consideration in the scale and complexity of the evaluation design is the implementation capacity of the evaluation team. Implementation issues can be very challenging, particularly in developing countries where there is little experience with applied research and program evaluations. The composition of the evaluation team is very important, as well as team members' experience with different types of methodologies and their capacity relative to other activities being carried out by the evaluation unit. This is particu-

larly relevant when working with public sector agencies with multiple responsibilities and limited staff. Awareness of the unit's workload is important in order to assess not only how it will affect the quality of evaluation being conducted but also the opportunity cost of the evaluation with respect to other efforts for which the unit is responsible. There are several examples of evaluation efforts that were derailed when key staff were called onto other projects and thus were not able to implement the collection of data on schedule at the critical point in time (such as a point during the school year or during agricultural season). Such situations can be avoided through coordination with managers in the unit responsible for the evaluation to ensure that a balance is achieved with respect to the timing of various activities, as well as the distribution of staff and resources across these activities. Alternatively, it can be preferable to contract a private firm to carry out the evaluation (discussed below).

Formation of the Evaluation Team

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

Among the core team is the evaluation manager, analysts (both economist and other social scientists), and, for evaluation designs involving new data collection, a sampling expert, survey designer, fieldwork manager and fieldwork team, and data managers and processors (for a comprehensive guide to designing and implementing surveys, see Grosh and Muñoz 1996). Depending on the size, scope, and design of the study, some of these responsibilities will be shared or other staffing needs may be added to this core team. In cases in which policy analysts may not have had experience integrating quantitative and qualitative approaches, it may be necessary to spend additional time at the initial team building stage to sensitize team members and ensure full collaboration. The broad responsibilities of team members include the following:

- Evaluation manager—The evaluation manager is responsible for establishing the information needs and indicators for the evaluation (which are often established with the client by using a logical framework approach), drafting terms of reference for the evaluation, selecting the evaluation methodology, and identifying the evaluation team. In many cases, the evaluation manager will also carry out policy analysis.

- Policy analysts—An economist is needed for the quantitative analysis, as well as a sociologist or anthropologist for ensuring participatory input and qualitative analysis at different stages of the impact evaluation. Both should be involved in writing the evaluation report.
- Sampling expert—The sampling expert can guide the sample selection process. For quantitative data, the sampling expert should be able to carry out power calculations to determine the appropriate sample sizes for the indicators established, select the sample, review the results of the actual sample versus the designed sample, and incorporate the sampling weights for the analysis. For qualitative data, the sampling expert should guide the sample selection process in coordination with the analysts, ensuring that the procedures established guarantee that the correct informants are selected. The sampling expert should also be tasked with selecting sites and groups for the pilot test and will often need to be paired with a local information coordinator responsible for collecting for the sampling expert data from which the sample will be drawn.
- Survey designer—This could be a person or team, whose responsibility is designing the data collection instruments, accompanying manuals and codebooks, and coordinating with the evaluation manager(s) to ensure that the data collection instruments will indeed produce the data required for the analysis. This person or team should also be involved in pilot testing and refining the questionnaires.
- Fieldwork manager and staff —The manager should be responsible for supervising the entire data collection effort, from planning the routes for the data collection to forming and scheduling the fieldwork teams, generally composed of supervisors and interviewers. Supervisors generally manage the fieldwork staff (usually interviewers, data entry operators, and drivers) and are responsible for the quality of data collected in the field. Interviewers administer the questionnaires. In some cultures, it is necessary to ensure that male and female interviewers carry out the surveys and that they are administered separately for men and women.
- Data managers and processors—These team members design the data entry programs, enter the data, check the data's validity, provide the needed data documentation, and produce basic results that can be verified by the data analysts.

In building up the evaluation team, there are also some important decisions that the evaluation manager must make about local capacity and the appropriate institutional arrangements to ensure impartiality and quality in the evaluation results. First is whether there is local capacity to implement the evaluation, or parts of it, and what kind of supervision and outside assistance will be needed. Evaluation capacity varies greatly from

country to country, and although international contracts that allow for firms in one country to carry out evaluations in another country are becoming more common (one example is the Progresa evaluation being carried out by the International Food and Policy Research Institute), the general practice for World Bank-supported projects seems to be to implement the evaluation using local staff while providing a great deal of international supervision. Therefore, it is necessary to critically assess local capacity and determine who will be responsible for what aspects of the evaluation effort. Regardless of the final composition of the team, it is important to designate an evaluation manager who will be able to work effectively with the data producers as well as the analysts and policymakers using the data and the results of the evaluation. If this person is not based locally, it is recommended that a local manager be designated to coordinate the evaluation effort in conjunction with the international manager.

Second is whether to work with a private firm or public agency. Private firms can be more dependable with respect to providing results on a timely basis, but capacity building in the public sector is lost and often private firms are understandably less amenable to incorporating elements into the evaluation that will make the effort costlier. Whichever counterpart or combination of counterparts is finally crafted, a sound review of potential collaborators' past evaluation activities is essential to making an informed choice.

And third is what degree of institutional separation to put in place between the evaluation providers and the evaluation users. There is much to be gained from the objectivity provided by having the evaluation carried out independently of the institution responsible for the project being evaluated. However, evaluations can often have multiple goals, including building evaluation capacity within government agencies and sensitizing program operators to the realities of their projects once these are carried out in the field. At a minimum, the evaluation users, who can range from policymakers in government agencies in client countries to NGO organizations, bilateral donors, and international development institutions, must remain sufficiently involved in the evaluation to ensure that the evaluation process is recognized as being legitimate and that the results produced are relevant to their information needs. Otherwise, the evaluation results are less likely to be used to inform policy. In the final analysis, the evaluation manager and his or her clients must achieve the right balance between involving the users of evaluations and maintaining the objectivity and legitimacy of the results.

Data Development

Having adequate and reliable data is a necessary input to evaluating project impact. High-quality data are essential to the validity of the evalua-

tion results. As discussed above, assessing what data exist is a first important step before launching any new data collection efforts. Table 2.1 links the basic evaluation methodologies with data requirements. Most of these methodologies can incorporate qualitative and participatory techniques in the design of the survey instrument, in the identification of indicators, and in input to the identification of controls, variables used for matching, or in instrumental variables.

Table 2.1 Evaluation Methods and Corresponding Data Requirements

<i>Method</i>	<i>Data requirement</i>		<i>Use of qualitative approach</i>
	<i>Minimal</i>	<i>Ideal</i>	
Experimental or randomized controls	Single project cross-section with and without beneficiaries	Baseline and follow-up surveys on both beneficiaries and nonbeneficiaries. Allows for control of contemporaneous events, in addition to providing control for measuring impact. (This allows for a difference-in-difference estimation.)	<ul style="list-style-type: none"> • Inform design of survey instrument, sampling • Identify indicators • Data collection and recording using <ul style="list-style-type: none"> – Textual data – Informal or semi-structured interviews – Focus groups or community meetings – Direct observation – Participatory methods – Photographs – Triangulation – Data analysis
Nonexperimental designs a) Constructed controls or matching	Large survey, census, national budget, or LSMS type of survey that oversamples beneficiaries	Large survey, and smaller project-based household survey, both with two points in time to control for contemporaneous events	<ul style="list-style-type: none"> – Participatory methods – Photographs – Triangulation – Data analysis
b) Reflexive comparisons and double difference	Baseline and follow-up on beneficiaries	Time series or panel on beneficiaries and comparable non-	

Table 2.1 (continued)

<i>Method</i>	<i>Data requirement</i>		<i>Use of qualitative approach</i>
	<i>Minimal</i>	<i>Ideal</i>	
c) Statistical control or instrumental variable	Cross-section data representative of beneficiary population with corresponding instrumental variables	beneficiaries Cross-section and time series representative of both the beneficiary and nonbeneficiary population with corresponding instrumental variables	

Sources: Adapted from Ezemenari, Rudqvist, and Subbarao (1999) and Bamberger

For evaluations that will generate their own data, there are the critical steps of designing the data collection instruments, sampling, fieldwork, data management, and data access. This section does not outline the step-by-step process of how to undertake a survey but rather provides a brief discussion of these steps. Some of the discussion in this section, notably regarding sampling and data management, is more relevant to evaluations based on the collection and analysis of larger-scale sample surveys using quantitative data than for evaluations using qualitative data and small sample sizes.

Deciding What to Measure. The main output and impact indicators should be established in planning the evaluation, possibly as part of a logical framework approach. To ensure that the evaluation is able to assess outcomes during a period of time relevant to decisionmakers' needs, a hierarchy of indicators might be established, ranging from short-term impact indicators such as school attendance to longer-term indicators such as student achievement. This ensures that even if final impacts are not picked up initially, program outputs can be assessed. In addition, the evaluator should plan on measuring the delivery of intervention as well as taking account of exogenous factors that may have an effect on the outcome of interest.

Evaluation managers can also plan to conduct the evaluation across several time periods, allowing for more immediate impacts to be picked up earlier while still tracking final outcome measures. This was done in the Nicaragua School Reform evaluation, in which the shorter-term impact of the reform on parental participation and student and teacher

attendance was established and the longer-term impacts on student achievement are still being assessed.

Information on the characteristics of the beneficiary population not strictly related to the impact evaluation but of interest in the analysis might also be considered, such as their level of poverty or their opinion of the program. In addition, the evaluator may also want to include cost measures in order to do some cost-effectiveness analysis or other complementary assessments not strictly related to the impact evaluation.

The type of evaluation design selected for the impact evaluation will also carry data requirements. These will be specific to the methodology, population of interest, impact measures, and other elements of the evaluation. For example, if an instrumental variable approach (one of the types of matched-comparison strategies) is to be used, the variable(s) that will serve as the instrument to separate program participation from the outcome measures must be identified and included in the data collection. This was done for the Bolivian Social Investment Fund impact evaluation, in which knowledge of the social fund and the presence of NGOs were used as instrumental variables in assessing the impact of social fund interventions.

It can be useful to develop a matrix for the evaluation, listing the question of interest, the outcome indicators that will be used to assess the results, the variable, and the source of data for the variable. This matrix can then be used to review questionnaires and plan the analytical work as was done in the evaluation of the Nicaragua Emergency Social Investment Fund (see Annex 6).

Developing Data Collection Instruments and Approaches. Developing appropriate data collection instruments that will generate the required data to answer the evaluation questions can be tricky. This will require having the analysts involved in the development of the questions, in the pilot test, and in the review of the data from the pilot test. Involving both the field manager and the data manager during the development of the instruments, as well as local staff—preferably analysts who can provide knowledge of the country and the program—can be critical to the quality of information collected (Grosh and Muñoz 1996.) It is also important to ensure that the data collected can be disaggregated by gender to explore the differential impact of specific programs and policies.

Quantitative evaluations usually collect and record information either in a numeric form or as precoded categories. With qualitative evaluations, information is generally presented as descriptive text with little or no categorization. The information may include an individual's responses to open-ended interview questions, notes taken during focus groups,

or the evaluator's observations of events. Some qualitative studies use the precoded classification of data as well (Bamberger, 2000). The range of data collection instruments and their strengths and weaknesses are summarized in table 2.2—the most commonly used technique being questionnaires.

The responses to survey questionnaires can be very sensitive to design; thus it is important to ensure that the structure and format are appropriate, preferably undertaken by experienced staff. For example, the utility of quantitative data has often been severely handicapped, for simple mechanical reasons, such as the inability to link data from one source to another. This was the case in a national education assessment in one country where student background data could not be linked to test score results, which made it impossible to assess the influence of student characteristics on performance or to classify the tests scores by students' age, gender, socioeconomic status, or educational history.

For both qualitative and quantitative data collection, even experienced staff must be trained to collect the data specific to the evaluation, and all data collection should be guided by a set of manuals that can be used as orientation during training and as a reference during the fieldwork. Depending on the complexity of the data collection task, the case examples show that training can range from three days to several weeks.

Pilot testing is an essential step because it will reveal whether the instrument can reliably produce the required data and how the data collection procedures can be put into operation. The pilot test should mimic the actual fieldwork as closely as possible. For this reason, it is useful to have data entry programs ready at the time of the pilot to test their functionality as well as to pilot test across the different populations and geographical areas to be included in the actual fieldwork.

Sampling. Sampling is an art best practiced by an experienced sampling specialist. The design need not be complicated, but it should be informed by the sampling specialist's expertise in the determination of appropriate sampling frames, sizes, and selection strategies. (The discussion on sampling included here refers primarily to issues related to evaluations that collect quantitative data from larger, statistically representative samples.) The sampling specialist should be incorporated in the evaluation process from the earliest stages to review the available information needed to select the sample and determine whether any enumeration work will be needed, which can be time consuming.

As with other parts of the evaluation work, coordination between the sampling specialist and the evaluation team is important. This becomes particularly critical in conducting matched comparisons because the sampling design becomes the basis for the "match" that is at the core of the

Table 2.2 Main Data Collection Instruments for Impact Evaluation

<i>Technique</i>	<i>Definition and use</i>	<i>Strengths</i>	<i>Weaknesses</i>
Case studies	Collecting information that results in a story that can be descriptive or explanatory and can serve to answer the questions of how and why	<ul style="list-style-type: none"> - Can deal with a full variety of evidence from documents, interviews, observation - Can add explanatory power when focus is on institutions, processes, programs, decisions, and events 	<ul style="list-style-type: none"> - Good case studies are difficult to do - Require specialized research and writing skills to be rigorous - Findings not generalizable to population - Time consuming - Difficult to replicate
Focus groups	Holding focused discussions with members of target population who are familiar with pertinent issues before writing a set of structured questions. The purpose is to compare the beneficiaries' perspectives with abstract concepts in the evaluation's objectives.	<ul style="list-style-type: none"> - Similar advantages to interviews (below) - Particularly useful where participant interaction is desired - A useful way of identifying hierarchical influences 	<ul style="list-style-type: none"> - Can be expensive and time consuming - Must be sensitive to mixing of hierarchical levels - Not generalizable
Interviews	The interviewer asks questions of one or more persons and records the respondents' answers. Interviews may be formal or informal, face-to-face or by telephone, or closed- or open-ended.	<ul style="list-style-type: none"> - People and institutions can explain their experiences in their own words and setting - Flexible to allow the interviewer to pursue unanticipated lines of inquiry and to probe into issues in depth 	<ul style="list-style-type: none"> - Time consuming - Can be expensive - If not done properly, the interviewer can influence interviewee's response

- Particularly useful where language difficulties are anticipated
- Greater likelihood of getting input from senior officials

Observation

Observing and recording situation in a log or diary. This includes who is involved; what happens; when, where, and how events occur. Observation can be direct (observer watches and records), or participatory (the observer becomes part of the setting for a period of time).

- Provides descriptive information on context and observed changes

- Quality and usefulness of data highly dependent on the observer's observational and writing skills
- Findings can be open to interpretation
- Does not easily apply within a short time frame to process change

Questionnaires

Developing a set of survey questions whose answers can be coded consistently

- Can reach a wide sample, simultaneously
- Allow respondents time to think before they answer
- Can be answered anonymously
- Impose uniformity by asking all respondents the same things
- Make data compilation and comparison easier

- The quality of responses highly dependent on the clarity of questions

- Sometimes difficult to persuade people to complete and return questionnaire
- Can involve forcing institutional activities and people's experiences into predetermined categories
- Can be time consuming

Written document analysis

Reviewing documents such as records, administrative databases, training materials, and correspondence.

- Can identify issues to investigate further and provide evidence of action, change, and impact to support respondents' perceptions
- Can be inexpensive

evaluation design and construction of the counterfactual. In these cases, the sampling specialist must work closely with the evaluation team to develop the criteria that will be applied to match the treatment and comparison groups. For example, in the evaluation of the Nicaragua school autonomy reform project, autonomous schools were stratified by type of school, enrollment, length of time in the reform, and location and matched to a sample of nonautonomous schools by using the same stratifications except length of time in the reform. This can be facilitated by having a team member responsible for the data collection work assist the sampling specialist in obtaining the required information, including data on the selected outcome indicators for the power calculations (an estimate of the sample size required to test for statistical significance between two groups), a list of the population of interest for the sample selection, and details on the characteristics of the potential treatment and comparison groups important to the sample selection process.

There are many tradeoffs between costs and accuracy in sampling that should be made clear as the sampling framework is being developed. For example, conducting a sample in two or three stages will reduce the costs of both the sampling and the fieldwork, but the sampling errors and therefore the precision of the estimates will be increased.

Once the outcome variables and population(s) of interest have been determined by the evaluation team, a first step for the sampling specialist would be to determine the power calculations (see Valadez and Bamberger 1994, pp. 382–84, for a discussion of the power calculation process). Since the power calculation can be performed using only one outcome measure, and evaluations often consider several, some strategic decisions will need to be made regarding which outcome indicator to use when designing the sample.

After developing the sampling strategy and framework, the sampling specialist should also be involved in selecting the sample for the fieldwork and the pilot test to ensure that the pilot is not conducted in an area that will be included in the sample for the fieldwork. Often initial fieldwork will be required as part of the sample selection procedure. For example, an enumeration process will be required if there are no up-to-date maps of units required for the sample (households, schools, and so forth) or if a certain population of interest, such as malnourished children, needs to be pre-identified so that it can be selected for the purpose of the evaluation.

Once the fieldwork is concluded, the sampling specialist should provide assistance on determining sampling weights to compute the expansion factors and correct for sampling errors and nonresponse. (Grosh and Muñoz 1996 provide a detailed discussion of sampling procedures as part of household survey work. Kish 1965 is considered one of the standard

textbooks in the sampling field.) And finally, the sampling specialist should produce a sampling document detailing the sampling strategy, including (a) from the sampling design stage, the power calculations using the impact variables, the determination of sampling errors and sizes, the use of stratification to analyze populations of interest; (b) from the sample selection stage, an outline of the sampling stages and selection procedures; (c) from the fieldwork stage to prepare for analysis, the relationship between the size of the sample and the population from which it was selected, nonresponse rates, and other information used to inform sampling weights; and any additional information that the analyst would need to inform the use of the evaluation data. This document can be used to maintain the evaluation project records and should be included with the data whenever it is distributed to help guide the analysts in using the evaluation data.

Questionnaires. The design of the questionnaire is important to the validity of the information collected. There are four general types of information required for an impact evaluation (Valadez and Bamberger 1994). These include

- Classification of nominal data with respondents classified according to whether they are project participants or belong to the comparison group;
- Exposure to treatment variables recording not only the services and benefits received but also the frequency, amount, and quality—assessing quality can be quite difficult;
- Outcome variables to measure the effects of a project, including immediate products, sustained outputs or the continued delivery of services over a long period, and project impacts such as improved income and employment; and
- Intervening variables that affect participation in a project or the type of impact produced, such as individual, household, or community characteristics—these variables can be important for exploring biases.

The way in which the question is asked, as well as the ordering of the questions, is also quite important in generating reliable information. A relevant example is the measurement of welfare, which would be required for measuring the direct impact of a project on poverty reduction. Asking individuals about their income level would not necessarily yield accurate results on their level of economic well-being. As discussed in the literature on welfare measurement, questions on expenditures, household composition, assets, gifts and remittances, and the imputed value of homegrown food and owner-occupied housing are generally

used to capture the true value of household and individual welfare. The time recall used for expenditure items, or the order in which these questions are asked, can significantly affect the validity of the information collected.

Among the elements noted for a good questionnaire are keeping it short and focused on important questions, ensuring that the instructions and questions are clear, limiting the questions to those needed for the evaluation, including a “no opinion” option for closed questions to ensure reliable data, and using sound procedures to administer the questionnaire, which may indeed be different for quantitative and qualitative surveys.

Fieldwork Issues. Working with local staff who have extensive experience in collecting data similar to that needed for the evaluation can greatly facilitate fieldwork operations. Not only can these staff provide the required knowledge of the geographical territory to be covered, but their knowledge can also be critical to developing the norms used in locating and approaching informants. Field staff whose expertise is in an area other than the one required for the evaluation effort can present problems, as was the case in an education evaluation in Nicaragua that used a firm specializing in public opinion polling to conduct a school and household survey. The expertise that had allowed this firm to gain an excellent reputation based on its accurate prediction of improbable election results was not useful for knowing how to approach school children or merge quantitative data sets. This lack of expertise created substantial survey implementation problems that required weeks of corrective action by a joint team from the Ministry of Education and the World Bank.

The type of staff needed to collect data in the field will vary according to the objectives and focus of the evaluation. For example, a quantitative impact evaluation of a nutrition program might require the inclusion of an anthropometrist to collect height-for-weight measures as part of a survey team, whereas the impact evaluation of an educational reform would most likely include staff specializing in the application of achievement tests to measure the impact of the reform on academic achievement. Most quantitative surveys will require at least a survey manager, data manager, field manager, field supervisors, interviewers, data entry operators, and drivers. Depending on the qualitative approach used, field staff may be similar with the exception of data entry operators. The skills of the interviewers, however, would be quite different, with qualitative interviewers requiring specialized training, particularly for focus groups, direct observation, and so forth.

Three other concerns are useful to remember when planning survey operations. First, it is important to take into consideration temporal

events that can affect the operational success of the fieldwork and the external validity of the data collected, such as the school year calendar, holidays, rainy seasons, harvest times, or migration patterns. Second, it is crucial to pilot test data collection instruments, even if they are adaptations of instruments that have been used previously, both to test the quality of the instrument with respect to producing the required data and to familiarize fieldwork staff with the dynamics of the data collection process. Pilot tests can also serve as a proving ground for the selection of a core team of field staff to carry out the actual survey. Many experienced data collectors will begin with 10 to 20 percent more staff in the pilot test than will be used in the actual fieldwork and then select the best performers from the pilot to form the actual data collection teams. Finally, communications are essential to field operations. For example, if local conditions permit their use, fieldwork can be enhanced by providing supervisors with cellular phones so that they can be in touch with the survey manager, field manager, and other staff to answer questions and keep them informed of progress.

Data Management and Access. The objectives of a good data management system should be to ensure the timeliness and quality of the evaluation data. Timeliness will depend on having as much integration as possible between data collection and processing so that errors can be verified and corrected prior to the conclusion of fieldwork. The quality of the data can be ensured by applying consistency checks to test the internal validity of the data collected both during and after the data are entered and by making sure that proper documentation is available to the analysts who will be using the data. Documentation should consist of two types of information: (a) information needed to interpret the data, including codebooks, data dictionaries, guides to constructed variables, and any needed translations; and (b) information needed to conduct the analysis, which is often included in a basic information document that contains a description of the focus and objective of the evaluation, details on the evaluation methodology, summaries or copies of the data collection instruments, information on the sample, a discussion of the fieldwork, and guidelines for using the data.

It is recommended that the data produced by evaluations be made openly available given the public good value of evaluations and the possible need to do additional follow-up work to assess long-term impacts by a team other than the one that carried out the original evaluation work. To facilitate the data-sharing process, at the outset of the evaluation an open data access policy should be agreed upon and signed, establishing norms and responsibilities for data distribution. An open data access policy puts an added burden on good data documentation and protect-

ing the confidentiality of the informants. If panel data are collected from the same informants over time by different agencies, the informants will have to be identified to conduct the follow-up work. This requirement should be balanced against the confidentiality norms that generally accompany any social sector research. One possible solution is to make the anonymous unit record data available to all interested analysts but ask researchers interested in conducting follow-up work to contact the agency in charge of the data in order to obtain the listing of the units in the sample, thereby giving the agency an opportunity to ensure quality control in future work through contact with the researchers seeking to carry it out.

Analysis, Reporting, and Dissemination

As with other stages of the evaluation process, the analysis of the evaluation data, whether quantitative or qualitative, requires collaboration between the analysts, data producers, and policymakers to clarify questions and ensure timely, quality results. Problems with the cleaning and interpretation of data will almost surely arise during analysis and require input from various team members.

Some of the techniques and challenges of carrying out quantitative analysis based on statistical methods are included in chapter 3. There are also many techniques for analyzing qualitative data (see Miles and Huberman 1994). Although a detailed discussion of these methods is beyond the scope of this handbook, two commonly used methods for impact evaluation are mentioned—content analysis and case analysis (Taschereau 1998).

Content analysis is used to analyze data drawn from interviews, observations, and documents. In reviewing the data, the evaluator develops a classification system for the data, organizing information based on (a) the evaluation questions for which the information was collected; (b) how the material will be used; and (c) the need for cross-referencing the information. The coding of data can be quite complex and may require many assumptions. Once a classification system has been set up, the analysis phase begins, also a difficult process. This involves looking for patterns in the data and moving beyond description toward developing an understanding of program processes, outcomes, and impacts. This is best carried out with the involvement of team members. New ethnographic and linguistic computer programs are also now available, designed to support the analysis of qualitative data.

Case analysis is based on case studies designed for in-depth study of a particular group or individual. The high level of detail can provide rich information for evaluating project impact. The processes of collecting and

analyzing the data are carried out simultaneously as evaluators make observations as they are collecting information. They can then develop and test explanations and link critical pieces of information.

Whether analyzing the quantitative or qualitative information, a few other general lessons related to analysis, reporting, and dissemination can also be drawn from the case examples in Annex 1.

First, analysis commonly takes longer than anticipated, particularly if the data are not as clean or accessible at the beginning of the analysis, if the analysts are not experienced with the type of evaluation work, or if there is an emphasis on capacity building through collaborative work. In the review of the case studies considered for this article, the most rapid analysis took approximately one year after producing the data and the longer analysis took close to two years. The case in chapter 3 illustrates some of the many steps involved in analysis and why it can take longer than anticipated.

Second, the evaluation manager should plan to produce several products as outputs from the analytical work, keeping in mind two elements. The first is to ensure the timing of outputs around key events when decisions regarding the future of the project will be made, such as mid-term reviews, elections, or closings of a pilot phase. The second is the audience for the results. Products should be differentiated according to the audience for which they are crafted, including government policymakers, program managers, donors, the general public, journalists, and academics.

Third, the products will have the most policy relevance if they include clear and practical recommendations stemming from the impact analysis. These can be broken into short- and long-term priorities, and when possible, should include budgetary implications. Decisionmakers will be prone to look for the “bottom line.”

Finally, the reports should be planned as part of a broader dissemination strategy, which can include presentations for various audiences, press releases, feedback to informants, and making information available on the Web. Such a dissemination strategy should be included in the initial stages of the planning process to ensure that it is included in the budget and that the results reach the intended audience.