CHAPTER 3

Not Enough Money: Addressing Budget Constraints

Step 2 of the RealWorld Evaluation (**RWE**) approach identifies five strategies for conducting an evaluation on a tight budget (see Figure 3.1). These strategies include simplifying the evaluation **design**, clarifying **client** information needs so as reduce the amount of data to be collected or the types of analysis required, making greater user of secondary data, reducing the sample size, and reducing the costs of data collection. Finally, we identify some of the common threats to validity and adequacy of the evaluation conclusions that occur when measures are taken to reduce costs.

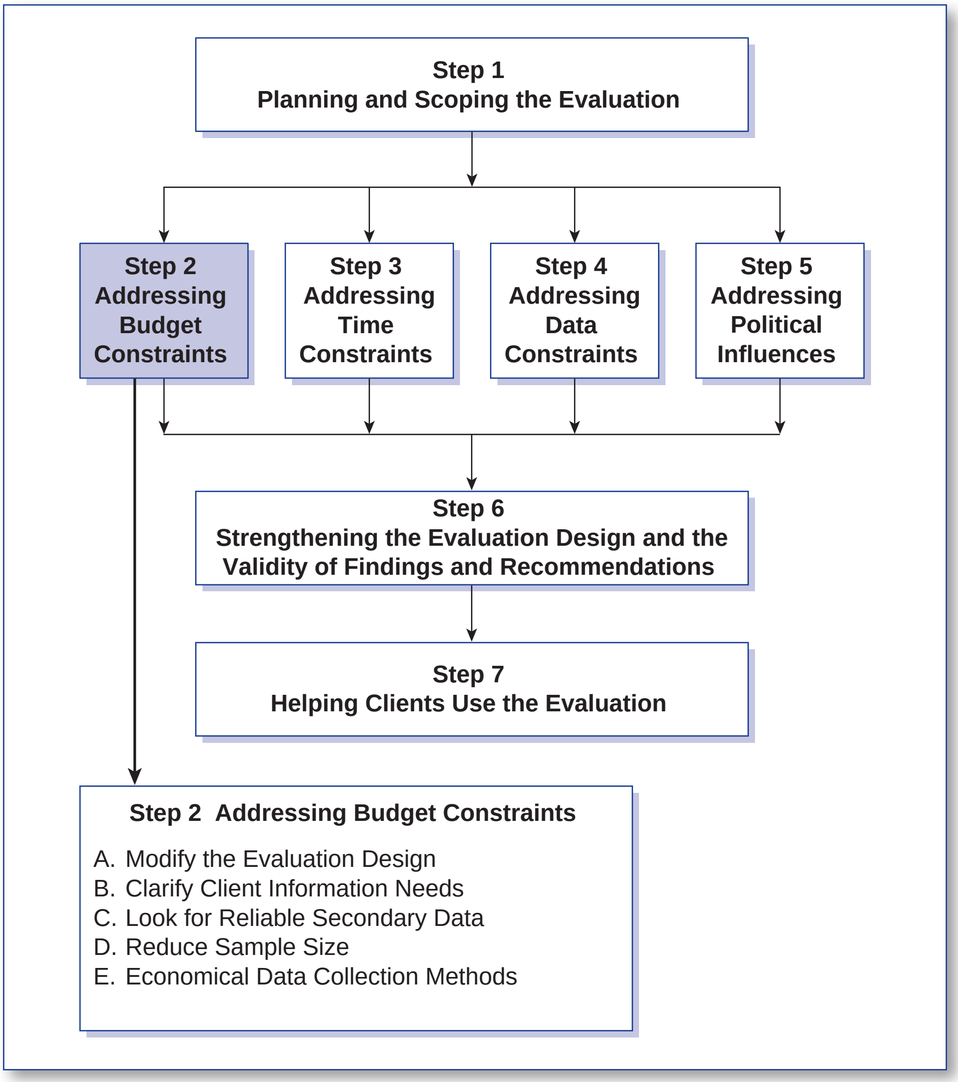
Often, **project** and program budgets include insufficient funds for evaluation, or by the time managers become concerned with **impact** issues, evaluation funds have been re-allocated to other activities. This chapter describes five strategies for addressing the budget constraints that evaluators often face (see Box 3.1 and Table 3.1):

1. Simplify the evaluation design (see also Chapter 11).
2. Clarify client information needs, seeking ways to cut out the collection of nonessential information (see also Chapter 2).
3. Look for reliable secondary data (see also Chapter 5).
4. Reduce the sample size (see also Chapter 15).
5. Use more economical data collection methods.

1. Simplifying the Evaluation Design

One way to significantly reduce the costs and time of the evaluation is to simplify the evaluation design. A review of the evaluation scenarios identified in Chapter 2 (Table 2.1) may suggest some ways in which the evaluation design could be simplified. For example, the purpose of the evaluation (points 1 and 2 in the table) and the methodological preferences of the client (points 12–14) might suggest that a relatively simple design would be acceptable.

Figure 3.1 Step 2: Addressing Budget Constraints


Step 1
Planning and Scoping the Evaluation
Step 2
Addressing
Budget
Constraints
Step 3
Addressing
Time
Constraints
Step 4
Addressing
Data
Constraints
Step 5
Addressing
Political
Influences
Step 6
Strengthening the Evaluation Design and the
Validity of Findings and Recommendations
Step 7
Helping Clients Use the Evaluation
Step 2 Addressing Budget Constraints
A. Modify the Evaluation Design
B. Clarify Client Information Needs
C. Look for Reliable Secondary Data
D. Reduce Sample Size
E. Economical Data Collection Methods

Of course, the analysis of the scenarios might also suggest that a more sophisticated and expensive design is required. When client preferences and the purpose of the evaluation permit simplification, this will often result in reducing the number of interviews or concentrating the interviews in a smaller number of physical locations so that travel time and cost can be reduced. The cost and time reduction strategies are easier to define for quantitative (QUANT) evaluations because interviews all have a similar duration and there is normally a standard cost for each interview. Consequently, it is easy to estimate the cost savings if the number of interviews is reduced. However, the cost reduction strategies are not so clearly defined for most qualitative (QUAL) evaluations because the time required to prepare a case study or conduct participant observation can vary greatly and is difficult to estimate in advance.

BOX 3.1
FIVE QUESTIONS TO HELP MAKE THE BUDGET GO FARTHER

1. Can we use a simpler and cheaper evaluation design?
2. Do we really need to collect all of this information?
3. Has someone already collected some of the information that we need?
4. Can we reduce the number of interviews, observations, cases, and so on without sacrificing the necessary precision?
5. Is there a cheaper way to collect the information?

Table 3.1 Reducing Costs of Data Collection and Analysis for Quantitative and Qualitative Evaluations

Quantitative Evaluations	Qualitative Evaluations
A. Simplifying the Evaluation Design[a]	
All these designs produce potential cost savings (see Table 3.3)[b]. • Truncated longitudinal design (Design 3): study starts at midterm • Pretest–posttest project group with posttest analysis of project and comparison groups (Design 4.1b): eliminates baseline comparison group • Posttest comparison of project and control group (Design 5): eliminates baseline • Pretest–posttest comparison of project group (Design 6): eliminates comparison group • Evaluation based on posttest data from project group (Design 7): eliminates comparison group and baseline project group	• Prioritize and focus on critical issues • Reduce the number of site visits or the time period over which observations are made • Reduce the amount and cost of data collection • Reduce the number of persons or groups studied
B. Clarifying Client Information Needs	
Prioritize data needs with the client to eliminate the collection of nonessential data.	

Quantitative Evaluations	Qualitative Evaluations
C. Using Existing Data	
• Census or surveys covering project areas • Data from project records • Records from schools, health centers, and other public-service agencies	• Newspapers and other mass media • Records from community organizations • Dissertations and other university studies (for both QUAL and QUANT)
D. Reducing Sample Size	
• Lower the level of required precision (lower precision = smaller sample) • Reduce types of disaggregation required (less disaggregation = smaller sample) • Use stratified sample designs (to reduce total interviews) • Use cluster sampling (lower travel costs)	• Consider critical or quota sampling rather than comprehensive or representative sampling • Reduce the number of persons or groups studied
E. Reducing Costs of Data Collection, Input, and Analysis	
• Self-administered questionnaires (with literate populations) • Direct observation—instead of surveys (sometimes saves money but not always) • Automatic counters and other nonobtrusive methods • Direct inputting of survey data through handheld devices • Optical scanning of survey forms and electronic surveys	• Decrease the number or period of observations • Prioritize informants • Employ and train university students, student nurses, and community residents to collect data (for both QUAL and QUANT) • Data input through handheld devices
Mixed-Method Designs	
• Triangulation to compensate for reduced sample size • Focus groups and community forums instead of household surveys • PRA and other participatory methods	

[a] See Table 3.2 for the listing of designs and Chapter 12 for a discussion of the designs.

[b] See the seven evaluation design models in Box 3.3 and Table 2.2.

1.1. Simplifying the Evaluation Design for Quantitative Evaluations

RWE approaches are used when the evaluation must be designed and implemented with budget as well as time, data, and political constraints. This means that many of the standard impact evaluation designs cannot be used. Chapter 11 reviews the principles of experimental and quasi-experimental designs and explains why the technically stronger QUANT designs cannot be used in many RWE contexts. Consequently, the RealWorld evaluator must make compromises on elements of the stronger designs because of the budget and other constraints and therefore must recognize the additional threats to the validity of the evaluation conclusions when weaker designs are used. Box 3.2 also points out that technically "robust" designs must be correctly implemented if they are not to lose their methodological strength.

BOX 3.2
"ROBUST" DESIGNS REQUIRE CORRECT DESIGN AND IMPLEMENTATION!

The fact that a "robust" design is selected does not guarantee methodologically sound conclusions unless a study is properly designed, implemented, and analyzed. If the sample is not properly selected, the survey instrument is not properly designed and administered, there is a high nonresponse rate, or if triangulation and other quality-control procedures are not used, then the conclusions may be of questionable validity. See Chapter 15 for a discussion of these issues.

Chapter 2 describes the seven basic impact evaluation design frameworks depending on when the evaluation begins (start, middle, or end of the project), at how many points in the project cycle data are collected (start, during implementation, at project completion, after project completion), and whether comparison group data are collected at the same time as data on the project population. Chapter 11 elaborates on this framework, showing that there are several design options for each of these frameworks depending on (a) whether the project and comparison groups are matched statistically (the strongest option) or judgmentally (a weaker option when secondary data are not available for statistical matching) and (b) whether baseline data are collected at the start of the project (the strongest option) or reconstructed when the evaluation does not start until late in the project (usually a weaker option). Table 11.3 (pp. 220–222) shows that the combination of the seven design frameworks with these two additional factors produces a total of 19 impact evaluation design options. Although there are a number of other factors that also affect methodological rigor (e.g., how carefully the evaluation is administered, the adequacy of the data collection instrument for collecting key information, the ability to identify and interview all sectors of the target population), it is possible to classify all of these designs into the following categories:

- The strongest statistical designs (Designs 1.1 and 2.1)
- Strong statistical designs (Designs 2.2 and 2.3)
- Acceptable statistical designs (Designs 2.3, 2.4 and 2.5)
- Weaker statistical designs (Designs 3.1, 4.1, and 5.1)
- Nonexperimental designs (Designs 6.1, 6.2, 6.3, 6.4, and 7.1).These designs do not include a matched comparison group and, consequently, it is not possible to define a conventional counterfactual to answer the question "what would have been the condition of the project population if the project had not taken place?" So from the statistical perspective, these designs are very weak. However, there are many situations in which some of these designs can be the strongest methodological design available, as they are better suited than QUANT designs for evaluating complex programs that are implemented through many steps and produce complicated behavioral changes or that are subject to the influence of many contextual factors.

There is a direct relationship between the methodological soundness of the evaluation design and the number of observation points at which surveys are conducted or other forms of data are collected. In the strongest designs (Designs 1.1 and 2.1), information is collected on both the project population and a **comparison group** before the project begins (baseline or pretest) and when the project has ended (posttest). For Design 1.1, data are also collected during project implementation and after the project has been operating for some time.These two designs also use random assignment of subjects to the project and control groups, and when properly implemented, they provide robust estimates of whether there are statistically significant differences between the project and comparison groups with respect to the indicators of project impact.

Design 2.2 cannot use random assignment but is a strong design because it is possible to use statistical matching for the project and comparison groups. Design 2.3 (regression discontinuity) can also provide strong and unbiased statistical estimates. It has the advantage that much of the baseline data can be obtained from administrative records so that data collection costs are lower (see Appendix F, pp. 556–602, for an explanation of this design and an example of its application).

Designs 2.4 and 2.5 are considered acceptable statistical designs. Design 2.4 (pipeline design) has the advantages that it is cheaper because when projects are implemented in phases over time, the Phase 2 beneficiaries can be used as the comparison group for Phase 1 (if certain conditions, discussed in Chapter 11 and Appendix F, are satisfied). This eliminates the cost and time required to identify and interview an external comparison group. Design 2.5 is similar to Design 2.2 except that judgmental matching, rather than statistical matching, of the two samples is used.

The weaker statistical designs (Designs 3.1, 4.1, and 5.1) either involve the elimination of baseline data collection for one or both samples or delay in the collection of baseline data until the project has been operating for some time.

However, with the possible exception of the regression discontinuity design, the strongest statistical designs are also the most expensive because they require that information be collected on two groups (project and comparison group) and at two or more points in time (at least before and after the project's interventions). Each of the cheaper and methodologically weaker statistical designs eliminates collection of data at one or more of these four points.

Table 3.3 estimates the cost savings from using the weaker statistical designs. These are rough estimates based on the assumption that eliminating any one of the four data collection points (pre- and posttest for project and comparison groups) will reduce the costs of data collection by about 25%. However, there are still certain fixed costs such as instrument design, sample selection, and interviewer training that are still required. It can be seen that some designs, such as the posttest-only comparison, may reduce the cost of the evaluation by up to 40%. It is more difficult to estimate the cost savings from the use of nonexperimental designs (Designs 6.1 through 7.1), as there is much greater variation in how each design is used. For example, a longitudinal design might only involve three or four visits to a community over the course of the year, but other designs may require constant contact or even living in the community for some time. The only nonexperimental design (NED)in which it is easier to estimate cost savings is Design 7.1, as this is often the default design that is used when budget and time constraints do not permit the use of other more rigorous designs. Often, this design only allows a short visit to the project locations, so there are considerable cost savings (which can be as much as 80%).

When estimating cost and time requirements, it is important to remember there is an important trade-off between cost and time savings on the one hand and validity of the evaluation findings on the other. All these less robust designs eliminate or weaken one or more of the pretest or posttest observations on the project or **comparison group** and, consequently, increase vulnerability to the

Table 3.2 Overview of Most Common RWE Impact Evaluation Designs Classified by Their Statistical Strength

Design[a]		
Strongest statistical designs (randomized assignment to project and comparison groups)		
1.1	Longitudinal comparison groups randomized design	Data collected at start, during implementation, end, and some time after project completion
2.1	Pretest–posttest comparison group randomized design	Data collected at start and end of project
Strong statistical designs (quasi-experimental designs with different selection procedures for project and comparison groups)		
2.2 A+B[b]	Pretest–posttest comparison group design with statistical matching	Groups matched using techniques such as propensity score matching
2.3	Regression discontinuity	Groups just above (participants) and below (comparison) selection cutoff point are compared to detect change in intercept or slope of regression line
Acceptable statistical designs		
2.4	Pipeline comparison group design	Phase 2 beneficiaries used as comparison group for Phase 1
2.5 A+B	Pretest–posttest comparison group design with judgmental matching	Consultation with experts and key informants combined with rapid diagnostic studies to select best match for project group
Weaker statistical designs		
3.1	Truncated pretest–posttest comparison group design	Baseline not conducted until project implementation has started
4.1 A+B	Pretest–posttest design with no comparison group baseline	
5.1	Posttest-only comparison group design	Data only available for posttest comparison with no baseline reference
Nonexperimental designs (no comparison group)		
6.1	Single-case design	Pretest–posttest comparison of single case with repeated application of intervention
6.2	Longitudinal design	Continuous observation of the project group over the life of the project
6.3	Interrupted time series	When time series data are available before and after project intervention, analysis examines whether there is a change in the line at the project intervention point
6.4 A+B	Pretest-posttest project group design	
7.1	Posttest-only project group design	This is often the default design that is used when the evaluation must be conducted with significant budget and time constraints

NOTES:

[a] This table is a summary of Table 11.3 in Chapter 11. See Appendix F for examples of each design.

[b] Some designs have two options. Option A is when the evaluation is commissioned at the start of the project so that primary baseline data can be collected. Option B is when the evaluation is not commissioned until late in the project and baseline data are "reconstructed" using secondary data, recall, key informants, or the other techniques discussed in Chapter 5. Normally, Option B is methodologically weaker, as baseline estimates may be less reliable due to the passage of time.

Table 3.3 Estimated Cost Savings for Less Robust RWE Designs Compared With Design 2.2[a]

Design		Estimated Cost Saving Compared with Design 2.2[a]
	Quasi-experimental designs (QEDs)	
2.4	Pipeline design[b]	10–20%
2.5	Pretest–posttest with judgmental matching	0%
3.1	Truncated longitudinal design	5–10%
4.1	No comparison group baseline study	10–20%
5.1	No baseline study for either group	30–40%
	Nonexperimental designs (NEDs)[c]	
6.1	Single-case design	Variable
6.2	Longitudinal design	Variable
6.3	Interrupted time series	60–70%
6.4	Pretest–posttest project group design	30–40%
7.1	Posttest-only project group	60–80%

NOTES:

[a] The estimated cost savings are based on the percentage reduction in the total number of interviews, but take into account that there are fixed costs, such as questionnaire design and training.

[b] The potential cost reductions come from the lower cost of selecting and interviewing the comparison group because it is selected from the same community as the project group. It is also likely that respondents would be more willing to cooperate, as they will be receiving benefits in a later phase, so the response rate might be higher.

[c] There is much greater variation in the design of NEDs, ranging from rapid and economical designs to expensive designs lasting over periods of years in some cases. So it is very difficult to estimate average costs and potential savings. The exception is Design 7.1, which is usually conducted on a small budget and with very limited time in the field.

four types of "threats to validity" described in Chapter 7 and Appendixes 1 through 5. It should be noted, however, that even the two most robust designs (Designs 1.1 and 2.1) are subject to a number of threats to validity (see Chapters 7 and 11 and Shadish, Cook, and Campbell 2002 for a more extended discussion).

Whenever weaker evaluation designs are used, it is strongly recommended to try to budget some additional resources and time to strengthen the design through the use of some of the "essential design components" listed in Table 2.3 (Chapter 2). Some of these strategies such as incorporating a program theory model or using triangulation may not greatly increase cost or time but can make a major contribution to strengthening the validity and utility of the evaluation findings and recommendations.

1.2. Simplifying the Design for Qualitative Evaluations

As indicated earlier, while some QUAL methods follow precise implementation guidelines (e.g., for some focus group techniques or some observation methods), in most cases, the researcher is given much more flexibility in terms of how the methods are applied. In fact, many QUAL evaluators would not accept the concept of standard designs.

Table 3.1 gives examples of how some but not all QUAL designs can be simplified to reduce costs. One approach is to identify and prioritize the critical issues that must be addressed and then to integrate all the tools to focus on the critical questions. Another option is to use recall to reduce the number of visits or the time period over which the observations are made. For example, respondents can tell the researcher about the time they spent traveling to collect fuel and water without the researcher having to accompany them on the trips. The possibility can also be considered of reducing the number of members of the household or community to be included in the study. Finally, it may be possible to simplify the research hypotheses so as to reduce the amount of data and the collection costs. For example, if the hypothesis concerns only differences of behavior between women of different age groups in the same ethnic group, the study will be simpler (and perhaps cheaper) than if both age and ethnicity are being studied.

2. 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. Some of the ways to elicit this information were discussed in Chapter 2. The approaches and issues in the clarification of information needs are often but not always similar for QUANT and QUAL evaluations.

3. Using Existing Data

Often, secondary data can be identified that obviate or reduce the need for the collection of primary data (see Chapter 5). Typical examples include the following (also see Table 3.1):

- Census or survey data covering the project and comparison communities. Many governments conduct periodic national household surveys that usually contain information on the socioeconomic conditions of households and communities and include information of interest to the evaluation. If the results can be disaggregated to the specific population reached by a project with adequate statistical validity, such secondary data can be helpful in a project evaluation.
- Data from project monitoring records (e.g., household income, type of housing, school attendance, microloans approved).
- Records from schools (e.g., enrollment, attendance, test scores), health centers (e.g., number of patients, types of illness), and other public-service agencies (e.g., water supply and sanitation, public transport).
- Newspapers and other mass media often have extensive coverage on economic and social issues that projects address (e.g., quality and availability of schools, access to health and sanitation facilities, public transport, etc.).

- Records from community organizations (minutes of meetings, photographs, posters, etc.).
- Dissertations and other academic studies.

The identification and evaluation of the validity of secondary data are discussed in Chapter 5.

4. Reducing Costs by Reducing Sample Size

4.1. Adjusting the Sample Size to Client Information Needs and the Kinds of Decisions to Which the Evaluation Will Contribute

Often, sample sizes are defined by survey researchers without reference to the kinds of decisions to be made by clients and the level of precision[1] actually required. Many clients assume that sample size is a purely technical question and that the evaluator should tell the client what is the "right" sample strategy and size. When sample size is not appropriately related to the purpose of the evaluation, how the results are to be used, and the required level of precision, larger and more costly samples may be used than are really necessary. However, in other cases, the sample may not be large enough to support the kinds of analysis required by the client. *It is absolutely essential to involve the client in decisions on the size and structure of the sample.*

The role of the evaluator is to understand the client's information needs and how the evaluation **findings** are to be used. It is critical to understand whether very precise statistical estimates are required or whether this is an exploratory study in which only general estimates of potential impacts are required. The evaluator must also present the trade-offs between precision and statistical credibility of the findings on the one hand and cost to the client on the other and agree together on the best option to provide the required information within the available budget (and time). In many cases, it is possible to reduce costs by cutting out some kinds of information or analysis included in the initial **terms of reference**, but the decision to do so should be made jointly between the client and the evaluator.

Dane (2011:Chapter 5) provides a useful explanation for clients who are not research specialists on the differences between different kinds of samples and the benefits of each. This can be helpful when discussing the trade-offs between cost and statistical rigor.

4.2. Factors Affecting Sample Size for Quantitative Evaluations

The required sample size can vary greatly according to the characteristics of the population, the nature of the project intervention, and the purpose of the evaluation. Table 3.4 lists 11 factors that affect the required sample size. All these factors are discussed in more detail in Chapter 15 (Sections 3 and 4).

The most important concepts in determining sample size are the **effect size** and the **power of the test.** The effect size refers to the size of the change (effect) that the **program** produces.

[1] *Precision* refers to the level of statistical significance used to accept that an observed project effect does not occur by chance (that an effect as strong as the one observed is not due to *pseudo effects* caused by factors unrelated to the project). The convention is to accept a 95% confidence level that the observed effect is not simply due to *statistical noise* (spurious factors). Where a higher level of precision is required, the confidence level can be increased to 99% (or even higher), but this will involve a substantial increase in the size and cost of the sample. See Chapter 15, Section 4.6, for more details.

Table 3.4 Factors Affecting the Sample Size

Factor	Explanation	Influence on Sample Size
1. The purpose of the evaluation	Is this an exploratory study, or are very precise statistical estimates required?	The more precise the required results, the larger the sample.
2. Will a one- or two-tailed test be used? (Is the direction of the expected change known?)	If the purpose of the evaluation is to test whether positive outcomes have increased or negative ones have declined, then a one-tailed test can be used. If the purpose is to test whether there has been "a significant change" without knowing the direction, a two-tailed test is required (see Chapter 15, Section 4.5).	The sample size will be approximately 40% larger for a two-tailed test.
3. Is only the project group interviewed?	In some evaluation designs, only subjects from the project group are interviewed. This is the case if information on the total population is available from previous studies or secondary data. Normally, a comparison group must also be selected and interviewed.	The sample size will be doubled if the same number of people must be interviewed in both the project and comparison groups.
4. Homogeneity of the group	If there is little variation among the population with respect to the outcome variable, then the standard deviation will be small.	The smaller the standard deviation (i.e., variability), the smaller the required sample.
5. The effect size	Effect size is the amount of increase the project is expected to produce (see Chapter 15, Section 4.2).	The smaller the effect size, the larger the required sample.
6. The efficiency with which the project is implemented	When project administration is poor, different individuals or groups may receive different combinations of services. The quality of the services can also vary. This makes it difficult to determine if lower-than-expected outcomes are due to poor project design or to the fact that many subjects are not receiving all intended services.	The poorer the quality and efficiency of the project, the larger the required sample.
7. The required level of disaggregation	In some cases, the client requires only global estimates of impact for the total project population. In other cases, it is necessary to provide disaggregated results for different project sites, for variations in the package of services provided, or for different socioeconomic groups (sex, age, ethnicity, etc.).	The greater the required disaggregation, the larger the sample.
8. The sample design	Sampling procedures such as stratification can often reduce the variance of the estimates and increase precision.	Well-designed stratification may reduce sample size.
9. The level of statistical precision	"Beyond a reasonable doubt" is usually defined as meaning there is less than a 1 in 20 possibility that an impact as large as this could have occurred by chance (defined as the "0.05 confidence level"). If less precise results are acceptable, it is possible to reduce sample size by accepting a lower confidence level—for example, a 1 in 10 possibility that the result occurred by chance.	The higher the confidence level, the larger the sample.

Factor	Explanation	Influence on Sample Size
10. The power of the test	The statistical power of the test refers to the probability that when a project has "real" effect, this will be rejected by the statistical significance test. The conventional power level is 0.8, meaning that there is only a 20% chance that a real effect would be rejected. Where a higher level of precision is required, the power can be raised to 0.9 or higher (see Chapter 15, Section 4.4).	The higher the required power level, the larger the sample.
11. Finite population correction factor	The finite population correction factor reduces the required sample size by the proportion that the sample represents of the population (see Chapter 15, Section 4.6, p. 380).	The greater the proportion the sample represents of the total population, the smaller the sample.

Where possible, use a **standardized effect size**[2] measure so that comparisons can be made between different projects using the same treatment or between projects using different treatments to produce the same effect. To estimate required sample size, it is necessary to define the minimum acceptable effect size (**MAES**). This is the minimum change (effect) that the client requires the evaluation to be able to test. Table 15.1 (Chapter 15, p. 372) presents eight different criteria that can be used to specify the MAES. In some cases, MAES is simply the expected effect size, whereas in other cases, it is derived from a comparison with other similar programs ("the project must be able to achieve an effect size at least as great as Project X"), or it might be determined by a policy objective ("to reduce the number of families with incomes less than $X"). The MAES is agreed to in consultation between the client and the evaluator. Once MAES and the required power of the test (see below) have been defined, the required sample size can be estimated.

A key determinant of the sample size is that *the smaller the effect size, the larger the required sample.* Because many projects can be expected to produce only a relatively small effect size, the required sample size to test for this effect will often be much larger than clients might have wished. In some cases, it may be concluded that the required sample size cannot be afforded, and a decision may have to be made to revise the objectives of the evaluation or even not to conduct the evaluation. This example emphasizes the importance of conducting an **evaluability assessment** during the scoping phase of the evaluation to ensure that the stated evaluation objectives could be achieved within budget, time, and data constraints (see Chapter 2).

The second key concept is the statistical power of the test (see Chapter 15). Statistical power is defined as "the probability that an estimate will be statistically significant when, in fact it represents a real effect of a given magnitude" (Rossi, Lipsey, and Freeman 2004:309). Figure 15.2 (p. 377) shows that when the effect size is small, there is a high probability that the statistical test may fail to detect the project effect even when it is real. In this figure, the power of the test is only 0.4, meaning that if 100 samples were selected, in 60 of these, the statistical test would fail to detect that the project had

[2] The standardized effect size is defined as the mean of the sample minus the population mean divided by the population standard deviation (see Chapter 15, Section 4.2).

an effect—even though the effect was real. There are two important rules with respect to sample size:

1. The smaller the effect size, the lower the power of the test.
2. The power of the test can be raised by increasing the sample size.

Table 3.5 illustrates how the required sample size is affected by the minimum decrease in health expenditures that an evaluation must be able to detect (Gertler et al. 2011, p. 186, Table 11.2). In this example, the statistical power is set at 0.9, which is reasonably high. In order to be able to detect a reduction as small as $1, a total sample size (combining treatment and comparison group) of 2,688 would be required. However, in order to detect a reduction as small as $2, the required sample size would drop to 672, and for a minimum of $3, only 300 interviews would be required. If a lower level of statistical power (0.8) could be accepted, the respective sample sizes would drop to 2,008 (compared to 2,688 with power = 0.9), 502 (compared to 672), and 224 (compared to 300). This example clearly illustrates the need for the client to define clearly the size of the effect that it must be possible to detect, as this has a dramatic effect on the sample size.

Table 3.5 Sample Size Required for Various Minimum Detectable Effects (Decrease in Household Health Expenditures), Power = 0.9. No clustering

Minimum detectable effect	Treatment group	Comparison group	Total sample
$1	1,344	1,344	2,688
$2	336	336	672
$3	150	150	300

SOURCE: Gertler et al. (2011) Table 11.2.

Table 15.4 (pp. 385–386) illustrates how effect size and the power of the test affect sample size. Using standard assumptions discussed in Chapter 15, and setting the statistical power level at 0.8, the table shows that if the project was expected to produce a relatively large change (an adjusted effect size of 0.5), then a sample of only 19 subjects would be required for both the project and comparison groups (a total of 38 subjects). However, if the project was expected to produce only a small change (an adjusted effect size of 0.2), then the sample size for each group would increase to 154 (a total of 308 subjects). In this latter case, if the client indicated that a higher statistical power must be used (setting the power of the test at 0.9 instead of 0.8), the total sample size would increase from 308 to 424.

Effect of the Level of Disaggregation on the Required Sample Size

In many cases, clients require a comparison of project impacts on different sectors of the target population, such as different regions, male- and female-headed households, people of different socioeconomic levels, or people who have received different project options or combinations of services. Each additional level or type of disaggregation normally requires a corresponding increase in sample size. Consequently, in cases where some of the levels of disaggregation can be eliminated (e.g., estimating impact on the total project population rather than for each region), it is often possible to achieve significant reductions in sample size. When the levels of disaggregation are reduced because of money constraints, other methods, such as key informant interviews, if representative and reliable, may be used to obtain information on differential impacts on various sectors within the community with at least some degree of validity.

4.3. Factors Affecting the Size of Qualitative Samples

Because QUAL sampling has different objectives from QUANT, it is usually not possible to estimate the required sample size with the same degree of statistical exactitude. The following are some of the factors affecting sample size for QUAL evaluations:

- QUAL samples can be considered as having four dimensions: (1) the number of subjects or units of observation (schools, families, drug dealers), (2) the number of physical locations in which observation takes place (the home, the place of work, the street, the bar), (3) the period of time over which the observations take place, and (4) the frequency of observations (hourly, daily, weekly, etc.). Consequently, sample costs and time can be saved by reducing the number of subjects, reducing the number of physical locations (observe only in the street or only in the school), and the duration of the study or the number of time periods over which observations are made (every day for a week, every day for a month, once a week for a year).
- The required levels of disaggregation must be determined. The more categories (types of schools, ethnic groups, farming systems) that must be compared, the larger the required sample.

The decision on the number of subjects, locations, or duration of the study or units of analysis (communities, schools, prostitutes) usually depends on the professional judgment of the researcher, and there are usually no precise rules as to whether, for example, four or six families would be the appropriate number to study or whether the observations should continue over one week or one month. Researchers are often tempted to increase the number of subjects or the duration because each additional case or observation period offers added dimensions. If pressed, however, it is often (but not always) possible for the evaluator to reduce the number of cases or observations without compromising the purpose of the evaluation.

4.4. Factors Affecting the Size of Mixed-Method Samples

Mixed-method evaluation designs combine QUANT and QUAL data collection and analysis. Consequently, decisions on sample size will usually combine decisions on both QUANT and QUAL components discussed above. Chapter 15 discusses determinants of mixed-method sample size (Section 6).

Mixed-method approaches can sometimes reduce sample size by creative combining of different techniques of data collection and the use of triangulation to check for consistency. For example, instead of conducting a sample survey to estimate community travel patterns, these might be estimated by combining observation with focus groups and key informants. Triangulation is a key element of the strategy to check for consistency and to explore further if findings from different sources are inconsistent. On the other hand, mixed methods will often try to ensure the representativity of QUAL data, and this may require a larger sample of QUAL cases than might have been the case for a purely QUAL design.

4.5. Practical Tools for Working With Small Samples: The Example of Lot Quality Acceptance Sampling (LQAS)

Lot quality acceptance sampling (LQAS) is an example of an approach designed to work with small, economical, and easily administered samples that has recently been gaining in popularity (Valadez and Devkota 2002). Originally developed to assess the achievement of coverage targets for local health delivery systems, LQAS has now been used to assess immunization coverage, antenatal care, oral rehydration, growth monitoring, family planning, disease incidence, and natural disaster relief.

LQAS is used to assess whether *coverage benchmarks* have been achieved in particular project locations. An example of a benchmark would be ensuring that 80% of families have received oral rehydration kits and orientation or that 70% of farmers have received information on new seed varieties. A major advantage of this approach is that a sample of 19 (households, farmers, etc.) will normally be sufficient to estimate whether any level of benchmark coverage has been achieved with no greater than a 10% error (see Chapter 15, Section 7.1). The findings are very simple to analyze, as the number of required positive responses is defined for any given benchmark coverage level. For example, if the target coverage level is 80%, then the sample of 19 families (farms, etc.) must include at least 13 cases in which the service had been satisfactorily received. If the target coverage was 60%, then the sample must find at least nine satisfactory cases. So in addition to the advantage of very small samples, an LQAS study is very easy for health workers, agricultural extension workers, and other nonresearch specialists to administer and interpret.

5. Reducing Costs of Data Collection and Analysis

Considerable cost savings can often be achieved through reducing the length and complexity of the data-collection instruments. The elimination of nonessential information can significantly reduce the length of the data-collection instrument or the duration of the observation. Examples of areas in which the amount of information can often be reduced include (a) demographic information on each household member, (b) amount of information on agricultural production and food consumption in a community, and (c) information on urban or rural travel patterns. It is again important to define information requirements with the client and not to arbitrarily eliminate information simply to produce a shorter data-collection instrument. For many QUAL studies, the amount and type of information cannot be defined as easily as for QUANT surveys, and consequently, the list of questions or issues cannot be pruned quite so easily. While the original QUAL design—questions, issues, methods, instruments—is available for pruning from the start, the pruning process is more complicated. With emergent designs, issues and questions arise as the research progresses, and often, many of what prove to be the critical issues were not even included on the initial lists of questions. However, the following are examples of ways to reduce the amount of information to be collected:

- The range of topics can be reduced to those of greatest priority.
- The number of interviewees can be reduced.
- The number and types of documents to be analyzed can be reduced.
- The time period studied can be shortened.

A number of alternatives can significantly reduce the costs of data collection for both QUAL and QUANT evaluations (see Table 3.1). Examples include the following:

- Collect information on community attitudes, time use, access to and use of services, and the like through **PRA** (participatory rural appraisal) group interview methods and **focus groups** rather than through household surveys or interviews with individuals. It is important to note, however, that well-designed focus groups are in themselves time consuming and, in some cases, can be more costly than surveys. Focus groups require identifying appropriate interviewees, arranging times that all members of the group can get together, preparing and field-testing the interview protocol, transcribing and validating the interview data, and conducting content analysis. In contrast, a survey requires only preparing, field-testing, administering the survey, and aggregating responses to items. The relative costs of the two approaches will, of course, depend on the proposed sample size for the survey.
- Replace surveys with direct observation—for example, to study time use, travel patterns, and use of community facilities. It is again important to note that although some types of observation can be quite rapid and economical (e.g., observation of pedestrian and vehicular travel patterns in areas with relatively few roads), in other cases, observing enough to ensure the validity of observation data and doing content analysis is not necessarily faster than a survey.
- Use key informants to obtain information on community behavior and use of services.
- Use self-administered instruments such as surveys, self-evaluations, reflection or response forms, diaries, and journals to collect data on income and expenditure, travel patterns, or time use.
- Make maximum use of preexisting data, including project records.
- Photography and videotaping can sometimes provide useful and economical documentary **evidence** on the changing quality of houses and roads, as well as on use of public transport services (Heath 2004; Kumar 1993; Patton 2002b).

Many of these suggestions involve methodological **triangulation** (Denzin 1989) to obtain conformational data in two or more ways or from two or more data sources. Triangulation is particularly important for RealWorld evaluators faced by budget and time constraints. The *triangulation by method and source* can help determine the accuracy of information when only limited amounts of data can be collected. Box 3.3 presents three illustrations of reducing the cost and time of data collection.

BOX 3.3
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.

SOURCE: Dimitrov (2005).

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 diaries, 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. Because the study covered only project participants, most of whom were very favorable toward 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, although with little analysis of external influences.

SOURCE: Valadez and Bamberger (1994:255–573).

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.

SOURCE: Dayal, van Wijk, and Mukherjee (2000).

6. Common Threats to Validity of Budget Constraints

Box 3.4 identifies some of the most common threats to validity and adequacy of evaluation conclusions that must be addressed when assessing the different approaches to budget constraints discussed in this chapter. Similar tables are included for reference in the following two chapters to identify the respective threats to validity when taking measures to reduce time or when working with limited data. It is recommended that readers who are not familiar with the concepts of threats to validity read Chapter 7 and then to return to this section.

BOX 3.4
THREATS TO ADEQUACY AND VALIDITY RELATING TO BUDGET CONSTRAINTS

NOTE. The numbers refer to the RWE "Integrated Checklist for Assessing the Adequacy and Validity of Quantitative, Qualitative, and Mixed-Method Designs" (see Appendix C). All the concepts are discussed and defined in Chapter 7.

Checklist 2. Internal Design Validity (Reliability and Dependability)

1. *How context rich and meaningful ("thick") are the descriptions?* Budget pressures often reduce the richness of the data collected.

3. *Did triangulation among complementary methods and data sources produce generally converging conclusions?* Budget constraints often reduce the use of triangulation because the application of different data-collection methods usually increases costs.

5. *Are areas of uncertainty identified? Was negative evidence sought, found?* Budget pressures can reduce the search for negative evidence.

8. *Were data collected across the full range of appropriate settings, times, respondents, etc.?* Budget pressures frequently result in the elimination of some groups—often, the most difficult to reach.

16. *History*. Budget pressures often constrain ability to control for historical differences between project and comparison areas.

24. *Use of less rigorous designs due to budget and time constraints.*

Checklist 3. Statistical Conclusion Validity

1. *The sample is too small to detect program effects.* Budget pressures often result in the sample size being reduced below the minimum size required to satisfy power analysis criteria (see Chapter 15, Section 4.6).

5 & 10. *Restriction of range and extrapolation from truncated/incomplete database.* Time pressures sometimes result in samples or secondary data with more limited coverage.

Checklist 4. Construct Validity

3. *Use of a single method to measure a construct (monomethod bias).*Budget pressures may limit the number of data collection methods or the number of independent indicators of key variables.

(Continued)

(Continued)

12. *Using indicators and constructs developed in other countries without pretesting in the local context.* Budget pressures often result in inadequate testing and customization of instruments.

Checklist 5. External Validity, Transferability, and Fittingness

7. These are often not adequately addressed when budget is a factor.

9. *Does the sample design theoretically permit generalization to other populations?* Simplifying sample design to save time can sometimes reduce representativity of the sample.

SUMMARY

- Five strategies can be considered for reducing costs of evaluation planning, data collection, and analysis. (It should be noted that each of these may reduce the validity of results obtained.)
- The first is to simplify the evaluation design, usually by eliminating the collection of data on the project or comparison group before the project begins (pretest) or on the comparison group after the project is implemented (posttest) (see Chapter 11). In the simplest design, when data are collected on only the posttest project group, the data-collection budget can be reduced by as much as 80%.
- The second is to agree with clients on the elimination of nonessential information from the data collection instruments.
- The third is to maximize the use of existing documentation (secondary data). See Chapter 5 for more details.
- The fourth is to reduce the sample size. Although this can produce significant savings, if the sample becomes too small, there is the danger of failing to detect statistically significant project effects even when they do exist. See Chapter 15 Section 4 for more details.
- The fifth is to reduce the costs of data collection through methods such as the use of self-administered questionnaires, direct observation (instead of surveys), automatic counters, inputting data through handheld devices, reducing the number of periods of observation, prioritizing informants, and hiring and training students, nurses, and other more economical data collectors. It should be noted, however, that although these methods may reduce the cost of data collection, they will not necessarily reduce or may even increase the costs of data analysis.
- Most of the above strategies for reducing costs involve trade-offs because they pose threats to the validity of the evaluation findings and recommendations. The chapter concludes with a brief introduction to the assessment of threats to validity discussed in more detail in Chapter 7.

FURTHER READING

Aron, A., E. Coups, and E. Aron. 2010. *Statistics for the Behavioral and Social Sciences: A Brief Course*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

This is a thorough but easily understandable review of all the statistical concepts discussed in this chapter.

Beebe, J. 2001. *Rapid Assessment Process: An Introduction*. Walnut Creek, CA: Altamira.

A clear overview of how to reduce the time required to conduct ethnographic studies of communities, programs, or organizations. Many of the techniques are also useful for reducing costs. Chapter 5 ("Trusting RAP") discusses some of the issues when conducting rapid evaluations but does not include a thorough discussion of threats to validity.

Bickman, L. and D. Rog. 2009. *The SAGE Handbook of Applied Social Research Methods. (2d ed.)* Thousand Oaks, CA: Sage.

A comprehensive review of evaluation research methodology. The following chapters are particularly useful for the present chapter: Chapter 1 (Applied Research Design), Chapter 5 (Randomized Control Trials), Chapter 6 (Quasi-Experimentation), and Chapter 7 (Designing a Qualitative Study).

Dane, F. 2011. *Evaluating Research: Methodology for People Who Need to Read Research.* Thousand Oaks, CA: Sage.

Explanation of the research designs discussed in this chapter from the perspective of people who read and use research. Chapters 7 (Experimental Research), 8 (Quasi-Experimental Research), and 12 (Evaluation Research) are particularly relevant for the present chapter.

Valadez, J. and B. R. Devkota. 2002. "Decentralized Supervision of Community Health Programs: Using LQAS in Two Districts of Southern Nepal" in *Community-Based Health Care: Lessons from Bangladesh and Boston*, edited by Raj Wyon. Management Sciences for Health.

Nontechnical explanation of how to use lot quality acceptance sampling, which is one of the techniques that is becoming popular for working with small samples.