Sampling Guide

January 1999

This publication was made possible through support provided by the Office of Health and Nutrition, Bureau for Global Programs, U.S. Agency for International Development, under the terms of Cooperative Agreement No. HRN-A-00-98-00046-00, the Food and Nutrition Technical Assistance Project (FANta), to the Academy for Educational Development. Additional support was provided by the Office of Food for Peace, Bureau for Humanitarian Response. Earlier drafts of the guide were developed with funding from the Food and Nutrition Monitoring Project (IMPACT) (Contract No. DAN-5110-Q-00-0014-00, Delivery Order 16), managed by the International Science and Technology Institute, Inc. (ISTI). The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development.

The U.S. Agency for International Development administers the U.S. foreign assistance program providing economic and humanitarian assistance in more than 80 countries worldwide.

Sampling Guide

January 1999

Acknowledgments

The Guide was written by Robert Magnani of the Tulane University School of Public Health and Tropical Medicine for the IMPACT project. The author wishes to thank Janet Rice, Tulane University for her helpful comments on the drafts. Eunyong Chung of the USAID Global Bureau's Office of Health and Nutrition provided useful insight and support for the development of this Guide. The Office of Food for Peace was instrumental in encouraging and supporting the Guide. Bruce Cogill, Anne Swindale, and Patrick Diskin of the IMPACT Project provided extensive comments and assistance. Special thanks to the efforts of the editor, Dorothy B. Wexler, and the layout advisor, Stacy Swartwood. The Cooperating Sponsors were essential to the development of the Guide. This Guide is dedicated to them.

Recommended Citation

Magnani, Robert. Sampling Guide. Arlington, Va.: Food Security and Nutrition Monitoring (IMPACT) Project, ISTI, Inc., for the U.S. Agency for International Development. January, 1999.

Copies of the Guide can be obtained from:

1. Food and Nutrition Technical Assistance Project (FANta), Academy for Educational Development, 1825 Connecticut Avenue, NW, Washington, D.C. 20009-5721. Tel: 202-884 8000. Fax: 202-884 8432. E-mail: fanta@aed.org. Homepage: www.fantaproject.org

2. Food Aid Management, 300 I Street, NE, Suite 212, Washington D.C., 20002. Tel: 202-544 6972. Fax: 202-544 7065. E-mail: fam@foodaid.org. Homepage: www.foodaid.org

Table of Contents

1. Purpose of Guide 1
2. Defining Measurement Objective 3
3. Determining Sample Size Requirements 7
4. Selecting the Sample 24
5. Analyzing the Data 40

Box

1. About this series 1

Figures

3-1. Illustrative informational needs for determining sample size, generic Title II infant and child feeding indicators 9
3-2. Values of Z_α and Z_β 11
3-3. Illustrative sample size calculations for indicators expressed as proportions 12
3-4. Sample sizes required for selected combinations of P_1 and changes or comparison-group differences to be detected (for α =.95 and β =.80) 13
3-5. Illustrative sample size calculation for an indicator expressed as a mean 15
3-6. Typical numbers of households to be contacted in order to find one reference individual for the generic Title II health indicators (assuming six persons per household) 17
3-7. Illustrative follow-up survey sample size calculations for indicators expressed as proportions 22
3-8. Illustrative follow-up survey sample size calculation for an indicator expressed as a mean 23
4-1. Steps in the selection of a systematic-random sample of clusters with PPS 27
4-2. Illustrative example — selection of a systematic-random sample of clusters with PPS 28
4-3. Steps in the selection of a systematic-random sample of clusters with equal probability 29
4-4. Illustrative example — selection of a systematic-random sample of clusters with equal probability 30
4-5. Steps in using the segmentation method to choose sample households 31
4-6. Example of a hypothetical cluster that has been divided into six segments 32
4-7. Map of hypothetical sample cluster showing possible starting points 33
4-8. Illustrative example of sample design and selection for a comparison area 36

5-1. Procedures for calculating sampling probabilities for sample elements (Pi) for selected two-stage cluster sampling designs 41
5-2. Illustrative computations of selection probabilities, sampling weights, and standardized sampling weights—hypothetical data 44

Appendix

Appendix 1. List of Generic Title II Indicators 46

1 Purpose of Guide

Box 1: About this series...

This series of Title II Generic Indicator Guides has been developed by the Food and Nutrition Technical Assistance (FANta) Project and its predecessor projects (IMPACT, LINKAGES), as part of USAID's support of the Cooperating Sponsors in developing monitoring and evaluation systems for use in Title II programs. These guides are intended to provide the technical basis for the indicators and the recommended method for collecting, analyzing and reporting on the generic indicators that were developed in consultation with the PVOs in 1995/1996.

Below is the list of available guides:

1. *Food Security Indicators and Framework for use in the Monitoring and Evaluation of Food Aid Programs* by Frank Riely, Nancy Mock, Bruce Cogill, Laura Bailey, and Eric Kenefick
2. *Infant and Child Feeding Indicators Measurement Guide* by Mary Lung'aho
3. *Agricultural Productivity Indicators Measurement Guide* by Patrick Diskin
4. *Sampling Guide* by Robert Magnani
5. *Anthropometric Indicators Measurement Guide* by Bruce Cogill
6. *Household Food Consumption Indicators Measurement Guide* by Anne Swindale and Punam Ohri-Vachaspati

In addition to the above categories, other guides are under preparation:

7. *Evaluation Design Guide* by Frank Riely
8. *Water and Sanitation Indicators Measurement Guide* by Pat Billig

The purpose of sampling is to reduce the cost of collecting data about a population by gathering information from a subset instead of the entire population. Sample surveys are often the most feasible means of gathering the data required for Title II program evaluations. This guide shows how to choose samples of communities, households, and/or individuals for such surveys in a manner that, when combined with appropriate indicators and evaluation study designs, will permit valid conclusions to be drawn as to the effectiveness of Title II programs. The guide emphasizes the use of probability sampling methods, which are deemed essential to ensure objectivity in program evaluations. Estimates of population characteristics derived from sample surveys

conducted following suggested guidelines may be expected to approximate the "true" population value within a specified margin of error with a known probability.

The guide was written for readers with a limited background in sampling. Knowledge of basic statistics will, however, come in handy in using the guide. Materials are presented step-by-step in the order likely to be followed in carrying out a Title II evaluation. Four principal phases are described:

1. Defining the measurement objectives of the survey. This addresses what the survey itself hopes to accomplish. It involves both the substance of study — i.e., what progress a target population has made in reaching project objectives — and the statistical issue of how precise the data needs to be.

2. Determining the sample size requirements. This explains how to calculate sample sizes after it has been decided *what* is being measured and *how precisely* it must be measured. The procedure is broken into three major steps. First, the total number of sample elements must be determined; for this, formulae are provided to identify how many individuals must sampled depending on whether progress is to be measured by changes in the *proportion* of the population that has a given characteristic or by changes in the *mean* of a given indicator (e.g., total calories consumed per capita per day). Second, the total number of elements must be converted into the number of households that must be contacted. Third, the total number of households need to be turned into the practical units (clusters and subjects within them) that will be visited by the survey team**.**

3. Selecting the sample. This defines probability sampling and explains why it is recommended. It then explains, step-by-step, various ways in which the clusters and elements mentioned above can be selected, depending on circumstances (particularly whether the size of the cluster is known or not). Suggestions on dealing with operational problems are also offered.

4. Analyzing the data. This addresses the statistical issues of calculating weights and standard errors as they arise as a result of the combination of methods used to select clusters and elements. Formulae are provided for weight calculation for several typical combinations.

This guide differs from other sampling guides intended for field-level personnel in that it provides brief explanations of the rationale for various sampling procedures and practices. The idea is that field personnel will be better equipped to adapt the procedures to local circumstances if they have an understanding of the underlying rationale for a given procedure. Illustrative examples of calculations and procedures are provided throughout the guide.

Although the guide was written to address sampling issues that are likely to arise during the course of Title II program evaluations, no document of this type can fully anticipate all of the nuances that might arise in actual applications. Accordingly, users of the guide should anticipate that they will, on occasion, need to consult with persons with sampling expertise.

2 Defining Measurement Objectives

The first step in designing a survey is defining measurement objectives. This is particularly important when the survey is to provide the primary data for program evaluations. When objectives are clearly specified, appropriate questions can be included in the survey protocols and a suitable sampling plan designed to accommodate them. When they are inadequately defined, both the survey and the program evaluation may get off track at the very outset.

Defining measurement objectives involves answering the following three questions:

- What is to be measured?
- From whom?
- At what level of precision?

What is to be measured?

The question "what is to be measured?" is usually answered in terms of *variables* or *indicators*.[1] Recommended indicators for various Title II programs are presented in other IMPACT Project guides and accordingly will not be considered at length here. For evaluations, they pertain to the types of results or effects that the program intends to produce: for example, reduction in levels of stunted children aged 6-59 months or increases in levels of exclusively breastfed children under 6 months.

Two other aspects of the question "what is to be measured" merit attention: (1) whether changes over time or changes in the target group as compared with a control group/area are to be measured, and (2) what potential confounding factors may arise when a control group/area is used.

The importance of the first issue (whether to measure changes over time or differences between project and control groups) is that the sample size requirements will vary considerably, depending on which is selected (see Chapter 3 for further discussion). The significance of the second is that any confounding factors that might skew the results in data analysis involving control groups/areas need to be identified at the outset. Naturally, every effort should be made to choose a control area that is as similar as possible as program area. But because there will almost

1. Ideally, what is to be measured would be defined by developing "dummy tables" for the final report. These tables would indicate how each data item measured in a given survey would be used in the analysis.

inevitably be some differences, any factors or variables that are thought likely to influence the outcome indicators for the evaluation need to be specified. This will enable them to be measured in the survey protocols. This issue is discussed in greater detail in the IMPACT Project Monitoring and Evaluation Guide.

From whom?

The question of *from whom* provides a basis for defining (1) the population to which the survey results may be validly extrapolated, and (2) the scope of the sampling and fieldwork operations to be undertaken.

Here, several issues need to be addressed.

Domains

A domain is a specific population or sub-group for which separate survey estimates are desired. For Title II project evaluations, domains will normally consist of either (1) the general population of the project target area or (2) the sub-population of project beneficiaries. When control groups or areas are used, this sub-population will constitute an additional domain.

Domains need to be defined at the outset since sample size requirements are determined on a per-domain basis (see Chapter 3). For example, if new project areas are to be compared with old project areas, these two groups would have to be identified as separate domains. This will ensure that the sample size in each group is sufficient to make meaningful comparisons. The process of dividing the population under study into separate sub-groups or domains is referred to as *stratification*.

> Designating a particular population or sub-group of interest as a domain is the only way to ensure that the sample size will be sufficient to reliably measure changes over time for the sub-group or differences between comparison groups.

Survey universe

The universe refers to the population and/or geographic area for which inferences may be made from the survey data. In this guide, the universe will normally be the population of the geographic area covered by the project being evaluated. If there is a control group, its universe would be the geographically defined population of non-beneficiaries from which the control sample has been chosen. If there are insufficient resources to cover the entire survey universe, a partial universe could be used — three of the five districts covered by a project, for example. This, however,

would limit the ability to generalize the survey data (and the evaluation results) to the smaller universe.[2]

Measurement units and respondents

Measurement units are the persons *to whom* the survey data refer, and respondents are the persons *from whom* the information is obtained. The measurement units are generally defined in the indicators. Indicators for Title II programs generally refer to households. An exception is infant feeding indicators in which infants/children less than 24 months of age are the units. Respondents and measurement units may or may not be the same. They are the same for indicators for which respondents are asked to report information about themselves. They differ when information is obtained from what are called *proxy* respondents. For example, information for infant feeding indicators is typically obtained from the mother or caretaker of each child. Household-level indicators are usually measured by obtaining information by a knowledgeable household informant.

Measurement units and respondents for each indicator need to be identified at the outset of the survey design process as they will affect both development of the sampling plan and the quality of the survey data gathered.

What level of precision is needed?

The level or degree of *precision* required for a survey refers to the magnitude of error in the survey estimates that is considered tolerable for a particular undertaking. Surveys may be designed to provide very precise estimates or only rough approximations. The degree will differ depending upon the available resources and the intended uses of the survey data.

For surveys designed to measure change over time or differences between comparison groups, precision is specified in terms of the *smallest* change or comparison group difference that can be measured reliably. What that is will be decided by the survey designer, based on the level of statistical significance and power desired (see Chapter 3, particularly Sections 1 and 3.6). It will also depend on program targets that have been specified for a particular indicator.

For surveys designed to measure change over time or differences between comparison groups, precision is specified in terms of the smallest change or comparison group difference that it is desired to be able to reliably measure.

2. Limited ability to generalize applies even if the districts are chosen randomly. This is because randomization requires a sufficiently large number of sampling units to yield unbiased estimates. A case might be made for using a sample of districts to "represent" the larger set of districts if it could be shown empirically that the sample districts were similar in terms of key characteristics and that program implementation did not favor the districts chosen for the evaluation.

The task of specifying survey precision requirements for Title II program evaluations is simplified when the program objectives are stated in terms of objectively verifiable indicators with performance targets. For example, a Title II program in Mozambique for the 1997-2001 period calls for improved on-farm storage and food processing by 2001 and specifies that this will be measured in part by whether there has been a 30 percent increase in adoption of improved storage techniques by the end of the project.

3 Determining Sample Size Requirements

1. Factors influencing sample size decisions

The sample size required for a given survey is determined by its measurement objectives. For surveys designed to measure either changes in indicators over time or differences in indicators between project and control areas, the required sample size for a given indicator for each survey round and/or comparison group depends on five factors. The first two are population characteristics and the last three are chosen by the evaluator or survey designer. They are as follows:

- the number of measurement units in the target population
- the initial or *baseline* level of the indicator
- the magnitude of change or comparison group differences expected to be reliably measured
- the degree of confidence with which it is desired to be certain that an observed change or comparison group difference of the magnitude specified above would not have occurred by chance (the level of statistical significance), and
- the degree of confidence with which it is desired to be certain that an actual change or difference of the magnitude specified above will be detected (statistical power).

To illustrate, for an evaluation designed to measure changes over time (i.e., a one-group pre-test/post-test design), assume that the desire is to measure a decrease of 20 percentage points in the proportion of children 6-59 months of age who are underweight with 95 percent confidence and 80 percent power. If an estimated 40 percent of the children were underweight at the time of the baseline survey, the objective would be to measure a change in the prevalence of underweight children from 40 percent to 20 percent and be (1) 95 percent confident that such a decline would not have occurred by chance and (2) 80 percent confident of detecting such a decline if one actually occurred (power). The sample size calculations would answer two questions: (1) How many children ages 6-59 months (the measurement unit) would be required to accomplish the above objectives?, and (2) How many households would have to chosen in order to find this number of children?

For a post-test only evaluation comparing project and control areas, assume that the same difference — 20 percentage points — is desired between the two for a specified indicator. The sample size would be set so as to ensure being able to reliably detect such a difference between the two areas. The same principle would hold for a pre- and post-test design with treatment and control areas, except that the sample size would be set to ensure reliable detection of the difference in the degree of change.

2. Initial informational needs

Before work can start on determining the sample size, information must be assembled on two matters:

- household composition, and
- the *expected* or normal levels or rates in the indicators to be measured.

Household composition refers to the proportion of total households that might have an individual(s) in whatever sub-group is at issue. (Use of this information is explained below in Section 3.1.2 on page 15 describing how to convert the number of elements needed into the number of households that must be contacted). For example, the Title II generic indicator "Percent of infants fed extra food for two weeks after diarrhea" will require estimates of the proportion of households that is likely to have children under 24 months of age. The usual source of information for household composition is the most recent population census. Ideally, data for the target area of the program being evaluated will be available. If not, data for the next level of aggregation (e.g., district, province, or region) would be used; if that is not available, national level data may be used.

Information on expected levels or rates for the various indicators to be measured will often be more difficult to come by. For percent of infants fed extra food after diarrhea, for example, two things must be determined: (1) what proportion of these children are likely to have experienced a diarrheal episode in the two weeks prior to the survey, and (2) what proportion are likely to have been given extra food following a diarrheal episode during the time period just prior to the survey. Possible sources of information are previous surveys that may have been conducted in the country or in a neighboring country, data from the Ministry of Health or other government agencies, or "guesstimates" from knowledgeable persons. Guidance on what should be done if no reliable source of information is available is provided in Section 3.3 below.

Figure 3-1 provides examples of the kind of preliminary data on the population in question that needs to be available before work can begin on establishing the sample size for a given survey for Title II infant and child feeding indicators.

Figure 3-1: Illustrative informational needs for determining sample size, generic Title II infant and child feeding indicators

A. Information on population composition:

1. Mean number of persons per household
2. Proportion of total population that are:
   a. Children under 0-59 months of age.
   b. Children under 24 months of age.
   c. Infants under 6 months of age.
   d. Infants between the ages of 6 and 10 months.

B. Information about *expected* levels or rates in the target population:

1. Proportion of children aged 6-59 months who are stunted.
2. Proportion of children aged 6-59 months who are underweight.
3. Proportion of infants under 6 months of age who were breastfed within 1 and 8 hours of birth.
4. Proportion of infants under 6 months of age who are breastfed only.
5. Proportion of infants 6-10 months of age who are fed complementary food.
6. Proportion of children less than 24 months of age who experienced a diarrheal episode during the 2 weeks prior to the survey.
7. Proportion of children less than 24 months of age experiencing a diarrheal episode during the 2 weeks prior to the survey who were given continued feeding.
8. Proportion of children less than 24 months of age experiencing a diarrheal episode during the 2 weeks prior to the survey who were given extra food.

3. Sample size computations

3.1 Calculating the number of sample elements and households

Two steps are involved in determining survey sample size requirements for a given survey:

1) calculating the number of sample elements required in order to satisfy the measurement requirements for a given indicator, and

2) calculating how many households would have to be contacted in order to find the number of elements needed in the first step.

Formulas for these calculations are presented in Sections 3.1.1 and 3.1.2 below.

3.1.1 Calculating the number of sample elements

Indicators may be expressed either as a proportion, a mean, or a total. In the first instance, an indicator may be expressed as a proportion of the population, as in the percentage of infants up to six months who are exclusively breastfed or who are stunted (see list of generic Title II infant and child feeding indicators listed in Figure 3-6). In the second, when they are expressed as mean or total, they may be stated in terms of the amount of a particular commodity, good, or total number of people, as, for example, the average daily per capita caloric intake in a given population.

3.1.1.1 Indicators expressed as proportions

The following formula (Basic Equation 1) may be used to calculate the required sample size for indicators expressed as a percentage or proportion. Note that the sample sizes obtained are *for each survey round or each comparison group*.

Basic Equation 1: Proportions

$$\mathbf{n = D\,[(Z_\alpha + Z_\beta)^2 * (P_1\,(1 - P_1) + P_2\,(1 - P_2))\,/(P_2 - P_1)^2]}$$

KEY:

- n = required minimum sample size per survey round or comparison group
- D = design effect (assumed in the following equations to be the *default* value of 2 — see Section 3.4 below)
- P_1 = the estimated level of an indicator measured as a proportion at the time of the first survey or for the control area

P_2 = the *expected* level of the indicator either at some future date or for the project area such that the quantity (P_2 - P_1) is the size of the magnitude of change it is desired to be able to detect

Z_α = the Z-score corresponding to the degree of confidence with which it is desired to be able to conclude that an observed change of size (P_2 - P_1) would not have occurred by chance (α — the level of statistical significance), and

Z_β = the z-score corresponding to the degree of confidence with which it is desired to be certain of detecting a change of size (P_2 - P_1) if one actually occurred (β — statistical power).

Z_α and Z_β have "standard" values depending on the reliability desired. These are provided below in Figure 3-2. Note that the higher the percentage, the more sure the program will be of measuring accurate results.

Figure 3-2: Values of Z_α and Z_β

α	Z_α	β	Z_β
.90	1.282	.80	0.840
.95	1.645	.90	1.282
.975	1.960	.95	1.645
.99	2.326	.975	1.960
		.999	2.320

The use of the formula to calculate indicators expressed as proportions is shown below in Figure 3-3. Standard parameters Z_α and Z_β taken from Figure 3-2 are incorporated.

Figure 3-3: Illustrative sample size calculations for indicators expressed as proportions

Example 1: Suppose an increase of 10 percentage points in the proportion of households exhibiting proper handwashing behavior is to be measured. Assume further that at the time of the first survey, about 50 percent of households were believed to be following proper hand washing practices. In this case, $P_1 = .50$ and $P_2 = .60$. Using standard parameters of 95 percent level of significance (α) and 80 percent power (β), values from Figure 3-2 of $Z_\alpha = 1.645$ and $Z_\beta = 0.840$ are chosen. Inserting these values in the above formula yields the following result:

$$n = 2\,[(1.645 + 0.840)^2 * ((.5)(.5) + (.6)(.4))] / (.6 - .5)^2$$

$$= 2\,[(6.175 * 0.49) / .10^2]$$

$$= 2\,[(3.02575) / .01] = 2\,(302.575) = 605.15,$$

or 606 households per survey round.

Example 2: Suppose that a child-feeding program expects to increase the proportion of infants under six months given only breastmilk by 20 percentage points over a five-year period. Assume further that, at the outset, about 60 percent of infants in the target population are thought to be breastfed exclusively for six months. Thus, $P_1 = .60$ and $P_2 = .80$. Because the program wants to be quite certain of being able to detect an increase of 20 percentage points if one actually occurred, a power of 90 percent is chosen, along with the standard 95 percent level of significance. Accordingly, $Z\alpha = 1.645$ and $Z\beta = 1.282$. The required sample size is thus:

$$n = 2\,[(1.645 + 1.282)^2 * ((.6)(.4) + (.8)(.2))] / (.8 - .6)^2$$

$$= 2\,[(8.567 * 0.40) / .20^2]$$

$$= 2\,[(3.4268) / .04] = 2\,(86.67) = 171.34,$$

or 172 infants per survey round.

Figure 3-4 enables the survey designer who knows the magnitude of change and degree of precision desired to choose sample sizes without having to perform the calculations shown above. The sample sizes shown were calculated using the basic formula in this section: values for initial levels of the indicator (P_1) range from .10 to .50 and changes/differences in a given indicator of specified magnitudes ($P_2 - P_1$) range from .05 to .30. The table is for values of $\alpha = .95$ and $\beta = .80$. A comparable table for $\alpha = .95$ and $\beta = .90$ may be found in Appendix A. Section 3.5 provides guidance on how to choose these parameters.

Figure 3-4: Sample sizes required for selected combinations of P_1 and changes or comparison-group differences to be detected (for $\alpha = .95$ and $\beta = .80$)

Change/difference to be detected ($P_2 - P_1$)						
P_1	.05	.10	.15	.20	.25	.30
.10	1,075	309	152	93	63	45
.15	1,420	389	185	110	73	52
.20	1,176	457	213	124	81	56
.25	1,964	513	235	134	57	60
.30	2,161	556	251	142	90	62
.35	2,310	587	262	147	92	62
.40	2,408	606	268	148	92	62
.45	2,458	611	268	147	90	60
.50	2,458	606	262	142	87	56

Note: Sample sizes shown assume a design effect of 2.0. For values of P_1 greater than .50, use the value in the table that differs from .50 by the same amount. For example, for $P_1 = .60$, use the value for $P_1 = .40$; for $P_1 = .70$, use the value for $P_1 = .30$. This would work because the normal distribution is symmetrical around the mean. For example, suppose it was desired to show a 10 percentage point drop in stunting, from 40 percent of stunted children to 30 percent, with a design effect of 2.0. The initial level of the indicator, or P_1, would then be .40 (see left hand column), and the sample size would be 606.

3.1.1.2 For indicators expressed as means or totals

The following formula may be used to calculate sample size requirements for indicators that are expressed as means or totals. The sample size must be calculated for each survey round or comparison group.

Basic Equation 2: Means or Totals

$$\mathbf{n = D\,[(Z_\alpha + Z_\beta)^2 * (sd_1^2 + sd_2^2) / (X_2 - X_1)^2]}$$

KEY:

- n = required minimum sample size per survey round or comparison group
- D = design effect for cluster surveys (use default value of 2, as discussed in Section 3.4)
- X_1 = the estimated level of an indicator at the time of the first survey or for the control area
- X_2 = the *expected* level of the indicator either at some future date or for the project area such that the quantity $(X_2 - X_1)$ is the size of the magnitude of change or comparison-group differences it is desired to be able to detect
- sd_1 and sd_2 = *expected* standard deviations for the indicators for the respective survey rounds or comparison groups being compared
- Z_α = the z-score corresponding to the degree of confidence with which it is desired to be able to conclude that an observed change of size $(X_2 - X_1)$ would not have occurred by chance (statistical significance), and
- Z_β = the z-score corresponding to the degree of confidence with which it is desired to be certain of detecting a change of size $(X_2 - X_1)$ if one actually occurred (statistical power).

The primary difficulty in using the above formula is that it requires information on the standard deviation of the indicator being used in the sample size computations. The preferred solution is to use values from a prior survey undertaken in the setting in which a program under evaluation is being carried out. If such data are not available, data from another part of the country or a neighboring country with similar characteristics may be used. Such data are often presented in survey reports.

The above formula is applied in Figure 3-5. The standard values for Z_α and Z_β provided in Figure 3-2 are used.

Figure 3-5: Illustrative sample size calculation for an indicator expressed as a mean

Suppose that the effects of a Title II program on daily per capita calorie consumption are to be measured. At the outset of the program, it is assumed that a mean of 1,700 calories per capita per day are being consumed by the project's target population. The program target is for an increase of 20 percent, or 340 calories per capita per day. Thus, $X_1 = 1{,}700$ and $X_2 = 2{,}040$. Recent survey data from a neighboring region indicate a mean daily per capita calorie consumption of 1,892 and a standard deviation of 1,136.

In calculating sample size requirements, a constant ratio of the standard deviation to the mean caloric intake in the two survey rounds is assumed. Using the data from the neighboring country, we can approximate the standard deviation by using the ratio of the neighboring country mean to the standard deviation and applying it to the estimated X_1 and X_2 means (1,700 and 2,040). Therefore, the estimated standard deviation (sd_1) is calculated by calculating the ratio of the reference or nearby population mean to standard deviation (1,892/1,136) or 1.6655 and applying it to the estimated X_1 mean: 1,700/1.6655 = 1,021. The second standard deviation (sd_2) is calculated by applying the same ratio of the neighboring country mean to standard deviation (1,892/1,136) or 1.6655 and applying it to the estimated X_2 mean: 2,040/1.6655 = 1,225.

Based on standard parameters of 95 percent level of significance and 80 percent power, values from Figure 3-2 of $Z_\alpha = 1.645$ and $Z_\beta = 0.840$ are chosen. Inserting these values into the above formula provides the following result:

$$n = 2[(1.645 + 0.840)^2 * (1021^2 + 1225^2) / (2{,}040 - 1{,}700)^2]$$

$$= 2[(6.175)(2{,}543{,}066) / (340)^2]$$

$$= 2[15{,}705{,}432 / 115{,}600] = 2(135.843) = 272$$

or 272 households per survey round.

3.1.2 Determining the number of households that need to be contacted

The above computations enable the survey designer to know how many sample *elements* need to be contacted to measure changes/differences in key indicators. Because not all households will have a member who fits into the category indicated in the indicator (e.g., children under six months of age or children who have experienced a diarrheal episode in the last two weeks), more households will normally need to be contacted during the survey fieldwork than the number of elements indicated. The next step, then, is to convert the sample size requirements expressed in terms of *elements* into a sample size expressed in terms of *households*. (Note that the two-step

sample size calculation procedure is not needed for indicators measured at the household level, since by definition household-level indicators may be measured for every household chosen in a given sample).

For example, for the indicator "Percent of infants/children <24 months breastfed within the first hour of life," normally only 6-8 percent of the project target area's population will consist of infants/children less than 24 months of age. (The range reflects the level of fertility in the population; the higher the level, the higher the proportion of infants and children in the population). Thus, in a population where the average number of persons per household is 6.0, it would be expected that 36-48 percent of households would contain infants/children less than 24 months of age.

The next step is convert this information into the number of households that must be contacted in order to find the required number of elements. Suppose in this case it had been determined that 300 infants/children under 24 months of age would be needed to yield the breastfeeding indicator used above. Assuming further that this is a high-fertility population, then the number of households would be calculated as 300/(.08 * 6) or 300/.48, equaling a total of n=625 households.

Note also that in the discussion above, it is assumed that *all* eligible subjects for a given indicator found in each sample household are to be included in the sample (e.g., all children 6-59 months of age). An alternative strategy would be to sample only one eligible subject per household (irrespective of how many are found). This, however, would result in more households having to be contacted, increasing survey fieldwork costs. The *take-all* strategy will also simplify the calculation of sampling weights since a within-household sampling fraction will not need to be calculated (see Figure 5-1, D, for formula for this calculation). The disadvantage of selecting all eligible subjects within sample households is that unless more complex methods of estimating sampling errors are used, the estimated levels of sampling error for survey estimates may be biased downward due to within-household clustering. As a practical matter, the magnitude of this bias is usually small, and unless a statistician and appropriate computer software are available during the analysis stage, this is a risk that is often accepted in survey undertakings (see Chapter 5, Section 3).

Sometimes, however, choosing only one subject per household will be necessary because sampling all eligible subjects will lead to survey interviews that are too lengthy. Unfortunately, detailed data on the proportion of households that contain subjects satisfying the criteria for the various indicators will rarely (if ever) be available, making it difficult to determine in advance the exact number of households that would have to be contacted in order to reach a given target sample size. Perhaps the best that can be done is calculate the number of households to be contacted as described above, and then add a *cushion* of an additional 20-25 percent of households to compensate because, in some cases, some of the eligible subjects within a household will not be chosen. It should be recognized, however, that this is only a crude approximation.

Figure 3-6 provides guidance on typical numbers of households that will have to be contacted to find a single measurement unit for each generic Title II health indicator.

Figure 3-6: Typical numbers of households to be contacted in order to find one reference individual for the generic Title II health indicators (assuming six persons per household)

Indicator	Reference Individual	Pct. of Population[1]	No. of Households Needed[1]
% stunted	children 6-59 months	14-19%	1.2/0.9
% underweight	children 6-59 months	14-19%	1.2/0.9
% breastfed within 8 hours of birth	infants < 24 months	6-8%	2.8/2.1
% breastfed only	infants < 6 months	1.5-2.0%	11.1/8.3
% fed complementary food	infants 6-10 months	1.0-1.5%	16.7/11.1
% who experienced a diarrheal episode in last 2 weeks	children < 24 months	6-8%	2.8-2.1
% experiencing a diarrheal episode given continued feeding	children < 24 months with diarrheal episode	1.5-2.0%[2]	11.1/8.3
% experiencing a diarrheal episode given extra food	children < 24 months with diarrheal episode	1.5-2.0%[2]	11.1/8.3

1. Ranges shown are for low/high fertility populations, respectively.
2. Assumes 25% frequency of diarrhea in preceding two weeks. Local estimates may be substituted.

3.2 Choosing indicators to determine sample size requirements

A number of indicators will be measured in a typical survey for a Title II program evaluation. Ideally, the requirements for each indicator would be considered in determining sample size needs for any given survey. Where the number of indicators to be measured is large, however, this would be cumbersome.

This problem is usually addressed in one of two ways. One option is to determine which of the indicators is likely to be the most demanding in terms of sample size and use the sample size required for that indicator. In doing so, the requirements of all other indicators will be satisfied. In most cases, this will be the indicator whose measurement unit is the most infrequently found in the target population. For example, in Figure 3-6, this will be "Percent of infants 6-10 months of age fed complementary foods," since only about 1.0-1.5 percent of the target population will be infants in this age range. The major advantage of this procedure is that it will automatically ensure an adequate sample size for all indicators to be measured. The downside is that a larger sample size will be chosen than is required for some/many indicators.

A second approach would be to identify a small number of indicators that are felt to be the most important for program evaluation purposes and limit sample size computations to these. This will ensure an adequate sample size for key indicators. The drawback is that an adequate sample size may not result for other indicators that may be more demanding in terms of sample size requirements. But since these have already been judged of secondary importance, this may constitute a reasonable compromise, especially where time and resources are limited.

Another compromise when financial and logistical considerations must be taken into account is to calculate sample size requirements for both the key indicators and for the most demanding indicator in terms of sample size and then to choose the largest feasible sample size between the two. This option ensures both adequate sample sizes for key indicators and the best estimates possible for the more demanding indicators given available resources. This approach is often used.

3.3 Choosing initial indicator values when they are unknown

The *baseline* value of an indicator expressed as proportions — that is, P_1 — is ideally informed by information available from other surveys that have been conducted in a given setting (i.e., as stated in Section 2 above, previous surveys that may have been conducted in the region or in a neighboring country; data from the Ministry of Health or other government agencies; or, when such information is unavailable, "guesstimates" from knowledgeable persons based on the best sources available). In choosing a value for P_1, it is best to lean toward a value of .50. The reason for this is that the variance of indicators that are measured as proportions reach their maximum as they approach .50. The safest course would be to always choose P_1=.5, as this will ensure an adequate sample size irrespective of what the actual value of P_1 is. This will, however, also result in samples that are larger than needed in the event that the actual value of P_1 is very different from .50. Thus, the recommended approach is to make the best guess based upon available

information, and lean toward selecting the value of P_1 closer to .50. For example, if it were thought that an indicator was in the .30-.40 range at the time of the baseline survey, .40 should be chosen.

For indicators expressed as means or totals, estimates are needed not only for the baseline at the time of the first survey or control area (X_1) but also for the standard deviation of X. The crucial parameter here is the standard deviation. Unfortunately, there is no simple rule of thumb in the event that data from other surveys are not available. The best advice is to be conservative and choose a value for the standard deviation that is a large proportion of assumed starting value of the indicator; for example, standard deviations that are 60-80 percent of the working estimate of X should be adequate in most cases.

3.4 Design effects

Both basic equations above include "D" for design effect. This provides a correction for the loss of sampling efficiency resulting from the use of cluster sampling instead of simple random sampling (see Chapter 4). It may be thought of as the factor by which the sample size for a cluster sample would have to be increased in order to produce survey estimates with the same precision as a simple random sample. The magnitude of D depends upon two factors: (1) the degree of similarity or homogeneity of elements within clusters, and (2) the number of measurement units to be taken from each cluster.

Ideally, an estimate of D for the indicators of interest could be obtained from a prior survey in a given setting. This will give some idea of the similarity or homogeneity among elements in the cluster. Short of this, *typical* values from surveys conducted elsewhere could be used. Unfortunately, such guidance is often not available and thus a default value of 2.0 is commonly used, especially for anthropometric and immunization surveys. Assuming that cluster sample sizes can be kept moderately small (see Section 4 below for further discussion of cluster sizes), the use of a standard value of D=2.0 should adequately compensate for the use of cluster sampling in most cases.

3.5 Significance and power

Statistical significance (α) and statistical power (β) can be thought of as analogous to false positives and false negatives: statistical significance guards against falsely concluding that a change has occurred, whereas statistical power guards against a false conclusion that nothing has happened as a result of a program. Of the two, perhaps more important to program evaluations is the power parameter, β, since it ensures that a program is not judged a failure when it in fact has had a positive result. Unless sample sizes are sufficient to be able to reliably detect changes or comparison-group differences of a specified size, the utility of surveys as a program evaluation tool is compromised. Insufficient power may lead to a false conclusion that there were no significant changes in indicators over time or differences between project and control groups, when in fact there were *real* changes/differences that were not detectable due to the insufficient

sample size used. To ensure sufficient power, a minimum value of β of .80 should be used, and .90 is preferable where resources permit.

For α, or the level of significance, the standard in most surveys is 95 percent; this is assumed sufficient to ensure that any change observed did not occur by chance. If resources do not permit, however, this parameter might be reduced to α =.90. This lower figure entails only a modest additional risk of falsely concluding that a change has occurred or that indicators for project and control groups are different. Going below this level of significance is not advised, however, and no equivalencies for a lower figure are provided in Figure 3-2.

3.6 Allowance for non-response

Non-response is a fact of life in surveys. Although efforts to minimize the level of non-response are strongly encouraged (see Chapter 4), there are practical limits to what can be done.

In order to ensure that target sample sizes for surveys are reached, allowances for non-response are customarily made during the calculation of sample size requirements. This normally involves increasing the sample size by a non-response *insurance* factor. Although this will vary somewhat from setting to setting, an allowance of 10 percent should prove adequate in most situations. Thus, if the sample size calculated for a survey called for n=1,000 households and a 10 percent cushion for non-response were to built in to the sample design, the revised target sample size for the survey would be n=1,100 households.

4. Determining the number of clusters and number of subjects per cluster to be chosen

Once overall sample size requirements have been determined, the final step in developing the sample design is to determine how many clusters and how many households per cluster should be chosen. This involves three primary considerations:

- The first is the magnitude of the cluster sampling design effect (D). The smaller the number of households per cluster, the less pronounced the design effect. This is because elementary units within clusters generally tend to exhibit some degree of homogeneity with regard to background characteristics and possibly behaviors. As the number of households per clusters increases, sampling precision is lost.
- Secondly, the numbers of households in a given cluster or site places a limit on how large the per-cluster sample could potentially be. The census listings or other materials that are to be used as a sampling frame should be carefully reviewed before deciding upon the cluster sample size to be used.
- Third, the resources available to undertake the survey fieldwork dictate what is feasible. Transporting and sustaining field staff and supervisors constitute the major costs of carrying out survey field work, and these tend to vary more or less directly with the number of clusters

to be covered. Accordingly, field costs are minimized when the number of clusters is kept small.

Because the latter two considerations are likely to vary substantially across applications and settings, only general guidance can be offered here. From a sampling precision point of view, smaller clusters are to be preferred over larger clusters. Thus, for a fixed target sample size (e.g., 600 households), a design with 30 clusters of 20 households each would be preferred to one with 20 clusters with 30 households, which is to be preferred over one with 10 clusters of 60 households. As a general rule, selecting no more than 40-50 households per cluster should be relatively safe. Of course, if resources will not permit clusters of this size, the "cluster take" could be increased, but it should be recognized that this will be at the cost of increased sampling error.

Although the use of 30 clusters in population-based surveys has become popularized, there is in fact no statistical justification for 30 as a minimum or ideal number. It nonetheless serves as a rough working guideline, representing a figure adequate to ensure that samples of target group members are sufficiently well spread across enough clusters that survey estimates are not unduly influenced by a handful of clusters.

Using 50 households as the standard sample size per cluster, the number of clusters is normally derived by dividing the sample size by 50. For example, if a sample size of n=2,000 households were required, the sample would be spread out over 40 clusters of 50 households each. If the sample size is too small (n=1,000, for instance), there will be too few clusters (in this case 20) and the sample size will not be sufficiently spread out. It would be advisable in such a situation to take 34 clusters of size 30 households (even though this results in the target sample size being exceeded by a small amount).

It is best that each cluster have the same number of sample elements. One reason is that this ensures roughly the same work load in each cluster, making operational control over the survey fieldwork somewhat easier. A second reason relates to avoidance of estimation bias by helping to ensure a self-weighting sample and is discussed in detail in Chapters 4 and 5.

5. Sample size requirements for follow-up surveys

The procedures for determining survey sample size described above are designed to take into account the requirements for a follow-up survey round of a program evaluation. In some cases, however, the sample size will need to be enlarged in the second round. In some cases, this will occur when the sample size for the first survey round did not take into account the requirements for measuring change over time; for example, if a sample size formula for a one-time survey had been used in determining the sample size for the baseline survey. In others, it will occur when an appropriate sample size formula was used, but the levels of the indicators observed in the baseline survey were different from those expected when the sample size calculations were made prior to the survey. Specifically, in cases when the indicator was measured as a proportion, the observed value in the baseline survey turned out to be substantially closer to .50 than had been expected; when the indicator was measured as a mean or total, the standard deviation of the indicator turned

out to be much larger than anticipated. In both of these cases, the sample size used for the baseline survey would be too small to satisfy the precision requirements for the evaluation effort if used for the follow-up survey.

Two ways are available to calculate how much larger the follow up sample needs to be. The first is to use one of the existing formulae, but this is problematic since these take into account statistical significance, but not power. Instead, the recommended solution is to (a) compute a revised estimate of sample size requirement using Basic Equation 1 or 2 taking into account the results of the baseline survey and (b) compensate for any shortcoming in sample size in the baseline survey by further increasing the sample size for the follow-up survey. This two-step procedure is illustrated in Figures 3-7 and 3-8.

Figure 3-7: Illustrative follow-up survey sample size calculations for indicators expressed as proportions

Suppose that it were desired to measure an increase in the proportion of households exhibiting proper hand washing behavior of 10 percentage points. At the time of the baseline survey, it was thought that about 30 percent of households followed proper hand washing practices. Thus, P_1 and P_2 were set to .30 and .40, respectively. Using the formula for indicators expressed as proportions (Basic Equation 1) and assuming standard parameters of 95 percent level of significance and 80 percent power, a sample size of n=556 is calculated.

The baseline survey, however, revealed that a much higher proportion of households (50 percent) already practiced proper hand washing behavior. The procedure recommended above calls for revising the sample size requirement using the baseline survey results — that is, P_1=.50, as follows:

$$n = D\left[(Z_\alpha + Z_\beta)^2 * (P_1 (1 - P_1) + P_2 (1 - P_2)) / (P_2 - P_1)^2\right]$$

$$n = 2\left[(1.645 + 0.840)^2 * ((.5)(.5) + (.6)(.4))\right] / (.6 - .5)^2$$

$$= 2\left[(6.175 * 0.49) / .10^2\right]$$

$$= 2\left[(3.02575) / .01\right] = 2\ (302.575) = 605.15,$$ or 606 households per survey round.

To compensate for the sample size shortfall in the baseline, the difference between the initial and revised sample size estimates (denoted n_1 and n_2 respectively) is added to the sample size to be used for the follow-up survey. Thus, the follow-up survey sample size would be:

$$n = 606 + (n_2 - n_1) = 606 + (606 - 556) = 606 + 50 = 656$$

Figure 3-8: Illustrative follow-up survey sample size calculation for an indicator expressed as a mean

In Figure 3-5, sample size computations for baseline and follow-up surveys for an effort to assess the effects of a Title II program on daily per capita calorie consumption were calculated. Suppose that instead of the expected mean daily per capita calorie consumption of $X_1 = 1{,}700$ and standard deviation of $sd_1 = 1{,}020$, values of $X_1 = 1{,}800$ and $sd_1 = 1{,}280$ were observed in the baseline survey. Following the recommended procedure, sample size requirements would be recomputed as follows (using standard parameters of 95 percent level of significance and 80 percent power and recalculating sd_2 using the ratio of X_1 / sd_1 observed in the baseline survey):

$$n = D\left[(Z_\alpha + Z_\beta)^2 * (sd_1^2 + sd_2^2) / (X_2 - X_1)^2\right]$$

$$n = 2\left[(1.645 + 0.840)^2 * (1{,}280^2 + 1{,}536^2) / (2{,}160 - 1{,}800)^2\right]$$

$$= 2\left[(6.175)(3{,}997{,}696) / (360)^2\right]$$

$$= 2[24{,}686{,}672 / 129{,}600] = 2(190.484) = 380.97,$$

or 381 households per survey round.

To compensate for the sample size shortfall in the baseline, the difference between the initial (n=272) and revised sample size estimates (denoted n_1 and n_2 respectively) would then be added to the sample size to be used for the follow-up survey. The target follow-up survey sample size would thus be:

$$n = 381 + (n_2 - n_1) = 381 + (381 - 272) = 381 + 109 = 490$$

4 Selecting the Sample

1. Overview of Sampling

1.1 Probability/non-probability sampling

Sampling procedures fall into two classes: formal or probability methods and informal or non-probability methods.

- **Formal sampling methods** are based on probability sampling theory. This requires two things: (1) that every sampling unit have a known and non-zero probability of selection into the sample, and (2) that random chance be the controlling factor in the selection of sampling units. Probability sampling also tends in practice to be characterized by (1) the use of lists or sampling frames to select the sample, (2) clearly defined sample selection procedures, and (3) the possibility of estimating sampling error from the survey data (discussed further in Chapter 5, Section 3).
- **Informal sampling methods** include a number of approaches that are based on other than probability principles. Although the general intent is often to make inferences to some larger population, methods of selection tend to be more subjective. In most cases, it is assumed that the person(s) making the sample selection is/are knowledgeable about the underlying dimensions on which the phenomena under study vary and are thus able to select the sample in such a way that these are appropriately *covered* (i.e., free from bias). It is intended/hoped that the sample is representative enough for the purposes of the survey, but this cannot be known with any measurable degree of certainty. *Quota* and *purposive* sampling are two examples of the several forms of informal or non-probability sampling (see below, Sections 3.2 and 3.5).

Probability sampling methods are strongly recommended for Title II program evaluations, despite their somewhat higher costs. Granted, both types of sampling methods could produce the same results. But because they are based on statistical theory, evaluations using probability sampling have a greater degree of credibility and are more easily defensible than those based on informal sampling methods, which may be vulnerable to questions over the whether the sample is a good representation of the population or whether it is biased.

The sampling procedures described in this guide are all based on probability principles. They have been adapted to take into account the types of difficulties that are often encountered in developing country settings.

1.2 Cluster sampling

Cluster sampling tends to be the most widely used type of probability sampling. It is normally chosen in preference to simple random sampling, though this is the most straightforward and well-known probability sampling method. Random sampling, which involves choosing the units individually and directly through a random process in which each unit has the same chance (probability) of being chosen, requires complete lists of elementary units (e.g., households in the project area). Because these are rarely available and are generally prohibitively expensive to create, random sampling is unlikely to be used as a "stand-alone" sampling method for Title II evaluations. Cluster sampling, by contrast, limits the scope of the sample frame construction and field work to a subset or sample of geographic areas to be covered and thus provides a way to control field costs.

A **cluster** is simply an aggregation of sampling units of interest for a particular survey that can be unambiguously defined and can be used as a sampling unit from which to select a smaller sub-sample. Ideally, clusters should meet four criteria. (1) They should have relatively clear physical boundaries to facilitate identification in the field. (2) They should be located somewhat close to one another; otherwise, costs will soar, defeating the major purpose of cluster sampling. (3) Clusters should not include too many people; this will help minimize the amount of sampling frame development that has to be done. (4) Information on the size of the cluster should ideally be available prior to sample selection. This will permit the use of sample selection procedures designed to improve sampling efficiency, or *probability-proportional-to-size* (PPS) selection (see Section 3.1 below). (The inability to obtain measures of cluster size prior to sample selection does not preclude the use of probability sampling).

Sampling frame development normally involves two steps: (1) selection of first-stage or *primary units* and (2) selection of elementary sampling units within the primary units. In many applications, for example, villages and/or city blocks will be chosen at the first stage and a sample of households from each at the second. In some cases, individuals from households may need to be selected, adding a third step to the process.

When the population of a sample unit is considered too large, selection of the *first-stage* or *primary* sampling units may involve two steps. In such cases, the selected cluster is divided into two or more smaller clusters, one of the smaller clusters is selected at random, and the sample frame development and sampling operations are performed within the selected smaller cluster. A potential problem is that bias may occur in the selection of the smaller cluster.

2. Sampling frames

A sampling frame is a list of potential sampling units. For the types of surveys considered in this chapter, sampling frames are typically lists of census enumeration areas, corresponding roughly to villages and city blocks, from the last population census. Where such lists do not exist, a list of villages and towns that covers all of the project's target population will suffice, but additional

stages of sample selection may be needed to produce small enough units to be workable for survey field operations.

When an adequate sampling frame does not exist, two alternatives are available. One is to undertake initial sample frame development for the area in question. For example, if a list of villages is not available for a project covering an entire province, a sample of districts located within the province could first be drawn up and a sampling frame of villages developed only in sample districts. This type of multi-stage cluster sampling is often used to compensate for sampling frame inadequacies. Some type of sample frame development or updating work is common in surveys and should be done in a way to involve minimal time and costs.

A second option would be to restrict the survey to that part of the survey universe for which a sampling frame exists. This, however, would limit the ability to generalize the evaluation findings to the portion of the project area covered in the evaluation exercise, which may introduce bias. This option should be chosen only as a last resort.

3. Sample selection procedures for programs with high levels of general population coverage

In some cases, a sizeable proportion of the population in the project target area may reasonably be expected to have been *exposed* to the intervention being evaluated (i.e., to have received, or at least have had an opportunity to receive, project benefits). Programs aimed at influencing infant feeding practices, for example, are generally targeted at the general population of a geographic area chosen for an intervention. In such cases, general population surveys represent a valid means of measuring program effects.

What follows are directions on how to select samples at each stage of the sampling process: (1) selecting sample clusters, (2) selecting the sample households, and (3) selecting individual survey subjects when the need arises. Different variants are described for each stage.

3.1 Procedure for selecting sample clusters

In the case of projects with general population coverage, there may be too many clusters listed in the sampling frame to allow for complete sampling. Therefore, a first step may be to reduce the number of clusters to be selected to a practicable number. The procedure recommended, **systematic-random sampling**, involves choosing one sample cluster at random and every i^{th} cluster thereafter from the series. (I^{th} refers to the number of each cluster chosen; i.e., in Figure 4-2, the first, fifth and eight clusters are chosen). This can be done two ways, depending upon whether information is available on the size of clusters in the target population. Both methods are described below.

When measures of cluster size are available

The statistically most efficient two-stage cluster design is one in which (1) clusters are selected with probability-proportional-to-size (PPS) at the first stage of sample selection and (2) a constant number of households is chosen from each cluster at the second stage.

The term *probability-proportional-to-size* (or PPS) means that larger clusters are given a greater chance of selection than smaller clusters. Use of the PPS selection procedure requires that a sampling frame of clusters with measures of size be available or developed in advance of sample selection. A *measure of size* is simply a count or estimate of a variable that is likely to be correlated with the number of survey subjects of interest in a given cluster; for example, both the total population of a village and the total number of households are likely to be highly correlated with the number of infants/children in the cluster and are thus good measures of size for surveys where infants/children will be important measurement units. Exact counts are not necessary: rough approximations or estimates will suffice. Any inaccuracies will be corrected in the second stage of sample selection, when specific numbers of households will be chosen (see Section 3.2 below).

Figure 4-1 lists the steps involved in selecting a sample of clusters using systematic sampling with probability-proportional-to-size. Figure 4-2 applies these steps to an illustrative example in which the objective is to select 40 clusters from a total of 170 clusters.

Figure 4-1: Steps in the selection of a systematic-random sample of clusters with PPS

(1) Prepare a list of first stage sampling units (i.e., clusters) with a corresponding measure of size for each (see column 2 in Figure 4-2).

(2) Starting at the top of the list, calculate the cumulative measure of size and enter these figures in a column next to the measure of size for each unit (see column 3).

(3) Calculate the sampling interval (SI) by dividing the total cumulative measure of size for the domain or stratum (M) by the planned number of units to be selected (a) — that is, SI = M/a.

(4) Select a random number (random start or RS) between 1 and (SI). Compare this number with the cumulative measure of size column. The unit within whose cumulative measure of size the number (RS) falls is the first sample unit (see column 4).

(5) Subsequent units are chosen by adding the sampling interval (SI) to the number identified in step (4); that is RS + SI, RS + SI * 2, RS + SI * 3, etc. (see column 4).

(6) This procedure is followed until the list has been exhausted. The resulting number of units should be approximately equal to the target number of clusters.

Figure 4-2: Illustrative example — selection of a systematic-random sample of clusters with PPS

Cluster No.	**Size - No. of Households**	**Cumulative Size**	**Sampling No.**	**Cluster Selected**
001	120	120	73	X
002	105	225		
003	132	357		
004	96	453		
005	110	563	503	X
006	102	665		
007	165	839		
008	98	937	934	X
009	115	1,052		
.	.	.		
.	.	.		
.	.	.		
170 (last)	196	17,219 (M in equation)		

Total Cumulative Size 17,219

Planned no. of clusters = 40 (a in equation)
Sampling interval = 17,219/40 = 430.475
Random start between 1 and 430.475 = 73
Clusters selected = 001, 005, 008, ...

Whenever possible, clusters should be chosen with probability-proportional-to-size in sample surveys. One reason for this is that this procedure is relatively efficient in terms of sampling

precision. A second is that, if an equal number of elements is chosen in each cluster at the second stage of sample selection, the end result will be a sample in which each household has the same overall probability of selection, or is *self-weighting*. This is a great advantage during data analysis (see Chapter 5 for further discussion).

When measures of cluster size are not available

A slightly different procedure should be used when measures of size for clusters are not available prior to sample selection. In this method, all clusters will have the same chance or probability of selection, or *equal probability*, rather than the probability being related to their size. The procedures for choosing a sample of clusters with equal probability are described in Figure 4-3, and an illustrative example is provided in Figure 4-4. In the example, the objective is again to select 40 clusters from a total of 170 clusters.

Figure 4-3: Steps in the selection of a systematic-random sample of clusters with equal probability

(1) Prepare a numbered list of sites or clusters, preferably ordered geographically (e.g., by areas of a city).

(2) Calculate the sampling interval (SI) by dividing the total number or clusters in the domain (i.e., target group) (M) by the number of clusters to be selected (a) — that is, SI = M/a.

(3) Select a random number (random start or RS) between 1 and SI. The cluster on the numbered list corresponding to this number will be the first sample cluster.

(4) Subsequent units are chosen by adding the sampling interval (SI) to the number identified in step (3); that is RS + SI, RS + SI * 2, RS + SI * 3, etc.

(5) This procedure is followed until the list has been exhausted.

Figure 4-4: Illustrative example — selection of a systematic-random sample of clusters with equal probability

Cluster No.	Selection	
001		Planned no. of clusters = 40
002	X	Sampling interval = 170/40 = 4.25
003		Random start between 1 and 4.25 = 2
004		Clusters selected = 002, 006, 011, ...
005		
006	X	
007		
008		
009		
010		
011	X	
.		
.		
.		
170 (last)		

Note that in selecting sample clusters, it is important that the decimal points in the sampling interval be retained. The rule to be followed is when the decimal part of the sample selection number is less than .5, the lower numbered cluster is chosen, and when the decimal part of the sample selection number is .5 or greater, the higher numbered cluster is chosen. In the above example, the sample selection number for the third sample cluster was 10.5, and thus cluster 011 was chosen for the sample.

Again, it is assumed that a fixed number of households is to be chosen from each sample cluster. In this case, however, because the probability of selecting a cluster was not based on the number of households it contains, the procedure leads to sample elements have differing overall probabilities of selection. In other words, the sample is *non-self-weighting*. This will complicate the situation during analysis (see Chapter 5).

3.2 Procedures for selecting sample households

Ideally, sample households should be selected by creating a list or sampling frame of all households located within each cluster and choosing a sample of units using either simple random or systematic sampling. Creating such lists of households is likely to be unacceptably costly and time consuming, however. As a short-cut, three alternatives are described below: *segmentation* and two variants of a *random-walk* method.

Segmentation method

The segmentation method involves dividing sample clusters into smaller segments of approximately equal size, choosing one at random from each cluster, and interviewing all households in the chosen segment. The size of the segment (i.e., the number of households in each) should be the same as the target number of sample households to be chosen per cluster (see Section 4 in Chapter 3). For example, if it had been determined that 30 clusters would be chosen for a given survey and 50 households would be chosen per cluster (yielding a sample size of n=1,500 households), the target segment size under the segmentation method would be 50 households.

Figure 4-5 sets forth the steps involved in using the segmentation method and Figure 4-6 provides a graphic representation of a hypothetical cluster that has been created using the segmentation method.

Figure 4-5: Steps in using the segmentation method to choose sample households

(1) Calculate the number of segments to be created. Divide the number of households recorded in the last census by the target segment size. The result will be the number of segments to be created in the field. For example, if the last census indicated that there were 250 households in the cluster and the target segment size was 40 households, 6 segments would need to be created. (Note that in performing this calculation, decimal numbers of segments should be rounded to the nearest whole number).

(2) Update the cluster map. Using a map of the cluster, verify/update the external boundaries of the cluster and enter any internal features that may be useful for dividing the cluster into easily recognizable segments.

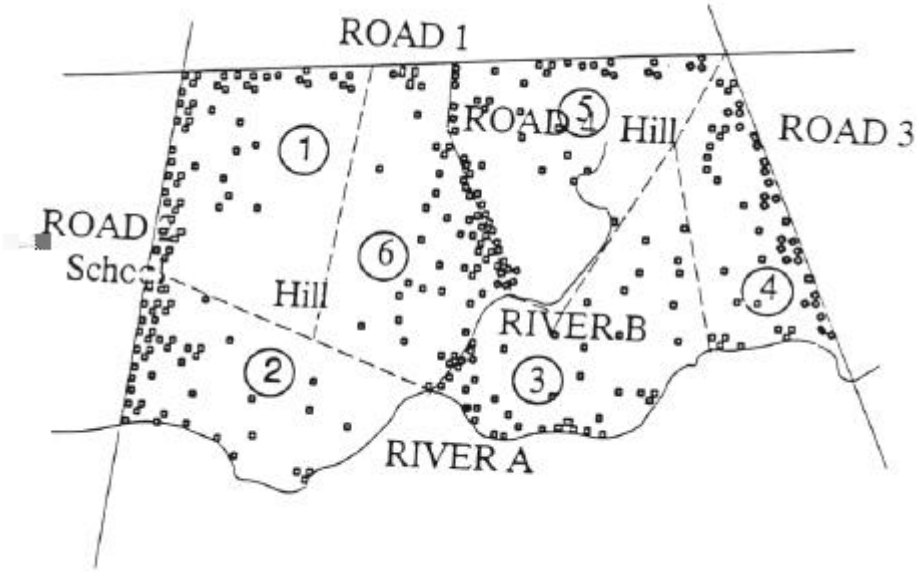
(3) Count and indicate the location of households located in the cluster on the map. This is intended to be a quick operation undertaken so that the cluster can be divided into segments with approximately equal numbers of households.

(4) Based on the cluster map, divide the cluster into equal-sized segments. The number of segments to be used is the number determined in Step 1 above.

(5) Choose one segment at random.

(6) Interview all households located within the boundaries of the randomly chosen segment.

Figure 4-6: Example of a hypothetical cluster that has been divided into six segments



Source: United Nations Children's Fund. 1995. *Monitoring Progress Toward the Goals of the World Summit for Children: A Practical Handbook for Multiple-Indicator Surveys*. New York: UNICEF.

Field work in a given sample cluster in the segmentation method is considered complete when all the households in the segment chosen for the survey have been interviewed (irrespective of how many study subjects were actually found).

Random-walk method

The random-walk method is used in EPI (expanded program of immunization) cluster surveys and thus is relatively widely known. The method entails (1) randomly choosing a starting point and a direction of travel within a sample cluster, (2) conducting an interview in the nearest household, and (3) continuously choosing the next nearest household for an interview until the target number of interviews has been obtained.

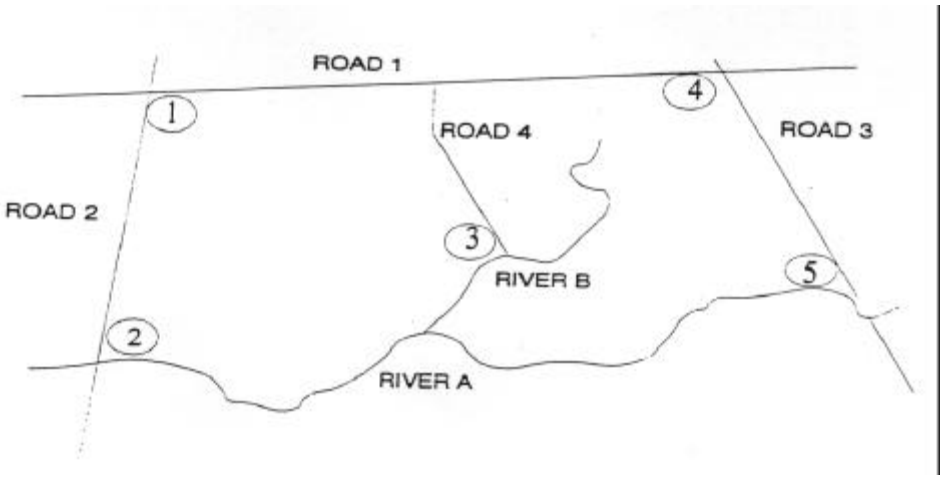
This approach can proceed in one of two ways, the only difference being in whether a map is available and how, as a result, the starting point is chosen. A summary is provided below.

Selecting the starting point from a boundary map

When a map of the sample cluster is available, a number of possible starting points should be selected at different, easily identifiable locations (see Figure 4-7 for an illustrative example), and from these a starting point should be randomly chosen. The advantage is that supervisory

personnel can choose the starting point before the field work begins, reducing any risk of bias that might arise when starting points are chosen on the basis of convenience as opposed to randomly.

Figure 4-7: Map of hypothetical sample cluster showing possible starting points



Source: United Nations Children's Fund. 1995. *Monitoring Progress Toward the Goals of the World Summit for Children: A Practical Handbook for Multiple-Indicator Surveys*. New York: UNICEF.

Selecting the starting point using the EPI method

For EPI surveys, it is assumed that no map of the cluster exists and that field staff will choose a starting location. To do so, they will follow instructions that call for (1) going to a central location in the cluster and selecting a travel direction at random by spinning a bottle, (2) moving in a straight line in that direction and counting all of the households until the edge of the cluster is reached, and (3) randomly choosing a number between 1 and the number of households counted as the starting point for the survey.

Although neither variant of the random-walk method calls for a measure of size of sample clusters, an estimate of the number of households located in each sample cluster should be obtained for Title II program evaluations when either approach is used. This allows for calculation of sampling probabilities (see Chapter 5 for further discussion). Counting or estimating number of households does not have to be terribly costly or time-consuming. In most instances, a knowledgeable local informant can provide a reasonable figure. If not, a quick tour of the cluster can usually provide an acceptably accurate count. This type of *quick count* procedure is often used in cluster sampling.

If a random-walk method is to be used, a *quota* will need to be selected, i.e., the field worker will need to continue to contact households until a predetermined number of study subjects (e.g.,

children under five years of age) has been located. The quota should be the target number of *households* to be chosen per cluster rather than numbers of *study subjects* needed for different indicators (e.g., children < 24 months of age, children experiencing a diarrheal episode in the last two weeks, etc).. Instead, if the target sample size per cluster is, for example, 50 households, the random-walk procedure would be followed until such time as this number of households had been interviewed. Quotas for households are tracked in preference to quotas for study subjects since keeping track of the latter would be difficult when multiple indicators are to be measured. As suggested above, quotas are not set for the segmentation method; rather field work in a given sample cluster is deemed complete when all the households in the segment have been interviewed (irrespective of how many study subjects were actually found).

The three methods described above vary considerably with regard to exposure to the risk of bias. The segmentation method comes the closest to approximating a conventional two-stage cluster sample and is thus less prone to bias. Because sketching a map of the community is required to use the method, however, it may not be feasible in all settings. The next best method with regard to bias is the variant of the random-walk method in which the starting sample household is chosen from a boundary map. The least preferred is the EPI random-walk method.

3.3 Procedure for selecting individual survey subjects

The general recommendation for most surveys is that all eligible subjects (e.g., children within certain age ranges) found within sample households be included in the sample. For certain types of surveys, however (e.g., when food balance sheets are to be used) and in communities where large extended families live together, including all adult females and their children in the sample would be prohibitively time-consuming and costly. The recommended procedure in such a case is to randomly choose one adult female from among those found in a given sample household and complete the interview for only that respondent and her family/children. This would constitute a third step in the sample selection process.

When one out of several possible subjects is selected from within a household, the computation of sampling weights must take this into account if unbiased survey estimates are to be obtained. This is a relatively straightforward undertaking: the surveyor need only record on the survey protocol the number of eligible potential respondents in each sample household from which the actual survey respondent was randomly chosen. This allows for calculation of the probability of having chosen the actual survey respondent. This probability is then incorporated into the calculation of an overall sampling probability (see Chapter 5, Figure 5-1 [D]).

3.4 Dealing with sampling operational problems

Implementation problems can arise in even the best-planned surveys. Typical of these are inaccessible clusters, non-response, and an insufficient number of households in a given cluster.

Inaccessible clusters

At times, it may be impossible to reach a sample cluster due, for example, to poor weather or impassable roads. Usually, the best approach is to replace the cluster with another randomly chosen cluster with similar characteristics. For example, if the cluster in question is located in the far northern part of the project area, it should be replaced with another cluster in the same general area, but one that *can* be reached during the period of survey fieldwork. To minimize the risk of bias, replacement clusters should be chosen from among similar clusters; convenience should not be an issue. Insofar as possible, supervisory personnel should make decisions on replacement clusters.

Survey non-response

Non-response is a problem common to all surveys. Typically, non-response is encountered when no one is home at sample households or when survey subjects refuse to be interviewed. Prescribed procedures exist for dealing with these problems:

Not at home: When there is no reply from a target household in a sample cluster, inquiries should be made from neighbors as to (1) whether the dwelling unit is inhabited and if so, (2) what time of the residents are usually home. If the dwelling unit is not occupied, no further action is required. If it is, at least one (and better still more) revisit(s) should be made, preferably at the time of day that the neighbor indicated that the residents were usually home.

Refusal: When the occupants of a target household refuse to be interviewed, at least one revisit, perhaps by another field team member or the team supervisor, should be made. The priority for revisits, however, should be for the not-at-homes.

Although some level of non-response is built into calculation of sample size requirements (see Chapter 3, Section 3.7), to the extent that it does occur, it can bias the survey results. This is because there are often systematic differences between people who choose to respond and those that do not and these differences may be reflected in the indicators that are being measured. The best way to deal with such possible non-response bias is to minimize non-response to the extent possible. Accordingly, field operational plans should allow sufficient time for follow-up of non-responders. Because some level of non-response was already anticipated during sample size calculations, the impact of non-response on the ultimate survey sample size should in most cases be tolerable as long as the level of non-response does not substantially exceed the expected level across a large number of clusters.

The cluster has an insufficient number of households to meet the target sample size

If the sample size computations have been performed correctly and the sizes of clusters in available sampling frames have been taken into account, there should be enough households to meet the target sample size. Furthermore, if a *cushion* has been built in to the sample size calculations as was recommended in Chapter 3, the effects on the ultimate survey sample size have already been compensated for. In the event that this situation does arise, field teams should be advised not to choose additional households from nearby clusters. Instead, they should concentrate their efforts on minimizing the number of non-response households, and then move on to the next assigned cluster.

3.5 Procedure for selecting samples in comparison areas

Comparison groups are normally expected to consist of populations of one or more nearby districts, municipalities, or other administrative units that have characteristics similar to those of the program being evaluated. The selection process normally consists of two stages. The first involves identifying groups that meet the criteria of similarity. The choice could be made *purposively* (i.e., characteristics of the group could be predefined and selection could be made according to the agreed upon criteria) unless several areas have profiles similar to the program area, in which case one could be chosen randomly. (See the Monitoring and Evaluation Guide for further guidance). Once the survey universe for the comparison area has been defined, it remains to select a sample of clusters and households to represent the comparison area. The sampling procedures are identical to those described above for general population surveys.

Figure 4-8 provides an illustrative example of sampling decisions for a comparison group.

Figure 4-8: Illustrative example of sample design and selection for a comparison area

Suppose that a program focused on improving agricultural production and processing was being conducted in a two districts of a country. A pre-test/post-test with comparison group design is being used for the program evaluation. For the comparison area, all districts adjacent to the two project districts are considered as possible candidates. Two of these are eliminated because they have different socioeconomic characteristics from the project districts and another is eliminated because it is the target of a comparable program being funded by another donor. Of the remaining adjacent districts, the two districts that are the most similar to the two project districts are chosen.

In each comparison-group district, a sample of 30 clusters is chosen using a systematic-random selection procedure with PPS, and 40 households are chosen from each sample cluster using the segmentation method in two survey rounds: a baseline survey and a follow-up survey.

4. Sample selection procedures for programs with limited population coverage

4.1 The sampling problem

Some Title II programs have quite limited levels of general population coverage. For example, the potential beneficiary population of a credit program for single mothers with children under five years of age might amount to less than five percent of households in the geographic area targeted for a program. In such cases, attempting to measure program impact at the general population level is inappropriate; it would be nearly impossible to detect changes in indicators at that level, even when sizeable changes have taken place for project beneficiaries. As a result, program evaluation efforts are often directed to assessing changes in indicators only among beneficiary households. Three alternative sampling strategies for such situations are presented below: restricting the survey universe; screening; and use of list frames.

4.2 Alternative sampling strategies

Restricting the survey universe

One approach is suitable when project beneficiaries are heavily concentrated in certain parts of the project area (e.g., in specific municipalities or villages). In such cases, the survey universe can be restricted to areas of program impact and the sampling frame would be limited to that subset of clusters. With this restriction, the sampling strategies described above could be used without further modification. This approach would not apply in most cases, since program beneficiaries are typically scattered throughout the geographic area targeted by a given project.

Screening

When beneficiaries are geographically dispersed throughout the project area, a commonly used approach is to introduce a *screening procedure*. At the start of the survey field operation, potential sample households are questioned as to whether they have benefited from the project. (Questions to this effect could be added at the beginning of the questionnaire). Field workers complete the survey interview only for households that are project beneficiaries, skipping households that are not. Alternatively, a small amount of information may be gathered from non-beneficiary households so that comparisons between beneficiaries and non-beneficiaries can be made. This approach would serve to ensure that resources are spent gathering information on project beneficiaries.

This strategy would require adding households to the target sample size. For example, if only about 20 percent of households in a target geographic area were believed to be project beneficiaries, only one in five households would be candidates to complete questionnaires. In this case, the number of sample households would need to be increased fivefold.

To illustrate, suppose that the sample size requirements for a given survey were determined to be n=1,500 households, and the sampling plan was to take 30 clusters of 50 households each. If only 20 percent of households in the project target area were thought to be project beneficiaries, then a total of n=7,500 (7,500 = 1,500 /.20) households would need to be contacted. If the target of 30 clusters were to be retained, this would require that n=250 households per cluster be contacted. In many cases, this would constitute the entire cluster, in which case the sampling strategy would be to interview all households in sample clusters. In many settings, however, the clusters in available sampling frames (i.e., census enumeration areas or other administrative units) are smaller than 250 households, requiring that a larger number of clusters be chosen. For example, if the average size of clusters in a given setting were to be 100 households, it would be necessary to choose a sample of 75 clusters and contact all households located within these clusters (7,500/100 = 75 clusters).

Use of list frames

Where lists of program beneficiaries are available, these offer a far more efficient way to choose a sample of beneficiary households than screening. Two approaches are possible here. One is to choose a sample of beneficiary households using simple random or systematic-random sampling. This, however, could result in a logistically difficult field operation, since sample households might be widely scattered over a large number of villages. Alternatively, beneficiaries on the list could be grouped geographically, and sampling performed on the resulting clusters of beneficiary households. The resulting sampling frame of clusters would be treated exactly the same as a sampling frame of more conventional clusters (i.e., census enumeration areas). The sampling procedures for general population surveys described above could then be used.

5. Design issues for follow-up surveys

5.1 Retention or replacement of sample clusters

One of the key design issues in any multi-round survey in which cluster sampling is used is whether to keep the same clusters in each survey or to choose a new sample. From a statistical point of view, the preferred strategy is retain the same clusters. This is because the background characteristics and behaviors of individuals in small geographic areas tend to be correlated over time. The effect of this correlation is to reduce somewhat the variability in survey estimates of change, making the task of detecting *real* change somewhat easier. Although the magnitude of the expected gains in sampling efficiency will vary across settings and indicators, there will almost always be some gain in measurement precision. Thus, the general recommendation is to retain the same sample of clusters while choosing a new sample of households in each cluster following the one of the procedures described above.

This strategy has one major pitfall: the temptation to allocate program resources disproportionately in favor of the areas that are known to be the focus of the program evaluation. If the program concentrates resources this way, these areas may show more positive results than may have occurred for the program area as a whole. Program managers must guard against skewing

resources toward evaluation target areas. If evaluators fear that this has occurred, the safest course would be to choose a new sample of clusters for the follow-up survey round. This could reduce sampling efficiency for the program evaluation somewhat but it would also reduce the risk of bias, which is a more important consideration in the larger scheme of things.

5.2 Effects of changes in survey methodology

Another problem is that *real* program effects may be distorted or *confounded* by changes in other factors from one round to the next. *Confounding factors* may be roughly divided into two broad categories: external and internal.

External confounding factors are changes in factors operating in the population under study over which programs and evaluators have little or no control. Examples include changes in population composition due to migration, changing economic environments, natural disasters, etc. Such factors are usually contended with in program evaluations by using control areas/groups, which help separate out the *real* program effects from those caused by other, non-program factors, and multivariate statistical methods during data analysis.

Internal confounding factors consist of changes in factors that are internal to a program evaluation effort and over which field staff responsible for the conduct of the evaluation have control. Examples are changes in the definition of the survey universe, questionnaire wording, sampling methodology, and the quality or timing (with regard to seasonality) of survey fieldwork. Especially when relatively modest sample sizes are used, evaluation findings can be quite sensitive to even modest changes in survey methodology and/or implementation. The lesson here is that survey methodology and implementation performance should be kept as constant as possible.

5.3 Sample attrition

Sometimes, plans to use the same sample of clusters in both the baseline and follow-up surveys must be scrapped because of natural disasters, security problems, etc. In such cases, the best course of action is follow the procedure recommended for inaccessible clusters and to replace the cluster for which it is not possible to obtain data with another randomly chosen cluster with similar characteristics (see Section 3.4 above). Use of random choice will minimize the risk of bias. The preference for similar characteristics is to try to limit sampling variability between survey rounds. Supervisory personnel should make the decisions on choice of replacement clusters.

If, however, the cluster is only temporarily inaccessible when the follow-up survey is scheduled, those undertaking the survey should wait until the cluster is once again accessible rather than selecting a replacement, with the proviso that this should not result in an unacceptable delay in the survey fieldwork.

5 Analyzing the Data

1. Overview

Once the survey data for a program monitoring and/or evaluation effort have been gathered and entered into a computer data base, what remains is to analyze the data. This will entail calculation of the various indicators measured in the survey(s) and assessment of the magnitude and statistical significance of changes over time and/or differences between comparison groups. Guidelines for analyzing Title II program evaluation data are provided in the companion Monitoring and Evaluation Guide. In this section, attention is limited to two analysis-related issues that are strongly influenced by the manner in which sampling was carried out: *sampling weights* and *calculation of standard errors of survey estimates*.

2. Weighting the data

Only one of the cluster sampling plans described in the previous chapter will result in self-weighting samples: that in which sample clusters are chosen with probability-proportional-to-size (PPS) and the segmentation method is used to select sample households. All of the others will result in non-self-weighting samples or samples in which sample households have unequal probabilities of selection. This must be compensated for at the data analysis stage. Failure to do so will result in estimation bias.

Weights compensate for unequal probabilities of selection. The standard method for correcting for these unequal probabilities is to apply sampling weights to the survey data during analysis by multiplying the indicator value by the weight. The appropriate sampling weight for each sample subject is simply the reciprocal of the probability of selection of that subject, or the inverse of the probability.

Basic Equation 3: Proportions

$$W_i = 1/P_i$$

KEY:

W_i = sampling weight for elements in the i^{th} cluster; and
P_i = probability of selection for elements in the i^{th} cluster.

In order to calculate sampling weights, probabilities of selection (P_i) must be calculated. The formula will depend on which variant of the two- (or three-)stage cluster sample design is used. Figure 4-1 shows calculations for sampling probabilities for four of the possible combinations: (A) PPS (first stage)/segmentation (second stage); (B) PPS/random-walk; (C) equal probability/random-walk; and (D) PPS/segmentation, but with one respondent. (Options not shown include equal probability/segmented; equal probability/random-walk, but with one respondent; and equal probability/segmentation, but with one respondent. Analysts can adapt the options to fit these other cases). Calculation of example A (probability-proportional-to-size/segmentation) will never be needed since, as demonstrated, this design results in a self-weighting sample.

Figure 5-1: Procedures for calculating sampling probabilities for sample elements (Pi) for selected two-stage cluster sampling designs

A. PPS at first stage, segmentation method at second stage:

$P_i = (m * M_i/M) * 1/S_i = m * C/M$

KEY

m = number of sample clusters chosen
M_i = measure of size for the i^{th} cluster
M = total measure of size for the survey universe (M = Mi)
S_i = number of segments created in the i^{th} cluster
C = standard (i.e., constant) segment size

Note that since this design results in a self-weighting sample, the application of sampling weights during analysis is not required.

Figure 5-1: Procedures for calculating sampling probabilities for sample elements (Pi) for selected two-stage cluster sampling designs (continued)

B. PPS at first stage, constant number of elements chosen at second stage using a random-walk method):

$P_i = (m * M_i/M) * k/N_i$

KEY

- m = number of sample clusters chosen
- M_i = measure of size for the i^{th} cluster from sample frame
- M = total measure of size for the survey universe (M = Mi)
- k = constant number of households chosen per cluster
- N_i = total number of households in the i^{th} cluster i.e., M_i updated in field with estimated or actual count of households

C. Equal probability at first stage, constant number of elements chosen at second stage using a random-walk method:

$P_i = (1/m) * (k/N_i)$

KEY:

- m = number of sample clusters chosen
- k = constant number of sample elements chosen per cluster
- N_i = total number of households in the i^{th} cluster

D. Design (B) above, but with one respondent (e.g., adult female) chosen per sample household

$P_i = (m * M_i/M) * (k/N_i) * 1/R_{ij}$

KEY:

- m = number of sample clusters chosen
- M_i = measure of size for the i^{th} cluster
- M = total measure of size for the survey universe (M = Mi)
- k = constant number of households chosen per cluster
- N_i = total number of households in the i^{th} cluster
- R_{ij} = total number of eligible respondents in the j^{th} household

A problem will arise when sampling weights are applied to data analysis performed using standard computer software packages (e.g., Epi-Info, SPSS). Applying weights to standard packages will inflate the number of sample cases and will thus imply a larger sample size than was actually the case.[3] As a result, statistical tests for differences and changes over time will be based upon incorrect sample sizes, and misleading conclusions as to the effects of programs might result. For example, changes or comparison-group differences that were not statistically significant based upon the actual sample size will appear to be significant based upon the weighted number of cases.

To compensate for this, *standardized weights* are often used. Standardized weights assign a weight to each sample observation that reflects its relative probability of selection in comparison with other sample observations, but do not change the overall survey sample size. Standardized weights (w_i') for sample elements in the i^{th} cluster are calculated as follows:

$$w_i' = w_i \, n_i \; / \; \sum w_i \, n_i$$

Since each element in a given cluster has the same probability of selection, each will also receive the same standardized weight. Figure 5-2 illustrates the computation of standardized weights using hypothetical survey data.

In order to make use of standardized sampling weights during data analysis, an appropriate weight variable will need to be included in the survey data file to be analyzed. The standardized weights could either be calculated by hand or using a spreadsheet and entered as a variable during data entry. Alternatively, the first- and second-stage selection probabilities could be entered and the weights calculated using appropriate SPSS or Epi-Info commands.

3. When the value of a variable or indicator is multiplied by the sampling weight, the result is the *appearance* that the sample size was larger than it actually was.

Figure 5-2: Illustrative computation of selection probabilities, sampling weights, and standardized sampling weights — hypothetical data

In this example, calculations of standardized weights are shown for the first five of a sample of clusters chosen in a hypothetical survey.

KEY:

n_i = the number of sample elements chosen in cluster i (which, since a constant number of elements was chosen per cluster, was always 50)
P_i = overall probability of selection for sample elements in cluster i
w_i = sampling weight for sample elements in cluster i
w_i' = standardized sampling weight for sample elements in cluster i

Cluster No.	n_i	P_i	w_i	$w_i n_i$	w_i'
1	50	.033	30.30	1515.00	.0167
2	50	.022	45.45	2272.50	.0251
3	50	.030	33.33	1666.50	.0184
4	50	.043	23.26	1163.00	.0125
5	50	.023	43.48	2174.00	.0240
.					
.					
.					
Total	2,500			90,526.28	

3. Estimating standard errors

Testing the statistical significance of observed changes or trends requires estimates of the magnitude of sampling error associated with the survey estimates, commonly referred to as *standard errors*. The estimation of sampling error can be a rather complex undertaking, depending upon the sample design used in collecting the data. As sample designs become more complex (e.g., when stratification, cluster sampling, and multiple stages of sample selection are used), the procedures for estimating standard errors become quite complicated. The estimation of standard errors for such designs is beyond what many Title II program evaluations may reasonably be expected to contend with without assistance from a statistician.

Unfortunately, standard statistical software packages such as SPSS and Epi-Info not provide an adequate solution to this problem. Although both packages will estimate the standard errors of observed changes and/or comparison group differences on indicators and perform appropriate

statistical tests, the standard errors produced by these software packages assume that simple random sampling was used in gathering the survey data. Since it is highly probable that cluster sampling will be employed in Title II program evaluation surveys, the estimated standard errors produced by these packages will usually be underestimated. The result is that some changes or trends are judged to be *real* changes or trends when in fact the changes are too small to be statistically significant.

Short of bringing more sophisticated computer software to bear on the problem, one option would be to compensate for the expected underestimation of standard errors by tightening the criteria used for judging statistical significance. For example, instead of using $p<.05$ as the cutoff point for judging an observed change or difference in an indicator to be significant, $p<.04$ or even $p<.03$ might be used. In this way, the danger of incorrectly judging an observed change to be significant is reduced without adding to the complexity of the statistical analyses of the survey data or having to use software on which local staff have no prior training or experience. Ultimately, however, where careful analyses of program evaluation data are needed, the use of software that are capable of taking complex sample designs into account (e.g., STATA or SUDAAN) is required.

Appendix 1
List of Generic Title II Indicators

Category	Level	Indicator
Health, nutrition, and MCH	Impact	% stunted children 24-59 months (height/age Z-score)
		% underweight children by age group (weight/age Z-score)
		% infants breastfed w/in 8 hours of birth
		% infants under 6 months breastfed only
		% infants 6-10 months fed complementary foods
		% infants continuously fed during diarrhea
		% infants fed extra food for 2 weeks after diarrhea
	Annual monitoring	% eligible children in growth monitoring/promotion
		% children immunized for measles at 12 months
		% of communities with community health organization
		% children in growth promotion program gaining weight in past 3 months
Water and sanitation	Impact	% infants with diarrhea in last two weeks
		liters of household water use per person
		% population with proper hand washing behavior
		% households with access to adequate sanitation (also annual monitoring)
	Annual monitoring	% households with year-round access to safe water
		% water/sanitation facilities maintained by community
Household food consumption	Impact	% households consuming minimum daily food requirements
		number of meals/snacks eaten per day
		number of different food/food groups eaten
Agricultural productivity	Impact	annual yield of targeted crops
		yield gaps (actual vs. potential)
		yield variability under varying conditions
		value of agricultural production per vulnerable household
		months of household grain provisions
		% of crops lost to pests or environment
	Annual monitoring	annual yield of targeted crops
		number of hectares in which improved practices adopted
		number of storage facilities built and used
Natural resource management	Impact	imputed soil erosion
		imputed soil fertility
		yields or yield variability (also annual monitoring)
	Annual monitoring	number of hectares in which NRM practices used
		seedling/sapling survival rate
FFW/CFW roads	Impact	agriculture input price margins between areas
		availability of key agriculture inputs
		staple food transport costs by seasons
		volume of agriculture produce transported by households to markets
		volume of vehicle traffic by vehicle type
	Annual monitoring	kilometers of farm to market roads rehabilitated
		selected annual measurements of the impact indicators