# MEL Tip Sheet: Sampling

*This tip sheet is for program managers and M&E staff developing sampling strategies for surveys and other M&E activities. Further pages detail key issues and methods, including examples.*

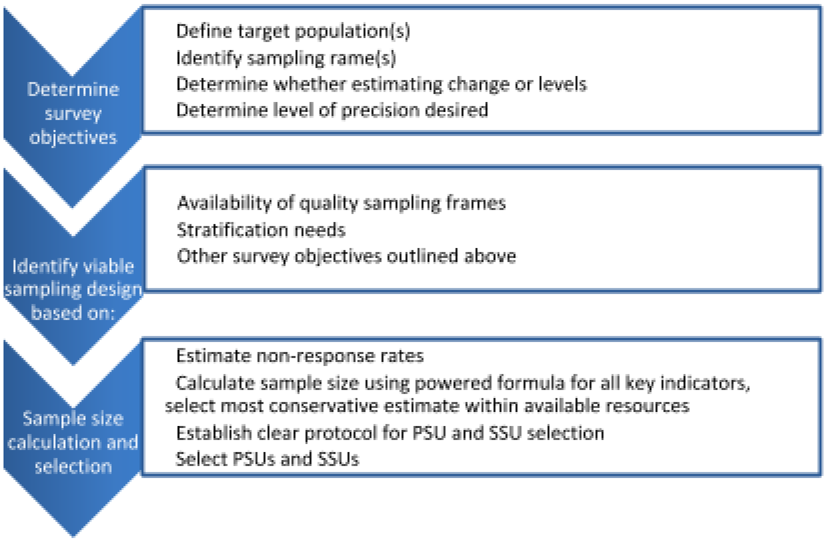
One essential component of monitoring and evaluation is the systematic collection, analysis and dissemination of data for various stakeholders. One key source of data is quantitative surveys that can either be census-based (collecting information from all households in a specified area) or sample-based surveys (collecting information from selected households in a specified area).[1] Censuses are rarely used by M&E practitioners because of their high costs. Thus, well-designed sample-based surveys are essential tools for M&E. Because information collected from sample-based surveys is limited to households and/or individuals selected for interviews, drawing conclusions about the entire population from which the households or individuals were selected from is subject to error. A survey designer's primary responsibility is ensuring adequate sample size to provide cost effective, representative information for the population being studied.

> *Sampling provides a means of gaining information about the population without the need to examine the population in its entirety."*

The information presented in this tip sheet is not exhaustive and users are encouraged to explore the references listed in the additional resources section. In an effort to ensure accessibility of additional information, this list includes only reference material available free of charge online. Users are encouraged to seek additional help and resources from the HQ-based MEL Team.

This document presents the key steps to take in determining a sampling plan (see Figure 1). It begins with the questions to ask when determining survey objectives, and follows with general descriptions of common sampling methodologies. Steps in calculating sample size are discussed, and case studies presented. Throughout the document, technical terms are defined, and users are encouraged to refer to the glossary for clarification.

Figure 1 - Steps in Sampling Design and Sample Selection


Determine survey objectives
Define target population(s)
Identify sampling rame(s)
Determine whether estimating change or levels
Determine level of precision desired
Identify viable sampling design based on:
Availability of quality sampling frames
Stratification needs
Other survey objectives outlined above
Sample size calculation and selection
Estimate non-response rates
Calculate sample size using powered formula for all key indicators, select most conservative estimate within available resources
Establish clear protocol for PSU and SSU selection
Select PSUs and SSUs

[1] Throughout this document the term "survey" is used interchangeably with "quantitative survey" and refers exclusively to quantitative surveys in this context. In addition, it is assumed that the surveyed elements are either households or individuals or both.

## Table of Contents

## 1. Determining survey objectives

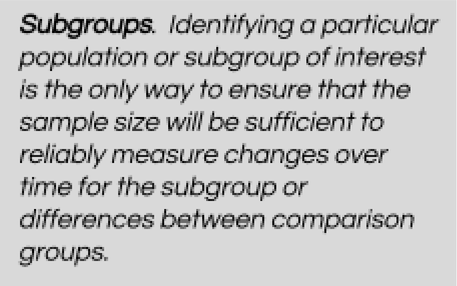
When conducting a sample-based survey, several key considerations must be taken into account. The results of these considerations can be thought of as defining survey objectives: What are we trying to measure? From whom? How will it be measured? What resources are available for the survey? It is *essential* that all sampling decisions be well documented in a research protocol. This documentation may evolve in implementing the fieldwork, but a clear and detailed record should be provided, including the rationale and justification of modifications. An outline is provided in Annex I.

### 1.1 Target populations

First, the target population(s) should be defined. Often surveys include several indicators covering multiple topics, for example agriculture, maternal and child health, food security, or water and sanitation. It is common to have multiple target populations within a single survey.

Several questions should be asked to help determine the target populations:

- Does the information obtained need to represent the entire population in the project area? Or only direct project beneficiaries? This may be subject to donor requirements. Beneficiary-based samples may be difficult to implement because beneficiary enrollment often happens throughout the project and thus the baseline sampling frame (defined below) could differ dramatically from the endline sampling frame, prohibiting accurate comparison between the two surveys.
- What types of individuals/institutions is the project targeting? There are several subgroups that the project may have targeted that should be considered if representative samples are desired for each subgroup. Examples of subgroups include:
    - Women of reproductive age
    - Children under 5 years of age; children under 2 years of age
    - Small holder farmers
    - Cooperative members


***Subgroups**. Identifying a particular population or subgroup of interest is the only way to ensure that the sample size will be sufficient to reliably measure changes over time for the subgroup or differences between comparison groups.*

In general, project documentation (results frameworks, logframes, and indicator plans) should provide most of this information.

### 1.2 Sampling frames

After identifying the target populations, sampling frames should be defined. Sampling frames are the source materials from which the sample is selected. Examples of sampling frames include household lists, women and child registries from health posts, village lists, and beneficiary lists. There are numerous potential sources for sampling frames, including but not limited to national statistics organizations, other NGOs, and previous surveys.

Ideally, the sampling frame should be *complete, accurate,* and *current*. Although in practice this is rarely possible, every effort should be made to have the best sampling frame practicable, with a clear understanding of any weaknesses (e.g., it is not current, or it is likely to exclude certain households). Existing sampling frames should be revised where needed to improve their completeness and accuracy (UN, 2005a). In some instances, complete sampling frames may not exist. For example, there may not be a list of all households with children under five in a given community. Alternatively, there may be a list of children who have received immunizations at the local health post, but this list would exclude children who have not received immunizations. If it is not possible to repair this sampling frame, the sample would be biased towards children who have received immunizations. This is an extreme example that should result in the creation of a new frame, but serves to illustrate an important point of identifying and understanding any biases in the sample frame.

Surveys often use more than one sampling frame. This is due to the fact that a comprehensive list of all households for all villages in the entire project area is often unavailable. In the rare event that a complete, accurate, and current list of households in the project area exists, a simple random sample may be drawn. This is only likely soon after a census has been conducted by the national statistics agency. More often there may be a list of villages and/or enumeration areas (EAs) available from the national statistics agency. This is suitable for the first stage of selection, while comprehensive household lists at the village level are suitable for subsequent stages.

### 1.3 Estimating change/difference versus levels

Initial surveys (assessments or baselines) estimate existing *levels*, or prevalence of a given indicator (such as literacy rates or acute malnutrition), while endline surveys or surveys comparing groups (panel survey waves, mid or end line surveys, control/treatment comparisons) estimate the change or difference from those initial levels over time. The sample sizes required to measure differences, particularly small differences, are large. This is an important consideration for M&E survey design. The expected changes for certain indicators may be small, either because the differences between comparison groups will be small or the fact that having large changes within the project lifetime may be impossible.

For example, a project may hope to reduce chronic malnutrition (stunting) in children aged 0-5 years from 40% to 35% in a five-year project. The sample size required to detect a change of five percentage points may be large and thus costly. Increasing the rate of reduction from five to ten percentage points (having a target of 30% instead of 35%) would reduce the sample size required to detect this change, but it may be unrealistic to expect the project to have such a large impact on stunting rates. This highlights the importance of early involvement of key M&E staff in project design, specifically with regards to project targets. Indicators with relatively small changes over the life of the project should be identified and discussed early. Some indicators may be critical to demonstrate project impact and thus resources should be committed to support sufficient sample sizes to measure these changes.

### 1.4 Precision and accuracy

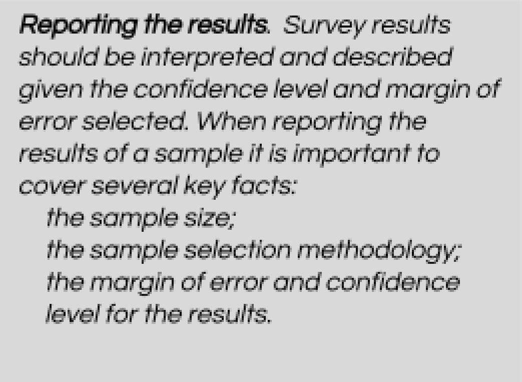
Precision and confidence levels influence sample size, and are important considerations in a sampling strategy. A key challenge of survey design is balancing precision and confidence levels while remaining within budgets.

Precision is also referred to as reliability, margin of error or confidence interval, and is related to the how well one can reproduce similar results (within the margin of error) if multiple measurements of an indicator were taken. This is not the same as accuracy, which measures how close an estimate for an indicator is to the "true" value. A common way of illustrating precision and accuracy is a target (see figure on right). The dots in cell A are dispersed (imprecise) and far from the bulls-eye (inaccurate). The dots in cell B are slightly dispersed, but clustered around the bulls-eye (accurate). Cell C is more precise as the dots are clustered closely together, but inaccurate as they are not in the bulls-eye. Cell D shows a highly precise and accurate collection of dots.

Precision is taken into account in calculating sample size (see section 5.2) when the margin of error is specified. Accuracy, however, is more challenging to account for because the "true" value of an indicator is always unknown – if it was known, a survey would not be necessary! Accuracy is ensured by reducing sampling and non-sampling errors (UN, 2005a).

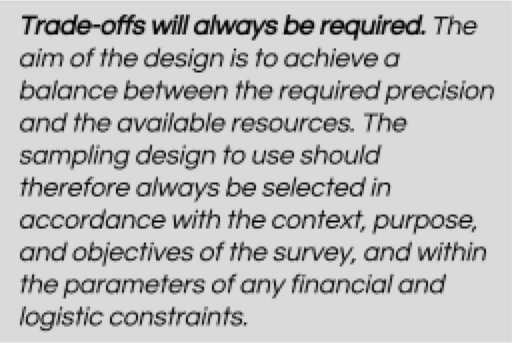
## 1.5 Statistical confidence and power

Sample size also increases with the level of statistical confidence (or "confidence level") for the margin of error around survey estimates. The confidence level indicates how confident researchers can be that the "true" population mean is within the margin of error. A confidence level of 95 percent means that if a population was sampled 100 times, in 95 samples the estimate of an indicator would be within the margin of error. For example, with a 95 percent confidence level, a survey might show that 32 percent of children are stunted with a margin of error of three percent. This means that there is only a five percent chance that the "true" population value for stunting is not between 29 and 35 percent. Lower confidence levels decrease sample size and can be used with narrower confidence intervals, but at the cost of reducing the confidence that the survey estimates contain the "true" population value.



***Reporting the results.*** *Survey results should be interpreted and described given the confidence level and margin of error selected. When reporting the results of a sample it is important to cover several key facts:*

*the sample size;*
*the sample selection methodology;*
*the margin of error and confidence level for the results.*



***Trade-offs will always be required.*** *The aim of the design is to achieve a balance between the required precision and the available resources. The sampling design to use should therefore always be selected in accordance with the context, purpose, and objectives of the survey, and within the parameters of any financial and logistic constraints.*

When measuring changes one must also take into account statistical power when calculating the sample size. Statistical power is referred to as guarding against "false negatives" (type II errors), such as concluding that a project has had no impact on a given indicator when, in fact, it has. Using greater statistical power to calculate sample size reduces the possibility of falsely concluding the project had no impact (FANTA, 1997). How this is used in sample size calculation is shown in section 4.2.2.

### 1.6 Non-response

Non-response occurs when respondents refuse to answer some or all of the survey questions, or when intended respondents are unavailable. If this is not taken into account, the actual number of completed interviews may be less than the needed sample size. Surveys can compensate by substituting households, or by increasing the number of households visited. In practice, the latter approach is less prone to non-sampling errors (which are more difficult to quantify than sampling errors) during implementation and is thus the recommended approach. A good general rule is to assume at least 10 percent non-response, but non-response rates vary significantly by country and should be determined on a case-by-case basis from previous experiences of other NGOs or national statistical agencies.

## 2. Common sampling designs

This section outlines typical sampling methods used in household surveys in developing countries. All of the techniques discussed here are probability-based and thus are based on statistical theory and capable of generating representative data about survey populations.[2]

### 2.1 Simple random sample (SRS)

SRS is the most basic form of a probability-based sample in which all *elements* (members of a population, such as individuals or households) have an equal probability of being selected. This requires a complete list of all elements in the population (the sampling frame). For example, in a national household survey (a survey in which the findings can be generalized to the entire nation), in order to draw a simple random sample, a comprehensive list of *all* households (*N*) would be needed to draw the sample (*n*).[3] The probability of selecting a single household "*a*" would be:

$$Prob_a = \frac{1}{N} \quad (1)$$

This example illustrates two key points: the basic computation of the probability of selection; and the rarity of SRS-based surveys. In most developing countries, having an accurate sampling frame of all households even at the sub-national level is unlikely. For geographically dispersed populations, SRS samples will result in high travel costs for data collection. While SRS-based surveys are not common, simple random sampling also has the lowest sample size, so whenever the opportunity presents itself, SRS will be the least cost option (assuming this is not offset by increased travel costs). There are many methods of drawing simple random samples, one of which is described in Annex II.

### 2.2 Two-stage cluster sampling

Two-stage cluster sampling is the most common form of sampling in developing countries. It does not require a comprehensive sampling frame of all households, and it groups sampled elements into small

[2] Non-probability based sampling techniques are not discussed here. Users interested in non-probability sampling are encouraged to explore section 3.2.2 of the UN Statistics Division sampling guide for a primer on these techniques.

[3] Throughout this guide the assumption of sampling without replacement is made as this is common practice for household surveys. Sampling without replacement means sampled elements can only be selected once (drawing numbers out of a hat without replacing them back into the hat after selection). The alternative (rarely used for household surveys) is sampling with replacement, where elements can be selected more than once (drawing numbers out of a hat and replacing them back into the hat after selection).

geographic areas (clusters), which results in a feasible distribution of households. There are two stages of sampling: selecting primary sampling units (PSUs), also known as the clusters; and selecting secondary sampling units (SSUs), which are generally households. Any selection process that happens after the SSUs have been selected constitutes a third stage (e.g., selecting one woman of reproductive age or one child under five from a household) and increases the design effect (see below).

***Will you cluster your sample?***
*Clustering a sample in two stages refers to first selecting clusters (such as villages or schools) and then selecting actual elements (households or school children) from within these clusters. Clustering a sample usually reduces the time required for field work and travel time, but requires an increased sample size to account for the error it introduces.*

The disadvantage of clustering is a decrease in reliability (i.e., increased error) of sample estimates because elements within a cluster are likely to have similar traits (e.g. people within a village are more similar to each other and are likely to be different from other villages within the sampling frame). This is the "clustering" effect and results in an increased sample size. The extent of the clustering effect is measured by the design effect (*deff*) which relates the extent to which two elements within a cluster are correlated (intra-class correlation or ICC). **Often the design effect is assumed to be 2.0** (FANTA, 1997).

### *2.2.1 Determining number of clusters and number of observations per cluster*

Determining the optimal number of clusters and number of observations per cluster is a function of statistical and logistical variables. From a statistical perspective, survey design should minimize the number of observations per cluster and increase the number of clusters (UN, 2005a)[4]. For example, if the total sample size is 100, it is preferable to interview 10 clusters of 10 households rather than 5 clusters of 20 households to allow for greater variation within the survey area. The general rule of thumb is that the number of elements per cluster should be between 10 and 25 (USAID, 2006).

From a logistical perspective, the optimal number of clusters and observations per cluster is related to the resources available for transportation, the number of enumerators, and the number of interviews enumerators can complete in one day (a function of questionnaire length and complexity). More clusters with fewer observations per cluster require more resources in terms of vehicles required and fuel consumed, as shown in the case study below.

[4] An in-depth discussion of intra-class correlation and design effects is beyond the scope of this document. Users wishing to further explore these issues should see Chapters VI and VII of UN (2005b), sections 3.3.5 and 3.5.4 of UN (2005a), and USAID 2006.

**Case Study 1: Determining optimal number of clusters and observations per cluster**

An M&E unit has calculated the sample size for a two-stage cluster survey to be 1,000 households. They are deciding on the number of clusters and observations per cluster. Based on prior experience and the questionnaire length, the M&E unit estimates that enumerators can complete five interviews per day, so the survey will take 200 enumerator days (1,000 households / 5 household interviews per day=200 days). They decide to hire 20 enumerators for 10 days. Five possible team compositions are outlined below.

| | # of teams | # of enumerators per team | # of interviews/cluster | # of clusters | n |
|---|---|---|---|---|---|
| A | 1 | 20 | 100 | 10 | 1,000 |
| B | 2 | 10 | 50 | 20 | 1,000 |
| C | 4 | 5 | 25 | 40 | 1,000 |
| D | 5 | 4 | 20 | 50 | 1,000 |
| E | 10 | 2 | 10 | 100 | 1,000 |

Reviewing each composition, the M&E unit eliminates options A and E. Option A involves the least travel as there are only 10 clusters and would thus be the cheapest option, but it would capture very little variation in the population. Option E would capture the most variation, but they do not have the resources to hire vehicles for all 10 teams. They review the remaining options B, C, and D, and settle on option C as it provides a nice balance of cost effectiveness in terms of the number of vehicles hired and clusters visited and capturing a sufficient amount of variation within the population.

### *2.2.2 PSU Selection Process*

The PSUs are generally selected using *probability-proportional-to-size* (PPS), a systematic sampling method in which larger clusters have greater probability of being selected. This is preferred to SRS or basic systematic selection of PSUs because PPS is a self-weighting design. See Annex IV for an example.

### *2.2.3 SSU Selection Process*

After selecting the PSUs, SSUs are selected, ideally from households lists constructed prior to fieldwork and using either SRS or systematic random sampling. The household lists constitute a sampling frame for the SSUs and thus should be complete, accurate, and current. This often requires significant effort in constructing new household lists for each survey, but is most likely to yield an unbiased sample at the cluster level. From these lists, random selection of sample households can be generated (see section 2.3 below). Alternative methods of SSU selection include the EPI-method (sometimes called *random walk*) and segmentation. These methods are described in detail in section 4.3.2 of FANTA (1997).

## 2.3 Systematic random sampling

Systematic random sampling is a selection method in which sampled households are selected using a sampling interval from an ordered list. Systematic random sampling orders the household population list and then selects households at regular intervals from that ordered list. Systematic random sampling involves a random start and then proceeds with the selection of every *k*th element. The sampling interval is calculated as follows:

$$k = \frac{N}{n}$$

(2)

Where:

*k*=sampling interval<br>
*N*=Total number of elements<br>
*n*=Total sample size

After determining the sampling interval, choose a random starting element from 1 to *k*. For example, if the sampling interval (*k*) is 47, the first element is a random number between 1 and 47. Each subsequent element is selected by adding *k* to the previous element. See Annex III for a detailed example.

## 2.4 Stratification

Stratification organizes a diverse population into independent, mutually exclusive (no overlap) groups (strata) which are internally similar. The advantage of stratification is increased precision in the overall sample by creating strata with small internal variation (because they are similar). Common examples of stratification are administrative divisions (districts), urban/rural, arid/rainy, and coastal/inland.

### *2.4.1 Allocation of observations to strata*

Determining the number of observations to allocate to each stratum depends on whether or not representative data for each stratum (stratum level estimates) are required. If stratum level estimates are required, sample sizes should be calculated for each stratum, particularly if the populations of the strata are small. Any sample estimates calculated for the entire survey area (all strata combined) will need to be weighted accordingly.[5]

If stratum level estimates are not required, the sample may still be stratified to ensure that important subgroups are represented. Using the example of a calculated sample size of 1,000 households with two strata (urban and rural), if urban households are 30 percent of households while rural households are 70 percent, proportional allocation would mean a sample of 300 urban households and 700 rural households. One could still calculate strata level estimates for urban and rural areas in this example, but they would be less precise and/or have a lower confidence level because the sample size is smaller.

[5] Please note that stratum level estimates are not the same as making statistical comparisons between strata (e.g. comparing groups). If statistical comparisons between groups are necessary, please refer to the sections regarding estimating change/differences

**Case Study 2: Stratification**

M&E staff are conducting a baseline survey for an education project that covers peri-urban and rural areas. Stakeholders suspect that there are large differences in education between peri-urban and rural areas as well as between male and female headed households. The M&E staff decides to employ a two-stage cluster survey, using the list of 200 schools with 200 children per school (population = 40,000) in the project area with corresponding enrollment numbers for the first stage of sampling and up-to-date student rosters for the second stage of sampling. They desire a confidence level of 95% and a margin of error of 10%.

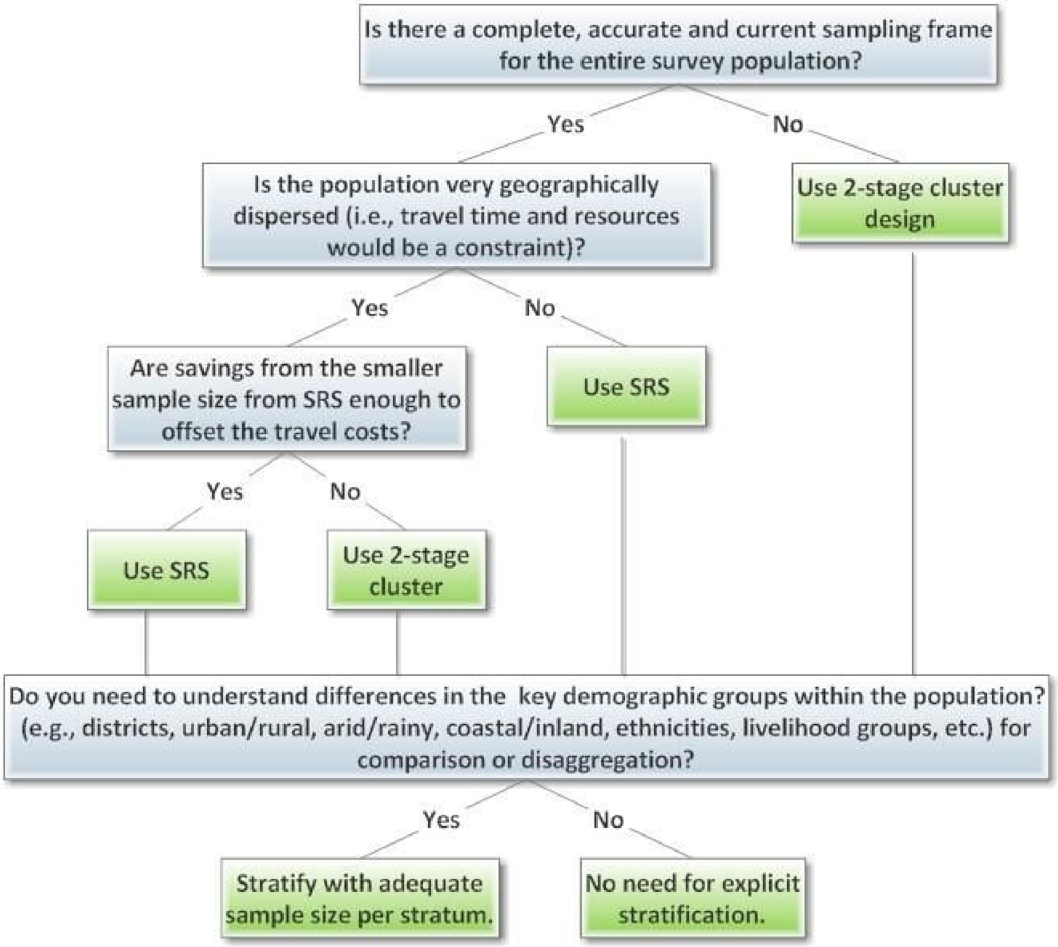
Twenty percent of the schools are in peri-urban areas, with the remainder located in rural areas. Secondary data shows that one-third of the households are headed by females. The table below shows the distribution of the sample for each stratum. Without stratum level estimates, the total sample size is 188, which can be distributed proportionally or equally. It is generally advised to distribute proportionally if no stratum level estimates are required. If stratum level estimates are required, the *total* sample size increases to 355 because stratum level estimates are required and a representative sample must be calculated for each stratum.

| Stratum | | *% in population* | *Without stratum level estimates* | | *With stratum level estimates* |
|---|---|---|---|---|---|
| | | | *Proportional allocation* | *Equal allocation* | |
| Location | Peri-urban | 20% | 38 | 94 | 168 |
| | Rural | 80% | 150 | 94 | 187 |
| | **TOTAL** | **100%** | **188** | **188** | **355** |
| Sex of HHH | Male | 66% | 125 | 94 | 185 |
| | Female | 33% | 63 | 94 | 179 |
| | **TOTAL** | **100%** | **188** | **188** | **364** |

## 3. Determining the appropriate sampling design

There are two basic designs for household surveys in developing countries:  two-stage cluster survey and simple random sample. Both of these designs may be stratified to allow for comparison between groups in the population. The decision tree below will help decide which design to use.

Figure 3: Decision tree for sample design determination


Is there a complete, accurate and current sampling frame for the entire survey population?
Yes
No
Is the population very geographically dispersed (i.e., travel time and resources would be a constraint)?
Use 2-stage cluster design
Yes
No
Are savings from the smaller sample size from SRS enough to offset the travel costs?
Use SRS
Yes
No
Use SRS
Use 2-stage cluster
Do you need to understand differences in the key demographic groups within the population? (e.g., districts, urban/rural, arid/rainy, coastal/inland, ethnicities, livelihood groups, etc.) for comparison or disaggregation?
Yes
No
Stratify with adequate sample size per stratum.
No need for explicit stratification.

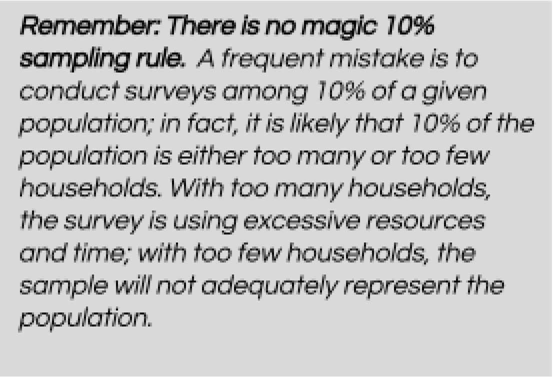
## 4. Sample size calculations

Now that general sampling concepts have been described, the calculation of sample size may be discussed. Ultimately the survey will have one total sample size, but to determine this, several decisions must be made.

### 4.1 Key questions to ask before calculating sample size

*1. Are you using two-stage cluster sampling or SRS?*

Using the decision tree above, the appropriate methodology (clustering or SRS) can be determined. Cluster samples usually require larger samples because of the design effect (see section 3.4 above).
In the equations below, the design effect is represented by the variable "*deff*".



***Remember: There is no magic 10% sampling rule.*** *A frequent mistake is to conduct surveys among 10% of a given population; in fact, it is likely that 10% of the population is either too many or too few households. With too many households, the survey is using excessive resources and time; with too few households, the sample will not adequately represent the population.*

*2. Are you estimating change or levels?*

M&E surveys generally try to answer one of two questions:

1) What is the current *level* of an indicator in the project area?
2) How have the levels of this indicator *changed* over the lifetime of the project? For example, if an indicator was 49% at the baseline, how much has it changed at the endline?

The answers to these questions help to determine the appropriate formula to use to calculate sample size, so it is important to understand whether levels or changes are being estimated. Levels are typically estimated for baselines or assessments; such surveys are conducted prior to any project activities when one-time estimates of the levels of key indicators are needed. Changes are typically estimated with mid or endline surveys, when M&E units want to measure how much indicators have changed over time to determine whether or not the project had the desired impact. Estimating changes, particularly small ones, generally require larger samples.

The formula for estimating change may be used for a baseline if the amount of change to be detected is already known, and project targets have already been defined for all indicators. Some donors may require this formula to be used for the baseline. This equation will usually yield a larger sample size.

*3. Are you stratifying?*

Are there important groups for which representative samples are required? If yes, sample sizes must be calculated using one of the formulas below for *each* stratum to ensure adequate sample size for all. The cumulative sample size for all strata can become quite large, so careful consideration of the value of stratification must be considered in light of the costs of the larger sample size.

*4. What is the current estimated level for your indicator(s) of interest?*

For any sample size calculation, estimated levels must be assumed for the indicators being measured. Estimates can be obtained from secondary sources such as Demographic and Health Surveys (DHS), census data, Living Standards Measurement Surveys (LSMS) or other recent surveys. If no recent or high quality data exist for the country, data from neighboring countries may be used *or* an estimated level of **50% may be used as this gives the highest sample size**. Extremely high (90%) or low (10%) estimated indicator levels yield smaller sample sizes than levels near 50%.

*5. What confidence level are you willing to assume*

The confidence level refers to the probability that the data from the sample estimate is close to the "true" population value. The more confident one wants to be, the larger the sample size. **Often, 95% confidence levels are used though under budget constraints 90% confidence also may be used.** This variable is "*z*" in the equations below.

*6. What degree of precision (margin of error) is sufficient?*

Precision refers to how accurate (close to the true value) the estimate from the sample is. More precise estimates (smaller margins of error) require larger sample sizes. One starting point for determining the margin of error is **multiplying the estimated level (see above) by 0.10**. So if the estimated level is 50%, the margin of error would be 50% x 0.10 = 5%. If this yields a sample size that is too large, one may increase the margin of error with the understanding that this will result in a less precise estimate. This variable is represented by “ $\varepsilon$ ” in the equation below.

*7. What is the non-response rate?*

The non-response rate refers to the percentage of respondents who are unwilling or unable to successfully complete an interview. As the non-response rate increases, the sample size will also increase. The non-response rate is determined from prior survey experience. If this is unavailable, **assume a non-response rate of 10%**. This variable is represented by “$r$” in the equations below.

*8. If you are estimating change/differences – how much change should you be able to detect?*

To detect small amounts of change a large sample size is required. For indicators where a project is likely to lead to small changes, careful consideration should be given to using these indicators as the basis for sampling because they will need large samples. If it is essential to demonstrate impact, additional resources should be committed from the outset of the project to ensure that large samples can be used.

*9. Is the indicator for a small subgroup of the population (women of reproductive age, children under five)?*

Some indicators are for subgroups of the population, such as children, women, or farmers. To ensure adequate sample sizes for these subgroups, it is common to adjust the total number of households sampled to ensure that representative samples are collected for each group. For example, if the calculated sample size is 500 farmers, and a good estimate (based on previous surveys or field experience) is that 75% of households are engaged in agriculture, to obtain at least 500 farmers, it will be necessary to visit at least 667 households (667 x 0.75 ≈ 500) (see section 4.2.3 below).

*10. Is the population from which the sample is being drawn relatively large or relatively small?*

When the population is small, the calculated sample size can be adjusted to account for the fact that the standard error of the estimate will be reduced because it is drawn from a small number of elements. If the population is large, no adjustment is necessary (UN, 2005a).

## 4.2 Sample size formulas

After answering *all* of the questions above, sample sizes for various indicators may be calculated using the appropriate formulas. The Excel-based sample size calculator has a “plug-and-play” method for calculations. If you need help choosing a formula or calculating your sample size, contact the MEL Team.

### *4.2.1 Formula for estimating levels*

The formula for a simple random sample to estimate the levels of an indicator is as follows (UN, 2005a):[6]

$$n = \frac{z^2(p*(1-p))}{\varepsilon^2(1-r)} \qquad (3)$$

Where:

*n* = sample size, *in number of elements to be sampled*
*z* = z-score of confidence level (either 1.96 or 1.645 corresponding to 95% and 90% confidence respectively)
*p* = proportion of the population exhibiting the characteristic of interest (estimated from secondary data)
*1-p* = proportion of the population not exhibiting the characteristic of interest
*r* = non-response rate (generally set to 10% but may change with context)
$\varepsilon$ = margin of error (to be determined by survey design team, general rule of thumb is 0.10x*p*= $\varepsilon$ )

This formula gives the sample size in terms of number of elements, e.g. if the indicator of interest is stunting among children under the age of five, *n* is equal to the number of children required for measurement, *not* the number of households required. The margin of error is in the denominator *and* is a squared term, so smaller margins of error have much larger sample sizes.

SRS surveys are rare. Two stage cluster surveys are common, and require adding the design effect (*deff*) to the above equation (UN, Designing Household Survey Samples: Practical Guidelines, 2005a).

$$n = deff\frac{z^2(p*(1-p))}{\varepsilon^2(1-r)} \qquad (4)$$

Where:

*n* = sample size
*deff* = design effect, usually assumed to be 2.0
*z* = z-score of confidence level (either 1.96 or 1.645 corresponding to 95% and 90% confidence respectively)
*p* = proportion of the population exhibiting the characteristic of interest (determined from secondary data)
*1-p* = proportion of the population not exhibiting the characteristic of interest
*r* = non-response rate (generally set to 10% but may change with context)
$\varepsilon$ = margin of error (to be determined by survey design team, general rule of thumb is 0.10x*p*= $\varepsilon$ )

[6] This is a version of the formula used for the raosoft.com online calculator. The raosoft.com calculator does *not* take into account design effect, non-response, or adjustment factors for subgroups, and sample sizes obtained from this calculator must be adjusted to account for these factors.

If the sample requires representative data at the *stratum* level, this formula should be calculated for *each* stratum. If data does not need to be disaggregated at the stratum level, only one sample size calculation is needed.

Whenever possible, M&E staff should investigate any known design effect value for an indicator of interest, as sometimes this effect can be high.  For example, the design effects for variables in the 2013 Nigeria Demographic and Health Survey (http://www.dhsprogram.com/publications/publication-fr293-dhs-final-reports.cfm) are calculated to be as high as 5.  If nothing is found for the variables of interest, 2 can be used for the design effect.

**Case Study 3: Conducting a multiple indicator baseline survey**

M&E staff want to design a sampling strategy for a baseline survey that provides representative data on key project indicators. The project focuses on nutrition and access to potable water in two districts that are located in different regions, and according to secondary data quite distinct from each other.

The three key indicators the project desires to measure levels for are:

1) Percentage of households with access to an improved water source
2) Percentage of women aged 15-49 that are underweight (BMI<18.5)
3) Percentage of children aged 0-59 months that are stunted (HAZ<-2SD)

After reviewing the indicator plan and project documents, the M&E team looks for sampling frames for the two districts. The national statistical agency conducted a census four years ago and a demographic and health (DHS) survey two years ago. There are population estimates for all EAs in the two districts, but no lists for women or children. The team decides a simple random sample is not feasible.

They are confident that a two-stage cluster sample is appropriate and have secured access to sampling frames used for the DHS. Next they begin calculating the sample size using equation 4 and the most recent secondary data from the DHS. They decide to use two margins of error (10% and 15%) to compare what the relative sample sizes would be. The table below shows their computations

| Indicator | *deff* | *z* | *p* | *1-p* | *r* | *margin of error* | *ε* | *s** | *κ** | *n* |
|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of households with access to potable water | 2.0 | 1.96 | 0.43 | 0.57 | 1.1 | 10% | 0.043 | 0.17 | 5.8 | **1,148** |
| | 2.0 | 1.96 | 0.43 | 0.57 | 1.1 | 15% | 0.065 | 0.17 | 5.8 | **511** |
| Percentage of women aged 15-49 that are underweight (BMI<18.5) | 2.0 | 1.96 | 0.18 | 0.82 | 1.1 | 10% | 0.018 | 0.20 | 5.8 | **3,353** |
| | 2.0 | 1.96 | 0.18 | 0.82 | 1.1 | 15% | 0.027 | 0.20 | 5.8 | **1,491** |
| Percentage of children aged 0-59 months that are stunted (HAZ<-2SD) | 2.0 | 1.96 | 0.43 | 0.57 | 1.1 | 10% | 0.043 | 0.13 | 5.8 | **1,501** |
| | 2.0 | 1.96 | 0.43 | 0.57 | 1.1 | 15% | 0.065 | 0.13 | 5.8 | **668** |

The table indicates that the most conservative (largest) sample would use the percentage of women underweight with a 10% margin of error. The maximum sample size they can afford is 1,500 households. Thus they accept the increased margin of error (15%) for this indicator and maintain a 10% margin of error for the other two indicators.

Here the M&E team adjusted only the margin of error. They could have also reduced the design effect (if they had stratified), the confidence level (if they were willing to accept the increased possibility that the actual population estimate was not within the confidence interval of their sample estimate), or assumed a lower non-response rate (if previous surveys provided evidence that this was possible). If they wanted to have representative data for each districts, they would possibly need to increase the sample size or use district level estimates for *p* (if available) to calculate the sample size for each district.

*See section 4.2.3 for information on $s$ and $\kappa$.

### *4.2.2 Formula for estimating changes/differences*

M&E staff must often make statistical comparisons between groups, such as measuring change over time (baseline group versus endline group), or control versus treatment group. For surveys with more than one round of data collection, additional considerations must be taken into account. There are three options for households interviewed for an endline:

1. Use the same households as the baseline (also known as a panel study)
2. Use the same clusters as the baseline (which may result in some overlap between baseline and endline households)
3. Draw an entirely independent sample (new clusters and households)

The advantages and disadvantages of each option are listed below. Option 2 is a good compromise when several rounds of surveying are anticipated. Option 1 is viable if only baseline/endline surveys are to be conducted, though the disadvantages should be weighed carefully.

| Option | Advantages | Disadvantages |
|---|---|---|
| **1. Use the exact same households from the baseline** (least sampling error, highest non-sampling error) | - Smallest estimated variance between baseline and endline samples<br>- Ability to observe distributed change in population, not just averages | - Need to adjust for attrition (non-response) which can be much higher than typical non-response rates<br>- Respondents become conditioned to survey instruments, creating bias<br>- Project may focus resources on these households to demonstrate greater impact |
| **2. Use the same clusters at baseline** | - Medium estimated variance between baseline and endline samples | - *Some* respondents become conditioned to survey instruments, creating bias<br>- Project may focus resources on these clusters to demonstrate greater impact |
| **3. Draw an entirely independent sample** | - Minimized non-sampling errors | - Greatest estimated variance between baseline and endline samples<br>- Observing change based on sampled averages may be skewed and |

| | | |
|---|---|---|
| (highest sampling error, least non-sampling error) | | misrepresent distributed change across targeted population. |

The formula used to estimate change differs from the formula used to estimate levels (FANTA, 1997):

$$n = deff\frac{(z_\alpha+z_\beta)^2 * (p_1(1-p_1)+p_2(1-p_2))}{(p_2-p_1)^2 * (1-r)} \qquad (5)$$

Where:

*n* = required minimum sample size per **survey round** or **comparison group**
*deff* = design effect, usually assumed to be 2.0
$p_1$ = 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 at some future date such that ($p_2$ - $p_1$) is the magnitude of change it is desired to detect
$z_\alpha$ = 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 (statistical significance)
$z_\beta$ = 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 (statistical power)
*r* = non-response rate (often set to 10% but may change with context)

Similar to the margin of error in equation (4), note that the smaller the difference between baseline and endline, the larger the sample size required.

**Case Study 4: Conducting a multiple indicator endline survey**

Five years later, the same M&E staff from Case Study 1 want to design a sampling strategy for an endline survey to measure the effect of the project.

Using targets defined in the project indicator plan, they begin the sample size calculation using equation 5 to assess change from the indicator baseline values. They decide to use two different magnitudes of change (10% and 5%) to compare the sample sizes. The table below presents their computations.

| Indicator | *deff* | $z_\alpha$ | $z_\beta$ | $p_1$ | $p_2$ | *% point change* | *r* | *s** | *κ** | *n* |
|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of households with access to potable water | 2.0 | 1.96 | 1.96 | 0.47 | 0.37 | 0.10 | 1.1 | 0.20 | 5.8 | **1,420** |
| | 2.0 | 1.96 | 1.96 | 0.47 | 0.42 | 0.05 | 1.1 | 0.20 | 5.8 | **5,802** |
| Percentage of women | 2.0 | 1.96 | 1.96 | 0.15 | 0.05 | 0.10 | 1.1 | 0.13 | 5.8 | **793** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| underweight (BMI<18.5) | 2.0 | 1.96 | 1.96 | 0.15 | 0.10 | 0.05 | 1.1 | 0.13 | 5.8 | **3,941** |
| Percentage of children aged 0-59 months that are stunted (HAZ<-2SD) | 2.0 | 1.96 | 1.96 | 0.42 | 0.32 | 0.10 | 1.1 | 0.13 | 5.8 | **2,089** |
| | 2.0 | 1.96 | 1.96 | 0.42 | 0.37 | 0.05 | 1.1 | 0.13 | 5.8 | **8,636** |

The table shows that the most conservative (largest) sample size would use a five percent reduction in child stunting, but this is prohibitively expensive. Rather than increasing the percentage change their sample would be able to detect they decide to decrease the level of confidence in their survey estimates to 90%. The resulting figures are shown in the table below.

| Indicator | *deff* | $z_\alpha$ | $z_\beta$ | $p_1$ | $p_2$ | *% point change* | *r* | *s** | *κ** | *n* |
|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of households with access to potable water | 2.0 | 1.64 | 1.64 | 0.47 | 0.37 | 0.10 | 1.1 | 0.20 | 5.8 | **994** |
| | 2.0 | 1.64 | 1.64 | 0.47 | 0.42 | 0.05 | 1.1 | 0.20 | 5.8 | **4,062** |
| Percentage of women aged 15-49 that are underweight (BMI<18.5) | 2.0 | 1.64 | 1.64 | 0.15 | 0.05 | 0.10 | 1.1 | 0.13 | 5.8 | **555** |
| | 2.0 | 1.64 | 1.64 | 0.15 | 0.10 | 0.05 | 1.1 | 0.13 | 5.8 | **2,759** |
| Percentage of children aged 0-59 months that are stunted (HAZ<-2SD) | 2.0 | 1.64 | 1.64 | 0.42 | 0.32 | 0.10 | 1.1 | 0.13 | 5.8 | **1,463** |
| | 2.0 | 1.64 | 1.64 | 0.42 | 0.37 | 0.05 | 1.1 | 0.13 | 5.8 | **6,047** |

Similar to the baseline, they decide to use a total sample of 1,500 households, which is sufficient to detect a 10 percent difference between baseline and endline for all indicators, with 90% confidence.

*See section 4.2.3 for information on s and $\kappa$.

### *4.2.3 Formula for adjusting sample sizes for sub-populations*

The formulas above give a calculated sample size (n) that refers to the number of *elements* that must be sampled, which is *not* necessarily households. For example, it could be children under five, women of reproductive age, or farmers. In cases where a small subgroup of the population is the basis for the indicator, the calculated sample size must be adjusted to ensure an adequate number of households are visited. The formula for the adjustment factor is:

$$n * \frac{1}{s\kappa} \qquad (6)$$

Where:

*n* = calculated minimum sample size
*s* = proportion of the total population accounted for by the target population
$\kappa$ = average household size

Multiplying the calculated sample size by this adjustment factor adjusts the sample to ensure enough households are visited to obtain the needed number of elements. For example, if children under five are 15% of the population and the average household size is five, to obtain a sample of 100 children under five the survey team must visit 100x(1/0.15x5) households, or 134 households because the average household has less than one child under five. Sometimes the number of households to be interviewed is *less* than the calculated sample size, if average households have more than one individual eligible for interview. For example, if women of reproductive age (aged 15-49) are 25% of the population and the average household size is five, households will on average have 1.25 (0.25x5) women in this age group.

### *4.2.4 Formula for adjusting for finite populations*

If the total population from which the sample is drawn is relatively small, the sample size can be adjusted using the finite population correction (FPC) factor. A rule of thumb to determine whether or not an FPC is required is calculating the percentage of the population being sampled, known as the sampling fraction. If this percentage is greater than five percent, the FPC can be used to adjust the sample size. If it is less than five percent, no FPC adjustment is necessary since the FPC is close to one. (UN, 2005a). The FPC is defined as follows:

$$fpc = \sqrt{\frac{(N-n)}{(N-1)}} \qquad (7)$$

Using the FPC is illustrated in the case study below.

**Case Study 5: Using a finite population correction**

The M&E staff have calculated the sample size of 450 for a survey of beneficiaries. The population from which the sample is to be drawn is 8,000 beneficiaries. They calculate the sampling fraction:

$$\frac{n}{N} = \frac{450}{8{,}000} = 5.6\%$$

They determine that the population is small enough that they can apply the FPC. Here the FPC is:

$$fpc = \sqrt{\frac{(N-n)}{(N-1)}} = \sqrt{\frac{(8{,}000-450)}{(8{,}000-1)}} = 0.972$$

The final sample size is calculated by multiplying *n* by 0.972.

$$n * fpc = 450 * 0.972 = 438$$

As the population increases, the sampling fraction approaches (but never reaches) zero. Likewise, the FPC approaches (but never reaches) one. For large populations, the FPC is effectively one, and thus has no effect on the calculated sample size.

### *4.2.5 Indicators that are not proportions*

For indicators that are not proportions, such as means or averages, contact the MEL Team for specific guidance on the appropriate calculations. See also the Connect posting, "Additional tips on sample size calculation" at: https://connect.mercycorps.org/discussion/20971.

## Frequently asked questions for sampling (FAQs)

**Question 1: What do we do when we don't have a sampling frame?**
Sampling frames are an essential component for probability based samples (see section 2.2 above). Often sampling frames need to be updated and revised, and it is uncommon to find a "ready to use" sampling frame from a secondary source. Developing a sampling frame is a difficult but essential task - without it, drawing a truly random sample is difficult if not impossible. In the unlikely event that no sampling frames (even old or incomplete ones) exist, one will have to be developed from scratch. If this is not possible, a statistically representative sample is not feasible with the methods outlined here.

**Question 2: Imagine in some places, the program functions very well and for that reason, you have almost all the households satisfied with the program, however, in other places households do not see the program as very useful, how should the sample be designed then?**
This issue could arise for many reasons and is a challenge faced by many implementing agencies. One way to manage this is to stratify by satisfied and unsatisfied communities. To do this, a clear means of assigning communities to strata must be defined. For example, it could be communities that only received one year of services versus 3-4 years (they were reached last by project) or the project was implemented by a consortium of agencies, one of which lacked the ability to implement effectively. In the former example, time in project could be the stratifying criteria while in the latter it would be geographic area covered by each implementing agency. Stratifying in this example could identify the extent of disparity in services and potential causes.

**Question 3: How do I take stratification into account when trying to calculate sample size? That is, how can I make sure that I have representative data for all the strata? Would it be accurate to come up with a random sample group first and then stratify it into subgroups?**
One can allocate observations to strata either proportionally or equally. Deciding between the two depends on whether or not representative data is required for each stratum or not. If not, proportional allocation is fine. If so, it may be necessary to do equal allocation in order to ensure adequate sample size for each stratum. When calculating the sample size using the formulae provided in section 5.0, it is necessary to calculate the sample size *for each strata* for the key indicators of interest.

**Question 4: Do we have to draw a random sample each time, or do a random initial sample and then track those individuals throughout the project?**
This is essentially asking whether or not to conduct panel survey. This is in part answered in section 5.2.2 and discussed below in the synopsis of survey designs. The primary advantage of conducting a panel survey is a smaller sample size associated with minimized variation between samples. This type also allows you to examine the distribution of change (e.g., what percent of individuals or households actually improved) rather than applying an average change to the entire population. The disadvantages are attrition (losing respondents), which can be quite high, the logistics of tracking respondents over time, and biases introduced by the "time in sample" effect, which has been shown to introduce bias over time.

**Question 5: What confidence level and margin of error is considered acceptable? 95% confidence level and 5% margin of error is very good, but budgets usually are limited. Will a lower confidence level and larger margin of error be ok?**

This is a fundamental question that M&E units always face when designing a survey. Most surveys strive to use the 95% confidence level and a margin of error equal to about 10% of the estimated level of the lead indicator used for sampling (i.e., if the estimated level is 50%, the margin of error would be 5%=50%x10%). Under budget constraints these values may be relaxed to a 90% confidence level with a margin of error of 15-20% of the estimated level of the lead indicator. Prior to doing this, the tradeoffs of lack of precision should be carefully evaluated by all stakeholders. In addition, budgeting well in advance of M&E activities can reduce the need to make these decisions.

**Question 6: Is it ok in general to use sample size calculators when it comes to individual surveys?**

One intention of this guidance is to give M&E staff an understanding of the formulas behind the online calculators so they can make informed decisions about when to use them and what if any adjustments should be made. It is important to note that the raosoft.com calculator does *not* take into account design effect, non-response, or adjustment factors for subgroups, and any sample sizes obtained from this calculator need to be adjusted to account for these factors.

**Question 7: What if my sample is larger than I need? Is there a risk in this case?**

The greatest risk in this case is wasted resources associated with additional travel, vehicles and fieldwork days. If a probability-based sampling method was followed and implementation was not prone to errors, statistically speaking there is no disadvantage to having a sample that is too large.

**Question 8: My project area covers a wide geographic scope suffering from conflict, insecurity, and lack of proper infrastructure like roads. Costs and risks associated with travel are extremely high. How should I design my sample?**

This question can be broken into two major issues: insecurity and high travel costs. In the case of extreme and common insecurity, one must carefully evaluate whether or not to conduct a quantitative survey at all. In these cases, other methods of data collection (e.g. qualitative) may prove a more cost-effective and safe solution. Implementing a successful quantitative survey in the middle of a war zone is extremely difficult and in some cases may not be worth the risk, particularly if compromises you need to make to the sampling method to ensure safety would jeopardize the statistical validity of findings anyway.

High travel costs should be taken into consideration before any M&E activities commence. They can be adjusted by 1) using two-stage cluster sampling to geographically organize the sample so less travel is required or 2) decreasing the number of clusters visited and increasing the number of observations per cluster to reduce the number of sites visited. If this option is chosen, please refer to section 3.4.1 for an in-depth discussion of the tradeoffs for this.

**Question 9: How do we tackle the issue of attribution? How can we determine what contribution the project has had on the community?**

This is another fundamental question for M&E staff. As discussed below in the synopsis of alternative sampling methods, the randomized control trial (RCT) is the gold standard for measuring the impact of an intervention since it controls for all other external factors.  However, RCTs are technically difficult to implement, very costly, and in many cases not feasible. Instead M&E practitioners are often limited to less rigorous quasi-experimental designs (see below), such as the often used “pre-post” comparison (i.e. baseline-endline comparison). While less rigorous, well designed quasi-experimental designs can provide compelling evidence of project impact, and while these results must always be carefully interpreted, it is often the best we can do.

**Question 10: What about indicators that are means or averages, and not proportions?**
This guidance sample size calculator on the DL also covers only the level/proportion. But I assume some indicators are means/averages, which would require a different formula. Is it that we always do sample size calculation with a primary outcome indicator based on percentage? I can see using a mean-based indicator (e.g., average HH income) for, say, a livelihood project.

## Synopsis of alternative sampling methods

| Method | Description | Application | Resources |
|---|---|---|---|
| 30x30 | Two-stage cluster survey composed of 30 clusters with 30 observations from each cluster for a total sample size of n=900 | Prescribed by WHO to measure global acute malnutrition, but has been widely used in other applications. Note the lack of definition for any parameter values (e.g. confidence level, margin of error, etc.), cluster number and number of observations per cluster | WHO (2000) The management of nutrition in major emergencies. Geneva: WHO |
| Lot Quality Assurance Sampling (LQAS) (variations: C-LQAS, MC-LQAS) | LQAS is essentially a stratified SRS that allows researchers to determine whether or not a certain level for an indicator is being achieved or not utilizing relatively small sample sizes (popular sample size is 19). Does *not* allow for point estimates (e.g. the prevalence of malaria is 38%).<br>Current research is being done to combine cluster sampling with LQAS (C-LQAS) to ease the rather strict and often impractical requirement of SRS.<br>In addition, multiple-category LQAS (MC-LQAS) is being explored to allow for three categorizations (e.g. low, medium, and high prevalence) rather than the traditional two (low or high prevalence). | Initially used in the American Industrial world of the 1920s to maintain quality control. Has been adopted extensively by the public health sector to assess immunization rates, HIV/AIDS, post-disaster assessment, women's health, etc.<br>Can be a very useful resource ***if*** SRS is feasible (see section above) and if M&E units and project managers are willing to accept relatively limited information (i.e. whether or not an indicator is above or below a threshold) rather than more informative (but also more costly) point estimates.<br>C-LQAS and MC-LQAS are exciting new variation on LQAS that are still being developed and refined. Extensive review of the literature should be conducted prior to considering these sampling options. | LQAS:<br>Lemeshow (1988) Sampling Techniques for Evaluation Health Parameters in Developing Countries: A working paper. Washington DC: National Academy Press.<br>Valadez, J.J., Weiss, W., Leburg, C., Davis, R. (2002) A Trainer's guide for Baseline Surveys and Regular Monitoring: Using LQAS for Assessing Field Programs in Community Health in Developing Countries.<br>C-LQAS:<br>Deitchler, M., Deconinck. H. & Bergeron, G. (2008) "Precision, time and cost: a comparison of three sampling designs in an emergency setting" Emerging Themes in Epidemiology 2 May 2008; 5:6<br>MC-LQAS:<br>Myatt, M. & Bennett, D. (2008) "A novel sequential sampling technique for the surveillance of transmitted HIV drug resistance by cross-sectional survey for use in low resource settings." Antiviral Therapy 13 Suppl 2 |
| Panel survey (aka longitudinal study) | Initial survey may utilize either SRS or multi-stage cluster sampling, but subsequent surveys collect data from the same households/individuals from the initial survey. | Used to measure trends in a panel of respondents over time, the primary advantage being small estimated variance between baseline and endline samples (see estimating change above) meaning a smaller sample size.<br>Keeping track of respondents and minimizing non-response (i.e. | UN (2005b) |

| | | | |
|---|---|---|---|
| | | attrition) is a major concern for panel studies. In addition, respondents can become "conditioned" to the survey and responses could become biased. | |
| Experimental designs (RCTs) | Experimental designs are also commonly known as randomized control trials (RCTs). These represent the gold standard in impact evaluation. In an RCT, two groups are randomly selected from the same population and one group receives a treatment (e.g. bednets, training, fuel efficient stoves, etc.) while the other group receives nothing and serves as a control, also known as the counterfactual. | RCTs have recently become very popular because they are able to clearly and (if done well) irrefutable evidence of the impact of an intervention. This is because the two samples are statistically *identical* to each other (remember they were both randomly selected from the same population) and the *only* difference between them is the treatment (intervention).<br>Some challenges of RCTs include the level of technical expertise and resources involved in designing and implementing an effective experiment. In addition, researchers need to consider ethical implications of denial of treatment to the control group, contamination of the control group (i.e. some control members receive the treatment), and external validity – i.e. how well the results can be replicated elsewhere. | http://www.povertyactionlab.org/methodology/what-randomization<br>Duflo, E., Glennerster, R., & Kremer, M. (2006) Using Randomization in Development Economics Research: A toolkit. JPAL: MIT |
| Quasi-experimental designs | Similar to RCTs, quasi-experimental designs seek to compare two groups, a control and a treatment group. Unlike RCTs, however, there is *no* random assignment between control and treatment group. There are several different types of survey designs that fall under this category, too numerous for adequate descriptions of all. | Quasi-experimental methods are generally used when RCTs are not feasible due to resource or capacity constraints, or logistical or ethical issues associated with randomization.<br>Done well, these can provide compelling evidence of impact of an intervention, though not with the same rigor of an RCT. | Glazerman, S., Levy, D., Myers, D. (2003) Non-experimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science September 2003589: 63-93<br>http://www.socialresearchmethods.net/kb/quasiexp.php |

## Glossary

| | |
|---|---|
| **Baseline** | Survey undertaken prior to any intervention. Typically used to estimate levels and establish a base to compare changes in key project indicators in over time. |
| **Census** | Survey of the entire population in a geographically defined region. Extremely resource intensive and rarely undertaken by NGOs. |
| **Cluster** | Geographically defined unit such as an EA or village. |
| **Confidence level** | Describes level of confidence with which precision or margin of error around the survey estimate is obtained, 95% generally being regarded as the standard. |
| **Design effect** | Describes the magnitude of the loss of efficiency utilizing cluster sampling rather than a simple random sample. |
| **Elements** | The fundamental unit of analysis in a survey. Elements could be households, women of reproductive age, children under five, etc. |
| **End-line** | Survey undertaken at the conclusion of an intervention. Typically used to estimate change in time compared to baseline values. |
| **Enumeration area (EA)** | Geographically well-defined cluster of elements often created for national surveys such as a census. |
| **Intraclass correlation (ICC)** | Degree to which two units in a cluster have the same value, compared to two units selected at random in the population. |
| **Level** | Estimate of the prevalence of a given indicator (such as literacy rates or acute malnutrition) collected or estimated at the beginning of a program. |
| **Non-sampling error** | Errors associated with survey implementation. Usually due to invalid or misapplied definitions (e.g. household, household head), unsatisfactory questionnaires, defective methods of data collection (e.g. non-calibrated scales), coding problems, inaccurate responses due to recall problems, etc. |
| **Statistical power** | Statistical power is referred to as guarding against "false negatives" also known as type II error. An example of a false negative is concluding the project has had no impact on a given indicator when in fact it has. Using greater statistical power in calculating the sample size reduces the possibility of falsely concluding the project had no impact when in fact it did. |
| **Panel Survey** | Type of longitudinal study where the same survey participants are studied over time, or surveying the same respondents as the baseline. |
| **Precision (reliability, margin of error, or confidence interval)** | Precision is also referred to as reliability, margin of error, or confidence interval and is related to the how well one can reproduce similar results (i.e. within the margin of error) if multiple measurements of a particular indicator were taken. |
| **Primary sampling unit (PSU)** | Geographically well-defined administrative unit selected at the first stage of sampling. |
| **Probability proportional to size (PPS)** | Systematic sampling method in which the probability of selection is directly related to the relative size of the element. |
| **Sample** | A carefully calculated selection of elements from a larger population based on probability. |
| **Sampling error** | Random error in survey estimates due to the fact that information is based on a sample, not the entire population. |
| **Sampling frame** | Source material from which a sample is drawn. It is typically a list of all those in a population who can be sampled, and may include individuals, households or institutions. |
| **Secondary sampling unit** | In two stage sampling this is the unit of analysis or element. |

| | |
|---|---|
| **Statistical power** | The degree to which one guards against having a “false negative” (also known as type II error) conclusion. |
| **Stratification** | Organizing the sampling frame into subgroups that are internally similar and externally distinct to ensure sample selection is spread across important subgroups. |
| **Target population** | Definition of population intended to be covered by the survey, e.g. project area. |

## Works Cited

FANTA. (2016). *Sampling Guide.* Washington, DC: FANTA. https://www.fantaproject.org/sites/default/files/resources/Sampling-Guide-Beneficiary-Based-Surveys-Feb2016.pdf

Airifin, Wan Nor. (2013). *Introduction to Sample Size Calculation.* Malaysia. Education in Medicine Journal (5) (2). https://www.researchgate.net/publication/272730017_Introduction_to_sample_size_calculation

Bullen, Piroska Bisits. *How to Choose a Sample Size (For the Statistically Challenged).* Tools4Dev.org http://www.tools4dev.org/resources/how-to-choose-a-sample-size/

Creative Research System. *Sample Size Calculator.* https://www.surveysystem.com/sscalc.htm

Conroy, Ronan. *Sample Size: A Rough Guide.* https://pdfs.semanticscholar.org/4781/878153e13322c028c7d8970e7f52fbaa102a.pdf

## Annex I: Research protocol outline

I. **Background and introduction**

   a. Provides background to project, context, etc.

II. **Objectives of the study**

   a. Outlines primary objectives of study (e.g. estimate change/levels)

   b. Describes key target population(s)

   c. Describes any stratification schemes

III. **Methodology**

   a. Describes general methodology and instruments

      i. Quantitative

      ii. Qualitative

IV. **Sample size calculation**

   a. Describes in depth indicators and target populations

   b. Presents sample size calculation formulae

   c. Presents table of sample sizes for key indicators

V. **Selection processes**

   a. Describes sampling frame(s)

   b. Clearly establishes protocol for PSU and SSU selection

VI. **Field work implementation**

   a. Team composition

   b. Estimated number of days of data collection

   c. Tentative field schedule

VII. **Budget**

VIII. **Annexes**

   a. Questionnaire

   b. Any reference material

## Annex II: Drawing a simple random sample in Excel

The steps below outline how to select a simple random sample using Excel. These assume basic functional knowledge of Excel (i.e. how to select and sort data with headers). This method below yields a simple random sample drawn without replacement.[7] In this example there are 20 elements in the sampling frame and the sample size is five. In practice sampling frames will be larger than this, but the principle is the same. Also notice that the random number column values will change when the data is sorted (i.e. the random number for Emily after the sorting is 0.230678996 while before the sorting it was 0.708375238); this is a function of Excel and does not affect the randomization of the list.

**Step 1:** Create sampling frame with a serial number column (A) and an element name column (B)
**Step 2:** Create a column (C) containing the formula "**=rand()**" – this creates a random number between 0 and 1
**Step 3:** Select all data and sort data by the random # column, this will randomize the list
**Step 5:** Select the first *n* elements, where *n* is the calculated sample size. In this case *n*=5

| A | B | C |
|---|---|---|
| **SN** | **Name** | **Random #** |
| 1 | Evan | 0.522187712 |
| 2 | Addison | 0.721172625 |
| 3 | Leah | 0.955778017 |
| 4 | Jordan | 0.80758306 |
| 5 | William | 0.695133887 |
| 6 | Emily | 0.708375238 |
| 7 | Nevaeh | 0.598052809 |
| 8 | Sophia | 0.946978987 |
| 9 | Benjamin | 0.041863475 |
| 10 | Logan | 0.744152558 |
| 11 | Isaiah | 0.331735652 |
| 12 | Aubrey | 0.623936624 |
| 13 | Samuel | 0.882505327 |
| 14 | Mason | 0.914833817 |
| 15 | Alexander | 0.049211495 |
| 16 | Joseph | 0.918567709 |

| A | B | C |
|---|---|---|
| **SN** | **Name** | **Random #** |
| 6 | Emily | 0.230678996 |
| 20 | Allison | 0.853756564 |
| 14 | Mason | 0.060315797 |
| 16 | Joseph | 0.204589465 |
| 11 | Isaiah | 0.419891075 |
| 2 | Addison | 0.860929987 |
| 7 | Nevaeh | 0.508474017 |
| 15 | Alexander | 0.351679481 |
| 13 | Samuel | 0.166642454 |
| 19 | Ella | 0.247629097 |
| 10 | Logan | 0.61861653 |
| 8 | Sophia | 0.647246434 |
| 12 | Aubrey | 0.116762208 |
| 5 | William | 0.523954646 |
| 17 | Gabriel | 0.305787708 |
| 3 | Leah | 0.126360711 |

[7] It should be noted that using the function "=randbetween(1,n)" may result in duplication since this uses sampling with replacement, not sampling without replacement. Sampling without replacement means sampled elements can only be selected once (e.g. drawing numbers out of a hat without replacing them back into the hat after selection). The alternative (rarely used for household surveys) is sampling with replacement, where elements can be selected more than once (e.g. drawing numbers out of a hat and replacing them back into the hat after selection).

| | | |
|---|---|---|
| 17 | Gabriel | 0.535387355 |
| 18 | Sofia | 0.10369085 |
| 19 | Ella | 0.141391749 |
| 20 | Allison | 0.144886266 |

| | | |
|---|---|---|
| 18 | Sofia | 0.25373812 |
| 9 | Benjamin | 0.365923959 |
| 1 | Evan | 0.425467795 |
| 4 | Jordan | 0.553570831 |

## Annex III: Drawing a systematic random sample in Excel

In the example below, five beneficiaries are systematically randomly selected from 32 beneficiaries for an interview. Notice that columns A and B are the same as above, but a column with random numbers between 0 and 1 is not necessary.

| A | B |
|---|---|
| **SN** | **Name** |
| 1 | Evan |
| 2 | Addison |
| 3 | Leah |
| 4 (**4**) | Jordan |
| 5 | William |
| 6 | Emily |
| 7 | Nevaeh |
| 8 | Sophia |
| 9 | Benjamin |
| 10 (**10.4**) | Logan |
| 11 | Isaiah |
| 12 | Aubrey |
| 13 | Samuel |
| 14 | Mason |
| 15 | Alexander |
| 16 | Joseph |
| 17 (**16.8**) | Gabriel |
| 18 | Sofia |
| 19 | Ella |
| 20 | Allison |
| 21 | Nathan |
| 22 | Henry |
| 23 (**23.2**) | Wyatt |
| 24 | Jonathan |
| 25 | Olivia |

| C | D |
|---|---|
| **S.I.** | **6.4** |
| **RS** | **4** |
| **Selected beneficiaries** | |
| **1** | **4** |
| **2** | **10.4** |
| **3** | **16.8** |
| **4** | **23.2** |
| **5** | **29.6** |

|  |  |
|---|---|
| 26 | Lily |
| 27 | Landon |
| 28 | Liam |
| 29 | Elijah |
| 30 (**29.6**) | Alexis |
| 31 | Abigail |
| 32 | Caleb |

**Step 1:** Calculate the sampling interval using the following formula: $.I. = \frac{N}{n}$ , where *N*=the total number of elements and *n*=the total number of sampled elements. Thus the sampling interval is computed to be 32/5=6.4. Note that the decimal is ***not*** rounded at this stage.

**Step 2:** Choose a random start between 1 and the sampling interval (e.g. 6.4). This can be done by entering the following formula in Excel "**=randbetween(1,S.I.)**", where S.I. is the sampling interval calculated above. In the second row in column D, a random start is chosen between 1 and 6.4 (4).

**Step 3:** Compute the selected serial numbers of beneficiaries using the following formula: RS, RS+S.I., RS+S.I.*2, … , RS+S.I.*(*n-1)*, where RS=the random start, S.I.=Sampling interval, and *n*=total number of elements selected. In this example:

|  | **RS** | **+S.I.*n** | **Selected beneficiary SN** |
|---|---|---|---|
| 1 | 4 |  | 4 |
| 2 | 4 | +6.4*1 | 10.4 |
| 3 | 4 | +6.4*2 | 16.8 |
| 4 | 4 | +6.4*3 | 23.2 |
| 5 | 4 | +6.4*4 | 29.6 |

**Step 4:** Now the decimals may be rounded and the selected beneficiaries are determined to be Jordan, Logan, Gabriel, Wyatt, and Alexis (highlighted in green, notice the corresponding SNs in parentheses).

## Annex IV: Drawing a PPS sample using Excel

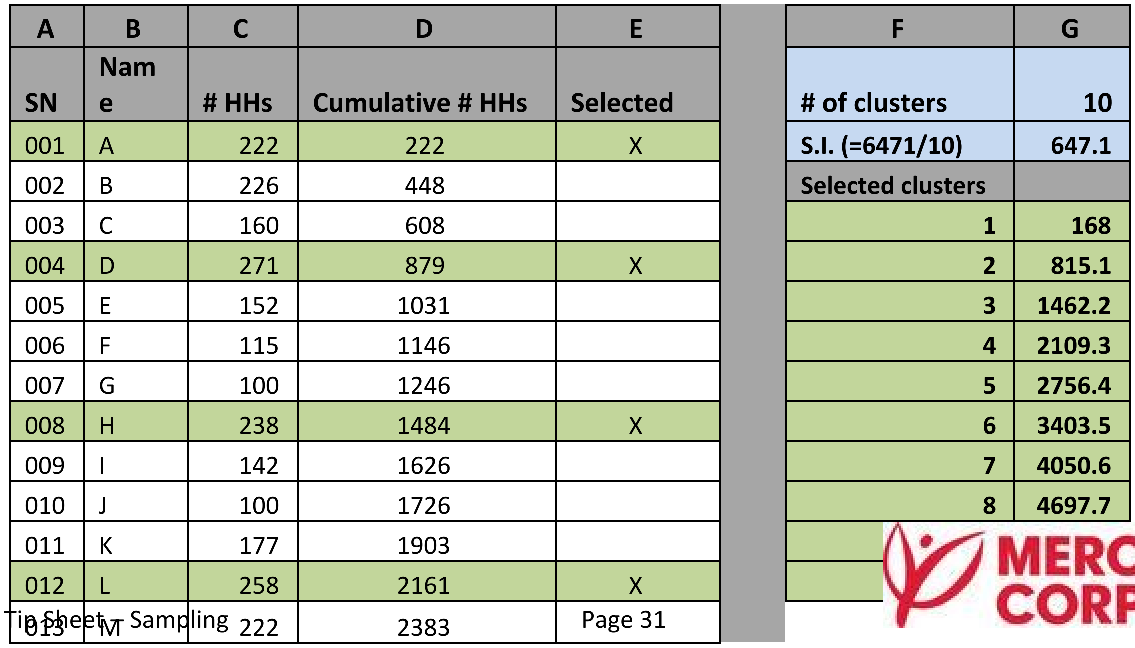
In the example below, a sampling frame of 32 clusters (e.g. villages) has been created and numbered (column A). Column (B) contains the name of the cluster and Column (C) contains the number of households (HHs) in the cluster. Ten clusters will be chosen using PPS.

**Step 1:** Create a column with cumulative size (D). Cumulative size means summing the households from previous clusters. For example, the cumulative size of clusters A and B is 448 since cluster A has 222 households and cluster B has 226 households. Notice that for cluster FF, the cumulative number of households represents the total number of households in the survey area (6,471). Essentially what this column is doing is creating a list of all of the households in the survey area.

**Step 2**: Calculate the sampling interval by dividing total cumulative size (6,471) by number of clusters selected (10) (see cells G1&2)

**Step 3:** Determine random start between 1 and the sampling interval (G4) using the excel formula "**=randbetween(1,S.I.)**", where S.I. is the sampling interval (in this case, 647.1)

**Step 4:** Calculate remaining selected clusters using the following formula: RS, RS+S.I., RS+S.I.*2, … , RS+S.I.*(*n-1)*, where RS=the random start, S.I.=Sampling interval, and *n*=total number of elements selected (see cells G5-13). Notice that unlike Annex 3 above, the numbers calculated in this step ***do not*** refer to the serial number of the cluster. Instead they refer to households within the cluster. For example, the random start is 168. This means the first randomly selected household is household number 168, which resides in cluster A since households 1-222 reside in cluster A. Similarly the second household selected is household number 815.1 which resides in cluster D, which contains households 609-879. Note that clusters with populations greater than the sampling interval may be selected more than once (see cluster X, #024).



| A | B | C | D | E |
|---|---|---|---|---|
| **SN** | **Nam e** | **# HHs** | **Cumulative # HHs** | **Selected** |
| 001 | A | 222 | 222 | X |
| 002 | B | 226 | 448 | |
| 003 | C | 160 | 608 | |
| 004 | D | 271 | 879 | X |
| 005 | E | 152 | 1031 | |
| 006 | F | 115 | 1146 | |
| 007 | G | 100 | 1246 | |
| 008 | H | 238 | 1484 | X |
| 009 | I | 142 | 1626 | |
| 010 | J | 100 | 1726 | |
| 011 | K | 177 | 1903 | |
| 012 | L | 258 | 2161 | X |
| 013 | M | 222 | 2383 | |

| F | G |
|---|---|
| **# of clusters** | **10** |
| **S.I. (=6471/10)** | **647.1** |
| **Selected clusters** | |
| 1 | **168** |
| 2 | **815.1** |
| 3 | **1462.2** |
| 4 | **2109.3** |
| 5 | **2756.4** |
| 6 | **3403.5** |
| 7 | **4050.6** |
| 8 | **4697.7** |

| | | | | |
|---|---|---|---|---|
| 014 | N | 282 | 2665 | |
| 015 | O | 105 | 2770 | X |
| 016 | P | 125 | 2895 | |
| 017 | Q | 150 | 3045 | |
| 018 | R | 157 | 3202 | |
| 019 | S | 291 | 3493 | X |
| 020 | T | 109 | 3602 | |
| 021 | U | 122 | 3724 | |
| 022 | V | 191 | 3915 | |
| 023 | W | 133 | 4048 | |
| 024 | X | 651 | 4699 | X,X |
| 025 | Y | 264 | 4963 | |
| 026 | Z | 206 | 5169 | |
| 027 | AA | 290 | 5459 | X |
| 028 | BB | 146 | 5605 | |
| 029 | CC | 172 | 5777 | |
| 030 | DD | 253 | 6030 | X |
| 031 | EE | 162 | 6192 | |
| 032 | FF | 279 | 6471 | |