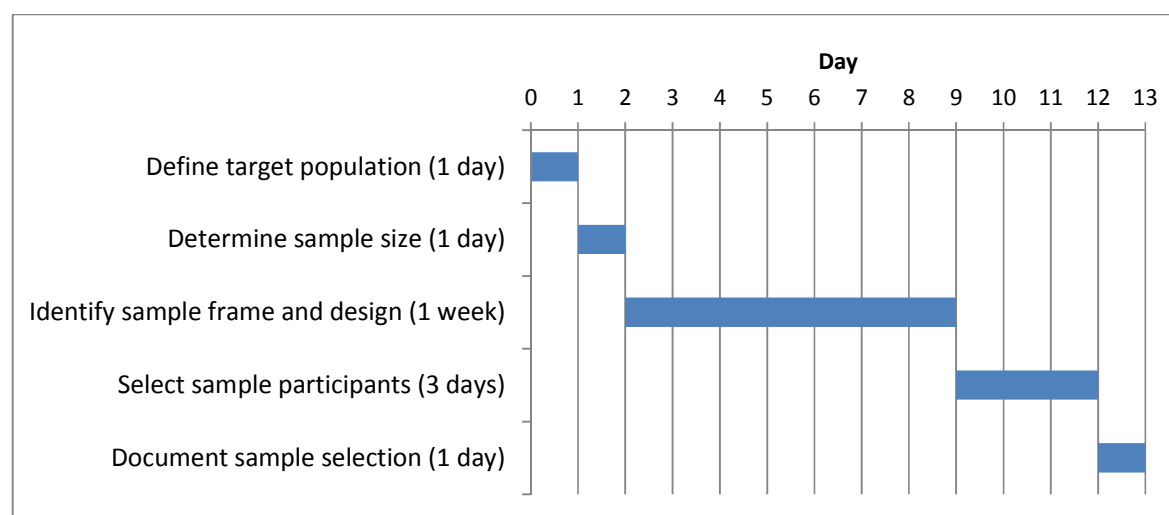


Section 2: Preparing the Sample

Overview

Introduction	This section covers the principles, methods, and tasks needed to prepare, design, and select the sample for your STEPS survey.
Intended audience	This section is primarily designed to be used by those fulfilling the following roles: • statistical adviser • STEPS Survey Coordinator • STEPS Coordinating Committee.
Tasks and timeframes	The sample is prepared as part of the process of planning and preparing the survey. This process should take between two days to one week, depending on the methods chosen and availability of information needed to draw the sample. The chart below lists the main tasks and timeframes covered in this section.



In this section This section covers the following topics:

Topic	See Page
Sampling Guidelines	2-2-2
Determining the Sample Size	2-2-3
Identifying the Sampling Frame	2-2-10
Choosing the Sample Design	2-2-12
Selecting the Sample	2-2-20
Documenting the Sample Design	2-2-24
Preparing Data Collection Forms	2-2-25

Sampling Guidelines

Introduction	High quality survey techniques can provide a good picture of risk factors for NCDs in a population by using a scientifically selected sample of that population. The sample will represent the entire target population if the sample is drawn correctly. High standards of sample design and selection are essential to achieve valuable and useful results from STEPS.
Reflecting the survey scope in the sample	To achieve a sample that reflects the scope of the survey it is essential to: • define a target population; • scientifically select a sample of the population that is representative of the target population; • plan ahead for reporting of survey results by sex and desired age groups.
Define the target population	Each country needs to define the target population for their STEPS Survey. To define the population, the purpose and use of the survey data need to be taken into account. For example, should the survey be representative of the entire population or a specific region? It is recommended that the target population for a STEPS NCD risk factor survey be at minimum all adults aged 18 to 69 years residing in the survey area. The age range may be expanded to include additional age groups, but it is not recommended to have a smaller age range.
Sample population	The sample population is a scientifically selected subset of the target population. Once the target population has been defined, the sample of participants within the target population will be selected.
Estimates for age-sex groups	The prevalence of most NCD risk factors tends to increase with age and vary by sex. Therefore it is recommended that survey results include estimates for specific age groups for each sex, in addition to the total survey population estimates, in order to provide a more nuanced picture of the prevalence of NCD risk factors in your target population. To ensure that precise estimates for each age-sex group can be calculated from the survey data, the total number of age-sex groups must be taken into consideration when calculating the sample size. Reporting estimates for a greater number of age groups will require a larger sample size. The STEPS recommended age groups are based on the Global Burden of Disease (GBD) age groups and are as follows: • 4 age groups per gender: 18-29, 30-44, 45-59, and 60-69 years • 3 age groups per gender: 18-29, 30-44, and 45-69 years • 2 age groups per gender: 18-44 and 45-69 years. If resources are extremely limited, estimates may be obtained only for the entire age span of the survey (e.g. 18-69). The next topic explains how to incorporate the number of age-sex groups into sample size calculation.

Determining the Sample Size

Introduction

In order to ensure a sufficient level of precision of the survey results, an adequate sample must be drawn from the target population. To calculate the sample size needed, the following factors must be taken into consideration:

- desired level of confidence of the survey results;
- acceptable margin of error of the survey results;
- design effect of the sampling methodology;
- estimated baseline levels of the behaviours or indicators to be measured.

Additionally, the sample size must be adjusted for:

- number of age-sex estimates
 - anticipated non-response.
-

Helpful Terminology

The following table provides a brief description of several key statistical terms. It is important to develop a good understanding of this terminology before proceeding to calculate the sample size.

Term	Description
Sample Mean / Prevalence	The estimated mean or prevalence of a given population parameter (e.g. mean number of days fruit was consumed in a given week) that is calculated from the survey data.
Population Mean / Prevalence	The true mean or prevalence of a given parameter for the entire target population. The sample mean is an estimate of the population mean.
Confidence Intervals	A range of values around the sample mean or prevalence in which the population mean or prevalence is likely to fall. For example, a 95% confidence interval indicates that for 95 out of 100 surveys, the population mean would fall into this range of values around the sample mean.

Continued on next page

Determining the Sample Size, Continued

Variables used for calculating sample size

The table below provides a description of the variables used in calculating the sample size as well as the recommended values for each variable.

Variable	Description	Recommended Value
Level of Confidence	- Probability value that is associated with a given confidence interval. - Describes the level of uncertainty in the sample mean or prevalence as an estimate of the population mean or prevalence. - The higher the level of confidence, the larger the sample size needed.	1.96 Note: 1.96 is the probability value associated with a 95% confidence interval.
Margin of Error	- The expected half-width of the confidence interval. - The smaller the margin of error, the larger the sample size needed.	0.05 Note: If the estimated baseline levels of the behaviours or indicators you wish to measure is very low (e.g. <0.10), then the Margin of Error should be decreased to 0.02 or smaller.
Design Effect (Deff)	- Describes the loss of sampling efficiency due to using a complex sample design. - The design effect for a simple random sample is 1.00. Sample designs more complex than a simple random sample require a larger sample to achieve the same level of precision in survey results as a simple random sample. Thus the design effect increases as the sample design becomes more complex.	1.50 Note: The value 1.50 is recommended for most STEPS surveys with complex sample designs. If design effect information is available from previous national surveys of a similar design to the proposed STEPS survey, it is recommended to use the previous estimates for design effect.
Estimated baseline levels of the behaviours or indicators we want to measure	- The estimated prevalence of the risk factors within the target population. - Values closest to 50% are the most conservative, requiring the largest sample size.	0.50, if no previous data are available on the target population. The value closest to 0.50, if previous data is available on the target population.

Equation for calculating sample size

The equation for calculating sample size is as follows:

$$n = Z^2 \frac{P(1-P)}{e^2}$$

where:

- Z = level of confidence
- P = baseline level of the indicators
- e = margin of error

Continued on next page

Determining the Sample Size, Continued

Example calculation

Using the above recommendations for each variable, the **initial** calculation for sample size would be:

$$n = 1.96^2 \frac{0.5 (1-0.5)}{0.05^2} = 384$$

However, this number **must** be adjusted to account for the design effect of the sample design, the number of age-sex estimates to be reported, and the anticipated non-response.

Adjusting for design effect

To adjust for the design effect of the sample design simply **multiply** the sample size by the design effect. For more information on choosing the sample design for your survey, see page 2-2-12.

Adjusting for number of age-sex estimates

As discussed previously, it is recommended that survey results be reported separately for specific age groups for each sex. In order to have an adequate level of precision for each age-sex estimate, the sample size must be **multiplied** by the number of age-sex groups for which estimates will be reported.

The number of age-sex estimates will vary according to the target age range of the survey and the resources available for the survey. For surveys covering the age range of 18-69, the number of age-sex estimates may be **8** (18-29, 30-44, 45-59, and 60-69 years for men and women), may be **6** (18-29, 30-44, and 45-69 years for men and women), or **4** (18-44 and 45-69 years for men and women).

If the age range of your survey extends beyond the recommended 18-69 years, the total number of age-sex estimates may need to be adjusted accordingly. For example, if the age range of 70+ years were also to be included in the survey, the total number of age-sex estimates would have to be increased accordingly.

Adjusting for anticipated non-response

To adjust for anticipated non-response **divide** by the anticipated **response rate**.

A response rate of 80% is the recommended rate to anticipate. This is a conservative estimate based on response rates of previous STEPS surveys. If response rates have been consistently higher in the country for similar household surveys, a less conservative (i.e. smaller) response rate may be used, such as 90%.

Example: For an anticipated response rate of 80%, divide the sample size by 0.80.

Continued on next page

Determining the Sample Size, Continued

Summary of sample size calculation

The table below provides a summary of the above steps to calculate sample size.

Step	Description
1	Determine the value of all variables needed to calculate sample size.
2	Use the level of confidence, margin of error, and baseline level of the indicators in the above equation to get an initial estimate for n (sample size).
3	Multiply n by the design effect and by the number of age-sex estimates.
4	Divide the result from step 3 by the anticipated response rate to attain the final sample size.

Sample Size Calculation Example 1 (4 age groups)

In this example, the recommended values for all parameters of the sample size equation will be used. Thus, the initial calculation proceeds as follows:

$$n = 1.96^2 * \frac{0.5 (1-0.5)}{0.05^2} = 384$$

This initial n is then multiplied by the design effect of 1.5 and, for example, 8 age-sex estimates desired for the survey results:

$$n = 384 * 1.5 * 8 = 4,608$$

Finally, n is divided by 0.80 to adjust for the anticipated 20% non-response rate:

$$n = 4,608 \div 0.80 = 5,760$$

5,760 is the final sample size.

Sample Size Calculation Example 2 (3 age groups)

In this example, the recommended values for all parameters of the sample size equation will be used and the initial calculation proceeds just as in the previous example:

$$n = 1.96^2 * \frac{0.5 (1-0.5)}{0.05^2} = 384$$

However, in this example the estimates will only be reported for 2 age groups for each sex as the sample size required for 4 age groups per sex is too large for the resources available. Thus, the initial n is then multiplied by the design effect of 1.5 and 3 age-sex estimates desired for the survey results:

$$n = 384 * 1.5 * 6 = 3,456$$

Finally, n is divided by 0.80 to adjust for the anticipated 20% non-response rate:

$$n = 3,456 \div 0.80 = 4,320$$

4,320 is the final sample size.

Continued on next page

Determining the Sample Size, Continued

Sample Size Calculation Example 3 (2 age groups)

In this example, the recommended values for all parameters of the sample size equation will be used and the initial calculation proceeds just as in the previous example:

$$n = 1.96^2 * \frac{0.5 (1-0.5)}{0.05^2} = 384$$

However, in this example the estimates will only be reported for 2 age groups for each sex as the sample size required for 4 age groups per sex is too large for the resources available. Thus, the initial n is then multiplied by the design effect of 1.5 and 4 age-sex estimates desired for the survey results:

$$n = 384 * 1.5 * 4 = 2,304$$

Finally, n is divided by 0.80 to adjust for the anticipated 20% non-response rate:

$$n = 2,304 \div 0.80 = 2,880$$

2,880 is the final sample size.

Sampling very small populations

When the target population is very small (appx. <50,000 people) the sample size can be reduced using a Finite Population Correction (FPC). The steps below describe how to check if the FPC is appropriate for a country and how to apply it to reduce the sample size.

Step	Description										
1	Complete only steps 1 and 2 in the preceding table to obtain the n for each estimate.										
2	Calculate the target population size for each estimate using available census data or a similar reliable data source. Example: If 8 age-sex groups will be the estimates, the number of individuals in each age-sex group (e.g. number of males aged 18-29) must be calculated.										
3	The FPC should only be applied when the sample to be drawn represents more than 10% of the target population. Thus for each estimate the n calculated in Step 1 must be divided by the target population size for that estimate to check to see if the FPC can be applied. Example: n has been calculated as 384. Eight age-sex estimates are desired. The table below shows the target population size for the first four estimates. <table border="1"> <thead> <tr> <th>Desired Estimates</th><th>Target Population Size</th></tr> </thead> <tbody> <tr> <td>Males, 18-29</td><td>2548</td></tr> <tr> <td>Females, 18-29</td><td>2641</td></tr> <tr> <td>Males, 30-44</td><td>3465</td></tr> <tr> <td>Females, 30-44</td><td>3356</td></tr> </tbody> </table>	Desired Estimates	Target Population Size	Males, 18-29	2548	Females, 18-29	2641	Males, 30-44	3465	Females, 30-44	3356
Desired Estimates	Target Population Size										
Males, 18-29	2548										
Females, 18-29	2641										
Males, 30-44	3465										
Females, 30-44	3356										

Continued on next page

Determining the Sample Size, Continued

3 (cont.)	Divide n by the target population for each estimate: $384/2548 = 0.15$ $384/2641 = 0.15$ $384/3465 = 0.11$ $384/3356 = 0.11$
4	If most or all of the quotients from step 3 are 0.10 or higher, then the FPC can be applied (continue to next step). Otherwise, return to step 3 in the preceding table and continue to calculate the total sample size using the n already calculated.
5	Apply the FPC to the n for each estimate using the following equation: $\text{new } n = \frac{n}{1 + \frac{n}{\text{population}}}$ where "population" refers to the target population for a given estimate, not the entire target population.
6	Sum all the "new n 's" together and multiply the sum by the design effect.
7	Divide the result from step 6 by the anticipated response rate to attain the final sample size.

Further modifications to sample size

There are a variety of situations which may require an adjustment to the sample size resulting from the calculations above. The table below describes some of these situations with directions on how to adjust the sample size. For any other situation not listed here, or if any other additional assistance is required, please contact the STEPS team.

If ...	Then ...						
Data for specific subgroups are required (e.g. ethnic groups, urban vs. rural dwellers).	There are two ways to proceed depending on the information desired: <table> <tr> <th>If ...</th><th>Then ...</th></tr> <tr> <td>Data will only be reported for all individuals in each subgroup.</td><td>Set the number of estimates to the larger of: - the number of age-sex estimates desired - the number of new subgroups. </td></tr> <tr> <td>Data will be reported for each age-sex group within each subgroup.</td><td>Multiply the number of age-sex groups by the total number of new subgroups (e.g. total number of ethnic groups) to determine the total number of estimates.</td></tr> </table> Note: It is important to keep these subgroups in mind when allocating the sample to ensure a sufficient number of participants can be drawn from each subgroup (see next topic).	If ...	Then ...	Data will only be reported for all individuals in each subgroup.	Set the number of estimates to the larger of: - the number of age-sex estimates desired - the number of new subgroups.	Data will be reported for each age-sex group within each subgroup.	Multiply the number of age-sex groups by the total number of new subgroups (e.g. total number of ethnic groups) to determine the total number of estimates.
If ...	Then ...						
Data will only be reported for all individuals in each subgroup.	Set the number of estimates to the larger of: - the number of age-sex estimates desired - the number of new subgroups.						
Data will be reported for each age-sex group within each subgroup.	Multiply the number of age-sex groups by the total number of new subgroups (e.g. total number of ethnic groups) to determine the total number of estimates.						

Continued on next page

Determining the Sample Size, Continued

Oversampling is desired for very small sub-populations.	Increase the overall n by increasing the n for the specific estimate(s) by 10%.
Oversampling is desired for specific sub-populations with higher than average non-response.	Increase the overall n by increasing the n for the specific estimate(s) by 10 to 20%.
Oversampling of the 60-69 year age group is desired because obtaining sufficient numbers of respondents from this age group is expected to be difficult due to high non-response and/or small size of this sub-population.	Increase the overall n by increasing the specific estimates for males and females in this age group by 10 to 20%. Oversampling 60-69 year olds within households can be done with the Android STEPS app.

Note: If oversampling is desired, adjustments usually must also be made when allocating the sample (see next topic). Often in addition to increasing the sample size, the sample allocation must take into consideration the location of hard-to-reach groups and allocate a greater proportion of the sample to these areas.

Sample Size Calculator

There is an Excel workbook, [sample_size_calculator.xls](#), that can assist in the calculations needed to determine the sample size for a survey. It is available on the STEPS website. The calculator allows to adjust all variables discussed here and also provides assistance in determining whether the Finite Population Correction (FPC) is applicable to a survey and, if so, how to correctly apply the FPC.

Smaller sample sizes

If the sample size calculations result in a sample size too large for the resources available, consider reducing the number of age-sex estimates desired for reporting of the results. Reducing the age-sex estimates can significantly reduce the sample size required for a survey.

Identifying the Sampling Frame

Introduction

A sampling frame is a list of units or elements that defines the target population. It is from this list that the sample is drawn. A sampling frame is essential for any survey.

Finding available sampling frames

To identify available sampling frames and determine which is best for a country, search for updated lists, databases, registers or other sources that give good coverage of the population to be surveyed. For example, look for population registers or census lists.

Various government departments and national bodies should be consulted to establish what frames exist in a country and, if suitable, whether they may be accessed for STEPS.

Enumeration areas (EAs)

Most often the sampling frame will use enumeration areas (EAs) which are small- to medium-sized geographic areas that have been defined in a previous census. Most countries have this information and it is usually preferable to incorporate this into the sampling frame.

Factors to consider

A sampling frame, or a collection of them, should cover all of the population in the surveyed country. Good coverage means that every eligible person in the population has a chance of being included in the survey sample.

Representativeness for all sub-populations should be considered when deciding which frame(s) to use, since there is a possibility that particular age, gender or ethnic groups or geographical areas are more or less likely to be included in the sampling frame. Bias will occur if there is poorer coverage for some groups.

Multiple Sampling Frames

Due to logistical and financial limitations, most national surveys employ multi-stage sampling, which is discussed in detail in the following topic. A multi-stage sample design will require a sampling frame for each stage of sampling.

Continued on next page

Identifying the Sampling Frame, Continued

Features of a good sampling frame

Some features of a good sampling frame are:

- it does not contain duplicates, or if present they can easily be identified and removed;
- it does not contain blanks, such as empty houses or a deceased individual;
- it contains information enabling all units to be distinguished from all others and to be easily located (e.g. a complete street address);
- at minimum, it contains information about the number of households or total number of individuals;
- it could be made accessible to the STEPS country team within a reasonable timeframe and at no large expense.

Note: Sampling frames must be assessed for all the above features, but particularly for **completeness** and **potential bias**.

Choosing the Sample Design

Introduction

The selection of the sample design is highly dependent on a variety of factors, most importantly the size of the population, the geography of the area to be covered, and the resources available for the survey. All factors must be kept in mind in selecting the sample design for the survey.

Stratification

Stratification is the process of dividing the sampling frame into mutually exclusive subgroups or strata. The sample is then drawn either proportionately or disproportionately from **all** strata. How the target population is stratified depends on the information that is available for the sampling frame and the information that is desired from the survey results.

Strata are often based on the physical location of the sampling units. Some examples of these types of strata are:

- enumeration areas (EAs) or other well-defined geographic regions
- urban vs. rural areas.

Less often, strata are based on the characteristics of the individuals in the sampling frame. This is less common in large national surveys due to a lack of precise data on all individuals in the target population and the difficulties of developing sampling frames for each strata. Some examples of these types of strata are:

- ethnicity
- socioeconomic status
- gender.

Stratification is not required but is recommended for the following reasons:

- increased precision of survey estimates
- guaranteed coverage of all strata
- administrative convenience.

Stratification can be applied in conjunction with other sampling strategies. This section discusses simple random sampling and multi-stage cluster sampling, both of which can be used along with stratification, as described later in this topic.

Stratification and sample allocation

If the decision has been made to stratify the population, it must then be decided whether to sample proportionately from all strata or to sample a larger proportion of individuals from some strata and a smaller proportion of individuals from other strata (disproportional allocation).

Continued on next page

Choosing the Sample Design, Continued

Stratification and sample allocation (cont.)

Proportional allocation means sampling the same proportion of individuals from each strata so that the resulting sample is distributed across the strata similarly to the underlying target population. This type of sample allocation is the appropriate method for surveys which will only be reporting data for all strata combined.

Disproportional allocation means sampling some strata at a higher rate than other strata. Often this is implemented by drawing an equal sized sample from each strata. This type of sample allocation is appropriate when survey results are desired for each individual strata. In this situation, a larger sample size is usually required to ensure adequate precision in the strata-specific estimates. The primary drawback to this method is a loss of sampling efficiency for the estimates for all strata combined.

Note: In some cases where very small strata exist, proportional allocation may be done but oversampling may be required for the very small strata.

Proportional Allocation Example

Because proportional allocation is more likely to be used for a STEPS survey, an example is provided here.

In this example, the sample size has been calculated to be 2,880. The target population has been divided into the 4 government districts of the country. These districts will serve as strata. The target population within each strata has been listed in the table below along with the proportion each comprises of the total target population.

Strata	Target Pop.	Proportion of Pop.
District 1	25,955	0.24
District 2	30,568	0.28
District 3	32,578	0.30
District 4	19,054	0.18
Total	108,155	1.00

$= 25,955 \div 108,155$

To compute the number of individuals from the total sample to be drawn from each strata, multiply the total sample size by the proportion for each strata.

Strata	Target Pop.	Proportion of Pop.	Sample
District 1	25,955	0.24	691
District 2	30,568	0.28	807
District 3	32,578	0.30	864
District 4	19,054	0.18	518
Total	108,155	1.00	2,880

$= 0.24 \times 2,880$

Continued on next page

Choosing the Sample Design, Continued

Simple random sampling

In a small number of settings simple random sampling may be feasible. For household surveys, the following characteristics generally should be met:

- small target population;
- small survey area, the entirety of which can be covered by the resources available;
- detailed sampling frame is available, listing, at minimum, all households in the survey area, or, at best, all eligible individuals in the survey area.

Simple random sampling can be combined with stratification. In stratified random sampling, the population is first stratified and then a random sample is drawn from each strata.

Note: If simple or stratified random sampling is deemed to be feasible in a country, a smaller sample size can be used. In the calculation for sample size a design effect of 1 should be used.

Multi-stage cluster sampling

Multi-stage cluster sampling is one of the most common sample designs for national surveys and it is the recommended method for most STEPS surveys.

"Multi-stage" indicates that sampling is done in several steps. First larger sampling units are selected then smaller sampling units are selected within the selected larger units. "Cluster" refers to the fact that the sampling units are subdivided into mutually-exclusive clusters and, unlike stratification, only a **sample** of these clusters is selected for the survey.

Why use multi-stage cluster sampling?

The table below highlights two primary reasons for using multi-stage cluster sampling. These are very common problems in national surveys that can be overcome with the use of multi-stage cluster sampling.

Problem	Solution
Detailed information does not exist for all households or individuals in the sample population and it is not feasible to create a detailed sampling frame for the entire survey area.	Multi-stage cluster sampling allows for the selection of larger sampling units (e.g. villages) that require less detailed information about the target population. It is only at the final stage of sampling (most often the selection of households) that detailed information needs to be available. However, because only a selection of clusters will be chosen at each stage of sampling, the detailed sampling frames are only needed for a subset of the entire target population.

Continued on next page

Choosing the Sample Design, Continued

Why use multi-stage cluster sampling? (cont.)

Problem	Solution
The survey area is too large and/or travel costs are too high to draw a sample from the entire country or all regions of interest.	Because the sample is only drawn from selected clusters, multi-stage cluster sampling allows for a reduced area to be surveyed while maintaining a sample that is nationally (or subnationally) representative. Note: Using multi-stage cluster sampling does not <i>guarantee</i> a representative sample. If done incorrectly, it will not result in a representative sample. The design of the clusters and the selection of clusters at every stage must be done carefully and consistently and must be documented in detail.

Preparing a Multi-stage Cluster Sample

In order to implement multi-stage cluster sampling, the population must be divided into clusters, each of which contain either a number of smaller clusters or, at the final stage, households or individuals.

The flowchart to the right is one example of the multiple sampling stages that could be defined for a country.

Most often the first stage uses enumeration areas (EAs) from census information. The intermediary stages, if any, may be comprised of existing geopolitical units (e.g. villages) or artificially-created units (e.g. a specified collection of city blocks).

Population



District



Village



Household



Individual

Important: The number of sampling units at the initial stage must be fairly numerous (i.e. >100) so at least 50-100 of them can be selected. Selecting a smaller number of sampling units at the initial stage of sampling results in more clustered data and a loss of precision in survey estimates.

A sampling frame will need to be constructed for all clusters in the first stage of sampling. At minimum these sampling frames must contain the total number of households or total number of target individuals in the cluster.

Sampling frames will only be needed for **selected** clusters at all subsequent stages of sampling, with detailed information (i.e. lists of households or eligible individuals) only needed for the sampling frames for the last stage of sampling.

Continued on next page

Choosing the Sample Design, Continued

Multi-stage Cluster Sampling Terminology

The table below describes some key terminology for multi-stage cluster sampling.

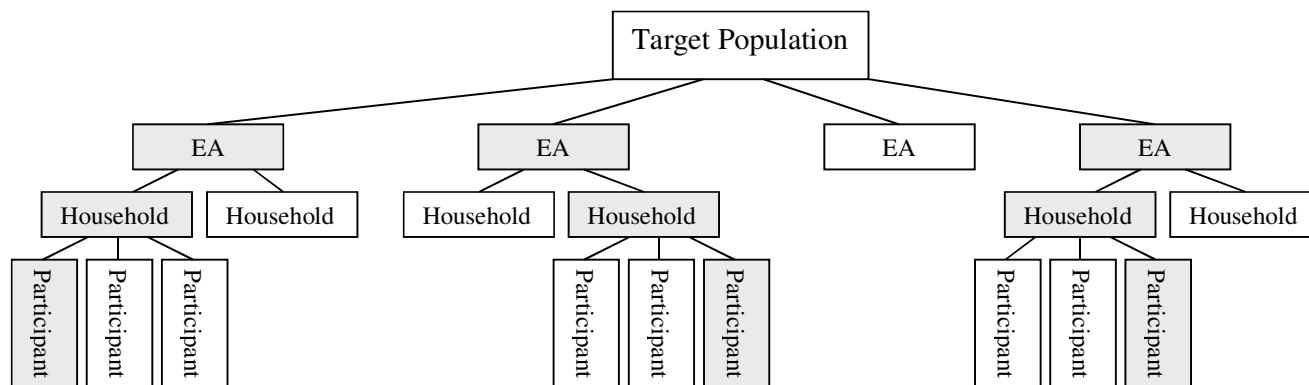
Term	Definition
Primary Sampling Unit (PSU)	These are the clusters that are selected first. Most often the PSUs are enumeration areas (EAs) from a recent census.
Secondary Sampling Unit (SSU)	The clusters that are selected second, separately within each selected PSU.
Tertiary Sampling Unit (TSU)	The clusters that are selected third, separately within each selected SSU.

The list of terms could be extended to describe more levels of sampling as needed.

Example 1

In the following example, there are three stages of sampling. EAs are serving as the PSUs. For each selected PSU, a sampling frame was created comprised of a list of households in the EA. Households were then selected within each PSU and then one participant was selected within each household.

Shaded boxes indicate that the cluster or participant was selected.



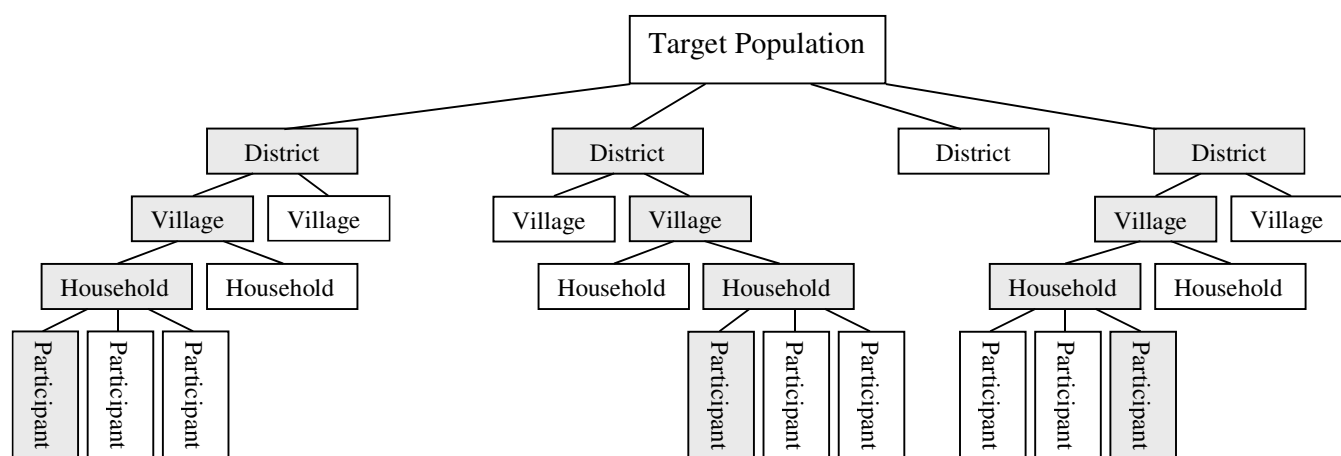
Continued on next page

Choosing the Sample Design, Continued

Example 2

In this example, there are four stages of sampling. Districts are serving as the PSUs. For each selected PSU, a sampling frame was created comprised of a list of all villages (the SSUs) with the target population of each village. For each selected village, a sampling frame was also created, comprised of a list of all households in the village. If a detailed list of all eligible individuals were available for any selected village, this list could be used in place of the household list and selection could proceed directly from the village level to the participant level.

Shaded boxes indicate that the cluster or participant was selected.



Qualities of a Good Multi-stage Cluster Design

One very important check to perform on the multi-stage cluster design is that every individual in the target population is included in only one sampling unit per stage. This means that the clusters at each level of sampling must cover the entire target population and be mutually exclusive (non-overlapping).

Additionally, it is important to check the characteristics of the PSUs. The first two items in the table below can be used to check the SSUs, TSUs, etc. as well, but given the nature of multi-stage cluster designs, these checks are most critical for the PSUs.

If ...	Then ...
PSUs exist that are very small.	Combine these PSUs with a neighboring PSU before selecting the sample.
PSUs exist that are very large.	Split these PSUs into two or more smaller PSUs that are more similar in size to other PSUs.
Total number of PSUs is small (i.e. <100).	Begin sampling at the SSU level (the SSUs would then become PSUs) or subdivide the existing PSUs to ensure that at least 50-100 PSUs can be selected.

Continued on next page

Choosing the Sample Design, Continued

Sample Allocation and Multi-stage Cluster Design

Once the sampling units to be used for PSUs, SSUs, etc. have been determined, the allocation of the sample must be decided. That is, the total number of PSUs to be selected, the total number of SSUs to be selected per PSU, etc. must be determined.

The table below describes the steps to take to determine how to allocate the sample.

Step	Description
1	Calculate the total sample size.
2	Assess the resources available and determine the total number of PSUs to be sampled, keeping in mind that at least 50 to 100 PSUs should be selected.
3	Divide the total sample size by the number of PSUs to be sampled to determine the number of individuals to be sampled per PSU.
4	Continue subdividing the sample size at each stage of sampling according to the number of sampling units to be selected at each stage.

Note: As stated previously, stratification can be combined with a multi-stage cluster design. The total number of PSUs would be allocated proportionately or disproportionately (depending on the requirements of the survey results) across all strata and sample allocation would continue within each strata following the steps above.

Example

For this example, assume that the total sample size has been calculated to be 2,880 individuals. It has also been decided that regions will serve as PSUs, villages will serve as SSUs, and then households will be selected in each village. Resources will allow for 72 PSUs to be selected, meaning that 40 ($= 2880/72$) individuals will be selected per PSU.

There is some flexibility in how the 40 individuals per PSU are allocated. At this point it would be worthwhile to consider a few scenarios and select the one that is feasible yet provides a good distribution of individuals across the PSU (i.e. not too many or too few of the 40 individuals drawn from a given village). Two scenarios are presented below:

Continued on next page

Choosing the Sample Design, Continued

Example (cont.)

Scenario	Description
1	10 individuals will be selected per village, meaning that 4 villages (= 40/10) must be selected per PSU. <u>Sample allocation:</u> 72 regions x 4 villages/region x 10 individuals/village = 2880
2	5 individuals will be selected per village, meaning that 8 villages (= 40/5) must be selected per PSU. <u>Sample allocation:</u> 72 regions x 8 villages/region x 5 individuals/village = 2880

In terms of resources, the key difference between the above scenarios is the number of villages that would need to be visited within each PSU. This number will likely be a deciding factor in the allocation of the sample, keeping in mind that having a high number of individuals selected from only a few villages would result in greater clustering of survey data and a potential loss of precision in survey estimates.

Example with stratification

For this example, assume again that the total sample size has been calculated to be 2,880 individuals and that regions will serve as PSUs, villages will serve as SSUs, and then households will be selected in each village. Resources will allow for 80 PSUs to be selected. However, the survey designers wish to ensure that the sample is drawn proportionately across the 4 islands that comprise the country.

The table below shows the proportion of the total underlying population that each island represents. The right-most column shows how the number of PSUs would be proportionately allocated across these 4 islands or strata.

Island	Proportion of Total Pop.	PSUs
A	0.50	40
B	0.175	14
C	0.125	10
D	0.20	16
Total	1.00	80

Thus, 40 regions (PSUs) will be picked out of all regions on island A, 14 regions will be picked out of all regions on island B, and so on. Once the PSUs are selected per island, sample allocation continues just as in the preceding example, with the same number of villages being selected in each PSU, regardless of the island on which the PSU is located.

Selecting the Sample

Introduction	Once the sample design is selected and the sampling frame has been prepared, sample selection can start. This section provides instructions for the various stages of sampling.
Available tools	There is an Excel workbook entitled STEPSSampling.xls that includes spreadsheets for every stage of the sample selection. STEPSSampling.xls will: • provide probability proportional to size (PPS) sampling (see description below) for primary and secondary sampling units as needed; • randomly select households or individuals; • provide information for weighting the data. The spreadsheet is available on the STEPS website (www.who.int/ncds/steps).
Probability proportional to size (PPS) sampling	Probability proportional to size (PPS) sampling is a method for selecting a sampling unit in which the probability of selection for a given sampling unit is proportional to its size (most often the number of individuals or households within the sampling unit). PPS sampling is appropriate for use when sampling units are of markedly different size. In these situations, were random sampling to be used to select sampling units, those individuals in the larger sampling units would have a much smaller chance of selection than those individuals in the smaller sampling units. PPS sampling corrects this problem, therefore reducing bias in survey estimates.
Instructions for PPS sampling	The table below outlines the steps required to perform PPS sampling on a list of sampling units. Before beginning, a list of sampling units and their corresponding sizes (in number of households or in population) must be compiled. It is recommended that this list be organized geographically, meaning that sampling units located near each other are also near each other on the list. Additionally, the number of sampling units (clusters) to be selected must be decided. The STEPSSampling.xls tool will automatically perform Steps 3 through 8 in the table below. The instructions worksheet inside the file explains how to perform PPS sampling using either the PSU or SSU worksheet in the file.

Continued on next page

Selecting the Sample, Continued

Instructions for PPS sampling (cont.)

Step	Action
1	Create a list of all sampling units with their size (either number of households or population). If possible, order this list geographically, placing sampling units that are physically adjacent near each other on the list.
2	Determine the number of sampling units to be selected from the list.
3	Create a new column containing the cumulative size of the sampling units. The final total should match the total population across all sampling units.
4	Divide the total cumulative population size (N) by the number of sampling units to be selected (n) to obtain the sampling interval (k). $k = N/n$
5	Choose a random number (r) that is between 1 and the sampling interval (k). $1 < r < k$
6	Start at the top of the list and select the first sampling unit whose cumulative population size includes the random number (r).
7	To select the second cluster, first add the sampling interval to the random number (r). Then begin counting from the previous cluster selected until the cumulative population size includes this sum (r+k).
8	Select the remaining clusters by adding the sampling interval, multiplied by 2, then 3 and so on, to the random number. Always start counting from the previous cluster selected not the start of the list. $r+(k*2)$ $r+(k*3)$ etc
9	Continue until the end of the list is reached. Do not stop as soon as n units have been selected. To avoid bias, all units selected must be used in the survey even if the number is slightly greater than n.

Using PPS sampling with a multi-stage cluster design

PPS sampling can be applied at all stages of a multi-stage cluster design except for the final stage in which households or individuals are selected.

The **STEPSsampling.xls** tool provides worksheets for selecting PSUs and SSUs using PPS sampling. The worksheet entitled PSU allows for the selection of up to 100 PSUs from an entered list of all PSUs. The worksheet entitled SSU allows for the selection of the SSUs within each selected PSU. Therefore, the SSU worksheet must be duplicated, one for each PSU that was selected, so that an independent selection of SSUs can be performed for each PSU.

Continued on next page

Selecting the Sample, Continued

Selection of households and/or individuals

The final stage of sampling, the selection of households and/or individuals, will depend on the type of information available. The table below describes the possible scenarios for the final stage of sampling and the sample selection process for each.

If ...	Then ...
A list of eligible individuals is available for the selected sampling unit (e.g. village).	First check that the list of eligible individuals meets the following requirements: • the list is up to date, for example, people who have moved away or who have died are not included in the list; • the list contains specific information allowing for each selected individual to be located by the interviewers. If both conditions are met, the selection of individuals can be done randomly from the list.
No or limited information is available about the individuals in the selected sampling unit but a list of households exists for the sampling unit.	First check that the list of households meets the following requirements: • the list is up to date and each household listed represents a single dwelling; • the list contains specific information allowing for each selected household to be located by the interviewers. If both conditions are met, the selection of households can be done randomly from the list. From the selected households, participants can be selected randomly using the STEPS Android app. If there is a concern that the list may be out of date, it is recommended that the field team first performs a quick mapping and household listing of the selected sampling units/clusters to update the list, noting abandoned/destroyed dwellings, new dwellings, or expanded dwellings (single family into multi-family).

Continued on next page

Selecting the Sample, Continued

Selection of households and/or individuals (cont.)

If ...	Then ...
The number of households is known for the sampling unit but there is no information about their location.	In this situation the sampling unit should be mapped to determine the location of the households. Please contact the STEPS team for more guidance on this method or other alternatives.

In the **STEPSsampling.xls** tool, the "RandHhold" worksheet can be used to randomly select the desired number of participants from a list of eligible individuals or the desired number of households from a list of households.

It is possible that some sampling units have more detailed information available than others. In this case, the above scenarios can be used on a case-by-case basis, meaning in some sampling units with more detailed information individuals may be selected directly while in other sampling units with less detailed information households may need to be selected first.

Note: In all STEPS survey designs, sampling is non-replacement, meaning that once a unit or person is selected they are not replaced with another person/unit. If non-respondents or persons who are not at home for the interview are replaced, a convenience sample will be performed and the results will only represent the people sampled and not the target population.

Selection of an individual within a household

In most STEPS surveys, the selection of an individual within a household is only done once the STEPS data collector is in the household. The Android STEPS app has an integrated random selection procedure, whereby eligible household members are entered and one participant from this list is randomly selected by the device.

Eligibility criteria for households and members of the households to be included will need to be defined by the STEPS Coordinating Committee in advance of the fieldwork.

Oversampling for 60-69 year olds

Depending on a country's population structure it may be difficult to obtain enough respondents from the 60-69 year old age group to get precise estimates for this age group. One possible solution to this problem is to oversample this age group at the household level. Therefore, at each household with adults aged 60-69, two adults will be selected.

Documenting the Sample Design

Introduction Once the sample design and methodology have been chosen, all aspects of the sample need to be clearly documented.

Purpose The purpose of documenting the sample design is primarily for the data analyst to understand how the sample was drawn in order to appropriately adjust the results to the target population. Additionally, an abbreviated version of the documentation should always accompany any presentation of the survey data to explain how the data were collected.

Recordkeeping during data collection Sufficient records must be kept **during data collection** to ensure that the data analyst can do all possible adjustments to make the results representative of the target population. Most importantly, the data analyst must know:

- the probability of selection of each sampling unit at every stage of sampling (i.e. probability of selection for each PSU, SSU, household, individual);
- the age and sex of any non-responders.

Thus, it is critical to keep a record of the following:

- all sampling frames used at each stage of sampling
 - sample selection method used at each stage of sampling
 - stratification design, if stratification is used
 - for each respondent, the PSU, SSU, etc. from which he/she was selected.
-

Future surveys Documenting the sample design and methodology is also important for future surveys when changes in risk factors over time are being examined, since methods chosen in future surveys may differ from this one and thus affect comparability.

Archiving documents It is important that all relevant sampling materials be archived. This includes the forms discussed in the next topic of this Section, "Preparing Data Collection Forms", as well as all information used to design and draw the sample.

If the sample is drawn by another government entity (e.g. the Statistics Bureau), be sure to obtain from them all materials and information that were used to draw the sample.

Preparing Data Collection Forms

Introduction

Once the sample has been drawn, the Interview Tracking Forms and the Step 3 Appointment Cards should be prepared for the data collection team. It is recommended that the field team supervisors and the statistical adviser collaborate on this task to ensure the forms are correctly filled out and properly organized for data collection.

Assigning Unique Identifiers and preparing stickers QR codes

Before preparing the data collection forms, ID Numbers must be assigned to all interviewers and to all selected clusters from which households and/or individuals will be selected. Additionally, all households and all participants to be selected should each be assigned a unique ID.

Due to the fasting requirement for Step 3 measurements, data collection for Step 1 and 2 generally takes place a day before data collection for Step 3. The unique ID of the participant will help match the Step 1 and 2 data with Step 3 data. In order to exclude errors during this matching process, it is recommended to also use Quick Response (QR) codes.

In preparation of the field work, the QR codes are printed on stickers. It is recommended to put one sticker on each container for urine collection or on each Step 3 Appointment Card as they are prepared, before handing them out to the interviewers. The Geneva STEPS team can help print the stickers.

During data collection, the QR codes are scanned for each participant with the Android device: once during data collection for Step 1 and 2, and once during data collection for Step 3.

The table below provides further instructions for assigning ID Numbers.

Continued on next page

Preparing Data Collection Forms, Continued

Variable	Description
Interviewer ID	Every interviewer should be assigned a unique ID number.
Device ID	Every Android should be assigned a unique ID number. If a device stops working during data collection, do not re-assign its Device ID to another device. It is recommended to assign the device the same ID as the interviewer using it.
Cluster ID	A unique number should be assigned to all selected sampling units from which households and/or individuals will be selected. Often these sampling units are villages, but could instead be city blocks, city districts, etc., depending on the sample design. Note: If household or individual selection is the <u>first or only</u> stage of sampling, it is not necessary to use Cluster IDs.
Household ID	All households to be visited should be assigned a unique ID. The Android STEPS app will automatically generate a unique household ID for each household visited based on a combination of the device ID and a unique number. If the app is not used and household IDs are generated manually, these numbers should be consecutive from 1 through the total number of households to be visited. If no interview is conducted at a selected household, the Household ID assigned to it is simply not used.
Participant ID	All participants should be assigned a unique ID. The Android STEPS app will automatically generate a unique participant ID for each participant based on the Household ID plus a unique number. If the app is not used and participant IDs are generated manually, these need not be consecutive and can be grouped by Cluster ID, where a sequence of participant IDs is associated with each Cluster ID (e.g. Participant IDs 101-120 are assigned to Cluster ID 1, Participant IDs 201-220 are assigned to Cluster ID 2, etc.). Note: In countries where no oversampling is done and IDs are generated manually, Participant and Household IDs can be the same.
QR code	In addition to the Participant ID, QR codes can be used to avoid errors, and to ensure easy matching of data collected for Step 1 and 2 with data collected for Step 3. The QR codes must be unique for each survey participant and are scanned using the Android device.

Continued on next page

Preparing Data Collection Forms, Continued

Assigning Unique Identifiers (cont.)

The following identifiers will also need to be assigned and made available to the data collection team as needed:

Variable	Description
Technician ID	If Step 2 and/or Step 3 will be implemented by someone other than the interviewer (e.g. specific Step 2 and/or 3 data collectors), these individuals should be assigned a Technician ID.
Device ID	Any equipment used for Step 2 and 3 should be assigned a unique Device ID.

Interview Tracking Form

All countries should use the Interview Tracking Form (see template in Part 6, Section 2) regardless of their sample design. This information is used for calculating the weights and response proportions for Step 1, Step 2, and Step 3.

Before data collection begins, Interview Tracking Forms should be completed for each Cluster and each interviewer who will conduct interviews in that Cluster.

Before data collection begins, the following should be completed on each Interview Tracking Form:

- Cluster ID
- Interviewer ID
- Household IDs (if not using auto-generated IDs from STEPS app)
- Participant IDs (if not using auto-generated IDs from STEPS app).

Note: If household or individual selection is the first stage of sampling (i.e. Cluster IDs are not used), then prepare the Interview Tracking Forms for each interviewer without assigning Cluster IDs.

Step 3 Appointment Card

The Step 3 Appointment Card (see template in Part 6, Section 2) that serves for arranging appointments at the Step 3 location should be partly filled in before the interviewers start data collection:

- Participant ID (if not using auto-generated IDs from STEPS app)
- Centre name

If the stickers with the QR codes for each participant are not put on the containers for urine samples, these can alternatively be put on the Step 3 Appointment Card.

