# Section 1: Creating the Final Dataset

## Overview

**Introduction**

This section covers all the tasks that need to be conducted to prepare the final STEPS dataset for analysis.

**Intended audience**

This section is designed for use by those fulfilling the following roles:

- Field team supervisors
- STEPS Survey Coordinator
- Data analyst.

**Overview of process**

Once data collection has been completed, one person should oversee the task of creating the final dataset. This task may be completed by the data analyst, but they may need assistance from the survey coordinator or field team supervisors to coordinate obtaining all data files from the devices used for data collection and amassing the Interview Tracking Forms.

The process for creating the final dataset is comprised of three stages:

- Downloading the data
- Cleaning the data
- Weighting the data.

While the first two stages should be able to be completed within a few hours, the time needed for weighting the data can vary from less than a day to several days or weeks, depending on the availability and cleanliness of the sampling information.

**In this section**

This section covers the following topics:

# Downloading the Data

**Introduction**

Prior to downloading the data from the online eSTEPS platform, all Android devices should be checked to ensure that all completed questionnaires have been uploaded to the server.

On each device, tap on "Submit Records" from the STEPS home screen Menu to check if there are any records still to be submitted. Only once this check has been done on all devices should the data be downloaded.

**Procedure**

Follow the instructions in the table below to download your STEPS data from the online eSTEPS platform.

| Step | Description |
|---|---|
| 1 | Log into the online eSTEPS platform using your user name and login. |
| 2 | Click on the link to your survey. This will take you to a list of all instruments associated with your survey.<br>**Note:** If there are two instruments for your survey (e.g. one for Step 1 and 2 data and one for Step 3 data) you will need to complete Steps 2-6 twice, once for each instrument. At the end of this process, you will have two separate datasets which will need to be merged together by matching records by Participant ID. |
| 3 | Click on the link to your instrument. |
| 4 | In the "Export data" box on the screen, choose the format in which you wish to download your data. Excel is the recommended format. |
| 5 | Click on "Show advanced export options" and make sure "Remove prefixed group names" is not selected. |
| 6 | Click on the "Prepare Excel" button just underneath the file type selection. |
| 7 | Wait for the file to be prepared. Once it is ready, you will see a "Download XLSX" button at the bottom of the Export data box on the screen. Click on this button to download your data. The file will automatically be named as follows: [your instrument file name]_[date]_[time].xlsx. Thus, the date and time of the data download are automatically included in the name of the file. |

**Household data**

While the downloaded data from the online eSTEPS platform already includes the household size information from the household listings (needed to weight the data), it is still important to download and review the household data for your survey as it contains important information pertaining to participant selection.

Log into the household database of the eSTEPS online platform using the Survey ID and password for your survey then click on "Download XLS" to download the household data to a csv file (can be read in Excel).

# Cleaning the Data

**Introduction**

While the STEPS Android app assures a very high level of data quality (i.e. skips have been properly followed and responses are internally consistent), there are still errors that can happen during data collection. Described here are a variety of checks that should be performed on your STEPS dataset.

**Participant ID**

Participant ID (PID) is automatically generated by the STEPS app when the participant is selected at the household level. If Step 3 data collection occurs during a follow-up visit to the household or at a nearby location, this PID will need to be entered by the Step 3 data collector. It is at this point that data entry errors may occur.

PID should be unique across all records and will serve to align the Step 1 and 2 with Step 3 datasets when Step 3 data is collected separately.

**Note:** It is possible to incorporate into the local STEPS Instrument a barcode or QR code as an additional means to label and match records. Contact the WHO Geneva STEPS team for more information.

**Location variables**

At the beginning of the STEPS Instrument, there are a few variables that identify the location of the survey. At minimum there is usually Cluster ID and Cluster Name, though the names of these variables may be modified in your local STEPS Instrument and additional variables may be added.

The location variables are critical for weighting the data. An error in Cluster ID (or the equivalent in your local STEPS Instrument) would mean the wrong sampling weight is assigned to a given record. Thus, location variables must be carefully reviewed to assure they are correct. It is also important that the IDs used to identify sampling units (e.g. villages) in your dataset match the IDs in the sampling documentation. You will need to review the sampling information to ensure such alignment.

**Open-ended questions**

Throughout the STEPS Instrument, there are open-ended questions, such as the number of manufactured cigarettes smoked per day. While the electronic STEPS Instrument should have included limits on these fields, these limits are typically quite generous. Therefore, the responses to these questions should be reviewed to identify any possible errors. Keep in mind that some responses may not seem questionable in isolation, but may seem very questionable when reviewed alongside the participant's responses to related questions. For example, a person who smokes 30 manufactured cigarettes a day may not seem unusual. But if the same person claims to also smoke 30 hand-rolled cigarettes per day and 30 cigars a day, then the response becomes questionable.

*Continued on next page*

**Resolving errors**

If possible, try to correct any errors found in the dataset. You can use existing survey documentation (e.g. sampling documentation, interview tracking forms) to double check location variables. For errors in questionnaire responses, it is best to follow up with the field team supervisor(s) to see if the relevant data collector can clarify. If possible, the participant can also be contacted to clarify their response.

Do not make any corrections to the dataset until you are certain you have the correct information. If you are unable to correct questionable data, it is recommended the data is excluded from the dataset. It is ok to exclude only part of an individual's record if the rest of their response does not appear to have any data entry errors.

# Weighting the Data

**Introduction**

If the data from your STEPS survey is analysed unweighted, the results are only representative of the sampled participants. In order to have results that are representative of your entire target population, your data must be weighted.

**What is a weight**

A weight is a value given to a data record to adjust the importance given to it in analysis. It may be thought of as the number of persons in the population that are represented by each individual in the sample. Weights are calculated to adjust for the following aspects of a survey:

- probability of selection (sample weight);
- non-response (non-response weight);
- differences between the sample population and target population (population weight).

**Sample weight**

The sample weight is comprised of the inverse of the probability of selection. For multi-stage sampling designs, this means calculating the probability of selection at every stage of selection and multiplying them all together. It requires knowledge of the probability of selection at all stages of sampling and is therefore the most difficult weight to calculate due to the amount of information needed.

While there are some tools available to help with the calculation of these weights, it is not possible to automate the process entirely due to differences in sample design between STEPS surveys.

If you used the STEPSsampling.xls file to draw your sample, you can use it to partially calculate the sample weights for your dataset. The worksheet "Info for Weighting" within the STEPSsampling.xls file contains directions for calculating the probability of selection up to the household or individual level (if individuals were selected directly).

If you used another means to draw your sample, it is recommended to create a summary table containing the probability of selection for each sampling unit in your sample. Contact the WHO Geneva STEPS team for help in developing a summary table.

For the probability of selection within the household (the final stage of sampling in most STEPS survey), the STEPS app automatically includes household size in the dataset.

**Non-response weight**

The non-response weight is calculated by taking the inverse of the response rate either for the overall survey or, more often, for each subset of the survey.

*Continued on next page*

**Non-response weight (cont.)**

Non-response weighting is typically done for age and sex, though it can also be done for any other variables, such as location. Whatever variable is used, it must fulfil the following requirements:

- the variable should be known to be somehow related to the risk factors (for example, hair colour likely has little to do with whether or not people eat enough fruits and vegetables);
- the variable must be known for BOTH responders and non-responders (for example, years of education would probably not be available from non-responders).

In surveys like STEPS, in which the age and sex is not known until the participant is selected during a household visit, it is often difficult to have complete age and sex information for all non-responders. Therefore a non-response rate correcting for varying response rates by age-sex group often cannot be done.

However, it may be of interest in many STEPS surveys to assess response rate by location (e.g. urban vs rural areas or high vs low socioeconomic areas). Risk factors may be expected to vary by location and the location of both responders and non-responders should be known – thus location fulfils the two criteria listed above. To calculate a non-response weight for location, the response rate for each location should be calculated and the inverse of this figure would be applied as the non-response weight for all records from the location.

**Population weight**

The population weight allows for the correction of over- or under-representation in the sample of the targeted age-sex groups.

In order to calculate the population weight, you can first count the total number of participants in each of the age-sex groups covered by your survey. Use the sample weights to attain weighted counts for each age-sex group. You will then need to use recent Census data or similar to get these same counts for your underlying target population. Use this information to create a table like the one below, in which the population weight is shown in the last column. The columns labelled A and B show the proportion of each age-sex group in the target population or sample. These are calculated by taking the number of individuals in that age-sex group and dividing it by the total number of individuals. The population weight is derived from the ratio of these two proportions.

*Continued on next page*

**Population weight (cont.)**

| *Example table to calculate population weight* | Target population | Proportion of target population (A) | Sample (sum of sample weights) | Proportion of sample (B) | Population Weight = A/B |
|---|---|---|---|---|---|
| Males, 18-29 | 2000 | 0.13 | 1181 | 0.08 | 1.78 |
| Males, 30-44 | 1760 | 0.12 | 2214 | 0.15 | 0.78 |
| Males, 45-59 | 1440 | 0.10 | 1919 | 0.13 | 0.77 |
| Males, 60-69 | 1600 | 0.11 | 2214 | 0.15 | 0.71 |
| Females, 18-29 | 2000 | 0.13 | 1476 | 0.10 | 1.33 |
| Females, 30-44 | 1200 | 0.08 | 1919 | 0.13 | 0.64 |
| Females, 45-59 | 3000 | 0.20 | 2214 | 0.15 | 1.33 |
| Females, 60-69 | 2000 | 0.13 | 1919 | 0.13 | 1.07 |
| Total | 15000 | | 15056 | | |

**Overall weight**

Once the sample, non-response (if needed), and population weights have been calculated and attached to your dataset, you will need to multiply these together to arrive at the overall weight for each Step of your survey. It is possible that non-response weight (if used) and population weights will vary slightly from each Step due to different response rates. It is thus recommended to calculate an overall weight for each Step of your survey. The Epi Info analysis programs provided by the WHO Geneva STEPS team (see Part 4, Section 2) have been designed so that there are different analysis weights for analyses of variables from each Step of the survey. These overall weights are named accordingly:

- WStep1
- WStep2
- WStep3.

Even if there is no difference in the overall weight for Step 1 versus Step 2, for example, you must create one analysis weight per Step in order to use the provided analysis programs.

*Continued on next page*

**Stratum and PSU**

If your sample design was anything other than a simple random sample, you will need to create variables that contain information about your sample design. These variables are conventionally named Stratum and PSU and their values depend on the sample design of your survey.

PSU typically contains the identifiers of the sampling units above the household level (e.g. villages, census blocks, or enumeration areas). PSU can usually be generated by copying the information from your Cluster ID variable.

Stratum allows you to identify a higher level of clustering in your sample design, such as province, region, or urban/rural. Using Stratum is optional. If you do not need it, simply create the Stratum variable and set it equal to 1 for all records.