

4 May 2016

The raw data sets are located in 3 folders called “Round 1”, “Round 2”, and “Round 3” corresponding to the 3 waves of the survey. This file is in each of those 3 folders. Below the following 3 topics are explained:

1. File naming
 2. Edits performed on raw data
 3. Removal of identifying variables
 4. Questionnaires
 5. Modules
-
1. File naming: files retained their original names except:
 - a. A prefix of “r1_” was added to round 1 data files
 - b. A prefix of “r2_” was added to round 2 data files
 - c. A prefix of “r3_” was added to round 3 data files
 - d. Suffixes were added to the following files to differentiate duplicated file names: “CMC.dta”, “COMMENTS.dta”, “TRANSLATIONS.dta”, “FoodItem.dta”, “FoodUnit.dta”, “HHMember.dta”. Duplicated file names (within a survey round) occurred because the raw data were located in different subfolders. Adding a unique suffix became necessary once all data files were moved to a single folder (per round).
 2. Edits to raw data
 - a. In round 1 data, the village identifying variable “clid” was renamed “CLID” for consistency with the other 2 rounds.
 - b. In the file “r2_TrackingForm” 5 corrupted observations were removed
 3. Removal of identifying variables: potentially identifying variables (villages names, individuals’ names, village facilities’ names, and comments that included names) were removed in two ways:
 - a. Removal: identifying variables were simply removed. The identifying variable was recoded to “confidential: removed from data” A unique identifier (“anonymous_data_ID”) was added to each data set that had any variable recoded without replacement. This unique identifier is used to restore the original values of the identifying variables. The added variable “anonymous_data_ID” has been constructed to be unique in every data set.
 - b. Removal with replacement: to preserve anonymity, villages were given an alternative identifying code so that merging across data sets is possible. The identifying variable was recoded to “confidential: replaced by X_anonymous” (where “X” is the identifying variable). The variable “X_anonymous” follows, immediately, after the recoded variable.
 4. Questionnaires
 - a. Round-specific questionnaires are included in each folder. In round 2 the questionnaire was in an Excel format. In some cases the Swahili version was also available.
 5. Modules
 - a. Each questionnaire consists of the modules listed below. Each module has a corresponding dataset or set of datasets. These dataset names are identified in the questionnaires. In the baseline (round 1) survey, the dataset names are listed in grey text underneath the question number. In the midline (round 2) survey, they are listed in the column titled “Table.” In the endline (round 3) survey, they are listed in bright blue text to the right of the question number.