How to Use Mongolia MICS Plus 2020-2021 Datasets

MICS Plus is a longitudinal household survey that collects information from a representative sample of households through interviews on direct phone calls. The same households are interviewed multiple times over an extended period of one year in regular intervals. Depending on the frequency of calls, up to 12 waves of calls may be made to each sampled household.

Details of the Mongolia MICS Plus 2020-2021 survey methodology are provided in a separate document. This note summarizes some of the methodological features with respect to the use of micro data sets and accompanies the publicly shared datasets.

Mongolia MICS Plus 2020-2021 datasets are being periodically released for public use after every three waves.

1

Survey Implementation

The Population and Household Registration Database (PHRD) of Mongolia, managed and updated by the National Statistics Office of Mongolia, was used as the sample frame of the Mongolia MICS Plus 2020-2021 survey. The PHRD was last updated in January 2020. The sample size for the Mongolia MICS Plus 2020-2021 was determined as 2,200 households at the outset, with the hope that around 2,000 households would be captured with completed interviews.

The PHRD included one or more phone numbers for each household. However, to establish the effective sample for the survey through validation and substitution, as described below, the first wave of calls was completed in 4 stages:

Stage 1

2,200 households were randomly selected from the PHRD. Calls were placed to these households, using the phone numbers available in the PHRD.

Stage 2

421 out of 2,200 households (19 percent) could not be reached during the first stage[1]. For these households, local governors were contacted to validate the information available from the 2020 PHRD, and to recover new phone numbers that could be used. For 206 households, the validation process provided either one or more phone numbers for households currently living at the sampled address. Call attempts were made to those 206 households in the 2nd stage.

[1] Households could not be reached for the following reasons: no phone number available for the household, invalid phone number, household could not be reached after repeated calls, phone number does not belong to sampled household.

Stage 3

At the end of the stage 2, there were 346 households with incomplete interviews. These households were substituted with other households from the 2020 PHRD by applying a model-based, conditional substitution method that employs Euclidian Distance analysis (available under Nearest Neighbor Analysis in SPSS). 344 of these households were successfully substituted and calls were made to the substitute households at this stage. For 2 households, it was not possible to identify households that were similar enough in terms of their selected background characteristics and these household were therefore not substituted. These household remained as non-responding households in the datasets.

Stage 4

At the end of Stage 3, 156 of the 344 households could not be reached, for whom the validation process with local governors was repeated. The validation process provided additional/revised information for 123 of the 156 households. Calls were placed to the 123 households, using the validated information. By the end of the 4th Stage, interviews were successfully completed in 1,987 households.

After the completion of Wave 1, the final effective sample size for the subsequent waves was established at 2,154 households (refer to Section 7, Number of Cases in Datasets below for a visual description of the four stages and for other details). The above-mentioned validation and substitution processes apply only to the first wave of calls and are not repeated in the subsequent waves.

2

Typology of Questions

MICS Plus interviews are administered to one respondent. Defined as a knowledgeable adult household member, the respondent answers all questions regardless of whether the question is about himself/herself, the household, the dwelling, or a specific household member. While it is common to have the same respondent across waves, note that different respondents may be interviewed across different waves.

One questionnaire is administered to the selected households of MICS Plus surveys. Questions are organized in modules. There are 3 main types of modules, resulting in different units of analysis.

- The module "Call Attempts Panel" captures information related to the calls made to the selected households. A household has as many records in the Call Attempts dataset as the number of call attempts made to that household, regardless of whether the call resulted in a completed interview or not.
- The "List of Household Members" captures key characteristics of each household member in the interviewed households.
- Each wave of questionnaires includes several "Sectoral" modules that capture information on various topics of interest. While the Call Attempts Panel and the List of Household Members are included in all waves, the Sectoral Modules may change from one wave to another. Examples of such modules include "COVID-19", "Education", and "Nutrition".

Sectoral Modules can have different groups of individuals of interest. For example, the "Education" module targets children age 2-17 years. In the case of this module, one child age 2-17 years is randomly selected from the List of Household Members; questions in this module target the selected child only. Random selection and subsequent application of sample weights (see sub-Section 5 "Sample weight variables") ensures that the data collected is representative of the child population age 2-17 years in Mongolia.

In general, a question can be about the household or dwelling, household members, the respondent, or a randomly selected individual as described above. The table below shows the possible typology of questions in MICS Plus surveys.

Type of question	Example	Representative of
About household	**CH4**. Does your household have access to internet at home?	Households in Mongolia
About household members	**CVH9.** Have you or any member of your household been tested for CORONAVIRUS?	Household population of Mongolia
About respondent	**CV2.** Since the last (***day of the week***), have you been able to keep distance from people when in public places? Would you say: always, very often, sometimes, rarely or never?	Knowledgeable adult household members who were interviewed. Since respondents are not selected randomly, such questions do not yield representative data on Mongolia's adult population's characteristics, opinions, and behavioural patterns.
About a randomly selected individual	**ED4.** Since the last (***day of the week***), did (***name***[2]) watch any TV lessons?	Population age 2-17 years in Mongolia

3

Datasets

Due to the different units of analysis, MICS Plus data have a hierarchical structure. The Mongolia MICS Plus digital data collection system was developed using CSPro, a software programme that well handles hierarchical datasets. For analysis, CSPro datasets are exported to the following three SPSS datasets:

HH - Household dataset

Unit of analysis: Households

Includes: Information on household characteristics, such as the type of accommodation, energy use, water and sanitation, as well as sectoral modules such as the COVID-19, Distance Learning/Education, and Nutrition. Variables corresponding to questions on the household and dwelling, as well as those sectoral modules that target randomly selected individuals are included in this dataset.

[2] Name of the randomly selected individual. In this example, name of the selected child age 2-17 years.

HL - Household members dataset

Unit of analysis: Household members

Includes: Characteristics of individual household members such as date of birth, age, sex, relationship to household head.

CA - Call attempts dataset

Unit of analysis: Call attempts

Includes: Information on all calls made to households, regardless of the interview result, such as date and time of call, call outcome, and consent for interview.

4

Dataset Naming Conventions

MICS Plus datasets in SPSS format are distributed in a compressed WINZIP folder and are uniquely named using the following naming conventions:

[CCC] 3-digit Standard ISO Country Code
[DD] Dataset type: HH (Household), HL (Household members), CA (Call attempts)
[W##] Wave number (W01 – first wave, W02 – second wave…)

For example, SPSS datasets of the first wave are stored in the WINZIP file named "MNGMICSPLUSW01.ZIP" and are named as MNGHHW01.SAV, MNGHLW01.SAV, and MNGCAW01.SAV. In addition to the SPSS datasets, the WINZIP files also include a technical note with contact details.

CSPro datasets are not shared publicly.

5

Variable Naming Conventions, Variable Construction

Contents of datasets correspond to questionnaire contents for each wave.

The most common correspondence is the "one question-to-one variable" form. For example, a variable named "EU1" in the dataset represents the question numbered "EU1" in the questionnaire. Other specific types of variables are explained below.

Identification (ID) variables

Each dataset has case-identifiers that uniquely identify each unit of analysis (household, household member, or call attempt) and allow merging of datasets when the relationship is logically possible. Details on merging datasets can be found below (Section 9 "Merging Datasets").

ID variable in the Household dataset - HHID, which is created by combining the Wave Number (HH0), Stratum (HH1), Household Number (HH2), and Wave 1 Stage Number[3] (HH0A), as shown below:

HHID	HH0	HH1	HH2	HH0A
110121	1	1	12	1
110141	1	1	14	1
110143	1	1	14	3

ID variables in the Household Members dataset – HHID and HHMEMID, which are combinations of Wave Number (HH0), Stratum (HH1), Household Number (HH2), Wave 1 Stage Number[3] (HH0A), and Line Number of Household Member (HL1)

HHID	**HHMEMID**	HH0	HH1	HH2	HH0A	HL1
290030	29003001	2	9	3	0	1
290030	29003002	2	9	3	0	2
290030	29003003	2	9	3	0	3

ID variables in the Call Attempts dataset – HHID and CAID, which are combinations of Wave Number (HH0), Stratum (HH1), Household Number (HH2), Wave 1 Stage Number[3] (HH0A), and Call Attempt Number (CA1).

HHID	**CAID**	HH0	HH1	HH2	HH0A	CA1
382990	38299001	3	8	299	0	1
382990	38299002	3	8	299	0	2
382990	38299003	3	8	299	0	3

Variables Specific to Wave 1

Variable "SampleType" (Whether household was initially selected or a substitute household) has been created to indicate whether a household was initially selected (1) or was a substitute household (2).

[3] Only the Wave 1 household dataset has up to 4 stages due to the specific implementation approach of this wave, as explained in Section 1 "Survey Implementation". For other waves, the variable "HH0A" (Wave 1 Stage Number) is set to "0 – Not applicable".

When analyzing Wave 1 datasets, one needs to select only the relevant records, depending on whether the analysis is intended to be based on households before or after substitution. For this purpose, variables "aftersub" (For analysis – household after substitution) and "beforesub" (For analysis – household before substitution) have been constructed, with values of "0 – No" and "1 - Yes". Therefore, one must select only households with value "1" on the variable "aftersub", if analysis will be based on households after substitution – these are the households that the published survey results are based on. For analysis based on households before substitution, households with value "1" on the variable "beforesub" should be selected.

There is a handful of variables that have the letter "B" in their names and the text "before substitution" in their labels. These are the variables that need to be used for analyses based on households before substitution. For example, variable "HH17" is the result of interview for households to be included in analyses after substitution (the final survey sample), while variable "HH17B" is the result of interview for those households to be included in analyses before substitution. The same applies to variables "hhweight" and "hhweightB" – sample weight variables (see sub-Section 5 "Sample weight variables").

The following is an example of SPSS code that selects households for 1) analysis after substitution and 2) analysis before substitution and applies sample weights.

```
* 1) Example of selecting households for analysis after substitution and applying sample
weight.

* open the household dataset.
get file = "MNGHHW01.sav".

* select only the households for analysis after substitution.
select if (aftersub = 1).

* select only the interviewed households.
select if (HH17 = 1).

* applying sample weight.
weight by hhweight.
```

```
* 2) Example of selecting households for analysis before substitution and applying sample
weight.

* open the household dataset.
get file = "MNGHHW01.sav".

* select only the households for analysis before substitution.
select if (beforesub = 1).

* select only the interviewed households.
select if (HH17B = 1).

* applying sample weight.
weight by hhweightB.
```

Multiple Response Questions

Multiple response questions are those questions that allow the coding of multiple answers to a single question. For such questions, response categories are alphabetical. When a multiple response question is asked, the interviewer does not read the response categories; the respondent provides answers that fit one or more response categories. The interviewer then records the most appropriate response code and probes until the respondent has no more responses. For multiple response questions, each response category has a designated string variable in the relevant dataset. These variables are named ending with the letter that corresponds to the response category.

For example, variables "CH5A", "CH5B", …., "CH5X" represent, respectively, response categories "A", "B", …, "X" of the question "CH5" below. Possible values for each of these variables in the datasets are the letters corresponding to the response category (e.g., "A" for the variable "CH5A") and a blank space (when that response category is not selected). In addition to these variables, a variable with "NR" in the name (e.g., CH5NR) and with the value "?" is created to code cases when there is no response to such questions.

CH5. What equipment do the members of your household use to access internet? *Probe:* Anything else? *Multiple responses are allowed.* *Do not read out the response categories.*	DESKTOP COMPUTER ..... A LAPTOP ..... B TABLET ..... C SMART PHONE ..... D SMART TV ..... E OTHER (*specify*) X

For a household where the respondent has responded that household members use desktop computers, tablets, and smartphones to access the internet, the following values would apply for the corresponding variables:

CH5A	A
CH5B	[blank]
CH5C	C
CH5D	D
CH5E	[blank]
CH5X	[blank]

Multipart Questions

Multipart questions are questions that contain two or more sub-questions, grouped together as items, with a leading question. In such cases, the response categories are the same for all sub-questions, which are usually items in relation to the leading question. Response categories can be numeric or alphabetical. For such questions, each sub-question has a designated numeric/string variable in the relevant dataset and those variables are named ending with the letter that corresponds to the relevant sub-question.

CH6. Does your household have :		YES	NO
[A] A fixed telephone line?	FIXED TELEPHONE LINE	1	2
[B] A radio?	RADIO	1	2
[C] A couch?	COUCH	1	2
[D] A wardrobe?	WARDROBE	1	2
[E] A metal bed?	METAL BED	1	2
[F] A double size bed?	DOUBLE SIZE BED	1	2

For example, variables "CH6A", "CH6B", …., "CH6F" represent, respectively, sub-questions "A", "B", …, "F" of the question "CH6" above. Each of these variables have values of "1 - Yes" or "2 - No".

Recoded variables

In addition to variables that correspond to the questions in the questionnaires, each dataset also has a number of recoded or computed variables that are necessary for analysis. Names of such variables are as self-explanatory as possible. For example, the variable "area" represents "Area of residence"; variable "headage" stands for "Age of household head"; and the variable "windex5" is for "Wealth index quintiles". Most of these recoded/computed variables are used as background characteristics in the tabulations. More details on the construction of these variables can be found in Section 8 "Background characteristics".

Sample weight variables

MICS Plus survey samples are not self-weighting; therefore, datasets also contain sample weight variables, which need to be used while carrying out analysis (unless unweighted analysis is needed for a specific purpose). Separate sample weights are calculated for households and randomly selected individuals (see Section 2 "Typology of Questions"). The table below lists the sample weight variables for each wave given that different waves have different groups of individuals selected randomly.

Variable name	Variable description	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
hhweight	Sample weights of households	x	x	x				
ch217weight	Sample weights of randomly selected children age 2-17 years	x					x	
ch017weight	Sample weights of randomly selected children age 0-17 years		x					
ch02weight	Sample weights of randomly selected children age 0-2 years		x					
adultweight	Sample weights of randomly selected household members age 15 years or above			x				
ch117weight	Sample weights of randomly selected children age 1-17 years			x				
ch04weight	Sample weights of randomly selected children age 0-4 years				x			
ch1019weight	Sample weights of randomly selected adolescents age 10-19 years					x		
ch519weight	Sample weights of randomly selected children age 5-19 years							x

Sample weights of households for each wave are calculated by adjusting the basic sample weights (inverse of selection probabilities) that are calculated for each stratum[4] (HH1) by the non-response for that stratum in the wave. Sample weights of individuals are calculated by multiplying the sample weights of households by the number of individuals of interest. For example, ch217weight = hhweight * Number of children age 2-17 years in the household, ensuring that the data obtained from randomly selected individuals are expanded to the survey population in the selected group of interest.

6

Special Values

Various codes are used to describe special values that apply to a large set of variables, as a matter of convention.

"Not applicable" and "Missing" values

A "Not applicable" value occurs when a question is not supposed to be asked or should be skipped according to flow of the questionnaire, On the other hand, a "Missing" value for a question applies when the question is supposed to be asked, but either the respondent has refused to provide an answer, or the question was not asked due to a technical error. "Not applicable" values are treated as system-missing in SPSS datasets while "Missing" values are coded with a 9, 99, 999, or 9999 depending on the field length of the variable (or with question marks for a string variable) and are treated as user-missing values (different statistical software may handle the "not applicable" and "missing" values differently). It is important to note that the "Not applicable" and "Missing" values can be included or excluded in analyses depending on the indicator of interest. Hence, one needs to pay careful attention to the selection of the denominator and the treatment of the "Missing" values for matching the results of MICS Plus surveys.

Other special values

In addition to the "Not applicable" and "Missing" values, there are often other special values that are usually pre-coded in the questionnaire with the following conventions:

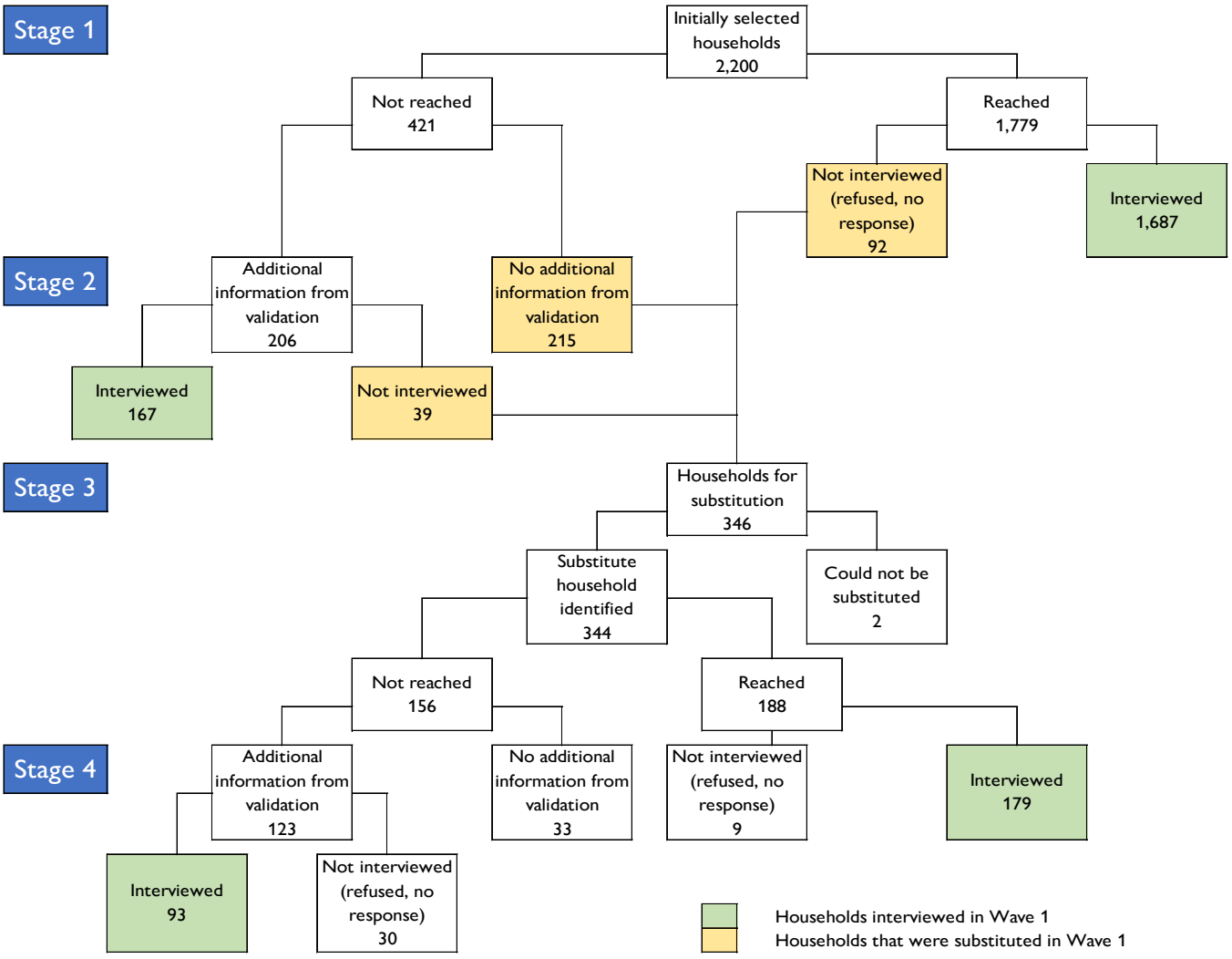
	Numeric variable	String variable
Other	6, 96, 996, 9996, etc.	"X"
Don't know	8, 98, 998, 9998, etc.	"Z"

[4] In total, 9 strata (HH1) are formed by the urban/rural (HH7A) areas of four regions, plus Ulaanbaatar (HH3)

7

Numbers of Cases in Datasets

As explained in Section 1 "Survey Implementation", Wave 1 was implemented in 4 stages. Each record in the Wave 1 household dataset (HH) represents a single stage for a specific household. As illustrated in the diagram below, there were 2,200 initially selected households (Stage 1); of which 206 underwent additional validation process (Stage 2), 344 were substituted (Stage 3) and 123 were subjected to the final validation (Stage 4). As a result, there are a total of 2,873 cases in the Wave 1 household dataset (HH). Similarly, out of 2,200 initially selected households 1,687 were successfully interviewed in the Stage 1, 167 in the Stage 2, 179 in the Stage 3, and 93 in the Stage 4, yielding a total number of 1,854 interviewed households before substitution, and 2,126 interviewed households after substitution.



In the datasets, each household is assigned two result codes, one before substitution, and one after substitution, irrespective of the number of the implemented stages. Variables "HH17" and "HH17B" represent the interview result before and after substitution, respectively. They store the final result code of the last carried out stage for each household, separately before and after substitution.

The following table presents the frequency distribution of households by the result of interview after substitution (HH17) at the end of Wave 1.

Total	**2,200**
Interviewed	2,126
Refused	17
No eligible respondent	2
Phone number(s) does not belong to sampled household	8
Phone number(s) inactive	1
Respondent busy / postponed	3
No response after repeated call attempts or phone(s) turned off	22
No phone number available for sampled household	6
Address vacant	15

Upon completion of Wave 1, the final effective sample size for the consecutive waves was established at 2,154 households by deciding to exclude households with the following interview result codes: "Refused", "Phone number(s) does not belong to sampled household", "No phone number available for sampled household", and "Address vacant". Therefore, there are a total of 2,154 cases in all household datasets starting from Wave 2.

The call attempts datasets (CA) contain one record for every call attempt made for each selected household with an available phone number.

The household members datasets (HL) have one record for each household member listed in the interviewed households. The same line number of each household member is kept across different waves. These datasets include information on household members who are currently living in the household, as well as those who are reported as migrated or deceased during a particular wave. Any analysis pertaining to the current household composition (e.g., estimating number of household members) should only include household members who are reported as currently living in the household.

8

Background characteristics

Results of the Mongolia MICS Plus 2020-2021 survey are presented in the form of tabulations that include disaggregation of indicators by background characteristics. Variables for these background characteristics are created by recoding other existing variables. Below are the most common background characteristics.

Area

Variable name: area (Area)

Relevant waves: all

This variable indicates whether the household's address at the time of the interview was from an urban or rural settlement. In Mongolia, each of the four regions (Western, Khangai, Central, and Eastern) have "Aimag centers", "Soum centers", and "Rural areas". An "Aimag center" is classified as an "Urban" settlement while "Soum centers" and "Rural areas" are classified as "Rural" areas. The capital city "Ulaanbaatar" is classified as urban.

Given the possibility that households can change their addresses between waves, in each wave, the interviewed households are asked whether they are still living at the same address as they were living during the previous wave (variable "CH0B"). If not, the new address is recorded. This variable refers to area of residence of the households at the time of interview of a particular wave, while the variable "HH7A" is the area of residence of the household at the time of the sample selection (before Wave 1).

Region

Variable name: region (Region)

Relevant waves: all

Similar to the variable "area", this variable refers to the region of the household during a particular wave, while the variable "HH3" is the region of the household at the time of sample selection (before wave 1).

Sex and Age of Household Head

Variable name: headsex (Sex of household head), headage (Age of household head)

Relevant waves: all

In Wave 1, demographic information on all household members (date of birth, age, sex, relationship to household head) were collected. In all subsequent waves, the respondent is asked whether there has been a change in the composition of the household (variable "HL0"). If there has been changes, a question is asked to ascertain whether the household head recorded in the previous wave is still the household head (variable "HL8"), and if so, who the household head is at the time of interview of that particular wave.

Sex and Age of Respondent

Variable name: respsex (Sex of respondent), respage (Age of respondent)

Relevant waves: all

MICS Plus surveys aim to have a knowledgeable adult household member living in the household as the respondent the questionnaire. While it is common to have the same respondent across waves, note that different respondents may be interviewed across different waves. This means that the variable respsex may not be the same in different waves.

Wealth Index Quantile

Variable name: windex5 (Wealth index quintile)

Relevant waves: Wave 2 and consecutive waves

The wealth index is a composite indicator of wealth. Starting from Wave 2, questionnaires have included questions on ownership of consumer goods (variables "CH6A" to "CH8J"), energy use (variables "EU1" to "EU12") and water and sanitation (variables "WS1" to "WS5"). Based on these and a few more variables (accommodation type, persons per room, access to internet, etc.) that are thought to be related to household wealth, the wealth index is constructed[5].

Note that the questions for constructing the wealth index were asked to the respondents only during Wave 2. In the subsequent waves, these questions are administered only to 1) households who have changed either their address of living and/or accommodation type between the waves and 2) households who are interviewed for the first time in that particular wave. Wealth index construction is repeated for each wave.

[5] The wealth index is a composite indicator of wealth. To construct the wealth index, principal components analysis is performed by using information on the ownership of consumer goods, dwelling characteristics, water and sanitation, and other characteristics that are thought to relate to the household's wealth, to generate weights (factor scores) for each of the items used. First, initial factor scores are calculated for the total sample. Then, separate factor scores are calculated for households in urban and rural areas. Finally, the urban and rural factor scores are regressed on the initial factor scores to obtain the combined, final factor scores for the total sample. This is carried out to address the urban bias in the wealth index values. Each household in the total sample is then assigned a wealth score based on the assets owned by that household and on the final factor scores obtained as described above. The survey household population is then ranked according to the wealth score of the household they are living in and is finally divided into 5 equal parts (quintiles) from lowest (poorest) to highest (richest). The wealth index is assumed to capture the underlying long-term wealth through information on the household assets and is intended to produce a ranking of households by wealth, from poorest to richest. It does not provide information on absolute poverty, current income, consumption, or expenditure levels. The wealth scores calculated are only applicable for the particular data set they are based on.

9

Merging Datasets

For analysis purposes, it is possible to combine two or more MICS Plus datasets from different waves, if needed. It is also possible to match different types of datasets within a specific wave. This is only necessary when variables required for the analysis are not present in one file but are present in another. It should be noted that care has been taken to add the number of variables that are considered important for the analysis from one dataset to another. For example, variables on household and sample characteristics from the Household dataset (HH) are already included in the Household members dataset (HL). Nonetheless, there are occasions when data users have to merge different datasets to obtain the variables they need for particular analysis. This section provides more details and examples on how to accomplish that task.

When merging datasets, the correct use of ID variables and identification of key variables are critical (refer to Section 5 "Variable Naming Conventions, Variable Construction" for more details on identification variables). Key variables are common variables between all source datasets, which link the observations of one data file to those of the other. The key variables must have the same names in all data files that are being merged. If names are not the unique, renaming of key variables in one or more datasets is required.

Another important step when merging MICS Plus datasets, is to determine the type of relationship between two files, as well as to define desired unit of analysis. For example, a relationship between households and household members is such that one entity (household) relates to several others (members of the household). There may be one or more household members for each household. This is a "one to many" relationship. On the other hand, in a "one to one" relationship, one dataset's entity is associated with one and only one entity in another dataset. For example, the household dataset from Wave 2 and the household dataset from Wave 3 have a "one to one" relationship.

Merging Datasets from Different Waves

There are two ways of merging datasets from different waves. One is to combine files containing the same variables but different cases. This is particularly useful when analyzing the same questions included in the questionnaires from different waves. With this type of merging all cases from different datasets are concatenated, and in the resulting dataset all cases from one file are added to the end of all cases from another file.

The following example of SPSS code concatenates cases from Wave 2 and Wave 3 "HH" datasets. The resulting file contains all cases from the Wave 2 "HH" dataset and all cases from Wave 3 "HH" dataset.

```
* Example of combining Wave 2 and Wave 3 variables on internet access from household
datasets.

* open the Wave 2 household dataset.
get file = "MNGHHW02.sav".

* sort data by ID variables.
sort cases by HHID.

* save working dataset and keep only the variables of interest; Household ID (HHID), wave
number (HH0), result of interview (HH17), access to internet (CH4A, CH4B, CH4C, CH4Y,
CH4NR), and sample weight variable (hhweight).
save outfile = "tmphhW2.sav"
    /keep HHID HH0 HH17 CH4A CH4B CH4C CH4Y CH4NR hhweight.

* repeat above steps for the Wave 3 household dataset.
get file = "MNGHHW03.sav".
sort cases by HHID.
save outfile = "tmphhW3.sav"
    /keep HHID HH0 HH17 CH4A CH4B CH4C CH4Y CH4NR hhweight.

* open the working Wave 2 dataset.
get file = "tmphhW2.sav".

* combine files and note that it is not necessary to identify the ID variable.
add files
    /file = *
    /table = "tmphhW3.sav".

* save the combined dataset that now has cases from both waves.
save outfile = "tmphhW2and3.sav".

* erase temporarily created files.
erase file = "tmphhW2".sav
erase file = "tmphhW3".sav
```

For certain analytical purposes, datasets from different waves can be merged by combining files that contain the same cases but different variables. For example, one might want to merge information on "Method of protection against COVID-19" (variables "CV1A" to "CV1NR") from Wave 1 household "HH" dataset to Wave 2 and Wave 3 household "HH" datasets. With this type of merging, key variables between datasets must be identified and used to match observations between them. For example, Stratum Number (HH1) and Household Number (HH2) are key variables that indicate which case from one household data file corresponds to which case from another. Similarly, Stratum Number (HH1), Household Number (HH2) and Line Number of Household Member (HL1) are key variables that match cases between household members (HL) datasets from different waves. Furthermore, the relationship between respective datasets from different waves is "one-to-one". This means that the new, merged dataset contains added variables of interest and has the same number of cases as the original dataset.

The SPSS syntax below demonstrates how to merge data from Wave 1 Household "HH" dataset onto the Wave 2 Household "HH dataset.

```
* Example of merging information on “Method of protection against COVID-19” from Wave 1
household dataset onto Wave 2 household datasets.

* open the Wave 1 household dataset.
get file = "MNGHHW01.sav".

* select only households after substitution, as they are present in consecutive waves
select if (aftersub = 1).

* sort data by key variables.
sort cases by HH1 HH2.

* save working dataset and keep only the variables of interest; the key variables (HH1,
HH2), methods of protection (CV1A to CV1NR).
save outfile = "tmphhW1.sav"
  /keep HH1 HH2 CV1A CV1B CV1C CV1D CV1E CV1F CV1G CV1H CV1I
        CV1J CV1K CV1L CV1M CV1X CV1Z CV1NR.

* open the Wave 2 household dataset.
get file = "MNGHHW02.sav".

* sort data by key variables.
sort cases by HH1 HH2.

* merge files and note that it is critical to identify the key variables.
match files
    /file = *
    /table = " tmphhW1.sav"
    /by HH1 HH2.

* save updated Wave 2 dataset with added information on methods of protection against
COVID-19.
save outfile = "tmphhW2.sav".

* erase temporarily created files.
erase file = "tmphhW1.sav".
```

Merging Datasets for the Same Wave

As described above, there will be instances when different types of datasets will have to be merged to obtain the variables that meet analysis needs. In the MICS Plus context that would usually relate to adding household level information from the Household “HH” dataset onto the Household Members “HL” dataset, or to the Call Attempts “CA” dataset. It is also possible to merge aggregated information from one of the datasets onto another one. For example, information on the number of household members per household can be added from the Household Members “HL” dataset onto the Household “HH” dataset, first by aggregating information on the number of household members for each household from the “HL” dataset, and second by adding that information onto the “HH” dataset.

When merging household members’ and call attempts datasets with their households, one needs to use Household ID (HHID) as a key variable. Since there is a “one-to-many” relationship between households and household members, as well as between households and call attempts, it is possible to merge the “HH” dataset onto “HL” or “CA” datasets, but not the other way around.

The SPSS code that merges Wave 2 "HH" dataset onto the Wave 2 "HL" dataset is provided in the following example.

```
* Example on how to merge variable "Main source of drinking water" (WS1) from the household
dataset onto the household members dataset.

* open the household dataset.
get file = "MNGHHW02.sav".

* select only the interviewed households.
select if (HH17 = 1).

* sort according to the ID variables.
sort cases by HHID.

* save the working dataset temporarily by keeping only the variables of interest.
save outfile = "tmphh.sav"
    /keep HHID WS1.

* open the household members dataset.
get file = "MNGHLW02.sav".

* sort according to the ID variables.
sort cases by HHID HHMEMID.

* perform the matching by the ID variable that is common to both (tmphh and HL) datasets.
match files
    /file = *
    /table = "tmphh.sav"
    /by HHID.

* save the household members dataset that now has the variable WS1 added.
save outfile = "MNGHLW02.sav".

* erase temporarily created files.
erase file = "tmphh.sav".
```