Global Labor Database Harmonization Guidelines

Jobs Group

Social Protection and Jobs

The World Bank

Contents

[1. Introduction 4](#_Toc71552006)

[1.1. Global Labor Database (GLD) Project 4](#_Toc71552007)

[1.2. File Organization 4](#_Toc71552008)

[1.3. Do-file Organization and Guidelines 5](#_Toc71552009)

[2. Survey & ID module 9](#_Toc71552010)

[2.1. ID Guidelines 9](#_Toc71552011)

[2.2. Mapping and Description of Variables 9](#_Toc71552012)

[2.3. Lessons Learned and Challenges 10](#_Toc71552013)

[2.4. Tabular Overview of Variables 12](#_Toc71552014)

[3. Geography 13](#_Toc71552015)

[3.1. Mapping and Description of Variables 13](#_Toc71552016)

[3.2. Lessons Learned and Challenges 15](#_Toc71552017)

[3.3. Tabular Overview of Variables 16](#_Toc71552018)

[4. Demography 17](#_Toc71552019)

[4.1. Mapping and Description of Variables 17](#_Toc71552020)

[4.2. Lessons Learned and Challenges 19](#_Toc71552021)

[4.3. Tabular Overview of Variables 21](#_Toc71552022)

[5. Migration 22](#_Toc71552023)

[5.1. Mapping and Description of Variables 22](#_Toc71552024)

[5.2. Lessons Learned and Challenges 22](#_Toc71552025)

[5.3. Tabular Overview of Variables 23](#_Toc71552026)

[6. Education 24](#_Toc71552027)

[6.1. Mapping and Description of Variables 24](#_Toc71552028)

[6.2. Lessons Learned and Challenges 24](#_Toc71552029)

[6.3. Tabular Overview of Variables 25](#_Toc71552030)

[7. Training 26](#_Toc71552031)

[7.1. Mapping and Description of Variables 26](#_Toc71552032)

[7.2. Lessons Learned and Challenges 26](#_Toc71552033)

[7.3. Tabular Overview of Variables 27](#_Toc71552034)

[8. Labour 28](#_Toc71552035)

[8.1. ILCS Definitions 28](#_Toc71552036)

[8.2. Mapping and Description of Variables 28](#_Toc71552037)

[8.2.1. Labour status, 7-day reference period 28](#_Toc71552038)

[8.2.2. Primary Employment, 7-day reference period 29](#_Toc71552039)

[8.2.3. Secondary Employment, 7-day reference period 36](#_Toc71552040)

[8.2.4. Other Employment, 7-day reference period 41](#_Toc71552041)

[8.2.5. Total Employment Earnings, 7-day reference period 42](#_Toc71552042)

[8.2.6. Labor status, 12-month reference period 42](#_Toc71552043)

[8.2.7. Primary Employment, 12-month reference period 44](#_Toc71552044)

[8.2.8. Secondary Employment, 12-month reference period 50](#_Toc71552045)

[8.2.9. Other Employment, 12-month reference period 55](#_Toc71552046)

[8.2.10. Total Employment Earnings, 12-month reference period 56](#_Toc71552047)

[8.2.11. Total Labor Income 56](#_Toc71552048)

[8.3. Lessons Learned and Challenges 57](#_Toc71552049)

[8.4. Tabular Overview of Variables 59](#_Toc71552050)

# Introduction

This document serves as a guideline on the harmonization of the Global Labor Database (GLD), explaining what the GLD is, what it harmonizes and how. It follows the structure of the GMD Guidelines and, due to the aimed interoperability, borrows heavily from it (e.g., the description of variables that are both in GMD and GLD are identical in both documents).

## Global Labor Database (GLD) Project

The GLD project harmonizes official, nationally representative household surveys with significant labour market information to a common dictionary that makes the surveys comparable across time and place. It is the sister project to the Global Monitoring Database (GMD), which aims to do the same with household surveys used to calculate poverty estimates. Despite some exceptions to the rule, it helps to think broadly that GMD harmonizes household surveys with income and expenditure information (which most often also contain labour market data), while GLD covers all other surveys with labour market modules (most often labour force surveys – LFS).

Both the GLD and the GMD are the follow up to the International Income Distribution Database (I2D2) which was the World Bank’s earlier effort at creating a single large database of surveys with a common variable standard.

## File Organization

Since GLD is a collaborative effort where different individuals work on the harmonization of different surveys, it is essential that everyone follows the same file structure and do-file organization to allow for easy access and understanding.

### Where to find the original raw data files

The first step for each harmonization is to obtain the original raw data files. The original data files used for the GLD harmonization should be stored on the GLD server ([\\wbntpcifs\gld](file:///\\wbntpcifs\gld)) cataloged according to the rules of the Microdata Library folder and file naming structure. We assume further that the raw survey data fulfills all relevant data sharing and data custody arrangements. We aim to publish all information on the central Microdata Library catalog (<http://microdatalib/>). We assume that the depositing of the original microdata in the Microdata Library is a collective responsibility shared by all World Bank staff working in all the World Bank’s regions, irrespective of global practice, since the dialogue and data acquisition through the NSO and line ministries often take place in a decentralized manner.

### Do-file and data file naming guidelines

Data sets and do-files have identical file naming conventions. Do-files used to create a data file must have the same file name and follow the naming guidelines below.

Example: The following files are the do file and STATA file for the GLD harmonization using the 2016 Philippines Labor Force Survey.

PHL\_2016\_LFS\_v01\_M\_v01\_A\_GLD\_ALL.do  
PHL\_2016\_LFS\_v01\_M\_v01\_A\_GLD\_ALL.dta

Table

|  |  |  |
| --- | --- | --- |
| **Generic Structure: CCC\_YYYY\_SurveyName\_vnn\_M\_vmm\_A\_GLD\_mod** | | |
| **Component** | **Description** | **Values** |
| CCC | Country Code | Example: KAZ, GEO, TJK |
| YYYY | Year of the Survey | The *starting* year of the survey |
| SurveyName | Abbreviation of the survey | Example: HBS, KIHS, LFS, EUSILC  Character length can vary. |
| v*nn* | stands for the version of the  master file | *nn*=01, 02, … |
| v*mm* | stands for the version of the harmonization, as there can be  revisions after the first release | *mm*=01, 02, … |
| mod | Stands for the modules of information captured in the survey. Except for special cases that would be specifically instructed the harmonization should aim to cover all modules and thus will be “ALL”.  A special case may be only information at HH level called HH. | *mod = ALL, HH.* |

## Do-file Organization and Guidelines

### Do-files: header guidelines

Each do file should have a header that gives relevant context information (see Box 1, next page) that helps understand the harmonization process not only for other users but for the harmonizers themselves at later points in time. It contains four blocks:

* Information on the code, the harmonizer, and the creation date.
* Details on the survey context
* Details on the versions of the standard classifications used
* Version control history, detailing the date and the contents of any changes performed.
  + An example would be the fact that a previous harmonization was using the wrong industry classification code, distorting the industry category variables. Then we would enter the date of the change and an explanation of why this was necessary and what changes were done.

Box 1 - GLD Harmonization Template Preamble

|  |
| --- |
| /\*%%=================================================================  0: GLD Harmonization Preamble  ===================================================================\*/  /\* -----------------------------------------------------------------  <\_Program name\_> [Name of your do file] </\_Program name\_>  <\_Application\_> [Name of your software (STATA) and version] <\_Application\_>  <\_Author(s)\_> [Name(s) of author(s)] </\_Author(s)\_>  <\_Date created\_> YYYY-MM-DD </\_Date created\_>  ------------------------------------------------------------------  <\_Country\_> [Country\_Name (CCC)] </\_Country\_>  <\_Survey Title\_> [SurveyName] </\_Survey Title\_>  <\_Survey Year\_> [Year of start of the survey] </\_Survey Year\_>  <\_Study ID\_> [Microdata Library ID if present] </\_Study ID\_>  <\_Data collection from\_> [MM/YYYY] </\_Data collection from\_>  <\_Data collection to\_> [MM/YYYY] </\_Data collection to\_>  <\_Source of dataset\_> [Source of data, e.g. NSO] </\_Source of dataset\_>  <\_Sample size (HH)\_> [#] </\_Sample size (HH)\_>  <\_Sample size (IND)\_> [#] </\_Sample size (IND)\_>  <\_Sampling method\_> [Brief description] </\_Sampling method\_>  <\_Geographic coverage\_> [To what level is data significant] </\_Geographic coverage\_>  <\_Currency\_> [Currency used for wages] </\_Currency\_>  --------------------------------------------------------------------  <\_ICLS Version\_> [Version of ICLS for Labor Questions] </\_ICLS Version\_>  <\_ISCED Version\_> [Version of ISCED used to code] </\_ISCED Version\_>  <\_ISCO Version\_> [Version of ISCO used to code] </\_ISCO Version\_>  <\_OCCUP National\_> [Version of national occupation code] </\_OCCUP National\_>  <\_ISIC Version\_> [Version of ISIC used to code] </\_ISIC Version\_>  <\_INDUS National\_> [Version of national industry code] </\_INDUS National\_>  ---------------------------------------------------------------------  <\_Version Control\_>  \* Date: [YYYY-MM-DD] - [Description of changes]  \* Date: [YYYY-MM-DD] - [Description of changes]  </\_Version Control\_>    -------------------------------------------------------------------\*/ |

### Do-files: variable and variable note tagging guidelines

Harmonization sections and all variables in the .do files will be tagged using a standardized approach.

The sections will be tagged according to the following rules (see example in Box 2):

* The section marker starts with /\*%% followed by the equal sign to pad out the line
* The section title is indented, starting with a number, colon, and section title
* The section marker closes with a line of equal signs ended by %%\*/ (inverse of start)

Box 2 - Section header example

|  |
| --- |
| /\*%%========================================================================  1: Setting up of program environment, dataset  ========================================================================%%\*/ |

All harmonized variables in the in each section in the data will be tagged according to the following convention (see example in Box 3 - Variable tagging example):

* The beginning of the code relating to a harmonized variable should be proceeded by \*<\_*var*\_>
* The end of the code relating to the variable creation should read \*</\_var\_> where “var” is the harmonized variable being created.
* Variables that are already named (e.g. if "hhid" is already defined) should be noted when the file when opened, using the same convention as above. Between the "open" and "close" codes, a starred outline should read:"\*'var' brought in from 'source'"
* If a variable is created more than once (for example, hhid is created from several sources, then used to merge), it should be tagged only once.
* For example, the variable “lfstatus” is created only for individuals with age of 15 and above, then one can put the variable-specific note as follow: \*<\_lfstatus\_note\_> Only for individuals with age of 15 and above \*</\_lfstatus\_note\_>.

Box 3 - Variable tagging example

|  |
| --- |
| \*<\_hhid\_>  /\* <\_hhid\_note>  The variable should be a string made up of the elements to define it, that is psu code, ssu, ...  Each element should always be as long as needed for the longest element. That is, if there are 60 psu coded 1 through 60, codes should be 01, 02, ..., 60. If there are 160 it should be 001, 002, ..., 160.  </\_hhid\_note> \*/  egen hhid = concat( [Elements] )  label var hhid "Household ID"  \*</\_hhid\_>  \*<\_pid\_>  gen pid =  label var pid "Individual ID"  \*</\_pid\_> |

There are two useful purposes of tagging the harmonization variables: (1) tagging is useful when cross checking the definitions of harmonized variables overtime, and when comparing the comparability of such variables with different countries; (2) tagging will improve the automated updating of the DDI by adding the block of codes used for generating the harmonized variables in the variable description of the DDI. Tagging will also improve the transparency of the metadata DDI for basic users in the Microdata Library.

Tagging should be done for one variable at a time, not a group of variables.

# Survey & ID module

## ID Guidelines

The Survey & ID module is a file at individual level with basic household and individual identifiers. The household identifier is hhid, and the individual identifier is pid.

## Mapping and Description of Variables

### countrycode

countrycode is a string variable that specifies the 3-character country code used by the World Bank to identify each country. Although there are different naming conventions, it is necessary to use those specified to ensure that the data for each country is appropriately labeled.

### survname

survname codes the acronym of the survey.

### survey

survey codes the type of survey (e.g., LFS for Labor Force Survey).

### icls\_v

Underlying version of the International Conference of Labour Statisticians that is being used in the survey to code [concepts of work and employment](http://documents1.worldbank.org/curated/en/468881598538973944/pdf/Who-Is-Employed-Evidence-from-Sub-Saharan-Africa-on-Redefining-Employment.pdf).

### isced\_version

Underlying version of the [International Standard Classification of Education](http://uis.unesco.org/en/isced-mappings) (ISCED) used in the survey. Acceptable values are either isced\_1997 or isced\_2011.

### isco\_version

Underlying version of the [International Standard Classification of Occupations](https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/) (ISCO) used in the survey. Acceptable values are either isco\_1988 or isco\_2008

### isic\_version

Underlying version of the [International Standard Industrial Classification of All Economic Activities](https://unstats.un.org/unsd/classifications/Econ/isic) (ISIC) used in the survey. Acceptable values are either isic\_2, isic\_3, isic\_3.1, or isic\_4.

### year

year is a numeric variable that denotes the year in which the implementation of the household survey was begun. For example, if a survey was implemented during October 2018 and September 2019, the year would be 2018.

### vermast

vermast codes the version of the master file (original data) being used in the harmonization.

### veralt

veralt codes the version of the harmonization.

### harmonization

harmonization codes the kind of harmonization (GLD or GMD). For GLD surveys this will always be GLD.

### int\_year

int\_year is a numeric variable that specifies the year when the survey questionnaire was administered to the household.

### int\_month

int\_month is a numeric variable that specifies the month when the survey questionnaire was administered to the household.

### hhid

hhid specifies the unique household identification number in the data file. The original format, string or numeric, of original data should be kept. If there is Household ID in the original data, hhid and hhid\_orig should be the same. If hhid\_orig is missing, it is constructed by "variable names in raw data" variables.

### pid

This variable allows identification of individuals. Variable will vary in length depending on how the identification code was constructed in each country. Depending on individual countries, this variable may be a concatenation of several variables in the raw data file. Keep format (string or numeric) of original data. If there is Personal ID in the original data, pid and pid\_orig should be the same. If pid\_orig is missing, it is constructed by "variable names in raw data" variables.

### weight

weight contains household weights, typically inversely proportional to the probability of the household being selected for the sample, that should be applied to all analysis to make the results representative of the population.

### psu

Primary sampling unit code.

### ssu

Secondary sampling unit code (if present).

### strata

Unit defining the first stage stratification strategy.

### wave

In case of the survey being rolled out over several waves (e.g. quarterly), codes the information of the iteration of the survey.

## Lessons Learned and Challenges

### Creating and checking IDs

Avoid using the sequential index of the observation as the ID (i.e. gen hhid = \_n). This is dangerous as the order of each observation may be different, even across vintages of the same file sorted by to different variables.

|  |
| --- |
| \* Create hhid like this:  gen hhid = psu \* 100 + hh  \* Note like this:  gen hhid = \_n |

When creating hhid and pid, especially from string variables or from group(varlist) or concat(varlist) functions, users should try to create them from roster data files first where all information or observations are available. In addition, the order of the variables in the varlist option above must be the same across the files. Across the data files, the order and the sort on the variables in the varlist must be done in the same way across files.

When the hhid and pid are in numeric format but less precision, it is recommended to bring them the accurate precision level so it can be used in the merging correctly. For example, the value of the hhid for an observation might be 100021210121 (a long number), users should format the variable by “format %15.0g hh”.

In case a household survey is conducted more than once per year – e.g. quarterly HH surveys – you may want to use this as panel data, in which case the household ID can remain as is. However, if you want to use the data as cross-sectional, then new HHIDs can be constructed for each HH for each quarter.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Quarter** | **Quarter 1** | **Quarter 2** | **Quarter 3** | **Quarter 4** |
| hhid\_orig | hhid=1 | hhid=1 | hhid=1 | hhid=1 |
| hhid | hhid=1Q1 | hhid=1Q2 | hhid=1Q3 | hhid=1Q4 |

hhid should never be missing and if there is any missing this variable should be checked.

|  |
| --- |
| assert missing(hhid) |

It is recommended to check the uniqueness level of the data files with identifier variables at the corresponding level of the data (i.e. household vs individual level data).

hhid and pid need to be unique in the database.

|  |
| --- |
| isid hhid pid  cap destring pid, replace duplicates report hhid pid local n=r(unique\_value)  `N'!= `n' |

### Checks on the country code and year

Ensure that country is a three-letter country code.

|  |
| --- |
| cap confirm str3 var country \_rc!=0 |

Harmonizers should also ensure that country codes are updated according to the [ISO country codes](https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3). Some common adjustments include the following:

|  |
| --- |
| replace countrycode="XKX" if countrycode=="KSV"  replace countrycode="TLS" if countrycode=="TMP"  replace countrycode="PSE" if countrycode=="WBG"  replace countrycode="COD" if countrycode=="ZAR" |

Furthermore, harmonizers should check that the years used are in an appropriate range.

The year needs to be a four-digit number in the range of 1980 to the current year (assumed here to be 2020).

|  |
| --- |
| assert missing(hhid) |

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable label | Variable name | Notes |
| Survey & ID | ISO 3 Letter country code | countrycode |  |
| Survey & ID | Survey acronym | survname | No spaces, no underscores, split sections by "-" (e.g. "ETC-II") |
| Survey & ID | Survey long name | survey | Possible names are: LFS, LSMS, … [I am unsure of this difference, some surveys contain either this or the previous variable, have yet to see one with both] |
| Survey & ID | Version of the ICLS followed | icls\_v | Defines the labor force definitions used according to the rules set out by the nth International Conference of Labour Statisticians. |
| Survey & ID | Version of ISCED used | isced\_version |  |
| Survey & ID | Version of ISCO used | isco\_version |  |
| Survey & ID | Verstion of ISIC used | isic\_version |  |
| Survey & ID | Year of survey start | year |  |
| Survey & ID | Master (Source) data version | vermast |  |
| Survey & ID | Alternate (Harmonized) data version | veralt |  |
| Survey & ID | Kind of harmonization | harmonization |  |
| Survey & ID | Year of interview start | int\_year | For HH and Individual interviews in that HH earliest possible date |
| Survey & ID | Month of interview start | int\_month | For HH and Individual interviews in that HH earliest possible date |
| Survey & ID | Household ID | hhid |  |
| Survey & ID | Personal ID | pid |  |
| Survey & ID | Survey weights | weight |  |
| Survey & ID | Primary sampling unit | psu |  |
| Survey & ID | Secondary sampling unit | ssu |  |
| Survey & ID | Stratification (of PSU) | strata |  |
| Survey & ID | Wave of the survey (e.g., Q1 for quarter 1) | wave |  |

# Geography

## Mapping and Description of Variables

### urban

urban is a dummy variable that specifies the location type – urban or rural - of the household. This variable is country specific as each country uses its own criterion to distinguish urban from rural areas. In many cases there is no clear division between urban and rural areas, and areas are classified as “semi-urban” or “mixed”. Harmonizers are advised to classify such categories as “urban.”

Urban categories:

1 = Urban

0 = Rural

### subnatid1

subnatid1 refers to a subnational identifier at the highest level within the country’s administrative structure. This is typically a province or state. The variable is string and country-specific categorical. Numeric entries are coded in string format using the following naming convention: “1 – Hatay”. That is, the variable itself is to be string, not a labelled numeric vector.

### subnatid2

subnatid2 refers to a subnational identifier at which survey is representative at the second highest level within the country’s administrative structure. This is typically a district. The variable is string and country- specific categorical. Numeric entries are coded in string format using the following naming convention: “1 – Hatay”. That is, the variable itself is to be string, not a labelled numeric vector.

### subnatid3

subnatid3 refers to a sub-national identifier at which survey is representative at the third level within the country’s administrative structure. This is typically a sub-district. The variable is string and country-specific categorical. Numeric entries are coded in string format using the following naming convention: “1 – Hatay”. That is, the variable itself is to be string, not a labelled numeric vector.

### subnatid4

subnatid4 refers to a sub-national identifier at which survey is representative at the lowest level within the country’s administrative structure. In some countries, this is effectively a village. The variable is string and country-specific categorical. Numeric entries are coded in string format using the following naming convention: “1 – Hatay”. That is, the variable itself is to be string, not a labelled numeric vector.

### subnatidsurvey

subnatidsurvey is a string variable that refers to the lowest level of the administrative level at which the survey is representative. In most cases this will be equal to “subnatid1” or “subnatid2”. However, in some cases the lowest level is classified in terms of urban, rural or any other regional categorization cannot be mapped to subnatids.

The below example shows how to code subnatidsurvey for a survey that is representative at the rural/urban level of the province (subnatid1).

|  |  |  |
| --- | --- | --- |
| subnatid1 | urban | subnatidsurvey |
| “1 – Province A” | 1 | “1 – Province A urban” |
| “1 – Province A” | 1 | “1 – Province A urban” |
| “1 – Province A” | 1 | “1 – Province A urban” |
| “1 – Province A” | 1 | “1 – Province A urban” |
| “1 – Province A” | 0 | “1 – Province A rural” |
| … | … | … |
| “1 – Province B” | 0 | “1 – Province B rural” |
| “1 – Province B” | 0 | “1 – Province B rural” |
| “1 – Province B” | 0 | “1 – Province B rural” |
| “1 – Province B” | 0 | “1 – Province B rural” |
| “1 – Province B” | 1 | “1 – Province B urban” |
| “1 – Province B” | 1 | “1 – Province B urban” |

While the below is the example of a survey representative nationally, nationally at urban / rural level and at province level.

|  |  |  |
| --- | --- | --- |
| subnatid1 | urban | subnatidsurvey |
| “1 – Province A” | 1 | “1 – Province A” |
| “1 – Province A” | 1 | “1 – Province A” |
| “1 – Province A” | 1 | “1 – Province A” |
| “1 – Province A” | 1 | “1 – Province A” |
| “1 – Province A” | 0 | “1 – Province A” |
| … | … | … |
| “1 – Province B” | 0 | “1 – Province B” |
| “1 – Province B” | 0 | “1 – Province B” |
| “1 – Province B” | 0 | “1 – Province B” |
| “1 – Province B” | 0 | “1 – Province B” |
| “1 – Province B” | 1 | “1 – Province B” |
| “1 – Province B” | 1 | “1 – Province B” |

The variable would contain survey representation at lowest level irrespective of its mapping to subnatids.

### subnatid1\_prev

subnatid1\_prev is coded as missing unless the classification used for subnatid1 has changed since the previous survey. In that case, it refers to the subnatid1 code used in the previous survey. This provides a way of tracking splits. For example, if province “32 – West Java” split into province “32 – West Java” and “36 – Banten” since the most recent survey, subnatid1\_prev would be "32 – West Java” for cases when subnatid1 is either “32 – West Java” or “36 – Banten”.

### subnatid2\_prev

subnatid2\_prev is coded as missing unless the classification used for subnatid2 has changed since the previous survey. In that case, it refers to the subnatid2 code used in the previous survey.

### subnatid3\_prev

subnatid3\_prev is coded as missing unless the classification used for subnatid3 has changed since the previous survey. In that case, it refers to the subnatid3 code used in the previous survey.

### subnatid4\_prev

subnatid4\_prev is coded as missing unless the classification used for subnatid4 has changed since the previous survey. In that case, it refers to the subnatid4 code used in the previous survey.

### psu

primary sampling unit (psu) refers to sampling units that are selected in the first (primary) stage of multi- stage sample design. These sampling units typically correspond to a number of large aggregate units (clusters), each of which contains sub-units. For example, a primary sampling unit can represent the set of all housing units contained in a well-defined geographic area, such as a municipality or a group of contiguous municipalities. Primary sampling units are numeric and country-specific. A unique identifier is created for each primary sampling unit. In Stata, users are advised to specify the primary sampling unit through the svyset command.

### strata

strata refer to the division of the target population – typically the census sample frame -- into subpopulations based on auxiliary information that is known about the full population. Sampling is conducted separately for each strata. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Sampling strata need to be considered when constructing the variance (or confidence intervals) of population estimates. strata is needed for the correct calculation of standard deviation for each sample design. Strata is numeric and country-specific. A unique identifier is created for each strata. In STATA, users are advised to specify strata through the svyset command. The variable is in string format with the following naming convention “code of stratum – stratum name”, for example: “1 – Dar-es-salaam”

### gaul\_adm1\_code

gaul\_adm1\_code is numeric and country-specific based on the GAUL database. It should be taken from the same data in the GAUL database (a copy of those codes is available at the D4G team, contact Minh Nguyen at [mnguyen3@worldbank.org](mailto:mnguyen3@worldbank.org) or David Newhouse at [dnewhouse@worldbank.org](mailto:dnewhouse@worldbank.org)) where the geographical area can be identified in the survey based on the name of the location/area. The number of unique values from the subnatid1 and the gaul\_adm1\_code could be different or the same. For example, in the case of a fictional country, if the highest-level representation is the state level (53 states) and Gaul also has 53 states, it is the same in this case. In a different example, the survey is representative at the level of statistical regions (7) while the identifiable GAUL code is at state level (53 states); with this setup, one can know how the seven statistical regions are constructed.

### gaul\_adm2\_code

gaul\_adm2\_code is numeric and country-specific based on the GAUL database. It should be taken from the same data in the GAUL database where the geographical area can be identified in the survey based on the name of the location/area.

## Lessons Learned and Challenges

* subnatid codes should reflect the most recent codes that pertain to that survey. subnatid\_prev codes can be used to track splits and new administrative units that have been introduced since the previous survey. It is important to ensure there is consistency in geographic variables across time. Sub- nationally representative units may be added in later additions of surveys, so names of subnational units must be consistent across time. This will allow analysts to make the current administrative units “backwards-compatible” with little additional effort.
* Harmonizers should ensure the subnatid1 through subnatid4 are string variables NOT categorical.
* The urban variable cannot be different from zero or one.

|  |
| --- |
| urban!= 1 & urban!= 0 |

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable name | Variable label | Notes |
| Geography | urban | Binary - Individual in urban area |  |
| Geography | subnatid[i] | Subnational ID - [ith] level | Subnational ID at the ith level, listing as many as available |
| Geography | subnatidsurvey | Lowest level of Subnational ID | subnatidsurvey is a string variable that refers to the lowest level of the administrative level at which the survey is representative. In most cases this will be equal to “subnatid1” or “subnatid2”. However, in some cases the lowest level is classified in terms of urban, rural or any other regional categorization cannot be mapped to subnatids. The variable would contain survey representation at lowest level irrespective of its mapping to subnatids. |
| Geography | subnatid[i]\_prev | Subnatid previous - [ith] level | Previous subnatid if changed since last survey |
| Geography | [n]su | Nth sampling unit |  |
| Geography | strata | Strata |  |
| Geography | gaul\_adm[i]\_code | GAUL ADM[i] code | See en.wikipedia.org/wiki/Global\_Administrative\_Unit\_Layers |

# Demography

## Mapping and Description of Variables

### hsize

hsize codes the size of the household. This should not include individuals who may be living in the household but do not form an economic unit with the other household members (e.g., a live-in maid is not part of the household).

### age

age refers to the interval of time between the date of birth and the date of the survey. Every effort should be made to determine the precise and accurate age of each person, particularly of children and older persons. Information on age may be secured either by obtaining the date (year, month, and day) of birth or by asking directly for age at the person’s last birthday. In addition, in the case of children aged less than or equal to 60 months, variable age should be expressed in the number of completed years and months in decimals. For example, If the interview of a 4 years old was in December and he was born in June, his age should be recorded as 4.5. Lastly, if the information on age is not available, it should be coded as missing rather than some other value such as “99” or “999”.

### male

male is a dummy variable that specifies the sex – male or female – of an individual within a household. While constructing this variable, it is important to make sure that all relevant values are included. Variable values coded as ‘98’ or other numeric characters should be excluded from the values of the `male’ variable. Sex of household member, two categories after harmonization:

1 = male

0 = female

### relationharm

relationharm is a string variable that indicates a relationship to the reference person of household (usually the head of household). Variable values coded as ‘98’ or other numeric characters should be excluded from the values of relationharm variable.

Relationship to head of household, six categories after harmonization:

1=head

2=spouse

3=children

4=parents

5=other relatives

6=non-relatives

Note: In cases where head is missing or a migrant, we assign spouse as the head of the household. If spouse is also not available, then we will use oldest member of the household as the head and recode all the relations to head accordingly.

### relationcs

relationcs is a country-specific categorical variable that indicates the relationship to the head of the household. The categories for relationship to the head of the household are defined according to the region or country requirements.

### marital

Marital is a categorical variable that refers to the personal status of each individual in relation to the marriage laws or customs of the country. The categories of marital status to be identified should include at least the following: (a) single (in other words, never married); (b) married; (c) married but separated;

(d) windowed and remarried; (e) divorced and not remarried. In some countries, category (b) may require a subcategory of persons who are contractually married but not yet living as man and wife. In all countries, category (c) should comprise both the legally and the de facto separated, who may be shown as separate subcategories if desired. The marital variable should not be imputed but rather calculated only for those to whom the question was asked (in other words, the youngest age at which information is collected may differ depending on the survey). The consistency between age and marital needs to be cross-checked. In most countries, there are also likely to be persons who were permitted to marry below the legal minimum age because of special circumstances. To permit international comparisons of data on marital status, however, any tabulations of marital status not cross-classified by exact age should at least distinguish between persons under 15 years of age and over. If it is not possible to distinguish between married and living together, then it should be assumed that the individual is married. Variable values coded as ‘98’ or other numeric characters should be excluded from the values of the ‘marital’ variable.

Marital status, five categories after harmonization:

1=married

2=never married

3=living together

4=divorced/separated

5=widowed

### eye\_dsablty

eye\_dsablty is a numerical variable that indicates whether an individual has any difficulty in seeing, even when wearing glasses. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

### hear\_dsablty

hear\_dsablty is a numerical variable that indicates whether an individual has any difficulty in hearing even when using a hearing aid. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

### walk\_dsablty

walk\_dsablty is a numerical variable that indicates whether an individual has any difficulty in walking or climbing steps. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

### conc\_dsord

conc\_dsord is a numerical variable that indicates whether an individual has any difficulty concentrating or remembering. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

### slfcre\_dsablty

slfcre\_dsablty is a numerical variable that indicates whether an individual has any difficulty with self-care such as washing all over or dressing. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

### comm\_dsablty

comm\_dsablty is a numerical variable that indicates whether an individual has any difficulty communicating or understanding usual (customary) language. Categories after harmonization:

1 = No – no difficulty

2 = Yes – some difficulty

3 = Yes – a lot of difficulty

4 = Cannot do at all

## Lessons Learned and Challenges

Data sets that are harmonized incorrectly can lead to skewed and/or incorrect data analysis. Harmonizers should run a series of checks to ensure data is harmonized properly, including the following:

Check to make sure that age is an integer since 5 years old.

age/int(age)!= 1 & age!= . & age > 5

age cannot have negative or extreme values (>120)

(age < 0 | age>120) & age<.

Age cannot be missing

age==.

Male variable can only take one of two values 0 or 1 (or missing).

male!=. & male!= 1 & male!= 0

Check if male is missing.

male==.

Check to make sure that there is variation in male

egen sdmale = sd(male) // sdmale should be 0

relationharm must be an integer in the range [1,6].

relationharm<1 & relationharm>6 & mod(relationharm, 1) == 1

marital must be an integer in the range [1,5].

marital<0 & marital>5 & mod(marital, 1) == 1

Children are “Never married” and should be coded as so even though it may be perceived as obvious. The marital status of individuals should be harmonized for all individuals. Harmonizers should check to make sure children are not systematically left with missing values for marital.

tab age marital, missing

weight cannot be missing

weight==.

Additionally, harmonizers should ensure that the household size variable is calculated correctly. Not all the individuals reported in a household that form the raw data are current household members. For example, for the EU-SILC survey, a household contains the current member, but also the members of the previous survey who have left the household for reasons such as death or migration.

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable name | Variable label | Notes |
| Demography | hsize | Household size |  |
| Demography | age | Age in years |  |
| Demography | male | Binary - Individual is male |  |
| Demography | relationharm | Relationship to head of household harmonized across all regions | GMD - Harmonized categories across all regions. Same as I2D2 categories. |
| Demography | relationcs | Relationship to head of household country/region specific | country or regionally specific categories |
| Demography | marital | Marital status |  |
| Demography | eye\_dsablty | Difficulty seeing | See "Recommended Short Set of Questions" on https://www.cdc.gov/nchs/washington\_group/wg\_questions.htm |
| Demography | hear\_dsablty | Difficulty hearing |
| Demography | walk\_dsablty | Difficulty walking / steps |
| Demography | conc\_dsord | Difficulty concentrating |
| Demography | slfcre\_dsablty | Difficulty w/ selfcare |
| Demography | comm\_dsablty | Difficulty communicating |

# Migration

## Mapping and Description of Variables

### migrated\_mod\_age

Codes the minimum age the migration module questions of the survey apply to (e.g., if the migration questions are for all 5 years and above this would be 5).

### migrated\_ref\_time

Codes the reference period the migration questions cover. If the migration questions only apply after an introductory time window questions (e.g., have you moved in the past five years) and then questions are only asked for those who fall within the time window, code the length of that window (e.g., 5 in the example). If migration questions are posed regardless of time (i.e., no time window) code 99.

### migrated\_binary

Binary question coding whether the individual has ever migrated (within the reference time set out above).

### migrated\_years

Number of full years since the last migration. Often surveys ask how long a person has lived at their current domicile since the last migration. Both questions cover the same information.

### migrated\_from\_urban

Codes whether the individual migrated to their current domicile from an urban area.

1 = Yes (i.e., came from urban area)

0 = No (i.e., came from rural area)

### migrated\_from\_cat

If the survey contains information on the area from where the person migrated, use the [concept of administrative division](https://en.wikipedia.org/wiki/Administrative_division) to inform the migration pattern. The codes are:

1 = From same admin3 area

2 = From same admin2 area

3 = From same admin1 area

4 = From other admin1 area

5 = From other country

To exemplify the use, Spain is divided into Communities (admin1 level), Provinces (admin2 level) and municipalities (names change within provinces, but rough concept holds – admin3 level). A person moving within the municipality, for example, from one village to the next, codes 1. A person moving within the province, say from a rural municipality to the province capital codes 2. A person moving within the same community yet leaving their province codes 3. A person moving from one community to another, say from [Andalusia](https://en.wikipedia.org/wiki/Andalusia) to [Galicia](https://en.wikipedia.org/wiki/Galicia_(Spain)), codes 4. If the person moved from outside the country (regardless of their nationality) codes 5.

### migrated\_from\_code

Based on the logic set out in the *migrated\_from\_cat* variable, codify the areas of migration using the survey subnation id classification. For example, if a person migrated from one admin1 area to another, use the subnatid1 codes to inform from which admin1 area they migrated to their current residence (which is codified in subnatid1).

This only codifies information within the country.

Note that most surveys will only provide information of last residence to a higher administrative level (e.g., admin1 level). Codify the information up to the highest level possible. See an example in 5.2 below.

### migrated\_from\_country

Codes the country from where the person migrated from as a [three letter ISO country code](https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3) or a clear string for regions (“Other Europe”, “Other World”, …)

### migrated\_reason

Codifies the reason why a person migrated. The codes are:

1 = Family reasons

2 = Educational reasons

3 = Employment

4 = Forced (political reasons, natural disaster, …)

5 = Other reasons

## Lessons Learned and Challenges

### Codifying *migrated\_from\_\* questions*

In the Indian LFS from 1999 (NSS Schedule 10) there are two questions that allow us to codify the four migrated\_from\_[text] variables (migrated\_from\_urban, migrated\_from\_cat, migrage\_from\_code, and migrated\_from\_country).

Question 15 of Block 4 asks interviewer to enter the “location code” for the kind of migration the interviewees claim to have made. The codes are:

*location of last usual residence: same district: rural-1, urban-2; same state but another district: rural-3, urban-4; another state: rural-5, urban-6; another country-7*

Question 17 of Block 4 then asks for the state code of migration (as codified in subnatid) and adds additional codes for countries from where people commonly migrated into India.

With these two questions we can harmonize the two variables in the following way:

|  |  |
| --- | --- |
| \*<\_migrated\_from\_urban\_> | |
|  | gen migrated\_from\_urban = .  replace migrated\_from\_urban = 1 if inlist(B4\_q15,"2","4","6") & migrated\_binary == 1  replace migrated\_from\_urban = 0 if inlist(B4\_q15,"1","3","5") & migrated\_binary == 1  label de lblmigrated\_from\_urban 0 "Rural" 1 "Urban"  label values migrated\_from\_urban lblmigrated\_from\_urban  label var migrated\_from\_urban "Migrated from area" |
| \*</\_migrated\_from\_urban\_> | |
|  |  |
| \*<\_migrated\_from\_cat\_> | |
|  | gen migrated\_from\_cat = .  replace migrated\_from\_cat = 2 if inlist(B4\_q15,"1","2") & migrated\_binary == 1  replace migrated\_from\_cat = 3 if inlist(B4\_q15,"3","4") & migrated\_binary == 1  replace migrated\_from\_cat = 4 if inlist(B4\_q15, "5", "6") & migrated\_binary == 1  replace migrated\_from\_cat = 5 if inlist(B4\_q15, "7") & migrated\_binary == 1  label de lblmigrated\_from\_cat 1 "From same admin3 area" 2 "From same admin2 area" 3 "From same admin1 area" 4 "From other admin1 area" 5 "From other country"  label values migrated\_from\_cat lblmigrated\_from\_cat  label var migrated\_from\_cat "Category of migration area" |
| \*</\_migrated\_from\_cat\_> | |
|  |  |
| \*<\_migrated\_from\_code\_> | |
|  | destring B4\_q17, gen(helper\_var)  gen migrated\_from\_code = .  replace migrated\_from\_code = subnatid1 if inrange(migrated\_from\_cat,1,3)  replace migrated\_from\_code = helper\_var if migrated\_from\_cat == 4  label var migrated\_from\_code "Code of migration area as subnatid level of migrated\_from\_cat"  drop helper\_var |
| \*</\_migrated\_from\_code\_> | |
|  |  |
| \*<\_migrated\_from\_country\_> | |
|  | gen migrated\_from\_country = ""  replace migrated\_from\_country = "BGD" if migrated\_from\_cat == 5 & B4\_q17 == "51"  replace migrated\_from\_country = "NPL" if migrated\_from\_cat == 5 & B4\_q17 == "52"  replace migrated\_from\_country = "PAK" if migrated\_from\_cat == 5 & B4\_q17 == "53"  replace migrated\_from\_country = "LKA" if migrated\_from\_cat == 5 & B4\_q17 == "54"  replace migrated\_from\_country = "BTN" if migrated\_from\_cat == 5 & B4\_q17 == "55"  replace migrated\_from\_country = "Gulf countries" if migrated\_from\_cat == 5 & B4\_q17 == "56"  replace migrated\_from\_country = "Other Asian" if migrated\_from\_cat == 5 & B4\_q17 == "57"  replace migrated\_from\_country = "USA" if migrated\_from\_cat == 5 & B4\_q17 == "58"  replace migrated\_from\_country = "CAN" if migrated\_from\_cat == 5 & B4\_q17 == "59"  replace migrated\_from\_country = "Other Americas" if migrated\_from\_cat == 5 & B4\_q17 == "60"  replace migrated\_from\_country = "UK" if migrated\_from\_cat == 5 & B4\_q17 == "61"  replace migrated\_from\_country = "Other Europe" if migrated\_from\_cat == 5 & B4\_q17 == "62"  replace migrated\_from\_country = "African countries" if migrated\_from\_cat == 5 & B4\_q17 == "63"  replace migrated\_from\_country = "Other World" if migrated\_from\_cat == 5 & B4\_q17 == "64"  label var migrated\_from\_country "Code of migration country (ISO 3 Letter Code)" |
| \*</\_migrated\_from\_country\_> | |

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable name | Variable label | Notes |
| Migration | migrated\_mod\_age | Migration module application age |  |
| Migration | migrated\_ref\_time | Reference time applied to migration questions | If migrated\_ref\_time = 5 means questions about migration refer to any migration in the last 5 years |
| Migration | migrated\_binary | Individual has migrated |  |
| Migration | migrated\_years | Years since latest migration | Years since last migration is the same as how long lived at current location |
| Migration | migrated\_from\_urban | Migrated from area | No means migrated from rural area |
| Migration | migrated\_from\_cat | Category of migration area |  |
| Migration | migrated\_from\_code | Code of migration area |  |
| Migration | migrated\_from\_country | Code of migration country |  |
| Migration | migrated\_reason | Reason for migrating |  |

# Education

## Mapping and Description of Variables

### ed\_mod\_age

Codifies the minimum age for which education questions are asked. For example, if education information is only requested from those 4 years and older the variable should be set to 4.

### school

Codifies whether the person is currently (i.e., at the time of the survey) attending formal education. The codes are:

0 = No

1 = Yes

### literacy

Codifies whether person can read and write in at least one language. The codes are:

0 = No

1 = Yes

### educy

Codifies the number of years spent in education.

### educat7

Classifies the highest level of education attained by the respondent to seven levels. The codes are:

1 = No education

2 = Primary incomplete

3 = Primary complete

4 = Secondary incomplete

5 = Secondary complete

6 = Higher than secondary but not university

7 = University incomplete or complete

### educat5

Classifies the highest level of education attained by the respondent to five levels. The codes are:

1 = No education

2 = Primary incomplete

3 = Primary complete but secondary incomplete

4 = Secondary complete

5 = Some tertiary/post-secondary

### educat4

Classifies the highest level of education attained by the respondent to four levels. The codes are:

1 = No education

2 = Primary (complete or incomplete)

3 = Secondary (complete)

4 = Tertiary (complete or incomplete)

Note: Code as primary education anyone who has undergone some schooling but has not finished secondary education.

### educat\_orig

Original education code as in the raw survey data. If the original survey has a single variable coding the education information, simply copy (gen educat\_orig = survey\_education\_var).

If the survey has two or more variables that need to be used to codify education, leave this missing and – if possible – make a note both in the do code and in the survey introduction notes.

As an example, below is the codification in the Mexican LFS. Here, education is both a grade (part 1 of question 13) and the number of years spent in that grade (part 2). In such a case *educat\_orig* is to be left missing.



### educat\_isced

Code of the highest educational level attained as per the International Standard Classification of Education (ISCED). Note that the preamble to the harmonization code should record what version of ISCED is being used.

Moreover, the code should always be as long as the longest depth available for the ISCED version. For example, the latest version at the time of writing (ISCED 2011) has up to three digits. Where the first digit is the level, the first two digits are the category, and all three digits codify the sub-category.

As an example, level 2 codifies “Lower secondary education”, 24 “Lower secondary general education”, and 242 “Sufficient for partial level completion, without direct access to upper secondary education”. Every code should be three digits long. If we only know the level (here 2) add two zeroes after it (here: 200). If we only have the category information (here 24) add a zero to reach three digits (here 240).

## Lessons Learned and Challenges

Text

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable name | Variable label | Notes |
| Education | ed\_mod\_age | Education module minumum age |  |
| Education | school | Currently in school |  |
| Education | literacy | Individual can read and write |  |
| Education | educy | Years of education |  |
| Education | educat7 | Level of education 7 categories | No option for "Other", as opposed to I2D2, anything not in these categories is to be set to missing |
| Education | educat5 | Level of education 5 categories |  |
| Education | educat4 | Level of education 4 categories |  |
| Education | educat\_orig | Original education code | Code if there is a single original education variable (as is in most cases). If there are two or more variables, leave missing, make a note of it. |
| Education | educat\_isced | International Standard Classification of Education (ISCED A) | Codes are for example:   2 Lower secondary education  24 General  242 Partial level completion, without direct access to upper secondary education  Should be coded as 200, 240, and 242 respectively. |

# Training

## Mapping and Description of Variables

### vocational

Codifies whether the person ever attended vocational training. The codes are:

0 = No

1 = Yes

### vocational\_type

Codifies whether the vocational training took place within the enterprise or was administered by an external party. The codes are:

1 = Inside Enterprise

2 = External

### vocational\_length\_l

Codifies how long the training was in months. Divided into lower and upper in case the information is coded as a range (e.g., 0-3 months, 6-12 months, …). If it is an exact number, lower and upper length are equal.

### vocational\_length\_u

see *vocational\_length\_u*

### vocational\_field

This variable has not been defined at the time of writing.

### vocational\_financed

Text. The codes are:

1 = Employer

2 = Government

3 = Mixed Employer-Government

4 = Own funds

5 = Other

## Lessons Learned and Challenges

Text

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module Code | Variable name | Variable label | Notes |
| Education | vocational | Ever received vocational training |  |
| Education | vocational\_type | Type of vocational training |  |
| Education | vocational\_length\_l | Length of training, lower limit |  |
| Education | vocational\_length\_u | Length of training, upper limit |  |
| Education | vocational\_field | Field of training |  |
| Education | vocational\_financed | How training was financed | If funded with different sources, chose main source |

# Labour

## ILCS Definitions

## Mapping and Description of Variables

### minlaborage

This is the lowest age for which the labor module is implemented in the survey or the minimum working age in the country. For this reason, the lower age cutoff at which information is collected will vary from country to country.

* + 1. Labour status, 7-day reference period

### lstatus

lstatus is an individual’s labor status in the last 7 days. The Value must be missing for individuals less than the required age (minlaborage).

Three categories are used after harmonization:

1 = Employed

2 = Unemployed

3 = Not-in-labor force

All persons are considered active in the labor force if they presently have a job (formal or informal, i.e., employed) or do not have a job but are actively seeking work (i.e., unemployed).

1 = Employed

Employed is defined as anyone who worked during the last 7 days or reference week, regardless of whether the employment was formal or informal, paid or unpaid, for a minimum of 1 hour. Individuals who had a job, but for any reason did not work in the last 7 days are considered employed.

2 = Unemployed

A person is defined as unemployed if he or she is, presently not working but is actively seeking a job. The formal definition of unemployed usually includes being ‘able to accept a job.’ This last question was asked in a minority of surveys and is, thus, not incorporated in the present definition. A person presently not working but waiting for the start of a new job is considered unemployed.

3 = Not-in-labor force

A person is defined as not-in-labor force if he or she is, presently not working and it is not actively seeking a job during the last 7 days or reference week.

### potential\_lf

Codifies whether the person is not in the labour force over the past 7 days (lstatus=3, missing otherwise) but could potentially be they are i) available but not searching or ii) searching but not immediately available to work. The codes are:

0 = No (not potentially in the labour force)

1 = Yes

### underemployment

Codifies whether the person is in the labour force and working over the past 7 days (lstatus=1, missing otherwise) but would take on more jobs or more hours at their job if possible/available. The codes are:

0 = No (not underemployed)

1 = Yes

### nlfreason

nlfreason is the reason an individual was not in the labor force in the last 7 days. This variable is constructed for all those who are not presently employed and are not looking for work (lstatus=3) and missing otherwise.

Five categories after harmonization:

1= Student (a person currently studying.)

2= Housekeeper (a person who takes care of the house, older people, or children)

3= Retired

4 = Disabled (a person who cannot work due to physical conditions)

5 = Other (a person does not work for any other reason)

Fill this information for all people interviewed in the labor section of the questionnaire regardless of their age.

### unempldur\_l

unempldur\_l is a continuous variable specifying the duration of unemployment in months (lower bracket).

The variable is constructed for all unemployed persons (lstatus=2, otherwise missing). If it is specified as continuous in the survey, it records the numbers of months in unemployment. If the variable is categorical it records the lower boundary of the bracket.

Missing values are allowed for everyone who is not unemployed. Other missing values are also allowed.

### unempldur\_u

unempldur\_u is a continuous variable specifying the duration of unemployment in months (upper bracket).

The variable is constructed for all unemployed persons (lstatus=2, otherwise missing). If it is specified as continuous in the survey, it records the numbers of months in unemployment. If the variable is categorical it records the upper boundary of the bracket. If the right bracket is open a missing value should be inputted.

Missing values are allowed for everyone who is not unemployed. Other missing values are also allowed. If the duration of unemployment is not reported as a range, but as continuous variables, the unempldur\_l and unempldur\_u variables will have the same value. If the high range is open-ended the unempldur\_u variable will be missing.

* + 1. Primary Employment, 7-day reference period

### empstat

empstat is a categorical variable that specifies the main employment status in the last 7 days of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all individuals. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

The definitions are taken from the International Labor Organization’s Classification of Status in Employment with some revisions to take into account the data available.

Five categories after harmonization:

1 = Paid Employee

2 = Non-Paid Employee

3 = Employer

4 = Self-employed

5 = Other, workers not classifiable by status

1 = Paid Employee

Paid employee includes anyone whose basic remuneration is not directly dependent on the revenue of the unit they work for, typically remunerated by wages and salaries but may be paid for piece work or in-kind. The ‘continuous’ criteria used in the ILO definition is not used here as data are often absent and due to country specificity.

2 = Non-Paid Employee

Non-paid employee includes contributing family workers who hold a self-employment job in a market-oriented establishment operated by a related person living in the same households who cannot be regarded as a partner because of their degree of commitment to the operation of the establishment, in terms of working time or other factors, is not at a level comparable to that of the head of the establishment. All apprentices should be mapped as ‘non-paid employee’

3 = Employer

An employer is a business owner (whether alone or in partnership) with employees. If the only people working in the business are the owner and contributing family workers, the person is not considered an employer (as has no employees) and is, instead classified as self-employed.

4 = Self-employed

Own account or self-employment includes jobs where remuneration is directly dependent from the goods and service produced (where home consumption is considered to be part of the profits) and where one has not engaged any permanent employees to work for them on a continuous basis during the reference period.

Members of producers’ cooperatives are workers who hold a self-employment job in a cooperative producing goods and services, in which each member takes part on an equal footing with other members in determining the organization of production, sales and/or other work of the establishment, the investments and the distribution of the proceeds of the establishment amongst the members.

5 = Other, workers not classifiable by status

Other, workers not classifiable by status include those for whom insufficient relevant information is available and/or who cannot be included in any of the above categories.

### ocusec

ocusec is a categorical variable that specifies the sector of activity in the last 7 days. It classifies the main job's sector of activity of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire.

Four categories after harmonization:

1 = Public sector, Central Government, Army (including armed forces)

2 = Private, NGO

3 = State-owned

4 = Public or State-owned, but cannot distinguish

1 = Public Sector, Central Government, Army (including armed forces) Public sector

The part of economy run by the government.

2 = Private, NGO

Private sector is that part of the economy which is both run for private profit and is not controlled by the state, it also includes non-governmental organizations

3 = State-owned enterprises

State-owned includes para-state firms and all others in which the government has control (participation over 50%).

4 = Public or State-owned, but cannot distinguish

Select this option is the questionnaire does not ask for State-owned enterprises, and only for Public sector.

Notes: Do not code basis of occupation (ISCO) or industry (ISIC) codes.

### industry\_orig

industry\_orig is a string variable that specifies the original industry codes in the last 7 days for the main job provided in the survey (the actual question) and should correspond to whatever is in the original file with no recoding. It will contain missing values for people below the working age. Other missing values are allowed. It classifies the main job of any individual with a job (lstatus=1) and is missing otherwise

### industrycat\_isic

Code (string variable) of the industry according to the International Standard Industry Classification (ISIC) in the last 7 days for the main job of any individual with a job (lstatus=1) and is missing otherwise. Note that the preamble to the harmonization code should record what version of ISIC is being used.

The code should always be as long as the longest depth available for the ISIC version. For example, the latest version at the time of writing (ISIC Rev 4, [available here](https://unstats.un.org/unsd/demographic-social/census/documents/isic_rev4.pdf)) codes industries by sections, divisions, groups, and classes, in decreasing order of hierarchy.

Figure 1 shows the classification structure for the manufacture of machinery and equipment. The letter C codes the Manufacturing *section*, while the code 28 represents “Manufacture of machines and equipment n.e.c” *division*. This division has two *groups* (281 and 282), containing one and three *classes* respectively.

Figure 1 - Example of ISIC classification



A single section will often cover several divisions. While D has only one division (35 – electricity, gas, steam and air conditioning supply) and could be potentially shortened to “3”, section C covers divisions 10 to 33.

If the information in the survey is only present at section level (or can only be translated from the national industry classification to section level) this variable should be a string with the letter coding the correct section.

In most cases, information will be coded as a set of digits. In this case, the codification should be a string of four digits with a zero padding before for division 1 through 9 (i.e., 01, to 09).

As an example, Figure 2, puts together a few excerpts from ISIC Rev.4. Note that, if we do not codify correctly, group 14 (Animal production) may be misunderstood for division 14 (Manufacture of wearing apparel).

If we only have information up to the group label, fill out the reminder of the digits with zeros. Hence the purple box in Figure 2 would be coded as “0140”. The red box, as we have all digits, including the zero padding at the start codes as “0142”.

The act of adding zeroes to the end is standard if the lower level hierarchy has no further distinctions and can be seen in the yellow box, where group 142 has no classes (or just one class) and thus is coded as “1420”.

Some groups do in fact have several classes, as can be seen for group 151. Again, if we only had information up to group level, we ought to code “1510”. If we have more detailed information, for example identifying the industry as “Tanning and dressing of leather; dressing and dyeing of fur” (green box) we would code “1511”.

Figure 2 - Examples of different ISIC codes



### industrycat10

industrycat10 is a categorical variable that specifies the 1-digit industry classification in the last 7 days for the main job of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire. The codes for the main job are given here based on the UN International Standard Industrial Classification. It classifies the main job of any individual with a job (lstatus=1) and is missing otherwise

Ten categories after harmonization:

1 = Agriculture, Hunting, Fishing, etc.

2 = Mining

3 = Manufacturing

4 = Public Utility Services

5 = Construction

6 = Commerce

7 = Transport and Communications

8 = Financial and Business Services

9 = Public Administration

10 = Other Services, Unspecified

Notes:

* In the case of different classifications (former Soviet Union republics, for example), recoding has been done to best match the ISIC codes.
* Category 10 is also assigned for unspecified categories or items.
* If all 10 categories cannot be identified in the questionnaire create this variable as missing and proceed to create industrycat4.
* Over the years, the different ISIC versions have changed. The original industrycat10 categories are largely based on ISIC Revision 2. The below table shows how to classify the different ISIC revision codes into industrycat10.

### 

### industrycat4

industrycat4 is a categorical variable that specifies the 1-digit industry classification in the last 7 days for the main job for Broad Economic Activities. This variable is either created directly from the data (if industry classification does not exist for ten categories) or created from industrycat10.

Four categories after harmonization:

1 = Agriculture

2= Industry

3 = Services

4 = Other

This variable is either created directly from the data (if industry classification does not exist for ten categories) or created from industrycat10.

### occup\_orig

occup\_orig is a string variable that specifies the original occupation code in the last 7 days for the main job. This variable corresponds to whatever is in the original file with no recoding.

### occup\_isco

Code (string variable) of the occupation according to the International Standard Classification of Occupations (ISCO) in the last 7 days for the main job of any individual with a job (lstatus=1) and is missing otherwise. Note that the preamble to the harmonization code should record what version of ISCO is being used.

The code should always be as long as the longest depth available for the ISCO) version. For example, the latest version at the time of writing (ISCO-08, [available here](https://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm)) codes occupations by Major, Sub-major, Minor, and Unit groups, in decreasing order of hierarchy.

ISCO code Major groups cover a single digit, running from 1 (Managers) to 9 (Elementary Occupations) with the additional category 0 (Armed Forces Occupations). Hence there is only need for zero-padding on the left side for Armed Forces Occupations. Figure 3 shows an example of the possible values that can be taken on.

Figure 3 - Example of values for ISCO-08



If we only had information at Major Group level, a person working as a Services and Sales Worker ought to be coded as the number 5000. If the information is at Sub-major Group level, it should be codified as the number 5100 for a Personal Services Worker, while a Travel Attendant, a Conductor, or a Guide (if information at Minor Group level) should be coded as the number 5110.

Information at the Unit Group level can be coded as is, since it already is at the maximum possible depth.

### occup\_skill

Categorical code for the broad skill level of workers at the main job in the last 7 days of any individual with a job (lstatus=1) and is missing otherwise. It follows from the ISCO classification as shown in Figure 4.

Figure 4 - ISCO broad skill level classification



Thus, the codes are:

3 = High

2 = Medium

1 = Low

. = Armed Forces and not elsewhere classified

### occup

occup is a categorical variable that specifies the 1-digit occupational classification for the main job in the last 7 days of any individual with a job (lstatus=1) and is missing otherwise. This variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. Most surveys collect detailed information and then code it, without keeping the original data, no attempt has been made to correct or check the original coding. The classification is based on the International Standard Classification of Occupations (ISCO). It classifies the main job of any individual with a job (lstatus=1) and is missing otherwise.

Eleven categories after harmonization:

1 = Managers

2 = Professionals

3 = Technicians and associate professionals

4 = Clerical support workers

5 = Service and sales workers

6 = Skilled agricultural, forestry and fishery workers

7 = Craft and related trades workers

8 = Plant and machine operators, and assemblers

9 = Elementary occupations

10 = Armed forces occupations

99 = Other/unspecified

### wage\_no\_compen

wage\_no\_compen is a continuous variable that specifies the last wage payment in local currency of any employed individual (lstatus=1) in its primary occupation at the reference period reported in the survey and it is missing otherwise. The wage should come from the main job, in other words, the job that the person dedicated most time in the week preceding the survey. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* For all those with self-employment or owners of own businesses, this should be net revenues (net of all costs EXCEPT for tax payments) or the amount of salary taken from the business. Due to the almost complete lack of information on taxes, the wage from main job is NOT net of taxes.
* By definition, non-paid employees (empstat=2) should have wage=0.
* The reference period of the wage\_no\_compen will be recorded in the unitwage variable.

### unitwage

unitwage is a categorical variable that specifies the time reference for the wage\_no\_compen variable. It specifies the time unit measurement for the wages of any individual (lstatus=1 & empstat=1) and it is missing otherwise. Acceptable values include:

1 = Daily

2 = Weekly

3 = Every two weeks

4 = Every two months

5 = Monthly

6 = Quarterly

7 = Every six months

8 = Annually

9 = Hourly

10 = Other

### whours

whours is a continuous variable that specifies the hours of work last week for the main job of any individual with a job (lstatus=1) and is missing otherwise. The main job defined as that occupation that the person dedicated more time to over the past week. The variable is constructed for all persons administered this module in each questionnaire.

Notes:

* If the respondent was absent from the job in the week preceeding the survey due to holidays, vacation, or sick leave, then record the time worked in the previous 7 days that the person worked.
* Sometimes the questions are phrased as, “on average, how many hours a week do you work?”.
* For individuals who only give information on how many hours they work per day and no information on number of days worked a week, multiply the hours by 5 days.
* In the case of a question that has hours worked per month, divide by 4.3 to get weekly hours.

### wmonths

wmonths is a continuous variable that specifies the number of months worked in the last 12 months for the main job of any individual with a job (lstatus=1) and is missing otherwise. The main job is defined as that occupation that the person dedicated more time to over the past week. The variable is constructed for all persons administered this module in each questionnaire.

### wage\_total

wage\_total is a continuous variable that specifies the annualized wage payment (regular wage plus bonuses, in-kind, compensation, etc.) for the primary occupation in local currency of any individual (lstatus=1 & empstat=1) and is missing otherwise. The wage should come from the main job, in other words, the job that the person dedicated most time in the week preceding the survey. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. wage\_total should be equal to wage\_no\_compen in case there are no bonuses, tips etc. offered as part of the job. The variable is constructed for all persons administered this module in each questionnaire. The annualization of the wage\_total should consider the number of months/weeks the persons have been working and receiving this income. Harmonizer should not assume the person has been working the whole year. Box 4 shows the creation of wage\_total when there are no bonuses nor other compensations.

Box 4 - Example of wage\_total creation

|  |
| --- |
| gen double wage\_total=.  replace wage\_total=(wage\_no\_compen\*5\*4.3)\*wmonths if unitwage==1 //Wage in daily unit  replace wage\_total=(wage\_no\_compen\*4.3)\*wmonths if unitwage==2 //Wage in weekly unit  replace wage\_total=(wage\_no\_compen\*2.15)\*wmonths if unitwage==3 //Wage in every two weeks unit  replace wage\_total=(wage\_no\_compen)/2\*wmonths if unitwage==4 //Wage in every two months unit  replace wage\_total=( wage\_no\_compen)\*wmonths if unitwage==5 //Wage in monthly unit  replace wage\_total=( wage\_no\_compen)/3\*wmonths if unitwage==6 //Wage in every quarterly unit  replace wage\_total=(wage\_no\_compen)/6\*wmonths if unitwage==7 //Wage in every six months unit  replace wage\_total= wage\_no\_compen/12\*wmonths if unitwage==8 //Wage in annual unit  replace wage\_total=(wage\_no\_compen\*whours\*4.3)\*wmonths if unitwage==9 //Wage in hourly unit |

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### contract

contract is a dummy variable that classifies the contract status (yes/no) of any individual with a job (lstatus=1) and is missing otherwise. It indicates whether a person has a signed (formal) contract, regardless of duration. The variable is constructed for all persons administered this module in each questionnaire. Two categories after harmonization:

0 = No

1 = Yes

### healthins

healthins is a dummy variable that classifies the health insurance status (yes/no) of any individual with a job (lstatus=1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. However, this variable is only constructed if there is an explicit question about health insurance provided by the job. Two categories after harmonization:

0 = No

1 = Yes

### socialsec

socialsec is a dummy variable that classifies the social security status (yes/no) of any individual with a job (lstatus=1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. However, this variable is only constructed if there is an explicit question about pension plans or social security. Two categories after harmonization:

0 = No

1 = Yes

### union

union is a dummy variable that classifies the union membership status (yes/no) of any individual with a job (lstatus=1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. However, this variable is only constructed if there is an explicit question about trade unions. Two categories after harmonization:

0 = No

1 = Yes

### firmsize\_l

firmsize\_l specifies the lower bracket of the firm size. The variable is constructed for all persons who are employed in the last 7 days for the main job. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the lower boundary of the bracket.

### firmsize\_u

firmsize\_u specifies the upper bracket of the firm size. The variable is constructed for all persons who are employed in the last 7 days for the main job. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the upper boundary of the bracket. If the right bracket is open, this variable should be missing.

* + 1. Secondary Employment, 7-day reference period

### empstat\_2

empstat\_2 is a categorical variable that specifies employment status of the secondary job with reference period of last 7 days of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all individuals. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

The definitions are taken from the International Labor Organization’s Classification of Status in

Employment with some revisions to take into account the data available. Five categories after harmonization:

1 = Paid Employee

2 = Non-Paid Employee

3 = Employer

4 = Self-employed

5 = Other, workers not classifiable by status

1 = Paid Employee

Paid employee includes anyone whose basic remuneration is not directly dependent on the revenue of the unit they work for, typically remunerated by wages and salaries but may be paid for piece work or in-kind. The ‘continuous’ criteria used in the ILO definition is not used here as data are often absent and due to country specificity.

2 = Non-Paid Employee

Non-paid employee includes contributing family workers who hold a self-employment job in a market-oriented establishment operated by a related person living in the same households who cannot be regarded as a partner because of their degree of commitment to the operation of the establishment, in terms of working time or other factors, is not at a level comparable to that of the head of the establishment. All apprentices should be mapped as non-paid employee.

3 = Employer

Employer is a business owner (whether alone or in partnership) with employees. If the only people working in the business are the owner and ‘contributing family workers, the person is not considered an employer (as has no employees) and is, instead classified as own account.

4 = Self-employed

Own account or self-employment includes jobs are those where remuneration is directly dependent from the goods and service produced (where home consumption is considered to be part of the profits) and have not engaged any permanent employees to work for them on a continuous basis during the reference period.

Members of producers’ cooperatives are workers who hold a self-employment job in a cooperative producing goods and services in which each member takes part on an equal footing with other members in determining the organization of production, sales and/or other work of the establishment, the investments and the distribution of the proceeds of the establishment amongst the members.

5 = Other, workers not classifiable by status

Other, workers not classifiable by status include those for whom insufficient relevant information is available and/or who cannot be included in any of the above categories.

### ocusec\_2

ocusec\_2 is a categorical variable that specifies the sector of activity in the last 7 days. It classifies the secondary job's sector of activity of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire.

Four categories after harmonization:

1 = Public sector, Central Government, Army (including armed forces)

2 = Private, NGO

3 = State-owned

4 = Public or State-owned, but cannot distinguish

1 = Public Sector, Central Government, Army (including armed forces) Public sector is the part of economy run by the government.

2 = Private, NGO

Private sector is that part of the economy which is both run for private profit and is not controlled by the state, it also includes non-governmental organizations

3 = State-owned enterprises

State-owned includes para-state firms and all others in which the government has control (participation over 50%).

4 = Public or State-owned, but cannot distinguish

Select this option is the questionnaire does not ask for State-owned enterprises, and only for Public sector.

Notes: Do not code basis of occupation (ISCO) or industry (ISIC) codes.

### industry\_orig\_2

industry\_orig\_2 is a string variable that specifies the original industry codes for the second job with reference period of the last 7 days and should correspond to whatever is in the original file with no recoding. Do not put missing values for people below the working age. Other missing values are allowed. It classifies the main job of any individual with a job (lstatus=1) and is missing otherwise.

### industrycat\_isic\_2

Code (string variable) of the industry according to the International Standard Industry Classification (ISIC) in the last 7 days for the second job of any individual with a job (lstatus=1) and is missing otherwise.

See industrycat\_isic for the details.

### industrycat10\_2

industrycat10\_2 is a categorical variable that specifies the 1-digit industry classification that classifies the second job with reference period of the last 7 days of any individual with a job (lstatus=1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire. The codes for the second job are given here based on the UN International Standard Industrial Classification.

Ten categories after harmonization:

1 = Agriculture, Hunting, Fishing, etc.

2 = Mining

3 = Manufacturing

4 = Public Utility Services

5 = Construction

6 = Commerce

7 = Transport and Communications

8 = Financial and Business Services

9 = Public Administration

10 = Other Services, Unspecified

Notes:

* In the case of different classifications (former Soviet Union republics, for example), recoding has been done to best match the ISIC codes.
* Category 10 is also assigned for unspecified categories or items.
* For details on how to code different ISIC versions to industrycat10\_2 see the industrycat10 entry.

### industrycat4\_2

industrycat4\_2 is a categorical variable that specifies the 1-digit industry classification for Broad Economic Activities for the second job with reference period of the last 7 days. This variable is either created directly from the data (if industry classification does not exist for 10 categories) or created from industrycat10\_2.

Four categories after harmonization:

1 = Agriculture

2= Industry

3 = Services

4 = Other

This variable is either created directly from the data (if industry classification does not exist for 10 categories) or created from industrycat10.

### occup\_orig\_2

occup\_orig\_2 is a string variable that specifies the original occupation code in the last 7 days for the secondary job. This variable corresponds to whatever is in the original file with no recoding.

### occup\_isco\_2

Code (string variable) of the occupation according to the International Standard Classification of Occupations (ISCO) in the last 7 days for the second job of any individual with a job (lstatus=1) and is missing otherwise.

See occup\_isco for the details.

### occup\_skill\_2

Categorical code for the broad skill level of workers at the second job in the last 7 days of any individual with a job (lstatus=1) and is missing otherwise.

See occup\_skill for details.

### occup\_2

occup\_2 is a categorical variable that specifies the 1-digit occupation classification. It classifies the second job of any individual with a job (lstatus=1) and is missing otherwise. This variable is constructed for all persons administered this module in each questionnaire. Most surveys collect detailed information and then code it, without keeping the original data. No attempt has been made to correct or check the original coding. The classification is based on the International Standard Classification of Occupations (ISCO). In the case of different classifications, re-coding has been done to best match the ISCO.

Eleven categories after harmonization:

1 = Managers

2 = Professionals

3 = Technicians and associate professionals

4 = Clerical support workers

5 = Service and sales workers

6 = Skilled agricultural, forestry and fishery workers

7 = Craft and related trades workers

8 = Plant and machine operators, and assemblers

9 = Elementary occupations

10 = Armed forces occupations

99 = Other/unspecified

### wage\_no\_compen\_2

wage\_no\_compen\_2 is a continuous variable that specifies the last wage payment in local currency of any individual (lstatus=1 & empstat=1) in its secondary occupation and is missing otherwise. The wage should come from the second job, in other words, the job that the person dedicated the second most amount of time in the week preceding the survey. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* For all those with self-employment or owners of own businesses, this should be net revenues (net of all costs EXCEPT for tax payments) or the amount of salary taken from the business. Due to the almost complete lack of information on taxes, the wage from main job is NOT net of taxes.
* By definition, non-paid employees (empstat\_2=2) should have wage=0.
* The reference period of the wage\_no\_compen\_2 will be recorded in the unitwage\_2 variable
* Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### unitwage\_2

unitwage\_2 is a categorical variable that specifies the time reference for the wage\_no\_compen\_2 variable. It specifies the time unit measurement for the wages for the secondary job of any individual (lstatus=1 & empstat=1) and is missing otherwise.

Ten categories after harmonization:

1 = Daily

2 = Weekly

3 = Every two weeks

4 = Every two months

5 = Monthly

6 = Quarterly

7 = Every six months

8 = Annually

9 = Hourly

10 = Other

### whours\_2

whours\_2 is a continuous variable that specifies the hours of work in last week for the second job with reference period of the last 7 days of any individual with a job (lstatus=1) and is missing otherwise. The

second job defined as that occupation that the person dedicated the second most amount of time to over the past week. The variable is constructed for all persons administered this module in each questionnaire. The lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

Notes:

* If the respondent was absent from the job in the week preceeding the survey due to holidays, vacation, or sick leave, then record the time worked in the previous 7 days that the person worked.
* Sometimes the questions are phrased as, “on average, how many hours a week do you work?”.
* For individuals who only give information on how many hours they work per day and no information on number of days worked a week, multiply the hours by 5 days.
* In the case of a question that has hours worked per month, divide by 4.3 to get weekly hours.

### wmonths\_2

wmonths\_2 is a continuous variable that specifies the number of months worked in the last 12 months for the secondary job of any individual with a job (lstatus=1) and is missing otherwise. The secondary job is defined as that occupation in which the person dedicated less time than the primary job over the past week. The variable is constructed for all persons administered this module in each questionnaire.

### wage\_total\_2

wage\_total\_2 is a continuous variable that specifies the annualized wage payment (regular wage plus bonuses, in-kind, compensation, etc.) in local currency of any individual (lstatus=1 & empstat=1) in its secondary occupation and is missing otherwise. The wage should come from the secondary job, in other words, the job that the person dedicated the second most amount of time in the week preceding the survey. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. wage\_total\_2 should be equal to wage\_no\_compen\_2 in case there are no bonuses, tips etc. offered as part of the job. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* The annualization of the wage\_total\_2 should consider the number of months/weeks the persons have been working and receiving this income. Harmonizer should not assume the person has been working the whole year.
* For an example on how to annualize wage see 5.2.2 Primary Employment last 7-days variable wage\_total (Example: Creation of wage\_total when there are no bonuses nor other compensations).

### firmsize\_l\_2

firmsize\_l\_2 specifies the lower bracket of the firm size. The variable is constructed for all persons who are employed. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the lower boundary of the bracket.

### firmsize\_u\_2

firmsize\_u\_2 specifies the upper bracket of the firm size. The variable is constructed for all persons who are employed. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the upper boundary of the bracket. If the right bracket is open, a missing value should be inputted.

* + 1. Other Employment, 7-day reference period

### t\_hours\_others

t\_hours\_others is a continuous variable that specifies the hours of work in last 12 months in all jobs excluding the primary and secondary ones.

### t\_wage\_nocompen\_others

t\_wage\_nocompen\_others is a continuous variable that specifies the annualized wage in all jobs excluding the primary and secondary ones. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments.

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### t\_wage\_others

t\_wage\_others is a continuous variable that specifies the annualized wage in all jobs excluding the primary and secondary ones. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. wage\_others should be equal to wage\_nocompen\_ others in case there are no bonuses, tips etc. offered as part of any of the jobs.

* + 1. Total Employment Earnings, 7-day reference period

### t\_hours\_total

t\_hours\_total is a continuous variable that specifies the hours of work in last 12 months in all jobs including primary, secondary and others.

### t\_wage\_nocompen\_total

t\_wage\_nocompen\_total is a continuous variable that specifies the total annualized wage income in all jobs including primary, secondary and others. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments.

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### t\_wage\_total

t\_wage\_total is a continuous variable that specifies the total annualized wage income in all jobs including primary, secondary and others. This income includes tips, compensations such as bonuses, dwellings or clothes, and other payments. t\_wage\_total should be equal to t\_wage\_nocompen\_total in case there are no bonuses, tips etc. offered as part of any of the jobs. If the number of months worked in this job is missing the harmonizer could assumed that the person worked the whole year in this job.

* + 1. Labor status, 12-month reference period

### lstatus\_year

lstatus\_year is an individual’s labor status in the last 12 months. The Value must be missing for individuals less than the required age (minlaborage).

Three categories are used after harmonization:

1 = Employed

2 = Unemployed

3 = Not-in-labor force

All persons are considered active in the labor force if they presently have a job (formal or informal, i.e., employed) or do not have a job but are actively seeking work (i.e., unemployed).

1 = Employed

Employed is defined as anyone who worked during the last 12 months or reference week, regardless of whether the employment was formal or informal, paid or unpaid, for a minimum of 1 hour. Individuals who had a job, but for any reason did not work in the last 7 days are considered employed.

2 = Unemployed

A person is defined as unemployed if he or she is, presently not working but is actively seeking a job. The formal definition of unemployed usually includes being ‘able to accept a job.’ This last question was asked in a minority of surveys and is, thus, not incorporated in the present definition. A person presently not working but waiting for the start of a new job is considered unemployed.

3 = Not-in-labor force

A person is defined as not-in-labor force if he or she is, presently not working and it is not actively seeking a job during the last 12 months or reference week.

### potential\_lf\_year

Codifies whether the person is not in the labour force over the past year (lstatus\_year=3, missing otherwise) but could potentially be they are i) available but not searching or ii) searching but not immediately available to work. The codes are:

0 = No (not potentially in the labour force)

1 = Yes

### underemployment\_year

Codifies whether the person was in the labour force and working in the past 12 months (lstatus\_year=1, missing otherwise) but would take on more jobs or more hours at their job if possible/available. The codes are:

0 = No (not underemployed)

1 = Yes

### nlfreason\_year

nlfreason\_year is the reason an individual was not in the labor force in the last 12 months. This variable is constructed for all those who are not presently employed and are not looking for work (lstatus\_year=3) and missing otherwise.

Five categories after harmonization:

1= Student (a person currently studying.)

2= Housewife (a person who takes care of the house, older people, or children)

3= Retired

4 = Disabled (a person who cannot work due to physical conditions)

5 = Other (a person does not work for any other reason)

Do not put missing values for people below the working age, employed, and unemployed. Other missing values allowed.

### unempldur\_l\_year

unempldur\_l\_year is a continuous variable specifying the duration of unemployment in months (lower bracket).

The variable is constructed for all unemployed persons (lstatus\_year=2, otherwise missing). If it is specified as continuous in the survey, it records the numbers of months in unemployment. If the ariable is categorical it records the lower boundary of the bracket.

Missing values are allowed for everyone who is not unemployed. Other missing values are also allowed.

### unempldur\_u\_year

unempldur\_u\_year is a continuous variable specifying the duration of unemployment in months (upper bracket).

The variable is constructed for all unemployed persons (lstatus\_year=2, otherwise missing). If it is specified as continuous in the survey, it records the numbers of months in unemployment. If the variable is categorical it records the upper boundary of the bracket. If the right bracket is open a missing value should be inputted.

Missing values are allowed for everyone who is not unemployed. Other missing values are also allowed. If the duration of unemployment is not reported as a range, but as continuous variables, the unempldur\_l\_year and unempldur\_u\_year variables will have the same value. If the high range is open- ended the unempldur\_u\_year variable will be missing.

* + 1. Primary Employment, 12-month reference period

### empstat\_year

empstat is a categorical variable that specifies the main employment status in the last 12 months of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all individuals. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

The definitions are taken from the International Labor Organization’s Classification of Status in

Employment with some revisions to take into account the data available. Five categories after harmonization:

1 = Paid Employee

2 = Non-Paid Employee

3 = Employer

4 = Self-employed

5 = Other, workers not classifiable by status

1 = Paid Employee

Paid employee includes anyone whose basic remuneration is not directly dependent on the revenue of the unit they work for, typically remunerated by wages and salaries but may be paid for piece work or in-kind. The ‘continuous’ criteria used in the ILO definition is not used here as data are often absent and due to country specificity.

2 = Non-Paid Employee

Non-paid employee includes contributing family workers who hold a self-employment job in a market-oriented establishment operated by a related person living in the same households who cannot be regarded as a partner because of their degree of commitment to the operation of the establishment, in terms of working time or other factors, is not at a level comparable to that of the head of the establishment. All apprentices should be mapped as ‘non-paid employee’

3 = Employer

An employer is a business owner (whether alone or in partnership) with employees. If the only people working in the business are the owner and contributing family workers, the person is not considered an employer (as has no employees) and is, instead classified as self-employed.

4 = Self-employed

Own account or self-employment includes jobs where remuneration is directly dependent from the goods and service produced (where home consumption is considered to be part of the profits) and where one has not engaged any permanent employees to work for them on a continuous basis during the reference period.

Members of producers’ cooperatives are workers who hold a self-employment job in a cooperative producing goods and services, in which each member takes part on an equal footing with other members in determining the organization of production, sales and/or other work of the establishment, the investments and the distribution of the proceeds of the establishment amongst the members.

5 = Other, workers not classifiable by status

Other, workers not classifiable by status include those for whom insufficient relevant information is available and/or who cannot be included in any of the above categories.

### ocusec\_year

ocusec\_year is a categorical variable that specifies the sector of activity in the last 12 months. It classifies the main job's sector of activity of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire.

Four categories after harmonization:

1 = Public sector, Central Government, Army (including armed forces)

2 = Private, NGO

3 = State-owned

4 = Public or State-owned, but cannot distinguish

1 = Public Sector, Central Government, Army (including armed forces)

Public sector is the part of economy run by the government.

2 = Private, NGO

Private sector is that part of the economy which is both run for private profit and is not controlled by the state, it also includes non-governmental organizations

3 = State-owned enterprises

State-owned includes para-state firms and all others in which the government has control (participation over 50%).

4 = Public or State-owned, but cannot distinguish

Select this option is the questionnaire does not ask for State-owned enterprises, and only for Public sector.

Notes: Do not code basis of occupation (ISCO) or industry (ISIC) codes.

### industry\_orig\_year

industry\_orig\_year is a string variable that specifies the original industry codes in the last 12 months for the main job provided in the survey (the actual question) and should correspond to whatever is in the original file with no recoding. It will contain missing values for people below the working age. Other missing values are allowed. It classifies the main job of any individual with a job (lstatus\_year =1) and is missing otherwise

### industrycat\_isic\_year

Code (string variable) of the industry according to the International Standard Industry Classification (ISIC) in the last 12 months for the main job of any individual with a job (lstatus\_year=1) and is missing otherwise.

See industrycat\_isic for the details.

### industrycat10\_year

industrycat10\_year is a categorical variable that specifies the 1-digit industry classification in the last 12 months for the main job of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire. The codes for the main job are given here based on the UN International Standard Industrial Classification. It classifies the main job of any individual with a job (lstatus\_year =1) and is missing otherwise

Ten categories after harmonization:

1 = Agriculture, Hunting, Fishing, etc.

2 = Mining

3 = Manufacturing

4 = Public Utility Services

5 = Construction

6 = Commerce

7 = Transport and Communications

8 = Financial and Business Services

9 = Public Administration

10 = Other Services, Unspecified Notes:

Notes:

* In the case of different classifications (former Soviet Union republics, for example), recoding has been done to best match the ISIC codes.
* Category 10 is also assigned for unspecified categories or items.
* If all 10 categories cannot be identified in the questionnaire create this variable as missing and proceed to create industrycat4\_year.
* For details on how to code different ISIC versions to industrycat10\_year see the industrycat10 entry.

### industrycat4\_year

industrycat4\_year is a categorical variable that specifies the 1-digit industry classification in the last 12 months for the main job for Broad Economic Activities. This variable is either created directly from the data (if industry classification does not exist for ten categories) or created from industrycat10\_year.

Four categories after harmonization:

1 = Agriculture

2= Industry

3 = Services

4 = Other

This variable is either created directly from the data (if industry classification does not exist for ten categories) or created from industrycat10\_year.

### occup\_orig\_year

occup\_orig\_year is a string variable that specifies the original occupation code in the last 12 months for the main job. This variable corresponds to whatever is in the original file with no recoding.

### occup\_isco\_year

Code (string variable) of the occupation according to the International Standard Classification of Occupations (ISCO) in the last 12 months days for the main job of any individual with a job (lstatus\_year=1) and is missing otherwise.

See occup\_isco for the details.

### occup\_skill\_year

Categorical code for the broad skill level of workers at the main job in the last 12 months of any individual with a job (lstatus\_year=1) and is missing otherwise.

See occup\_skill for details.

### occup\_year

occup\_year is a categorical variable that specifies the 1-digit occupational classification for the main job in the last 12 months of any individual with a job (lstatus\_year =1) and is missing otherwise. This variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. Most surveys collect detailed information and then code it, without keeping the original data, no attempt has been made to correct or check the original coding. The classification is based on the International Standard Classification of Occupations (ISCO). It classifies the main job of any individual with a job (lstatus\_year=1) and is missing otherwise.

Eleven categories after harmonization:

1 = Managers

2 = Professionals

3 = Technicians and associate professionals

4 = Clerical support workers

5 = Service and sales workers

6 = Skilled agricultural, forestry and fishery workers

7 = Craft and related trades workers

8 = Plant and machine operators, and assemblers

9 = Elementary occupations

10 = Armed forces occupations

99 = Other/unspecified

### wage\_no\_compen\_year

wage\_no\_compen\_year is a continuous variable that specifies the last wage payment in local currency of any individual (lstatus\_year =1 & empstat\_year =1) in its primary occupation at the reference period reported in the survey and it is missing otherwise. The wage should come from the main job, in other words, the job that the person dedicated most time in the 12 months preceding the survey. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* For all those with self-employment or owners of own businesses, this should be net revenues (net of all costs EXCEPT for tax payments) or the amount of salary taken from the business. Due to the almost complete lack of information on taxes, the wage from main job is NOT net of taxes.
* By definition, non-paid employees (empstat\_year=2) should have wage=0.
* The reference period of the wage\_no\_compen\_year will be recorded in the unitwage\_year variable.
* Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### unitwage\_year

unitwage\_year is a categorical variable that specifies the time reference for the wage\_no\_compen\_year variable. It specifies the time unit measurement for the wages of any individual (lstatus\_year =1 & empstat\_year =1) and it is missing otherwise. Acceptable values include:

1 = Daily

2 = Weekly

3 = Every two weeks

4 = Every two months

5 = Monthly

6 = Quarterly

7 = Every six months

8 = Annually

9 = Hourly

10 = Other

### whours\_year

whours\_year is a continuous variable that specifies the hours of work last week for the main job of any individual with a job (lstatus\_year =1) and is missing otherwise. The main job defined as that occupation that the person dedicated more time to over the past 12 months. The variable is constructed for all persons administered this module in each questionnaire.

Notes:

* Sometimes the questions are phrased as, “on average, how many hours a week do you work?”.
* For individuals who only give information on how many hours they work per day and no information on number of days worked a week, multiply the hours by 5 days.
* In the case of a question that has hours worked per month, divide by 4.3 to get weekly hours.

### wmonths\_year

wmonths\_year is a continuous variable that specifies the number of months worked in the last 12 months for the main job of any individual with a job (lstatus\_year =1) and is missing otherwise. The main job is defined as that occupation that the person dedicated more time to over the past 12 months. The variable is constructed for all persons administered this module in each questionnaire.

### wage\_total\_year

wage\_total\_year is a continuous variable that specifies the annualized wage payment (regular wage plus bonuses, in-kind, compensation, etc.) for the primary occupation in local currency of any individual (lstatus\_year =1 & empstat\_year =1) and is missing otherwise. The wage should come from the main job, in other words, the job that the person dedicated most time in the year preceding the survey. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. wage\_total\_year should be equal to wage\_no\_compen\_year in case there are no bonuses, tips etc. offered as part of the job. The variable is constructed for all persons administered this module in each questionnaire. The annualization of the wage\_total\_year should consider the number of months/weeks the persons have been working and receiving this income. Harmonizer should not assume the person has been working the whole year.

For an example on how to annualize wage see 5.2.2 Primary Employment last 7-days variable wage\_total (Example: Creation of wage\_total when there are no bonuses nor other compensations).

### contract\_year

contract\_year is a dummy variable that classifies the contract status (yes/no) of any individual with a job (lstatus\_year =1) and is missing otherwise. It indicates whether a person has a signed (formal) contract, regardless of duration. The variable is constructed for all persons administered this module in each questionnaire. Two categories after harmonization:

0 = No

1 = Yes

### healthins\_year

healthins\_year is a dummy variable that classifies the health insurance status (yes/no) of any individual with a job (lstatus\_year =1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. However, this variable is only constructed if there is an explicit question about health insurance provided by the job. Two categories after harmonization:

0 = No

1 = Yes

### socialsec\_year

socialsec\_year is a dummy variable that classifies the social security status (yes/no) of any individual with a job (lstatus\_year =1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. However, this variable is only constructed if there is an explicit question about pension plans or social security. Two categories after harmonization:

0 = No

1 = Yes

### union\_year

union\_year is a dummy variable that classifies the union membership status (yes/no) of any individual with a job (lstatus\_year =1) and is missing otherwise. Variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country. However, this variable is only constructed if there is an explicit question about trade unions. Two categories after harmonization:

0 = No

1 = Yes

### firmsize\_l\_year

firmsize\_l\_year specifies the lower bracket of the firm size. The variable is constructed for all persons who are employed in the last 12 months for the main job. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the lower boundary of the bracket.

### firmsize\_u\_year

firmsize\_u\_year specifies the upper bracket of the firm size. The variable is constructed for all persons who are employed in the last 12 months for the main job. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the upper boundary of the bracket. If the right bracket is open, this variable should be missing.

* + 1. Secondary Employment, 12-month reference period

### empstat\_2\_year

empstat\_2\_year is a categorical variable that specifies employment status of the secondary job with reference period of last 12 months of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all individuals. For this reason, the lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

The definitions are taken from the International Labor Organization’s Classification of Status in

Employment with some revisions to take into account the data available. Five categories after harmonization:

1 = Paid Employee

2 = Non-Paid Employee

3 = Employer

4 = Self-employed

5 = Other, workers not classifiable by status

1 = Paid Employee

Paid employee includes anyone whose basic remuneration is not directly dependent on the revenue of the unit they work for, typically remunerated by wages and salaries but may be paid for piece work or in-kind. The ‘continuous’ criteria used in the ILO definition is not used here as data are often absent and due to country specificity.

2 = Non-Paid Employee

Non-paid employee includes contributing family workers who hold a self-employment job in a market-oriented establishment operated by a related person living in the same households who cannot be regarded as a partner because of their degree of commitment to the operation of the establishment, in terms of working time or other factors, is not at a level comparable to that of the head of the establishment. All apprentices should be mapped as non-paid employee.

3 = Employer

Employer is a business owner (whether alone or in partnership) with employees. If the only people working in the business are the owner and ‘contributing family workers, the person is not considered an employer (as has no employees) and is, instead classified as own account.

4 = Self-employed

Own account or self-employment includes jobs are those where remuneration is directly dependent from the goods and service produced (where home consumption is considered to be part of the profits) and have not engaged any permanent employees to work for them on a continuous basis during the reference period.

Members of producers’ cooperatives are workers who hold a self-employment job in a cooperative producing goods and services in which each member takes part on an equal footing with other members in determining the organization of production, sales and/or other work of the establishment, the investments and the distribution of the proceeds of the establishment amongst the members.

5 = Other, workers not classifiable by status

Other, workers not classifiable by status include those for whom insufficient relevant information is available and/or who cannot be included in any of the above categories.

### ocusec\_2\_year

ocusec\_2\_year is a categorical variable that specifies the sector of activity in the last 12 months. It classifies the secondary job's sector of activity of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire.

Four categories after harmonization:

1 = Public sector, Central Government, Army (including armed forces)

2 = Private, NGO

3 = State-owned

4 = Public or State-owned, but cannot distinguish

1 = Public Sector, Central Government, Army (including armed forces)

Public sector is the part of economy run by the government.

2 = Private, NGO

Private sector is that part of the economy which is both run for private profit and is not controlled by the state, it also includes non-governmental organizations

3 = State-owned enterprises

State-owned includes para-state firms and all others in which the government has control (participation over 50%).

4 = Public or State-owned, but cannot distinguish

Select this option is the questionnaire does not ask for State-owned enterprises, and only for Public sector.

Notes: Do not code basis of occupation (ISCO) or industry (ISIC) codes.

### industry\_orig\_2\_year

industry\_orig\_2\_year is a string variable that specifies the original industry codes for the second job with reference period of the last 12 months and should correspond to whatever is in the original file with no recoding. Do not put missing values for people below the working age. Other missing values are allowed. It classifies the main job of any individual with a job (lstatus\_year=1) and is missing otherwise.

### industry\_isic\_2\_year

Code (string variable) of the industry according to the International Standard Industry Classification (ISIC) in the last 12 months for the second job of any individual with a job (lstatus\_year=1) and is missing otherwise.

See industrycat\_isic for the details.

### industrycat10\_2\_year

industrycat10\_2\_year is a categorical variable that specifies the 1-digit industry classification that classifies the second job with reference period of the last 12 months of any individual with a job (lstatus\_year =1) and is missing otherwise. The variable is constructed for all persons administered this module in each questionnaire. The codes for the second job are given here based on the UN International Standard Industrial Classification.

Ten categories after harmonization:

1 = Agriculture, Hunting, Fishing, etc.

2 = Mining

3 = Manufacturing

4 = Public Utility Services

5 = Construction

6 = Commerce

7 = Transport and Communications

8 = Financial and Business Services

9 = Public Administration

10 = Other Services, Unspecified

Notes:

* In the case of different classifications (former Soviet Union republics, for example), recoding has been done to best match the ISIC codes.
* Category 10 is also assigned for unspecified categories or items.
* For details on how to code different ISIC versions to industrycat10\_2\_year see the industrycat10 entry.

### industrycat4\_2\_year

industrycat4\_2\_year is a categorical variable that specifies the 1-digit industry classification for Broad Economic Activities for the second job with reference period of the last 12 months. This variable is either created directly from the data (if industry classification does not exist for 10 categories) or created from industrycat10\_year.

Four categories after harmonization:

1 = Agriculture

2= Industry

3 = Services

4 = Other

This variable is either created directly from the data (if industry classification does not exist for 10 categories) or created from industrycat10\_2\_year.

### occup\_orig\_2\_year

occup\_orig\_2\_year is a string variable that specifies the original occupation code in the last 12 months for the secondary job. This variable corresponds to whatever is in the original file with no recoding.

### occup\_isco\_2\_year

Code (string variable) of the occupation according to the International Standard Classification of Occupations (ISCO) for the second job of any individual with a job in the last 12 months (lstatus\_year=1) and is missing otherwise.

See occup\_isco for the details.

### occup\_skill\_2\_year

Categorical code for the broad skill level of workers at the second job in the last 12 months of any individual with a job (lstatus\_year=1) and is missing otherwise.

See occup\_skill for details.

### occup\_2\_year

occup\_2\_year is a categorical variable that specifies the 1-digit occupation classification. It classifies the second job of any individual with a job (lstatus\_year =1) and is missing otherwise. This variable is constructed for all persons administered this module in each questionnaire. Most surveys collect detailed information and then code it, without keeping the original data. No attempt has been made to correct or check the original coding. The classification is based on the International Standard Classification of Occupations (ISCO). In the case of different classifications, re-coding has been done to best match the ISCO.

Eleven categories after harmonization:

1 = Managers

2 = Professionals

3 = Technicians and associate professionals

4 = Clerical support workers

5 = Service and sales workers

6 = Skilled agricultural, forestry and fishery workers

7 = Craft and related trades workers

8 = Plant and machine operators, and assemblers

9 = Elementary occupations

10 = Armed forces occupations

99 = Other/unspecified

### wage\_no\_compen\_2\_year

wage\_no\_compen\_2\_year is a continuous variable that specifies the last wage payment in local currency of any individual (lstatus\_year =1 & empstat\_year =1) in its secondary occupation and is missing otherwise. The wage should come from the second job, in other words, the job that the person dedicated the second most amount of time in the week preceding the survey. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* For all those with self-employment or owners of own businesses, this should be net revenues (net of all costs EXCEPT for tax payments) or the amount of salary taken from the business. Due to the almost complete lack of information on taxes, the wage from main job is NOT net of taxes.
* By definition, non-paid employees (empstat\_year\_2 =2) should have wage=0.
* The reference period of the wage\_no\_compen\_year\_2 will be recorded in the unitwage\_2\_year variable.
* Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### unitwage\_2\_year

unitwage\_2\_year is a categorical variable that specifies the time reference for the wage\_no\_compen\_2\_year variable. It specifies the time unit measurement for the wages for the secondary job of any individual (lstatus\_year =1 & empstat\_year =1) and is missing otherwise.

Ten categories after harmonization:

1 = Daily

2 = Weekly

3 = Every two weeks

4 = Every two months

5 = Monthly

6 = Quarterly

7 = Every six months

8 = Annually

9 = Hourly

10 = Other

### whours\_2\_year

whours\_2\_year is a continuous variable that specifies the hours of work in last week for the second job with reference period of the last 12 months of any individual with a job (lstatus\_year =1) and is missing otherwise. The second job defined as that occupation that the person dedicated the second most amount of time to over the past year. The variable is constructed for all persons administered this module in each questionnaire. The lower age cutoff (and perhaps upper age cutoff) at which information is collected will vary from country to country.

Notes:

* Sometimes the questions are phrased as, “on average, how many hours a week do you work?”.
* For individuals who only give information on how many hours they work per day and no information on number of days worked a week, multiply the hours by 5 days.
* In the case of a question that has hours worked per month, divide by 4.3 to get weekly hours.

### wmonths\_2\_year

wmonths\_2\_year is a continuous variable that specifies the number of months worked in the last 12 months for the secondary job of any individual with a job (lstatus\_year =1) and is missing otherwise. The secondary job is defined as that occupation in which the person dedicated less time than the primary job over the past year. The variable is constructed for all persons administered this module in each questionnaire.

### wage\_total\_2\_year

wage\_total\_2\_year is a continuous variable that specifies the annualized wage payment (regular wage plus bonuses, in-kind, compensation, etc.) in local currency of any individual (lstatus\_year =1 & empstat\_year =1) in its secondary occupation and is missing otherwise. The wage should come from the secondary job, in other words, the job that the person dedicated the second most amount of time in the year preceding the survey. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. wage\_total\_2\_year should be equal to wage\_no\_compen\_2\_year in case there are no bonuses, tips etc. offered as part of the job. The variable is constructed for all persons administered this module in each questionnaire. For this reason, the lower age cutoff (and perhaps upper age cutoff) will vary from country to country.

Notes:

* The annualization of the wage\_total\_2\_year should consider the number of months/weeks the persons have been working and receiving this income. Harmonizer should not assume the person has been working the whole year.
* For an example on how to annualize wage see 5.2.2 Primary Employment last 7-days variable wage\_total (Example: Creation of wage\_total when there are no bonuses nor other compensations).

### firmsize\_l\_2\_year

firmsize\_l\_2\_year specifies the lower bracket of the firm size. The variable is constructed for all persons who are employed. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the lower boundary of the bracket.

### firmsize\_u\_2\_year

firmsize\_u\_2\_year specifies the upper bracket of the firm size. The variable is constructed for all persons who are employed. If it is continuous, it records the number of people working for the same employer. If the variable is categorical, it records the upper boundary of the bracket. If the right bracket is open, a missing value should be inputted.

* + 1. Other Employment, 12-month reference period

### t\_hours\_others\_year

t\_hours\_others\_year is a continuous variable that specifies the hours of work in last 12 months in all jobs excluding the primary and secondary ones.

### t\_wage\_nocompen\_others\_year

t\_wage\_nocompen\_others\_year is a continuous variable that specifies annual wage in all jobs excluding the primary and secondary ones. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments.

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### t\_wage\_others\_year

t\_wage\_others\_year is a continuous variable that specifies the annual wage in all jobs excluding the primary and secondary ones. This wage includes tips, compensations such as bonuses, dwellings or clothes, and other payments. t\_wage\_others should be equal to t\_wage\_nocompen\_ others in case there are no bonuses, tips etc. offered as part of any of the jobs.

* + 1. Total Employment Earnings, 12-month reference period

### t\_hours\_total\_year

t\_hours\_total\_year is a continuous variable that specifies the hours of work in last 12 months in all jobs including primary, secondary and others.

### t\_wage\_nocompen\_total\_year

t\_wage\_nocompen\_total\_year is a continuous variable that specifies the total annualized wage income in all jobs including primary, secondary and others. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments.

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### t\_wage\_total\_year

t\_wage\_total\_year is a continuous variable that specifies the total annualized wage income in all jobs including primary, secondary and others. This income includes tips, compensations such as bonuses, dwellings or clothes, and other payments. t\_wage\_total should be equal to t\_wage\_nocompen\_total in case there are no bonuses, tips etc. offered as part of any of the jobs.

* + 1. Total Labor Income

### njobs

njobs is a numeric variable that specifies the total number of jobs.

### t\_hours\_annual

t\_hours\_annual is a continuous variable that specifies the annual numbers of hours worked in all the jobs including primary, secondary and others regardless of their period of reference.

### linc\_nc

linc\_nc is a continuous variable that specifies the total annualized wage income in all the jobs including primary, secondary and others regardless of their period of reference. This excludes tips, bonuses, other compensation such as dwellings or clothes, and other payments.

Note: Use gross wages when available and net wages only when gross wages are not available. This is done to make it easy to compare earnings in formal and informal sectors.

### laborincome

laborincome is a continuous variable that specifies the total annualized individual labor income in all jobs including primary, secondary and others regardless of their period of reference. This income includes tips, compensations such as bonuses, dwellings or clothes, and other payments. This variable should be used as the total annual labor income of an individual.

## Lessons Learned and Challenges

For any variable not collected in a country, the variable should be created and left as missing (.) in the final harmonized file.

Variables in the data files must follow the sequence in which they appear in the manual.

Labor status during last 12 months only reflects if the person has worked during last year or not, as the vast majority of the surveys do not provide enough information to distinguish those individuals that are unemployed from those that are out of the labor force.

Individuals working in cooperatives are considered as “paid employee” in the employment status

variable.

Several checks should be conducted to ensure that the data is harmonized correctly. lstatus should be an integer in the range [1,3].

lstatus<0 & lstatus>3 & mod(lstatus, 1) == 1

If lstatus is classified as employed then the employment type needs to be defined.

lstatus==1 & empstat>5

minlaborage should be an integer.

mod(minlaborage,1)==1

The minimum age for employment should not be higher than 20.

minlaborage >20 & minlaborage <.

empstat should be an integer in the range [1,5].

mod(empstat, 1) != 0 & empstat<1 & empstat>5

If employment type is defined then labor force status should be employed.

empstat<=5 & lstatus!=1

industrycat10 should be an integer in the range [1,10].

industrycat10<0 & industrycat10>10 & mod(industrycat10, 1) == 1

industrycat4 should be an integer in the range [1,4].

industrycat4<0 & industrycat4>4 & mod(industrycat4, 1) == 1

There should not be a mismatch between industry and industrycat4.

((industrycat4==1 & industrycat10!=1 ) | (industrycat4==2 & (industrycat10 <2 | industrycat10 >5)) | (industrycat4==3 & (industrycat10 <6 | industrycat10 >9)) | (industrycat4==1 & industrycat10 !=1 ) ) & industrycat10 !=.

firmsize\_u should not be lower than firmsize\_l

firmsize\_u < firmsize\_l

Conversion of given wage to annual wage. Example: Creation of wage\_total when there are no bonuses nor other compensations

gen double wage\_total=.

\* Wage in daily unit

replace wage\_total=(wage\_no\_compen\*5\*4.3)\*wmonths if unitwage==1

\* Wage in weekly unit

replace wage\_total=(wage\_no\_compen\*4.3)\*wmonths if unitwage==2

\* Wage in every two weeks unit

replace wage\_total=(wage\_no\_compen\*2.15)\*wmonths if unitwage==3

\* Wage in every two months

replace wage\_total=(wage\_no\_compen)/2\*wmonths if unitwage==4

\* Wage in monthly unit

replace wage\_total=( wage\_no\_compen)\*wmonths if unitwage==5

\* Wage in every quarterly unit

replace wage\_total=( wage\_no\_compen)/3\*wmonths if unitwage==6

\* Wage in every six months unit

replace wage\_total=( wage\_no\_compen)/6\*wmonths if unitwage==7

\* Wage in annual unit

replace wage\_total= wage\_no\_compen/12\*wmonths if unitwage==8

\* Wage in hourly unit

replace wage\_total=(wage\_no\_compen\*whours\*4.3)\*wmonths if unitwage==9

## Tabular Overview of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | minlaborage | Labor module application age |  |
| Labor | lstatus | Labor status (7-day ref period) | In some GMD harmonizations (not in their dictionaries) this is given as lstatus\_7. In general, the 7 day reference ones are then for example ocusec\_7 and ocusec\_2. This is then not neat with ocusec\_year, ocusec\_year\_2. Either ocusec\_week, ocusec\_year or ocusec\_7, ocusec\_365. |
| Labor | potential\_lf | Potential labour force (7-day ref period) | A binary indicator taking a value only if the person is not in the labour force (missing if in LF or unemployed). Codes 1 if i) available but not searching or ii) searching but not immediately available to work. Codes 0 otherwise. |
| Labor | underemployment | Underemployment (7-day ref period) | A binary indicator taking value only if the person is in the labour force and working (missing if not or unemployed). Codes 1 if person would take on more jobs or more hours at their job if possible/available, 0 otherwise. |
| Labor | nlfreason | Reason not in the labor force  (7-day ref period) |  |
| Labor | unempldur\_l | Unemployment duration (months)  lower bracket (7-day ref period) |  |
| Labor | unempldur\_u | Unemployment duration (months) upper bracket (7-day ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | empstat | Employment status, primary job (7-day ref period) |  |
| Labor | ocusec | Sector of activity, primary job (7-day ref period) | NGOs were classified in I2D2 as public sector, switched to private. |
| Labor | industry\_orig | Original industry code, primary job (7-day ref period) |  |
| Labor | industrycat\_isic | ISIC code of the classification, primary job (7-day ref period) | Code of ISIC - 4 for example is  Q - Human health and social work activities  86 Human health activities  861  8610 Hospital activities  8620 Medical dental practice activities  The four options would be coded as "Q", 8600, 8610, 8620 respectively. |
| Labor | industrycat10 | 1-digit industry classification, primary job (7-day ref period) |  |
| Labor | industrycat4 | 4-category industry classification, primary job (7-day ref period) |  |
| Labor | occup\_orig | Original occupational classification, primary job (7-day ref period) |  |
| Labor | occup\_isco | ISCO code of the classification, primary job (7-day ref period) | ISCO-08 codes are  8 Plant and machine operators, and assemblers  81 Stationary plant machine operators  811 Mining and mineral processing plant operators  8111 Miners and quarriers  Given the level of detail available, this are to be coded as 8000, 8100, 8110, and 8111 respectively. |
| Labor | occup\_skill | Skill level based on ISCO standard | https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/ |
| Labor | occup | 1 digit occupational classification, primary job (7-day ref period) |  |
| Labor | wage\_no\_compen | Last wage payment, primary job, excl. bonuses, etc. (7-day ref period) |  |
| Labor | unitwage | Time unit of last wages  payment, primary job (7-day ref period) |  |
| Labor | whours | Hours of work in last week, primary job (7-day ref period) |  |
| Labor | wmonths | Months worked in the last 12 months, primary job (7-day ref period) | This definition may appear confusing since it is months out of the past 12 months of work for the 7 day recall and there is a wmonths\_year variable for the 12 month recall. It is not clearly defined in the guidelines, yet I would read it as the number of months in the main job for the 7 day recall job, which would be fewer than the wmonths\_year number if the person switched jobs say 2 months ago, had the previous one since over a year. |
| Labor | wage\_total | Annualized total wage, primary job (7-day ref period) |  |
| Labor | contract | Contract (7-day ref period) |  |
| Labor | healthins | Health insurance (7-day ref period) |  |
| Labor | socialsec | Social security (7-day ref period) |  |
| Labor | union | Union membership (7-day ref period) |  |
| Labor | firmsize\_l | Firm size (lower bracket), primary job (7-day ref period) |  |
| Labor | firmsize\_u | Firm size (upper bracket), primary job (7-day ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | empstat\_2 | Employment status, secondary job (7-day ref period) |  |
| Labor | ocusec\_2 | Sector of activity, secondary job (7-day ref period) |  |
| Labor | industry\_orig\_2 | Original industry code, secondary job (7-day ref period) |  |
| Labor | industrycat\_isic\_2 | ISIC code of the classification, secondary job (7-day ref period) | See industrycat\_isic |
| Labor | industrycat10\_2 | 1 digit industry classification, secondary job (7-day ref period) |  |
| Labor | industrycat4\_2 | 4-category industry classification, secondary job (7-day ref period) |  |
| Labor | occup\_orig\_2 | Original occupational classification, secondary job (7-day ref period) |  |
| Labor | occup\_isco\_2 | ISCO code of the classification, secondary job (7-day ref period) | See occup\_isco |
| Labor | occup\_skill\_2 | Skill level based on ISCO standard | https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/ |
| Labor | occup\_2 | 1 digit occupational classification, secondary job (7-day ref period) |  |
| Labor | wage\_no\_compen\_2 | wage payment, secondary job, excl. bonuses, etc. (7-day ref period) |  |
| Labor | unitwage\_2 | Time unit of last wages payment, secondary job (7-day ref period) |  |
| Labor | whours\_2 | Hours of work in last week, secondary job (7-day ref period) |  |
| Labor | wmonths\_2 | Months worked in the last 12 months, secondary job (7-day ref period) | See note on wmonths |
| Labor | wage\_total\_2 | Annualized total wage, secondary job (7-day ref period) |  |
| Labor | firmsize\_l\_2 | Firm size (lower bracket), secondary job (7-day ref period) |  |
| Labor | firmsize\_u\_2 | Firm size (upper bracket), secondary job (7-day ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | t\_hours\_others | Total hours of work in the last 12 months in other jobs excluding the primary and secondary ones |  |
| Labor | t\_wage\_nocompen\_others | Annualized wage in all jobs excluding the primary and secondary ones (excluding tips, bonuses, etc.). |  |
| Labor | t\_wage\_others | Annualized wage (including tips, bonuses, etc.) in all other jobs excluding the primary and secondary ones. (7 day ref). |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | t\_hours\_total | Annualized hours worked in all jobs (7-day ref period) |  |
| Labor | t\_wage\_nocompen\_total | Annualized wage in all jobs excl. bonuses, etc. (7-day ref period) |  |
| Labor | t\_wage\_total | Annualized total wage for all jobs (7-day ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | lstatus\_year | Labor status (12-mon ref period) |  |
| Labor | potential\_lf\_year | Potential labour force (12-mon ref period) | A binary indicator taking a value only if the person is not in the labour force (missing if in LF or unemployed). Codes 1 if i) available but not searching or ii) searching but not immediately available to work. Codes 0 otherwise. |
| Labor | underemployment\_year | Underemployment (12-mon ref period) | A binary indicator taking value only if the person is in the labour force and working (missing if not or unemployed). Codes 1 if person would take on more jobs or more hours at their job if possible/available, 0 otherwise. |
| Labor | nlfreason\_year | Reason not in the labor force (12-mon ref period) |  |
| Labor | unempldur\_l\_year | Unemployment duration (months) lower bracket (12-mon ref period) |  |
| Labor | unempldur\_u\_year | Unemployment duration (months) upper bracket (12-mon ref period) |  |
| Labor | empstat\_year | Employment status, primary job (12-mon ref period) |  |
| Labor | ocusec\_year | Sector of activity, primary job (12-mon ref period) |  |
| Labor | industry\_orig\_year | Original industry code, primary job (12-mon ref period) |  |
| Labor | industrycat\_isic\_year | ISIC code of the classification, primary job (12 month ref period) | See industrycat\_isic |
| Labor | industrycat10\_year | 1 digit industry classification, primary job (12-mon ref period) |  |
| Labor | industrycat4\_year | 4-category industry classification primary job (12-mon ref period) |  |
| Labor | occup\_orig\_year | Original occupational classification, primary job (12-mon ref period) |  |
| Labor | occup\_isco\_year | ISCO code of the classification, primary job (12 month ref period) | See occup\_isco |
| Labor | occup\_skill\_year | Skill level based on ISCO standard | https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/ |
| Labor | occup\_year | 1 digit occupational classification, primary job (12-mon ref period) |  |
| Labor | wage\_no\_compen\_year | Last wage payment, primary job, excl. bonuses, etc. (12-mon ref period) |  |
| Labor | unitwage\_year | Time unit of last wages payment, primary job (12-mon ref period) |  |
| Labor | whours\_year | Hours of work in last week, primary job (12-mon ref period) |  |
| Labor | wmonths\_year | Months worked in the last 12 months, primary job (12-mon ref period) |  |
| Labor | wage\_total\_year | Annualized total wage, primary job (12-mon ref period) |  |
| Labor | contract\_year | Contract (12-mon ref period) |  |
| Labor | healthins\_year | Health insurance (12-mon ref period) |  |
| Labor | socialsec\_year | Social security (12-mon ref period) |  |
| Labor | union\_year | Union membership (12-mon ref period) |  |
| Labor | firmsize\_l\_year | Firm size (lower bracket), primary job (12-mon ref period) |  |
| Labor | firmsize\_u\_year | Firm size (upper bracket), primary job (12-mon ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | empstat\_2\_year | Employment status, secondary job (12-mon ref period) |  |
| Labor | ocusec\_2\_year | Sector of activity, secondary job (12-mon ref period) |  |
| Labor | industry\_orig\_2\_year | Original industry code, secondary job (12-mon ref period) |  |
| Labor | industrycat\_isic\_2\_year | ISIC code of the classification, secondary job (12 month ref period) | See industrycat\_isic |
| Labor | industrycat10\_2\_year | 1 digit industry classification, secondary job (12-mon ref period) |  |
| Labor | industrycat4\_2\_year | 4-category industry classification, secondary job (12-mon ref period) |  |
| Labor | occup\_orig\_2\_year | Original occupational classification, secondary job (12-mon ref period) |  |
| Labor | occup\_isco\_2\_year | ISCO code of the classification, seconddary job (12 month ref period) | See occup\_isco |
| Labor | occup\_skill\_2\_year | Skill level based on ISCO standard | https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/ |
| Labor | occup\_2\_year | 1 digit occupational classification, secondary job (12-mon ref period) |  |
| Labor | wage\_no\_compen\_2\_year | last wage payment, secondary job, excl. bonuses, etc. (12-mon ref period) |  |
| Labor | unitwage\_2\_year | Time unit of last wages payment, secondary job (12-mon ref period) |  |
| Labor | whours\_2\_year | Hours of work in last week, secondary job (12-mon ref period) |  |
| Labor | wmonths\_2\_year | Months worked in the last 12 months, secondary job (12-mon ref period) |  |
| Labor | wage\_total\_2\_year | Annualized total wage, secondary job (12-mon ref period) |  |
| Labor | firmsize\_l\_2\_year | Firm size (lower bracket), secondary job (12-mon ref period) |  |
| Labor | firmsize\_u\_2\_year | Firm size (upper bracket), secondary job (12-mon ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | t\_hours\_others\_year | Annualized hours worked in all but primary and secondary jobs (12-mon ref period) |  |
| Labor | t\_wage\_nocompen\_others\_year | Annualized wage in all but primary & secondary jobs excl. bonuses, etc. (12-mon ref period) |  |
| Labor | t\_wage\_others\_year | Annualized wage in all but primary and secondary jobs (12-mon ref period) |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | t\_hours\_total\_year | Annualized hours worked in all jobs (12-mon ref period) | t\_hours\_total\_year |
| Labor | t\_wage\_nocompen\_total\_year | Annualized wage in all jobs excl. bonuses, etc. (12-mon ref period) | t\_wage\_nocompen\_total\_year |
| Labor | t\_wage\_total\_year | Annualized total wage for all jobs (12-mon ref period) | t\_wage\_total\_year |

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Variable name | Variable label | Notes |
| Labor | njobs | Total number of jobs | Total Labor income will be created based on either the 7 days or 12 months reference period variables or a combination of both. Harmonizers should make sure that all jobs are included and none of them are double counted. |
| Labor | t\_hours\_annual | Total hours worked in all jobs in the previous 12 months |  |
| Labor | linc\_nc | Total annual wage income in all jobs, excl. bonuses, etc. | Difference to t\_wage\_nocompen\_total\_year? |
| Labor | laborincome | Total annual individual labor income in all jobs, incl. bonuses, etc. | Difference to t\_wage\_total\_year? |