Global Labor Database User Manual

A guide to understanding, using, and interacting
with the Global Labor Database

WORLD BANK GROUP

May 15, 2024

Table of Contents

1. Introduction to the GLD

1.1. What is GLD?

The Global Labor Database (GLD) is part of the World Bank initiatives to harmonize labor force surveys and household surveys with a relevant labor module. Its mission is to create an open and transparent harmonization with sufficient background information to allow data analysts to use, alter, and expand the harmonization. In this sense, background information goes beyond code, questionnaires, and reports, and includes documenting survey details learned during harmonization which are not recorded elsewhere. An example of this documenting changes to the currency or the administrative divisions.

The GLD aims to be an open-source database, meaning that as much information should be accessible to as many people as possible. It also strives to be transparent, making all steps that create the harmonization traceable, from raw data acquisition to harmonized variable coding. Hence, all steps of the harmonization process are documented and made available, including the survey documentation, code and notes that allow users to fully comprehend the survey design and the choices made in the harmonization. The availability of the codes and documentation enables users to customize and add variables not in the GLD harmonization. Most harmonization efforts provide users with a "take it or leave it" option, but the GLD's open and transparent approach allows users to trace and deviate from the standard harmonization at any point, giving them a head start regardless of where they wish to jump in.

Finally, the GLD follows up and expands on the previous initiative to harmonized household surveys, the International Income Distribution Database (I2D2). The I2D2 was superseded by the Global Monitoring Database (GMD), which however focused on household budget surveys and did not harmonize labor force surveys. The GLD was created to remedy this gap in the survey type coverage and complement it, with a stronger focus on labor market information through an expanded dictionary and more rigorous validation of labor indicators.

1.2. What is the objective of GLD?

Labor force surveys represent a critical data source to generate key labor market indicators disaggregated by individual characteristics that policy makers monitor, target, and evaluate. Appended across time and space, they are used for comparison and benchmarking. The objective of GLD is to make the process of producing these estimates easier, traceable, and reproducible for World Bank staff and researchers worldwide.

One major issue when generating survey-based indicators is the cross-country comparability and the time-consuming process of harmonization, which requires reading both data files and survey materials in detail to understand what to code and how, a structured and consistent harmonization methodology as well as many steps of validation.

The first objective of GLD is thus to create a database of harmonized surveys with comprehensive and reliable labor market information that can be used in analytical work for cross-country and over time comparisons. By creating a harmonized output, this database can be fed into other products that automate analytical processes like country level jobs diagnostics.

The second objective of GLD is to allow users to go beyond the standard dataset, to support them in delving deeper into their analyses and comparisons to find deeper insights. GLD empowers such a customized approach by providing all codes and technical reports, as well as documenting all intricacies of the survey discovered during harmonization so that users can focus on answering the questions they need answering, not on figuring out in what year an administrative boundary was changed and how the sample size was thus affected.

1.3. Who is the intended audience?

Target users of GLD include researchers, data analysts and practitioners in the international development community, statistical offices, ministries of labor, of economy and planning and other relevant government agencies analyzing labor market data to monitor and analyze labor market outcomes, and to inform the design of labor policies. These users can exploit two kinds of uses of the GLD.

The first use is the "as-is" harmonization. This refers to the user taking the harmonized data files as prepared by the data team and using those variables (or combinations thereof) for their analysis.

The second use is the "amended" or "hacked" harmonization. This refers to the user wanting to go beyond the prepared harmonization. This may be, for example, because they are interested in another specific variable from the survey, present in the questionnaire but not harmonized as not common in most surveys. In this case, the user can still utilize the harmonization do file to standardize most variables (as concepts like education level or labor status are likely still going to be relevant) but in addition add other ones. This use entails editing the harmonization code and/or adding to it at specific points to serve the users purpose without them needing to process the survey entirely.

1.4. What are the principles guiding GLD?

GLD follows a set of principles to guide its development and maintenance. In this introduction we focus on (a) GLD coverage and expansion, (b) transparency and data access, (c) data quality and validation.

GLD coverage and expansion

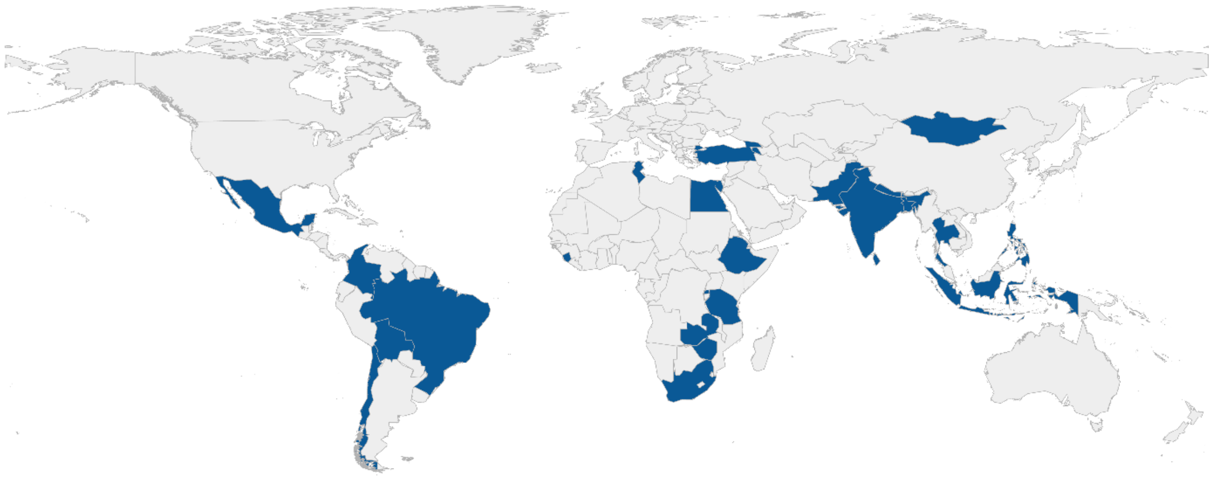
As of April 2024, the GLD holds 345 surveys from 24 countries (1 high-income countries, 9 upper medium-income, 11 lower middle-income, and 9 low-income countries). Table 1, gives an overview of the countries (by three digit ISO code), the number of surveys and years covered in GLD. Figure 1, below the table, shows the location of the countries in GLD.

Table 1 - GLD coverage by country, number of surveys, and time range

Country	Number of surveys	Range of years
ARM	9	2014 - 2022
BGD	5	2005 - 2016
BOL	6	2015 - 2021
BRA	37	1981 - 2022
CHL	13	1990 - 2017
COL	23	1996 - 2021
EGY	14	2006 - 2019
ETH	4	1999 - 2021
GEO	6	2017 - 2022
IDN	30	1989 - 2019
IND	15	1983 - 2022
LKA	23	1992 - 2021
MEX	16	2005 - 2020

MNG	17	2002 - 2022
NPL	3	1998 - 2017
PAK	15	1992 - 2020
PHL	23	1997 - 2019
RWA	5	2017 - 2021
SLE	1	2014 - 2014
THA	36	1985 - 2021
TUR	20	2000 - 2019
TUN	15	1997 - 2017
TZA	6	2000 - 2020
ZAF	13	2008 - 2020
ZMB	9	2008 - 2022
ZWE	5	2011 - 2022

Figure 1 - Map of the world with the GLD countries highlighted



Note: Boundaries shown on this map are not authoritative and should not be considered as an endorsement by The World Bank.

The initial choice of countries was driven by the availability of multiple LFS over time for the same countries. Thereafter, the GLD team has established selection guidelines to try to balance the GLD country coverage across income groups and regions, and to keep the GLD updated with the latest surveys.

Ensuring that the GLD is updated means to harmonize latest surveys once they become available for each country in the GLD. In general, we view a survey as in date if it is from the previous four years (e.g., from at least 2020 in 2024). Thus, both between regions and within regions, the choice in surveys to add to GLD should reflect an effort to not just be present at all income levels but to have up to date surveys for all.

However, acquiring new surveys is mostly determined by data availability, that is, whether it is possible to obtain new data or whether the National Statistical Office (NSO) of countries do not permit the sharing of survey data. Hence, if NSOs of a region share very little of their data, the GLD's imbalance due to a lack of surveys would not be correctable. Similarly, if NSOs do not run or only seldomly run a labor force survey, their countries' GLD entry would missing or small despite our best efforts.

Transparency and data access

GLD is designed to be as accessible and transparent as possible. Every step of our harmonization should be transparent and traceable. In addition, all the outputs produced by the GLD team (harmonization code and documentation of survey details, choices made during harmonization) are shared freely to all on GitHub, a web platform for collaborative software development and version control.

Access to the raw and harmonized microdata is restricted on a survey-by-survey basis in accordance with data license regulations. These limitations stem primarily from data privacy mandates issued by the National Statistics Offices, alongside other pertinent considerations. Such restrictions are imperative to ensure compliance with legal frameworks governing the confidentiality and usage of sensitive survey data. Adherence to these protocols not only upholds ethical standards but also safeguards the integrity and confidentiality of the information contained within the database.

GLD data is stored on a server managed by the GLD team. The team aims to use data sources we can share at least with World Bank colleagues whenever possible. Data is also accessible via datalibweb and the microdata library. Currently all GLD surveys are accessible to all World Bank staff except for data from Egypt, where the publisher of the data has requested that raw data (and thus harmonized data) be only accessed via their portal. More details on this in section 2.2 Data storage platforms and access rules section.

Data quality and validation

Central to the GLD objective of being a reliable source for cross country comparisons and benchmarking is ensuring data are of the highest quality. Only then is it possible to leverage large datasets and use GLD as an input into automated analytical workflows.

To validate the harmonization, the GLD team has three main tools. The first is the validation done with country office colleagues and NSO staff when harmonizing. GLD harmonizers are in touch with relevant colleagues with domain knowledge to understand the survey (knowledge they can then document and share) and ensure their mapping of variables from the raw data to the harmonized variables is sensible.

Once the harmonization is finalized, there are two automated quality check procedures. The first checks the survey for integrity and coherence with external sources (e.g., is the calculated labor force participation in line with what ILO, WDI report). The second checks a series of surveys in a country over time to detect any unexpected jumps in the series.

Finally, via direct exchange with the GLD team or on the online GitHub platform, users can alert the team of issues in the harmonization that had made it through, nonetheless. A process of updating the harmonization then kicks in to try to correct any issues as quickly as possible. For a more detailed description of all quality checks see section 4. Validation and quality checks.

1.5. Complementarities with similar data efforts

Within the World Bank, there are other two harmonization initiatives: (1) the I2D2 harmonization of both LFS and household surveys which has been active for more than a decade and has been recently discontinued; and (2) the Global Monitoring Database (GMD) which is harmonizing only household budget surveys used for poverty and inequality analysis primarily.

The GMD thus includes variables on household consumption and calculates certain income and consumption aggregates that are not present in GLD. On the other hand, GLD has more detailed labor variables, especially providing (wherever possible) industry and occupation information using ISIC and ISCO codes to the fullest depth possible. As both use a common set of variables, both can be used as inputs to automated analytical tools. This is the case, for example for the Jobs Indicators database (JOIN), which reads in data from GMD, GLD, and I2D2 to create country indicators. Moreover, the GLD offers a larger set of migration variables in its data dictionary than the GMD.

Finally, in trying to strike a balance between harmonization and information in the raw data, the GLD also stores more “original” variables from the raw data than GMD (or I2D2). That is, there are more variables that directly provide users with the information as is to be found in the survey. For example, in addition to `occup` and `occup_isco` (occupation information by 10 main categories or the ISCO codes), `occup_orig` contains the occupation information found in the survey. Thus, users can deviate from our codes without need to redo the harmonization and evaluate how the mapping from raw to harmonized variable was done directly.

Similarly, the ILO harmonizes LFSs to generate indicators published on the ILOSTAT data platform, though the underlying harmonized microdata and relative codes are not made public as in GLD. The ILOSTAT platform provides users with an extensive set of indicators, not just on labor markets, but also on other socioeconomic and sociodemographic information for more than 180 countries accessible in the ILO Survey Catalogue. However, the microdata are only accessible to ILO staff. Moreover, no harmonization codes are made available and thus what the user can access are indicators at the national level. Hence, while ILOSTAT is a great resource to obtain country level indicators for the most common topics, GLD can complement this via allowing users going deeper and leveraging the full microdata, running regressions or calculating indicators at sub-national level.

1.6. Sustainability

Maintaining the level of detail provided by the GLD is a significant undertaking that requires a serious investment of resources. The World Bank is the right organization to house such an effort due to the externalities generated by a public effort to create accessible data.

To maintain the costs of the effort low and ensure the sustainability of the project as it scales and requires more management, two things are necessary: 1) a strong collaboration with regional data teams and 2) the creation of a community of users on GitHub, a web platform for collaborative software development and version control, initially among World Bank staff with the objective to expand to all users through a collaborative yet still curated approach.

2. GLD content, storage, and access

This section discusses the kinds of information that can be found on GLD, the rules that govern each type, and the formats they are in.

2.1. Data and information collected in GLD

Raw microdata

The raw microdata are the individual level data as received or downloaded by the GLD Team. It may come directly from the National Statistical Offices (NSO), an aggregator (like the African Development Bank Microdata Catalog or the World Bank Microdata Catalog) or from colleagues. They are collectively referred to as the data sharers in this document.

The raw microdata reflects the original state of the data and its variables should mirror the questionnaire. In cases where no better alternative is found, the raw data the GLD team starts with may have been already processed by either the NSO or an intermediary. If this is the case, this should be described and noted in the background information (see the **Error! Reference source not found.** section for more details).

The data sharer will determine the data privacy rights of the data and thus the access. Data access rights can be broadly categorized into three buckets. The first bucket is public domain data. This data is freely accessible to all users. This would be the case of directly downloadable on the NSO website. The second bucket is World Bank *official use* data. This refers to data that can be used by and shared with World Bank colleagues without any restriction but cannot be shared outside the organization. The third buckets is limited release data. This describes data that are part of GLD yet cannot be shared freely. Access needs to be decided on a case-by-case basis and often requires requesting permission from the data sharer.

Raw microdata is taken in by the GLD Team in whatever format the data sharer choses (e.g., CSV, TXT, Stata, SPSS, …). IF the raw microdata is not in Stata .dta format, the GLD team will convert it to that format for further use. The code to do so is an example of the "other code" GLD produces (that is, GLD code that is not harmonization code – see the Other code section for more information).

Harmonized microdata

The harmonized microdata is the output of the harmonization process. It is individual level information where every row represents an individual. The GLD data dictionary variables are contained in the columns. The harmonized microdata will only contain those variables for which there are answers. That is, if a variable cannot be coded and thus would be missing for all individuals in the data it is dropped to keep the file size as small as possible.

The access rights are inherited from the raw microdata. Whatever the rights are for the raw microdata will be applied to the harmonized data.

The harmonized microdata is produced as a Stata ,dta file.

Harmonization code

The harmonization code is the set of instructions used to convert the raw microdata into the harmonized output and the comments explaining, wherever necessary, the rational for the specific coding instructions given. For a description of how to read the code please see section The structure of the harmonization code, for the definition of individual variables please see section The GLD data dictionary.

The harmonization code is created by the GLD team and shared openly and free of charge under the provisions of the MIT License rules (see our License details here).

The harmonization code is written as a Stata .do file.

Other code

In addition to the harmonization code and the bulk of the GLD code, various other code is written and made available. This code can be categorized into four types:

- **Raw Microdata Conversion Code:** Code written to convert raw microdata, not in Stata .dta format, into this format.

- **GLD Quality Check Code:** Code responsible for conducting quality checks on GLD data. For more details on these quality checks, refer to the Validation and quality checks section.

- **Code Templates:** Predefined code structures or templates to facilitate coding tasks within the GLD framework.

- **GLD Ecosystem Tools Code:** Code that builds the "GLD Ecosystem Tools," which are small software programs designed to address tasks commonly performed by GLD users on the GLD data. An example includes a tool supporting the conversion of ISIC and ISCO codes (classifications for industries and occupations) across different revisions of the classifications. All GLD tools can be found here.

All code of each category, like the harmonization code, is shared under an MIT License free of charge. The code may be written using any software program. Please refer to the individual tools to determine the software type used. Additionally, requests for modifications to existing tools or new tool suggestions can be made (see the How to correct and expand the GLD tools section)

Survey and other documentation

The survey documentation refers to all documents that are published by the NSO (or other institution running the survey) that serve to understand the survey. This includes but is not limited to the questionnaire, intermediate and final reports, as well as enumerator manuals. The survey documentation is treated as public information and thus shared freely. It is stored and distributed in whatever format the GLD team received it (most commonly a PDF or a spreadsheet format).

Background or contextual information refers to the information the GLD team has acquired in the process of harmonizing that is not (or not directly) in the survey documentation or cannot be embedded into the comments in the harmonization code itself.

A common occurrence is the change of the employment definition. In its 19th session the International Conference of Labour Statisticians (ICLS), under the auspices of the International Labour Organization (ILO) changed the definition (more information here). Therefore, work for own consumption (for example subsistence agriculture) was no longer considered employment. This change has been implemented differently at different times by different countries. Per the rules of the GLD harmonization each survey will be harmonized using the definition of employment that was applied to each survey and not unified (see the Defining the boundaries of GLD harmonization section for more details). To help users understand these changes, its implication, and – if possible – try to align definitions with earlier ones, the GLD team prepares documentation to detail the changes, show the relevant questionnaire passages and propose alternative coding. Information like this is collectively referred to as *Country Survey Details*.

The Country Survey Details are text files that accessible to all and may be freely shared. You may see them here online.

2.2 Data storage platforms and access rules

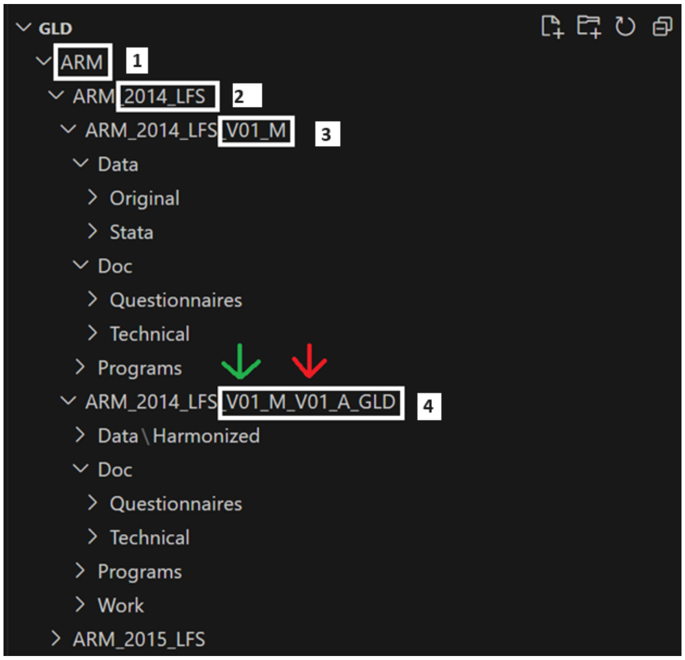
This section details which GLD information type is stored where and how it is organized. To see what types of information are stored on GLD please see Data and information collected in GLD section.

The World Bank GLD Server

The raw microdata, harmonization codes, harmonized output, as well as the documentation is stored on a dedicated server. It is organized following the file and folder naming convention of the World Bank Microdata Library (accessible here).

The image below exemplifies the structure with the case of Armenia. Within GLD the first level is the country, identified the country's three-letter code (number 1 in Figure 2, below). The surveys in a country follow the structure CCC_YYYY_[SurveyName], where YYYY is the year of survey start and SurveyName the name or acronym of the survey. In this case the Armenian Labor Force Survey from 2014 is ARM_2014_LFS (number 2 in the image).

Figure 2 - Example of GLD Server folder structure



```
∨ GLD
  ∨ ARM  1
    ∨ ARM_2014_LFS  2
      ∨ ARM_2014_LFS_V01_M  3
        ∨ Data
          > Original
          > Stata
        ∨ Doc
          > Questionnaires
          > Technical
        > Programs
      ∨ ARM_2014_LFS_V01_M_V01_A_GLD  4
        > Data\Harmonized
        ∨ Doc
          > Questionnaires
          > Technical
        > Programs
        > Work
    > ARM_2015_LFS
```

Inside the survey folder are the master data and the harmonized data folders. Both begin with the same name as the folder above in the nested logic, but the master only has the vintage of the master data (here V01_M, number 3 in the image, as this is the first (and only so far) version of the raw data we have). Master data may be updated if the NSO publishes, for example, a revision to the data.

The harmonized folder (number 4 in the image) starts like the master data folder it harmonizes from (see green arrow in the image above) but adds the vintage of the harmonization (red arrow). It also adds the initials of the collection we are harmonizing towards (GLD).

Inside the master data folder there are three folders: Data, Doc, and Programs. Data itself is divided into Original and Stata. The former contains all files as downloaded if the original download is not in Stata format (i.e., not a .dta file), the latter contains the Stata survey microdata.

Any code changing the raw data is stored in the Programs folder. For example, code converting raw data from other formats to dta, that is, reading from Data/Original, converting it, and storing it in Data/Stata, would be stored under Programs.

The last folder, Doc, contains all further documentation that is needed to work on the survey. It should be divided into two further folders: Questionnaires, containing the questionnaires and all other necessary document to understand the questionnaire and its flow; and Technical, containing all other technical information (e.g., reports, national occupation classifications, etc.).

As a best practice, it is advised to leave in the Doc folder a small Readme file (commonly titled "Where is this data from – ReadMe.txt") to give information about where the source material is from. This is important for future colleagues, so they can trace information establish the access policy.

The structure for each of the harmonization folders is roughly the same, only with added version numbering and collection name (GLD). The Data/Harmonized folder shall contain the harmonized output in '.dta' form (in this case ARM_2014_LFS_V01_M_V01_A_GLD.dta). The Data/Additional Data folder is an optional contains data not in the raw data that is needed to create the harmonization. For example, if the conversion of the national industry classification to the international version is done via merging in an extra file, this file would be placed under Data/Additional Data. If no such files were used the folder need not exist.

The harmonization code (i.e., the code that takes the Data/Stata input from the master folder system and saves output in Data/Harmonized) is stored in the Programs folder (and in this case would be ARM_2014_LFS_V01_M_V01_A_GLD_ALL.do).

The Doc folder contains any other documentation necessary to describe and understand the survey. Note this is the same content as in ARM_2014_LFS_V01_M/Doc in the example above. Content should be in both at the same time – a small price on duplication we believe is worth for ease of finding for the user.

The Work folder contains any output created during the harmonization that is not the final harmonization. For example, if you needed to create a subfile of the survey containing only households from a certain region for inspection or any other process you may need during your work, these outputs should be stored here. Data/Harmonized should only contain finalized files, here you may store any intermediate results.

The GLD server is closed to members of the GLD team and access cannot be granted other than for exceptional circumstances. To allow World Bank staff member access to the server structure, the GLD team has created a copy of the GLD server, called the *GLD WB Staff Server*, only containing the subset of the GLD surveys that can be shared.

The GLD WB Staff Server is a subset in two ways. Firstly, it only contains surveys whose raw microdata can be freely shared with World Bank colleagues. Secondly, it only contains the latest harmonized version. For example, while for the 2020 Indian Periodic Labour Force Survey (PLFS), GLD contains the raw data folder (IND_2020_PLFS_V01_M) and four vintages (IND_2020_PLFS_V01_M_V01_M_A_GLD to

IND_2020_PLFS_V01_M_V04_A_GLD), the GLD Staff server only contains the master and IND_2020_PLFS_V01_M_V04_A_GLD.

This reduces space on the server and ensures users are using the latest files are being used. If a user was running some code, calling from the GLD WB Staff Server the IND_2020_PLFS_V01_M_V0*3*_A_GLD files, it would not run and call an error, forcing them to update to the latest vintage.

Other than these two differences, the GLD WB Staff Server is organized like the full GLD server. Access to the GLD Staff Server can be requested and granted by the GLD Focal Point (please reach out to gld@worldbank.org).

Access is available to any staff member with an active World Bank email address and access to a World Bank laptop or Virtual Desktop. Once a user has mapped the server (see instructions here on mapping) they do not need to take further steps as the GLD team updates the GLD WB staff server. What is present on the server should always represent the latest vintage of any harmonization available.

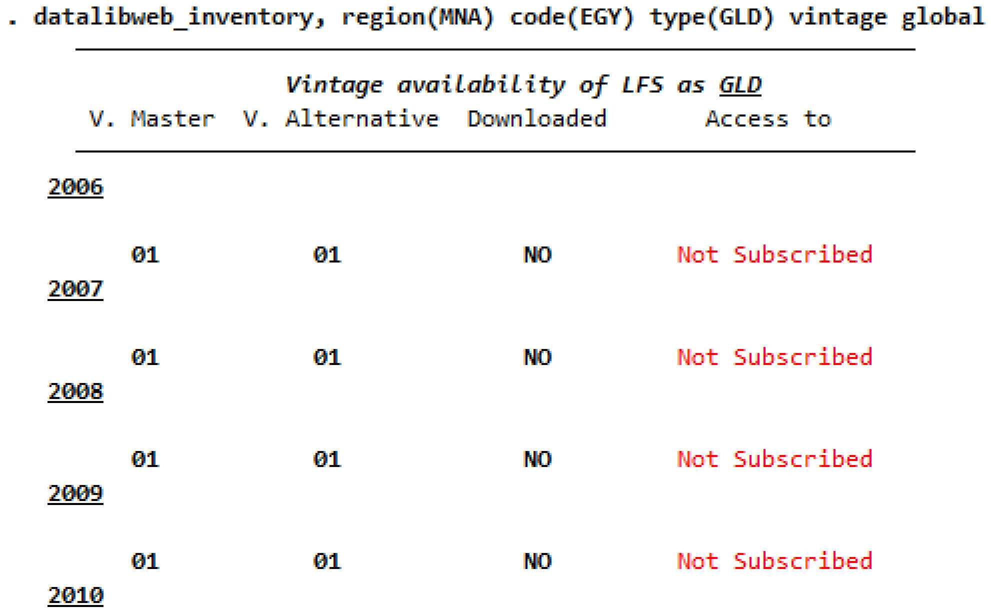
Datalibweb

datalibweb is an application programming interface (API) with two components. It has a website (internal to the World Bank) for data exploration and request and a set of API endpoints to securely access to granted microdata. Currently, the API endpoints are integrated with Stata through the datalibweb Stata package. For more details and information, please visit the datalibweb GitHub repository or, from the World Bank intranet, type in “datalibweb/” into your browser.

To integrate with datalibweb, GLD has granted its API access to the full GLD server. It thus contains the same information types (raw and harmonized microdata, harmonization codes, and documentation) and can monitor GLD for update. The datalibweb interface allows users to see what surveys are available either via a click and select menu system that appears upon entering “datalibweb” into the Stata command line or programmatically (using the Stata package’s syntax). . For details on how to use the Stata package please see the GitHub page; for questions on datalibweb kindly visit the intranet site (datalibweb/) or reach out to the team (datalibweb@worldbank.org). For details on the syntax and default behavior of datalibweb (e.g., unless requested otherwise it will load the latest vintage of a harmonization) see the help file by executing “help datalibweb” in the command line.

Access to datalibweb is controlled via the datalibweb intranet site. Uploaders to datalibweb (people who provide data to datalibweb, like the GLD Focal Point) can set data to public (accessible to all with access to datalibweb) or private. If the survey is set to private, the survey information will appear as “Not Subscribed” (see Figure 3 below for Egypt).

Figure 3 - Example of EGY GLD surveys accessed through datalibweb

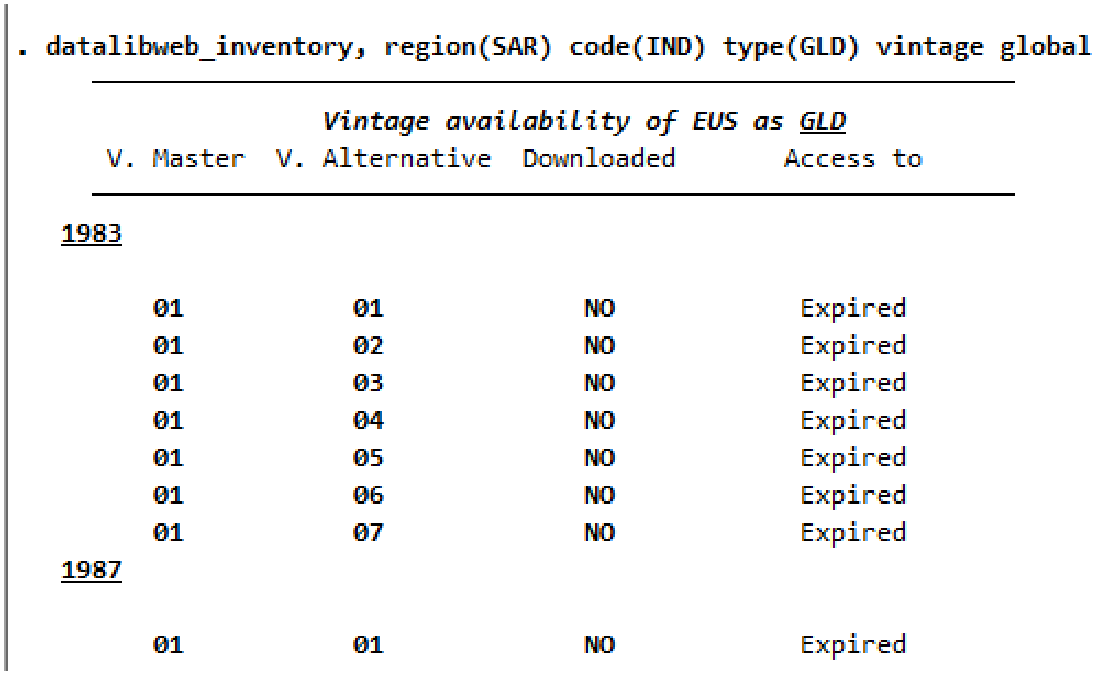


```
. datalibweb_inventory, region(MNA) code(EGY) type(GLD) vintage global

              Vintage availability of LFS as GLD
  V. Master  V. Alternative  Downloaded       Access to

2006

     01            01             NO        Not Subscribed
2007

     01            01             NO        Not Subscribed
2008

     01            01             NO        Not Subscribed
2009

     01            01             NO        Not Subscribed
2010
```

In such cases users can navigate the datalibweb intranet site (accessible by entering “datalibweb/” into their browser) and request access to it. This triggers an email to the data owner who can approve or deny the request.

Note that datalibweb may automatically set surveys that are public to expire. This is the case for the 1983 EUS shown below (Figure 4), for example.

Figure 4 - Example of IND GLD surveys accessed through datalibweb



```
. datalibweb_inventory, region(SAR) code(IND) type(GLD) vintage global

              Vintage availability of EUS as GLD
  V. Master  V. Alternative  Downloaded       Access to

1983

     01            01             NO           Expired
     01            02             NO           Expired
     01            03             NO           Expired
     01            04             NO           Expired
     01            05             NO           Expired
     01            06             NO           Expired
     01            07             NO           Expired
1987

     01            01             NO           Expired
```

This is because datalibweb automatically expires data access after a certain time. However, since the data is still declared as public, it can be accessed if requested directly. In the case of the latest vintage of the 1983 survey this would be:

```
datalibweb, country(IND) year(1983) type(GLD) vermast(01)
veralt(07) survey(EUS) module(ALL) clear
```

This command grants access to the data via datalibweb despite the expiration.

Microdata Library

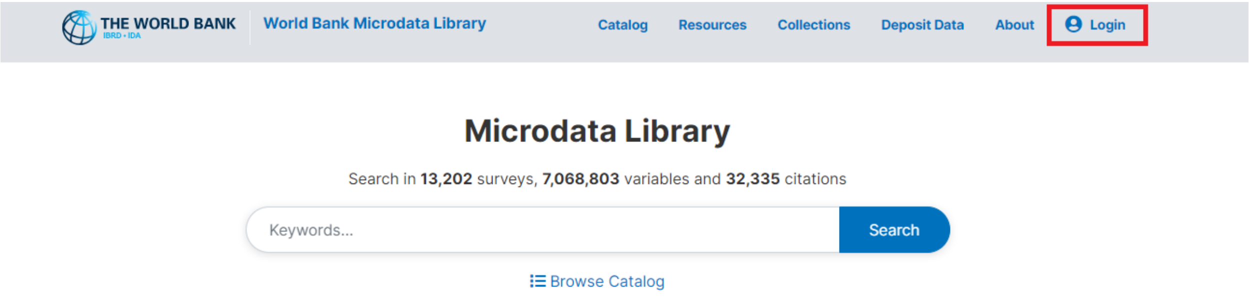
The Microdata Library (MDL) is the World Bank internal microdata catalog. Note that there is an equally named external looking website (its internet presence is microdata.worldbank.org as opposed to microdata**lib**.worldbank.org). In this manual we refer only to the internal site.

The microdata library contains both raw and harmonized surveys. Each type of data would be treated as a separate entry in the MDL. Each survey entry in turn contains additional resources, which most often consists of the survey documentation, but may include other documents, like a harmonization code file.

To see the MDL users are not obliged to login, but this will be necessary to download or request data (more details on how to access data on MDL at the end of this section).

The image below shows the homepage of the MDL and where to login (red box in Figure 5 below).

Figure 5 - Header of the Microdata Library

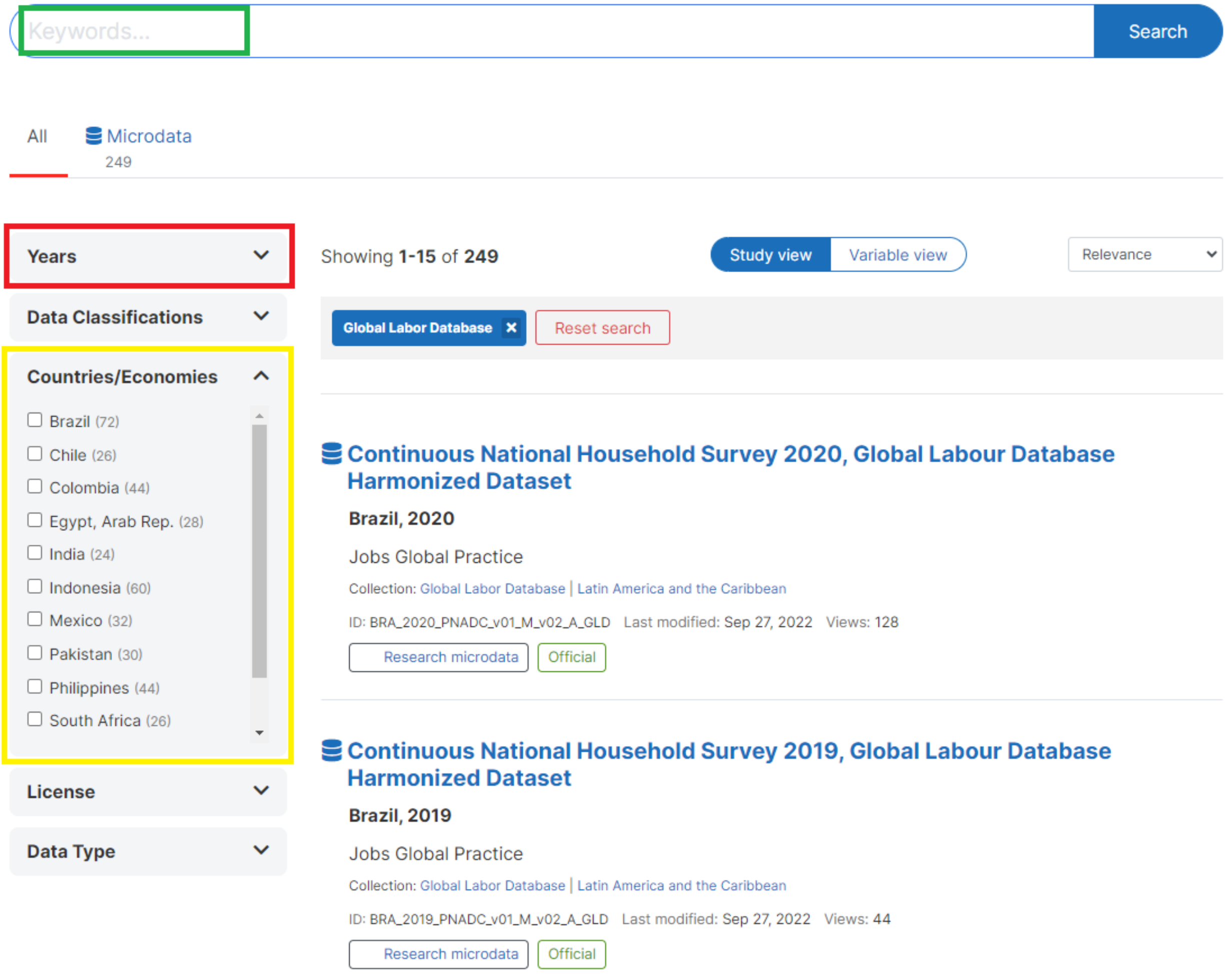


Microdata Library

Search in **13,202** surveys, **7,068,803** variables and **32,335** citations

The Microdata Library Catalog itself can be navigated using either by searching for keywords (green box in Figure 6 below) or using the filters on the right had side like the year (red box) or the country (yellow box).

Figure 6 - Navigating the Microdata Library catalog



Keywords... Search

All | Microdata 249

Years

Data Classifications

Countries/Economies

☐ Brazil (72)
☐ Chile (26)
☐ Colombia (44)
☐ Egypt, Arab Rep. (28)
☐ India (24)
☐ Indonesia (60)
☐ Mexico (32)
☐ Pakistan (30)
☐ Philippines (44)
☐ South Africa (26)

License

Data Type

Showing **1-15** of **249**

Study view | Variable view | Relevance

Global Labor Database ✕ | Reset search

Continuous National Household Survey 2020, Global Labour Database Harmonized Dataset

Brazil, 2020

Jobs Global Practice

Collection: Global Labor Database | Latin America and the Caribbean

ID: BRA_2020_PNADC_v01_M_v02_A_GLD Last modified: Sep 27, 2022 Views: 128

Research microdata | Official

Continuous National Household Survey 2019, Global Labour Database Harmonized Dataset

Brazil, 2019

Jobs Global Practice

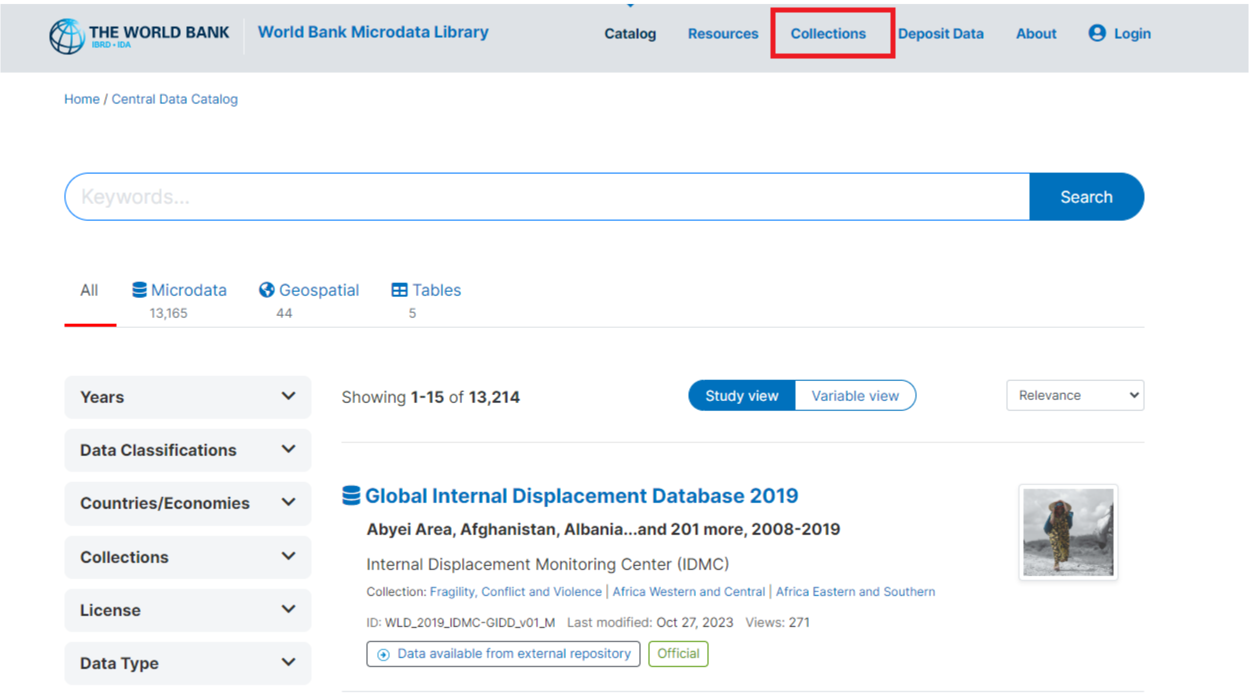
Collection: Global Labor Database | Latin America and the Caribbean

ID: BRA_2019_PNADC_v01_M_v02_A_GLD Last modified: Sep 27, 2022 Views: 44

Research microdata | Official

GLD data can be found directly via the purpose built GLD collection. To look for different harmonization collections users need to click on the option “Collections” on the top of the page (red box in Figure 7below).

Figure 7 - Accessing collections on the Microdata Library



Showing **1-15** of **13,214**

Global Internal Displacement Database 2019

Abyei Area, Afghanistan, Albania...and 201 more, 2008-2019

Internal Displacement Monitoring Center (IDMC)

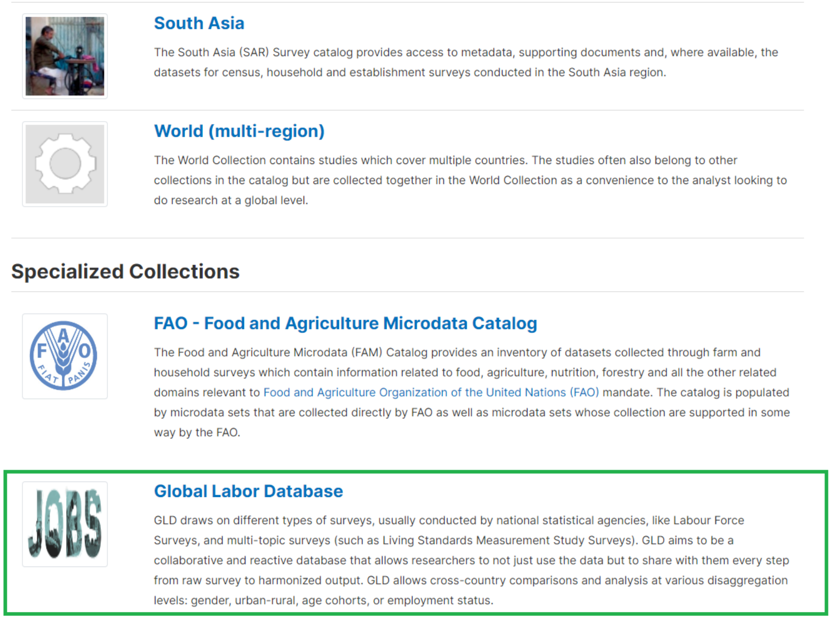
Collection: Fragility, Conflict and Violence | Africa Western and Central | Africa Eastern and Southern

ID: WLD_2019_IDMC-GIDD_v01_M Last modified: Oct 27, 2023 Views: 271

Data available from external repository Official

Once on the collections site, users need to scroll down to the "Specialized Collections" header, where GLD is present (see green box in Figure 8 below).

Figure 8 - Screenshot of collections available on the Microdata Library



South Asia

The South Asia (SAR) Survey catalog provides access to metadata, supporting documents and, where available, the datasets for census, household and establishment surveys conducted in the South Asia region.

World (multi-region)

The World Collection contains studies which cover multiple countries. The studies often also belong to other collections in the catalog but are collected together in the World Collection as a convenience to the analyst looking to do research at a global level.

Specialized Collections

FAO - Food and Agriculture Microdata Catalog

The Food and Agriculture Microdata (FAM) Catalog provides an inventory of datasets collected through farm and household surveys which contain information related to food, agriculture, nutrition, forestry and all the other related domains relevant to Food and Agriculture Organization of the United Nations (FAO) mandate. The catalog is populated by microdata sets that are collected directly by FAO as well as microdata sets whose collection are supported in some way by the FAO.

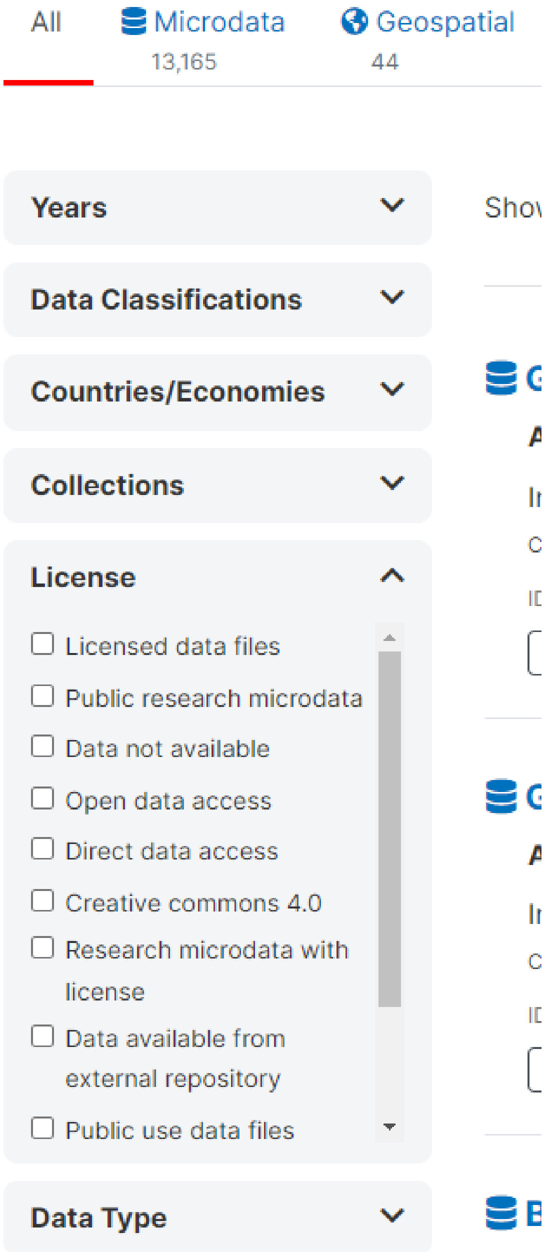
Global Labor Database

GLD draws on different types of surveys, usually conducted by national statistical agencies, like Labour Force Surveys, and multi-topic surveys (such as Living Standards Measurement Study Surveys). GLD aims to be a collaborative and reactive database that allows researchers to not just use the data but to share with them every step from raw survey to harmonized output. GLD allows cross-country comparisons and analysis at various disaggregation levels: gender, urban-rural, age cohorts, or employment status.

Finding surveys works the same way as finding surveys on the MDL in general, by using the search box or the panes explained above.

Access to surveys on the Microdata Library (MDL) is regulated by the data licenses. Figure 9, below, shows the license options users can filter for when searching for surveys.

Figure 9 - License options on MDL



Depending on the license type, users will be able to download the data directly, be asked to fill a form, or barred from accessing. All processes are internal to the MDL. Forms are sent to the MDL team and reviewed by them.

GitHub

Before delving into what information is store on GitHub we first introduce the reader to GitHub. If you are already familiar with GitHub please feel free to skip the next four paragraphs.

GitHub is a web-based platform that allows developers and programmers to store, manage, and collaborate on software projects. It is a central hub where people can upload, download, and work on code files, as well as track changes, report issues, and discuss ideas. GitHub is built on the Git version control system, which lets multiple people work on the same project simultaneously without overwriting each other's work.

The core feature of GitHub is the "repository," which is essentially an online folder that contains all the files and revision history for a particular software project. Users can create their own repositories to store their code, or contribute to existing repositories created by others. GitHub provides a user-friendly web interface, as well as desktop and mobile apps, to make it easy for people to access and manage their repositories.

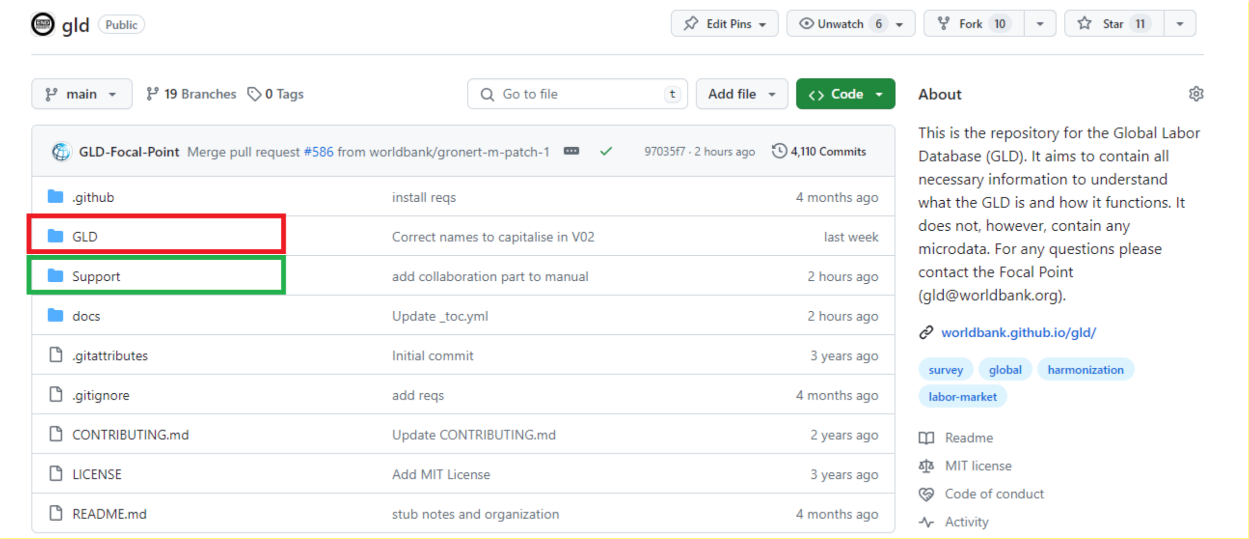
When you interact with a GitHub repository, you may "clone" it, which creates a local copy on your own computer. From there, you can make changes, add new files, and commit those changes back to the central repository. GitHub also supports "branching," which allows multiple versions of a project to be developed in parallel. Users can then submit "pull requests" to propose incorporating their changes into the main branch of the project.

Beyond just code storage and version control, GitHub enables collaboration by allowing users to report bugs, suggest improvements, and discuss the direction of a project through the repository's online tools. This collaborative nature has made GitHub an essential platform for open-source software development, as well as a valuable resource for teams working on a project concurrently.

The GLD team does not story any microdata on GitHub but holds the full set of harmonization codes for all GLD surveys, the templates, and the Country Survey Details as well as the validation’s quality check codes and the GLD ecosystem tools. The code to convert raw microdata received in non-Stata .dta format to .dta are not kept on GitHub either.

On the landing page (shown below in Figure 10) there are two relevant folders that lead to the information: GLD (red box) and Support folder (green box). The former contains the harmonization codes for the GLD surveys while the latter contains all additional information to understand and leverage both the GLD surveys and the GLD ecosystem (quality checks, Country Survey Details, Tools and GLD documentation like the data dictionary). The other folders contain information to make the repository work and are not further discussed here.

Figure 10 - GLD GitHub repository landing page



GLD-Focal-Point Merge pull request #586 from worldbank/gronert-m-patch-1 ✓ 97035f7 · 2 hours ago 4,110 Commits

.github	install reqs	4 months ago
GLD	Correct names to capitalise in V02	last week
Support	add collaboration part to manual	2 hours ago
docs	Update _toc.yml	2 hours ago
.gitattributes	Initial commit	3 years ago
.gitignore	add reqs	4 months ago
CONTRIBUTING.md	Update CONTRIBUTING.md	2 years ago
LICENSE	Add MIT License	3 years ago
README.md	stub notes and organization	4 months ago

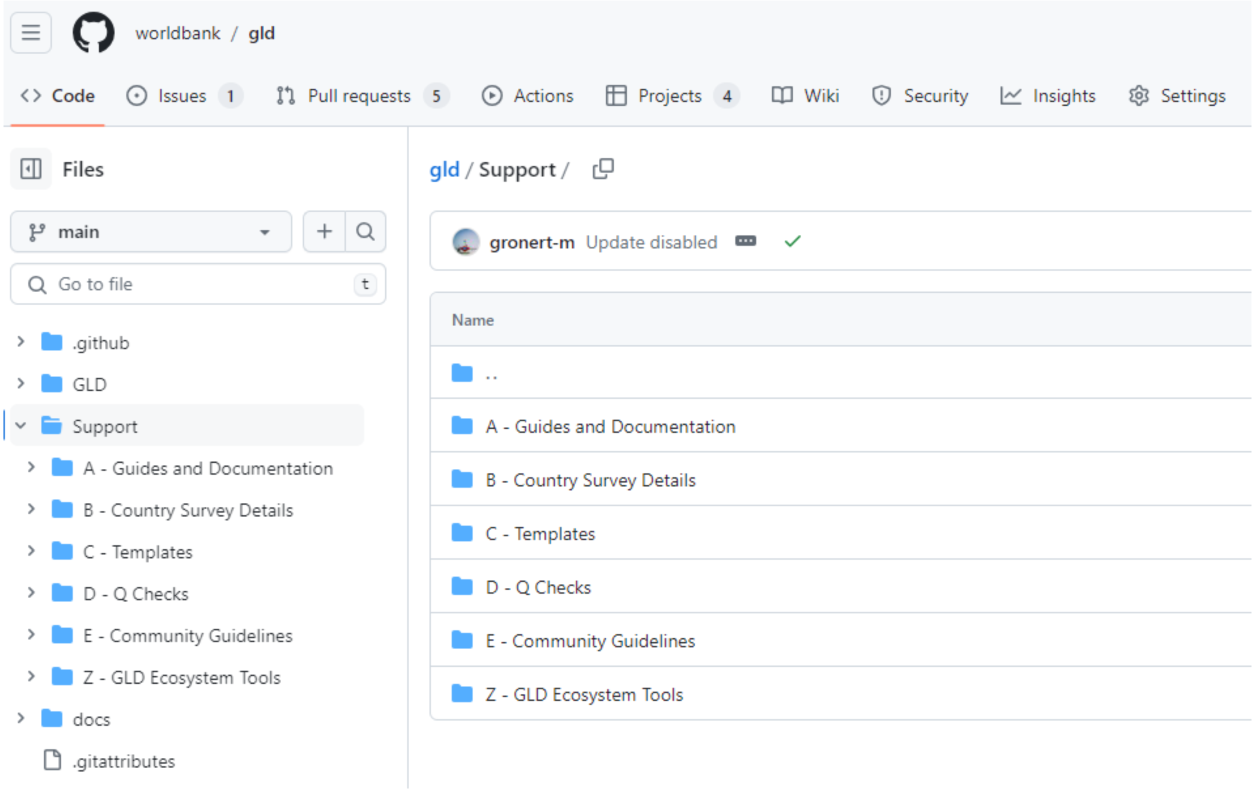
This is the repository for the Global Labor Database (GLD). It aims to contain all necessary information to understand what the GLD is and how it functions. It does not, however, contain any microdata. For any questions please contact the Focal Point (gld@worldbank.org).

worldbank.github.io/gld/

The GLD folder follows the logic described above for the GLD Server, with a `CCC/CCC_YYYY_Survey-Name/CCC …` structure. The only difference is that it only contains the harmonized folder (i.e., `_V##_M_V##_A_GLD`) and inside this folder there is only the Programs folder.

The Support folder has six subfolders as shown below (Figure 11). Subfolder A – Guides and Documentation contains individual documents that explain the functioning and rules of GLD. In part they are constituents of this manual (i.e., the information in them is also here) but are kept separate for easier sharing. For example, the next section will explain all the variables in GLD, but the data dictionary is available there as a standalone document for ease of review and sharing. In part these are documents referenced in this manual (e.g., the World Bank Micro Data Library folder and file naming convention).

Figure 11 - Structure of the GLD GitHub repository's Support folder



- .github
- GLD
- Support
  - A - Guides and Documentation
  - B - Country Survey Details
  - C - Templates
  - D - Q Checks
  - E - Community Guidelines
  - Z - GLD Ecosystem Tools
- docs
- .gitattributes

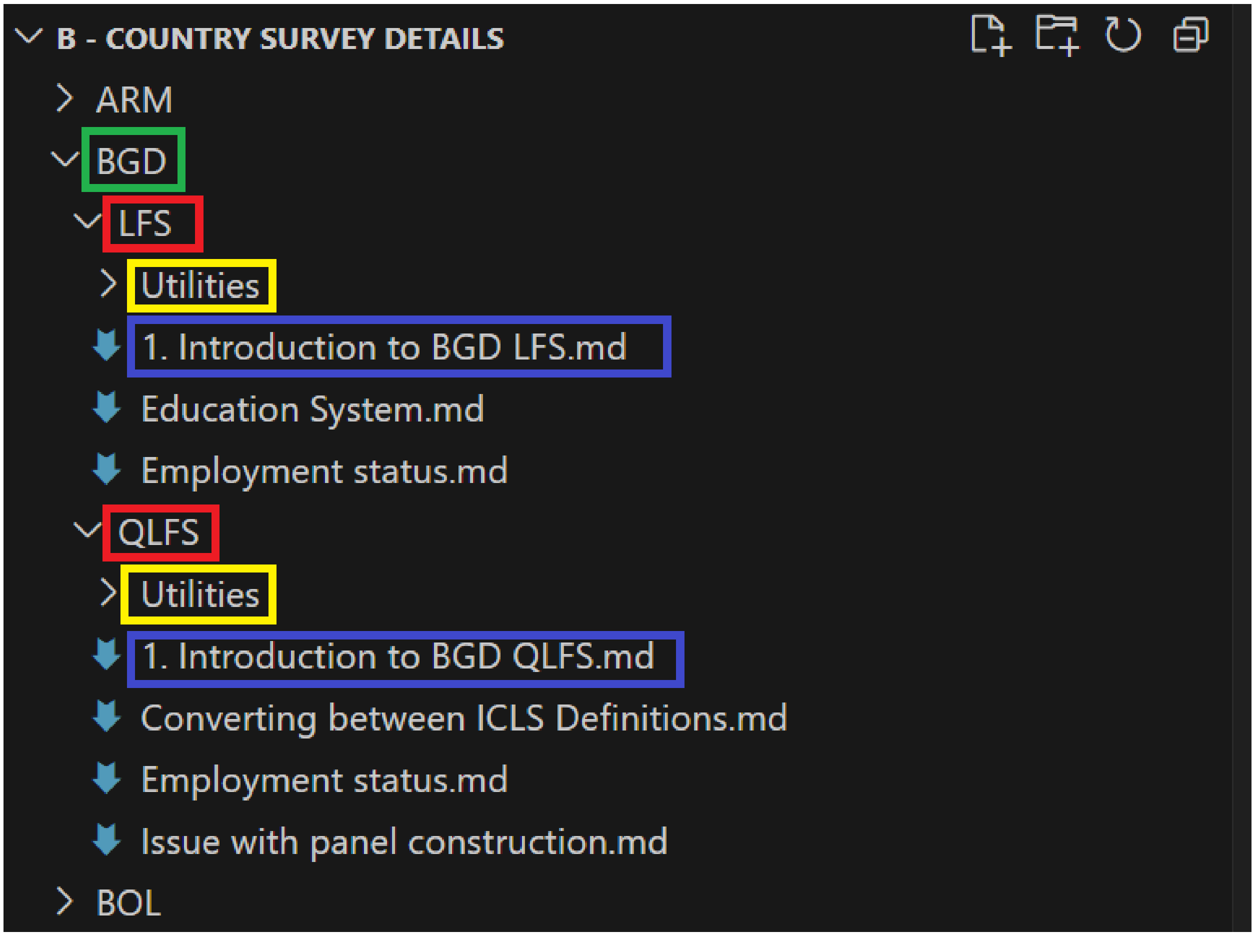
gld / Support /

Name
..
A - Guides and Documentation
B - Country Survey Details
C - Templates
D - Q Checks
E - Community Guidelines
Z - GLD Ecosystem Tools

The subfolder B – Country Survey Details contains the metainformation on surveys that cannot be coded into the harmonization or the harmonization code – or need more explaining to be understood. The structure of the folder starts with the three-letter code of the country, within which are the surveys, listed by their name or initials (e.g., LFS, ENE, SAKERNAS). Inside each survey folder there should always be at least one element: the introduction file. All other files and folders are optional and depend on the situation.

The example below (Figure 12) shows that the structure starts, like the GLD folder, with the three-letter code (green box). Inside of it are all the surveys (LFS and QLFS in this case, red boxes). Each survey will have a Utilities folder (yellow box) where images or documents referenced in the CSD are referred. All surveys have an introductory text (purple box) followed by additional texts that expand on the issues from the introduction (if necessary).

Figure 12 - Structure of the B - Country Survey Details folder



```
B - COUNTRY SURVEY DETAILS
  ARM
  BGD
    LFS
      Utilities
      1. Introduction to BGD LFS.md
      Education System.md
      Employment status.md
    QLFS
      Utilities
      1. Introduction to BGD QLFS.md
      Converting between ICLS Definitions.md
      Employment status.md
      Issue with panel construction.md
  BOL
```

The introduction file starts with "1." as GitHub orders files alphanumerically and this file should be at the top always. It should always read as Introduction to CCC (or country name) Survey. It contains the basic information every user should read before starting work on a survey. The first part is standardized and is the same for all surveys – further down the template for this document is introduced. The standardized part informs the user of what the survey is, where the information is from (and if public, where to get it), what the sampling procedure was and to what geographic level the information is statistically significant. The latter part called "Other noteworthy aspects" allows the harmonizers to expand on non-standard issues that are of relevance to that survey in that specific country.

If the introduction contains images or references some document, these are stored in the *utilities* folder, a sort of catch-all for the survey. Similarly, if a topic requires more in-depth discussion, it will be referenced in the introduction, but then discussed in detail in a separated file (e.g., how to deal with a break in the definition of employment over a series). Any images or documents referenced here should also be stored in the utilities folder.

The Support subfolder "C – Templates" contains the GLD templates. Currently, there are three. The harmonization code template, the code to create new GLD structured folders, and the template for the CSD introduction. These should allow new harmonizers to create new harmonizations using the same structures.

The subfolder “D – Q Checks” contains the GLD quality checks. There are two types of checks. The checks to be done to each survey (e.g., the 2019 Indian PLFS) and the checks to be done (once all are ready) to a series of years of surveys (e.g., years 2017 to 2022 of the Indian PLFS). The former are in the *Single survey checks* folder, while the latter are in the *Survey series checks* folder. The details of how the checks work is contained in the [Validation section](../GLD Manual/ GLD_Manual_Validation.md).

The subfolder “E – Community Guidelines” contains the guidelines to interact with the GLD team on GitHub. This refers to the conduct we expect from users as well as checklists to ensure any action is as impactful as possible. Further details of how interact with the team are contained in the Contributing to GLD’s quality and expansion section.

The GLD GitHub repository is an open access repository that can be used by any person without restrictions. Users can also contribute to the GLD. Contributions are vetted and reviewed before being included in the GLD.

Summary of data and information storage

Table 2 below provides an overview of what information is stored where and how accessible the information is. The rows refer to the GLD server, the GLD public server, datalibweb, the World Bank Microdata Library, and the GLD GitHub repository. The columns refer to the different information types.

A dark-green colored cell means the information is directly accessible and comprehensive of the surveys on GLD. A light-green colored cell means the information is directly accessible but not comprehensive of GLD (i.e., it represents a subset). A yellow-colored cells refers to information that may require requesting access but should (if properly updated in the case of the MDL) be comprehensive of the GLD. Lastly, grey colored cells mean that the information type is not available from that source (e.g., there are no microdata on GitHub).

Table 2 - Overview of data storage platforms

	Raw Microdata	Harm. Microdata	Documen-tation	Harm. Code	Other Code	CSD
Server	🟩	🟩	🟩	🟩	▨ `	⬜
Public S	▨	▨	▨	▨	⬜	⬜
DLW	🟨	🟨	🟨	🟨	⬜	⬜
MDL	🟨	🟨	🟨	🟨	⬜	⬜
GitHub	⬜	⬜	⬜	🟩	▨	🟩

3. The GLD harmonization methodology

For the purposes of this section, harmonization refers only to the act of converting the raw microdata into the variables of the GLD data dictionary.

3.1. Defining the boundaries of GLD harmonization

The scope of the GLD harmonization is the data dictionary. However, not always the information in the survey fits the concept in the data dictionary. For example, if the questionnaire asks respondents whether the respondent or their employer “contribute to social or private security” is that sufficient to code the *socialsec* variable? Similarly, with a view to the objective of comparison and benchmarking, how should the harmonization handle change to concepts over time, like changes to the definition of employment or administrative areas?

To define the boundaries of the harmonization, the GLD team follows two principles: (i) each survey is harmonized independently; and (ii) in unclear situations, users are empowered to take informed decisions, rather than having the GLD team making choices for users.

The principle that surveys are harmonized independently means that each survey is harmonized based on the standard and realities present at the time of its collection and processing. Changes introduced in subsequent surveys are not applied retrospectively. For example, if new geographical regions or occupation classifications are introduced, previous surveys will reflect the older versions without any retroactive adjustments.

The principle to allow users to make informed decisions means to provide comprehensive information about changes and document their impact on the data (for example, the updated occupation classifications in Pakistan and changes in the definition of employment in Tanzania). However, the GLD harmonization takes a conservative approach, making minimal assumptions and deferring significant decisions to users. This ensures data accuracy and transparency, allowing users to make well-informed choices based on their research objectives.

Unifying elements that have changed over time across multiple surveys or taking leaps on questions that fit the data dictionary only partially is outside the scope of the harmonization. Users are provided with the necessary context and information to bridge these gaps according to their requirements.

3.2. The structure of the harmonization code

The GLD harmonization code template begins with a header or preamble summarising key survey aspects (see Box 1, next page). It contains four blocks:

1. Information on the code, the author, and the creation date.
2. Details on the survey context
3. Details on the versions of the standard classifications used
4. Version control history, detailing the date and the contents of any changes performed

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_>

-------------------------------------------------------------------*/
```

After the box, the harmonization code is divided into 9 sections. Section 1 contains the codes to set up file and folder paths and assemble the dataset to harmonize from the raw data. It is at this step users would need to update folder and file paths and name if they are using a different storing system (e.g., their server is labelled “E” instead of “Y”).

Sections 2 to 8 cover variables of the different blocks of the data dictionary (e.g., geography, socio-demographic, education, …). Section 9 does the final clean up to keep only variables in the dictionary that have values (i.e., we do not keep variables that have missing values for all respondents), in the correct order. Data is also compressed and unused labels are discarded so the final output is as size-efficient as possible.

Each section is 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
====================================================================%%*/
```

Within the "variable" sections 2 to 8, all harmonized variables in the in each section in the data are 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_> where "var" is the harmonized variable being created.

- The end of the code relating to the variable creation should read *</_var_>.

- Variables that are already named (e.g. if "hhid" exists in raw data file) should be noted. Between the "open" and "close" tags, a starred outline should read:"*'var' present in 'source'".

- If a variable requires more extensive or explicit comments, these should be written between note tags. The note tag is the same as the variable tag only followed by "_note" (e.g., "note_var"). For example, the variable "lstatus" is created in an uncommon or unexpected way, then harmonizers can add the variable-specific note as follow: *<_lstatus_note_> Text explaining issues with variable, why which choices made *</_lfstatus_note_> (see Box 3 below).

Box 3 - Variable tagging example

```
*<_wage_no_compen_>

/*<_wage_no_compen_note_>

The wage questions in the questionnaire are organized into two parts:
the first part asks for the specific number of income if the given
respondent could recall and was willing to answer; the second part
provides different categories of income range if the given respondent
could not recall or was not willing to answer the first part.
```

```
The general logic here is to impute wage values for people who only
answered an income range. We used industry, occupation, income
categories and gender to estimate their specific income values.

21.86% of total non-missing wage values were imputed using this method.
*<_wage_no_compen_note_>*/

     * Overall --> wage info
     * Set values of 0 to missing
       gen wage14=E14_1+E14_2
     replace wage14=. if wage14==0

     * First replace outliers by
     winsor2 wage14, suffix(_w) cuts(1 99)

     * Create salary categories based on winsor values
     gen salary_cat=.
     replace salary_cat=1 if inrange(wage14_w, 1, 55000)

[…]

     * Keep only for employed employees, label
     replace wage_no_compen=. if lstatus!=1|empstat==2
     label var wage_no_compen "Last wage payment primary job 7 day
recall"
*</_wage_no_compen_>
```

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.

3.3. The GLD data dictionary

This section defines one by one each variable in the data dictionary and how they should be harmonized. It is divided into blocks as is done in the harmonization code. Each block section then also contains some lessons learned if any and a tabular overview of the variables.

Survey & ID module

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

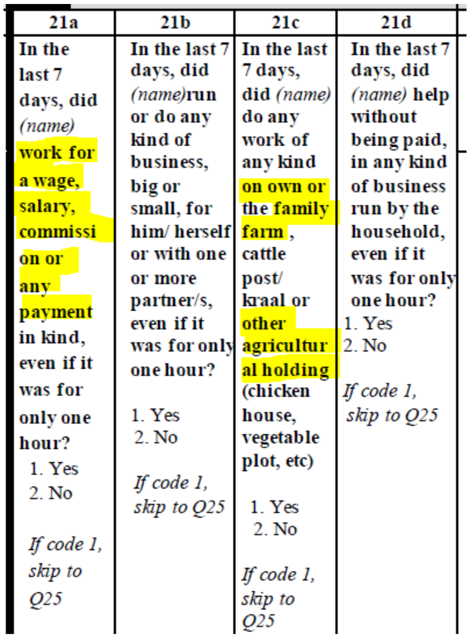
Underlying version of the International Conference of Labour Statisticians that is being used in the survey to code concepts of work and employment.

Most commonly, surveys harmonized to GLD will either follow ICLS-13 or ICLS-19, that is, the directives set out during the 13th or the 19th conference, especially pertaining employment.

In ICLS-13 all work – other than household work – is seen as employment. Thus, subsistence farmers are as employed as the CEO of an international conglomerate.

The below screenshot (Figure 13) is from the questionnaire of the Zimbabwean 2014 LFS, where any yes answer will skip to questions on main employment (Q25). As highlighted, work for a wage is treated in the same than work on any agricultural holding. This survey follows ICLS-13.

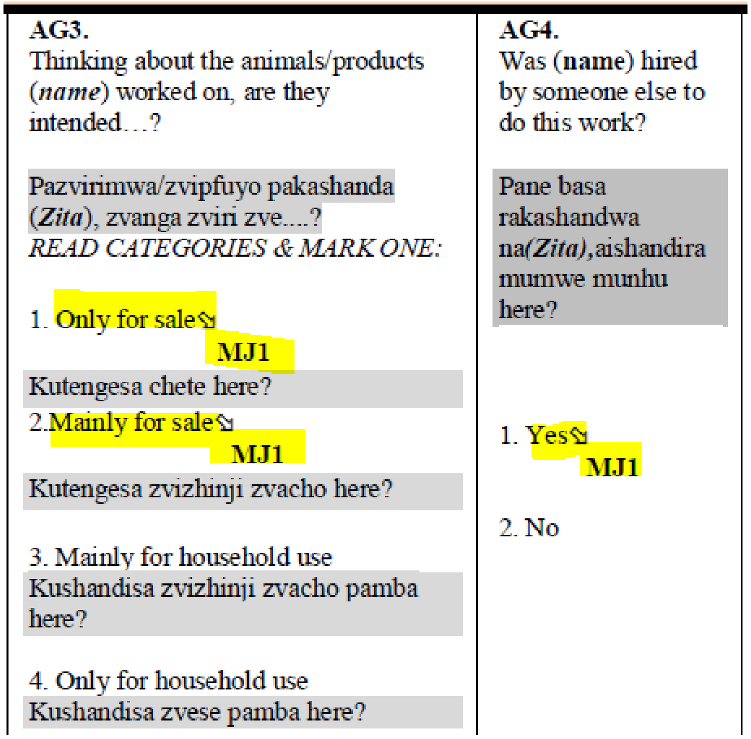
Figure 13 - Example of 2014 ZWE LFS questionnaire



21a	21b	21c	21d
In the last 7 days, did *(name)* work for a wage, salary, commission or any payment in kind, even if it was for only one hour? 1. Yes 2. No *If code 1, skip to Q25*	In the last 7 days, did *(name)*run or do any kind of business, big or small, for him/ herself or with one or more partner/s, even if it was for only one hour? 1. Yes 2. No *If code 1, skip to Q25*	In the last 7 days, did *(name)* do any work of any kind on own or the family farm , cattle post/ kraal or other agricultural holding (chicken house, vegetable plot, etc) 1. Yes 2. No *If code 1, skip to Q25*	In the last 7 days, did *(name)* help without being paid, in any kind of business run by the household, even if it was for only one hour? 1. Yes 2. No *If code 1, skip to Q25*

Five years later, the Zimbabwean statistics office, ZimStat, switched to ICLS-19. In ICLS-19, only work for market exchange is considered employment (treating subsistence farming in the same way as household labour). Thus, an additional question is added to differentiate what kind of farming is taking place. The below screenshot (Figure 14) is part of the set of agriculture questions. If the agricultural work on the own agricultural holding is only or mostly for market exchange (codes 1 and 2) the individual should be asked about their first main job (MJ1). If agricultural production is only or mainly for own consumption, then the questionnaire continues, here asking whether they work for others for hire.

Figure 14 - Example of 2019 ZWE LFS questionnaire



AG3. Thinking about the animals/products (*name*) worked on, are they intended…? Pazvirimwa/zvipfuyo pakashanda (*Zita*), zvanga zviri zve….? *READ CATEGORIES & MARK ONE:* 1. Only for sale↘ **MJ1** Kutengesa chete here? 2.Mainly for sale↘ **MJ1** Kutengesa zvizhinji zvacho here? 3. Mainly for household use Kushandisa zvizhinji zvacho pamba here? 4. Only for household use Kushandisa zvese pamba here?	**AG4.** Was (**name**) hired by someone else to do this work? Pane basa rakashandwa na(*Zita*),aishandira mumwe munhu here? 1. Yes↘ **MJ1** 2. No

If the survey asks questions to understand what kind of farming takes place (subsistence or market exchange) and defines a skip pattern to lead to employment questions based on that, the survey questionnaire follows ICLS-19, otherwise it follows ICLS-13.

isced_version

Underlying version of the International Standard Classification of Education (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 (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 (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.

weight_m

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 for each month. To be added only if present in the raw data and survey reports estimate results per month.

weight_q

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 for each quarter. To be added only if present in the raw data and survey reports estimate results per quarter.

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.

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.

panel

A string variable denoting which panel the individual belongs to. A panel is defined as all individuals who entered a survey at the same time (e.g., Q3 of 2020) and are scheduled to exit at the same time after a fixed number of survey waves (e.g., after four quarters).

Note that due to attrition not all intakes may exit at the same time. This variable is only to be coded if the concept is present in the raw data already.

visit_no

A numeric variable denoting the visit number (e.g., first visit coded as 1, second visit as 2, …) within a panel. This variable is only to be coded if the concept is present in the raw data already.

Lessons Learned and common challenges

Coding IDs correctly is integral to allow for analytical tools to leverage the information. The coding of the identifications should follow from the survey structure and should not be built via a sequential index (e.g., `gen hhid = _n`). Since observations may be ordered differently, possibly even across different vintages of the same file, the coding may lead to different outcomes.

```
        * 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 pidlocal
        n=r(unique_value)
        `N'!= `n'
```

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. 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)
```

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	
Survey & ID	Panel the individual belongs to	panel	Only code if concept already in survey
Survey & ID	Visit number in panel order	visit_no	Only code if concept already in survey

Geography

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 variable *subnatid1* or *subnatid2*. However, in some cases the lowest level is classified in terms of urban, rural (i.e., variable *urban*) or any other regional categorization cannot be mapped to *subnatid#*.

The below example (Table 3) shows how to code *subnatidsurvey* for a survey that is representative at the rural/urban level of the province (*subnatid1*).

Table 3 - Example of a survey significant at subnatid1 and urban/rural level

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”
…	…	…
“2 – Province B”	0	“2 – Province B rural”
“2 – Province B”	0	“2 – Province B rural”
“2 – Province B”	0	“2 – Province B rural”
“2 – Province B”	0	“2 – Province B rural”
“2 – Province B”	1	“2 – Province B urban”
“2 – Province B”	1	“2 – Province B urban”

While the below (Table 4) is the example of a survey representative nationally, nationally at urban / rural level and at province level.

Table 4 - Example of a survey significant at subnatid1 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”
…	…	…
“2 – Province B”	0	“2 – Province B”
“2 – Province B”	0	“2 – Province B”
“2 – Province B”	0	“2 – Province B”
“2 – Province B”	0	“2 – Province B”
“2 – Province B”	1	“2 – Province B”
“2 – Province B”	1	“2 – 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.

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 stratum. 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. The strata variable 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 stratum. 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 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 codescan be used to track splits and new administrative units that have been introduced since the previoussurvey. 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
```

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	strata	Strata	
Geography	gaul_adm[i]_code	GAUL ADM[i] code	See en.wikipedia.org/wiki/Global_Administrative_Unit_Layers

Demography

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.

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

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 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 to Galicia, 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. Set to missing if migrated_from_cat is 5.

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 0 below.

migrated_from_country

Codes the country (if migrated_from_cat is 5) from where the person migrated from as a three letter ISO country code 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_>
```

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

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

The concept of secondary complete includes all students who have had attended at least one year (complete or incomplete) of upper secondary education (as defined in the ISCED Mappings of UNESCO). That is attendance and completion of "junior high" shall be coded as secondary incomplete while attendance to "senior high school" even if for one year, will be coded as secondary 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)

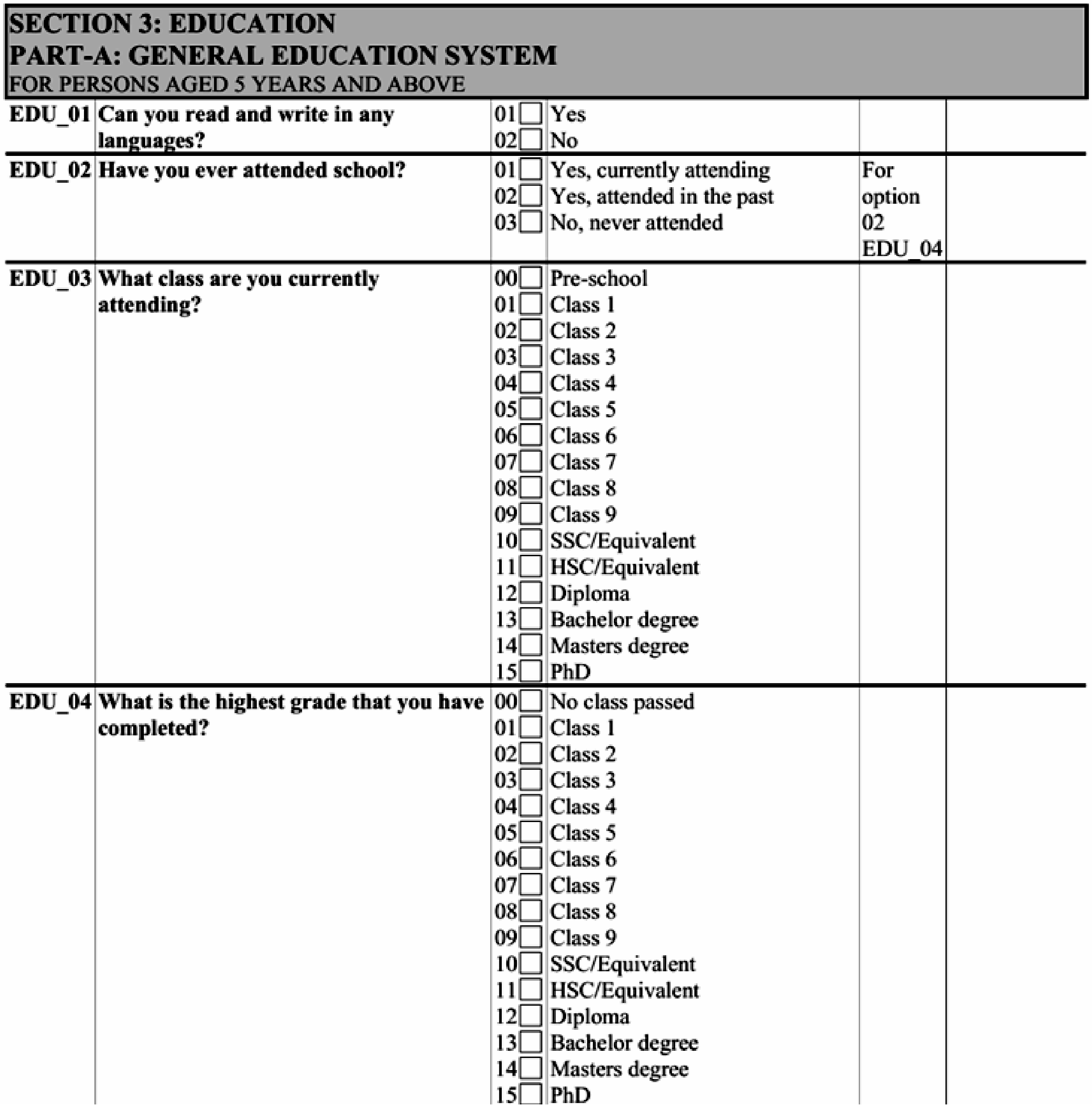
Note: Code as primary education anyone who has undergone some schooling but has not finished secondary education.

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 splits the respondents into different groups (commonly a question for people attending school and a different one for those no longer in education), then educat_orig should be made up of both variables. Figure 15 below, shows the example from the 2022 Bangladesh LFS. The skip pattern from variable EDU_02 is not clear, but in the data people who are attending (EDU_02 code 01) answer question EDU_03 and skip EDU_04, while those who state that they attended in the past but are not currently enrolled (EDU_02 code 02) skip EDU_03 and answer EDU_04. People who never attended school (EDU_02 code 03) skip both EDU_03 and EDU_04. In this case educat_orig should be coded as the union of the information of either group (i.e., educat_orig ought to contain the information of both EDU_03 and EDU_04).

Figure 15 - Education variables in the 2022 Bangladeshi LFS

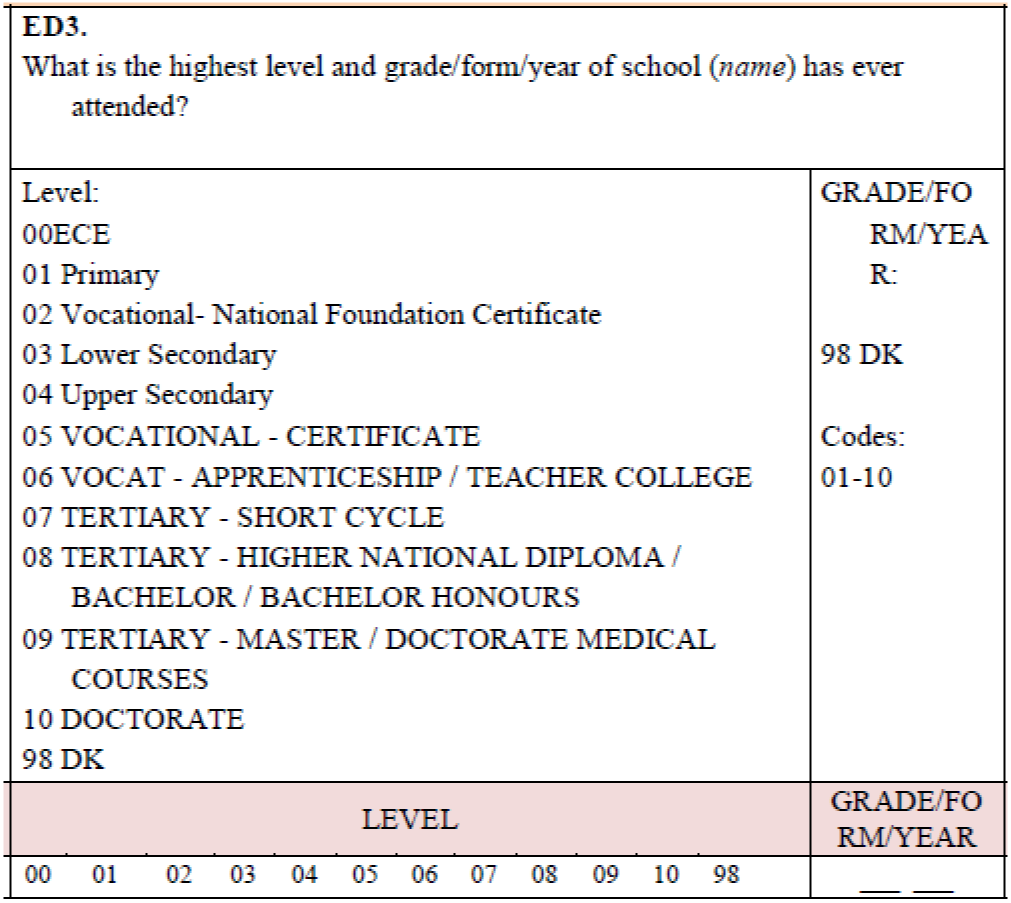


SECTION 3: EDUCATION
PART-A: GENERAL EDUCATION SYSTEM
FOR PERSONS AGED 5 YEARS AND ABOVE

EDU_01	**Can you read and write in any languages?**	01□	Yes		
		02□	No		
EDU_02	**Have you ever attended school?**	01□	Yes, currently attending	For option 02 EDU_04	
		02□	Yes, attended in the past		
		03□	No, never attended		
EDU_03	**What class are you currently attending?**	00□	Pre-school		
		01□	Class 1		
		02□	Class 2		
		03□	Class 3		
		04□	Class 4		
		05□	Class 5		
		06□	Class 6		
		07□	Class 7		
		08□	Class 8		
		09□	Class 9		
		10□	SSC/Equivalent		
		11□	HSC/Equivalent		
		12□	Diploma		
		13□	Bachelor degree		
		14□	Masters degree		
		15□	PhD		
EDU_04	**What is the highest grade that you have completed?**	00□	No class passed		
		01□	Class 1		
		02□	Class 2		
		03□	Class 3		
		04□	Class 4		
		05□	Class 5		
		06□	Class 6		
		07□	Class 7		
		08□	Class 8		
		09□	Class 9		
		10□	SSC/Equivalent		
		11□	HSC/Equivalent		
		12□	Diploma		
		13□	Bachelor degree		
		14□	Masters degree		
		15□	PhD		

In other cases, the information may be split into two sub-questions. Figure 16, below, shows the question on education from the 2021 Zimbabwean LFS. Here the information is split into two variables: a level (which we will refer to as ED3a) and a grade within that level (ED3b).

Figure 16 - Education question in the 2021 ZWE LFS



ED3.
What is the highest level and grade/form/year of school (*name*) has ever attended?

Level: 00ECE 01 Primary 02 Vocational- National Foundation Certificate 03 Lower Secondary 04 Upper Secondary 05 VOCATIONAL - CERTIFICATE 06 VOCAT - APPRENTICESHIP / TEACHER COLLEGE 07 TERTIARY - SHORT CYCLE 08 TERTIARY - HIGHER NATIONAL DIPLOMA / BACHELOR / BACHELOR HONOURS 09 TERTIARY - MASTER / DOCTORATE MEDICAL COURSES 10 DOCTORATE 98 DK	GRADE/FORM/YEAR: 98 DK Codes: 01-10
LEVEL	GRADE/FORM/YEAR
00 01 02 03 04 05 06 07 08 09 10 98	___ ___

If it is possible to combine the information (e.g., by concatenating both variable) this should be tried and explicitly commented in the code. If it is not possible to confidently convey the information split over two (or more) variables, the harmonizer shall leave the educat_orig variable 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).

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

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_orig

Information on the field of training as stored originally in the survey. This variable is to be a string variable. If numeric convert to string while preserving its original structure.

vocational_financed

Text. The codes are:

1 = Employer

2 = Government

3 = Mixed Employer-Government

4 = Own funds

5 = Other

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_orig	Original field of training information	
Education	vocational_financed	How training was financed	If funded with different sources, chose main source

Labour

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.

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' (i.e., passively seeking a job). This last question is not asked in all surveys but should be included if present.. 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.

Over the past decade, a significant alteration has occurred in survey methodologies regarding the definition of employment, specifically affecting labor status. This change stems from the deliberations of the International Conference of Labour Statisticians (ICLS), a forum for standardization in labor statistics convened by the International Labour Organization (ILO) every five years. During the 19th session of the ICLS in 2013, delegates passed a resolution pertaining to the categorization of work, employment, and labor underutilization.

In essence, the ICLS-19 resolution redefines employment as work performed for others in exchange for pay or profit, excluding activities such as subsistence agriculture or self-housing construction, which were previously classified as employment. This revision necessitates a careful adjustment in coding practices for variables associated with labor status, such as the "lstatus" variable.

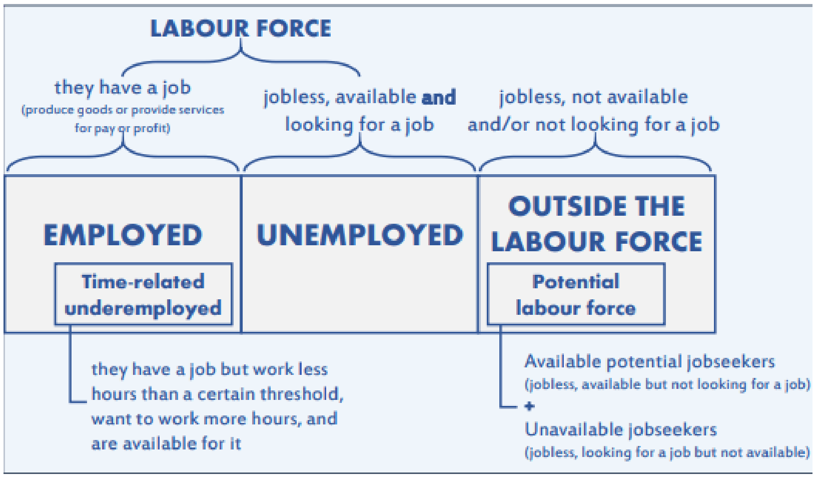
To ensure accurate coding, harmonizers should discern the underlying ICLS version utilized in each survey through thorough examination of questionnaire text, skip patterns, and survey reports. Consequently, harmonizers must code the "lstatus" variable in accordance with the pertinent ICLS version (and annotate the version used via the "icls_v" variable).

It is imperative to emphasize that each survey undergoes independent harmonization as explained in section Defining the boundaries of GLD harmonization. Therefore, any disparities arising from the adoption of different employment definitions in previous surveys should not impede the harmonization process for the current survey under consideration.

potential_lf

The ILO defines unemployment (as stated above) as seeking *and* available for a job. The potential labour force is formed by the “available potential job seekers”, who are available but not looking for a job and the “unavailable job seekers”, that is those looking but not available. The below image (Figure 17 – source here) shows the different definitions.

Figure 17 - Definition of different labor force status


LABOUR FORCE
they have a job
(produce goods or provide services for pay or profit)
jobless, available and looking for a job
jobless, not available and/or not looking for a job
EMPLOYED
Time-related underemployed
UNEMPLOYED
OUTSIDE THE LABOUR FORCE
Potential labour force
they have a job but work less hours than a certain threshold, want to work more hours, and are available for it
Available potential jobseekers
(jobless, available but not looking for a job)
+
Unavailable jobseekers
(jobless, looking for a job but not available)

The variable potential_lf thus 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.

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

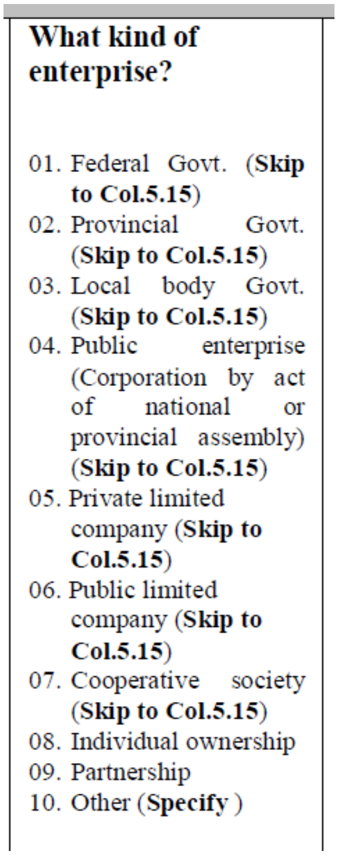
State-owned includes para-state firms and all others in which the government has control (participation over 50%).

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.

Additionally, recall the fact that, in common English usage, a public company (often denoted as public limited company or PLC) are private companies in the private sector but whose ownership is organized via stocks tradeable in a *public* market, i.e., accessible to all, not run by the public sector. Figure 18 below is an example of the relevant question in the 2010 Pakistani LFS:

Figure 18 - Excerpt of PAK LFS questionnaire



What kind of enterprise?

01. Federal Govt. (**Skip to Col.5.15**)
02. Provincial Govt. (**Skip to Col.5.15**)
03. Local body Govt. (**Skip to Col.5.15**)
04. Public enterprise (Corporation by act of national or provincial assembly) (**Skip to Col.5.15**)
05. Private limited company (**Skip to Col.5.15**)
06. Public limited company (**Skip to Col.5.15**)
07. Cooperative society (**Skip to Col.5.15**)
08. Individual ownership
09. Partnership
10. Other (**Specify**)

Here code 4 represents a public enterprise (explicitly mentioned as a corporation created by a legislative body) and code 6 a public limited company (a private sector company whose shares can be bought by the general public). *Code 6 should not be considered part of the public sector*.

Coding would then be (assuming the original question is called var_sector):

```
gen ocusec = .

replace ocusec = 1 if inrange(var_sector,1,3)

replace ocusec = 2 if inrange(var_sector,5,9)

replace ocusec = 3 if var_sector == 4

replace ocusec = 4 if var_sector == 10
```

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

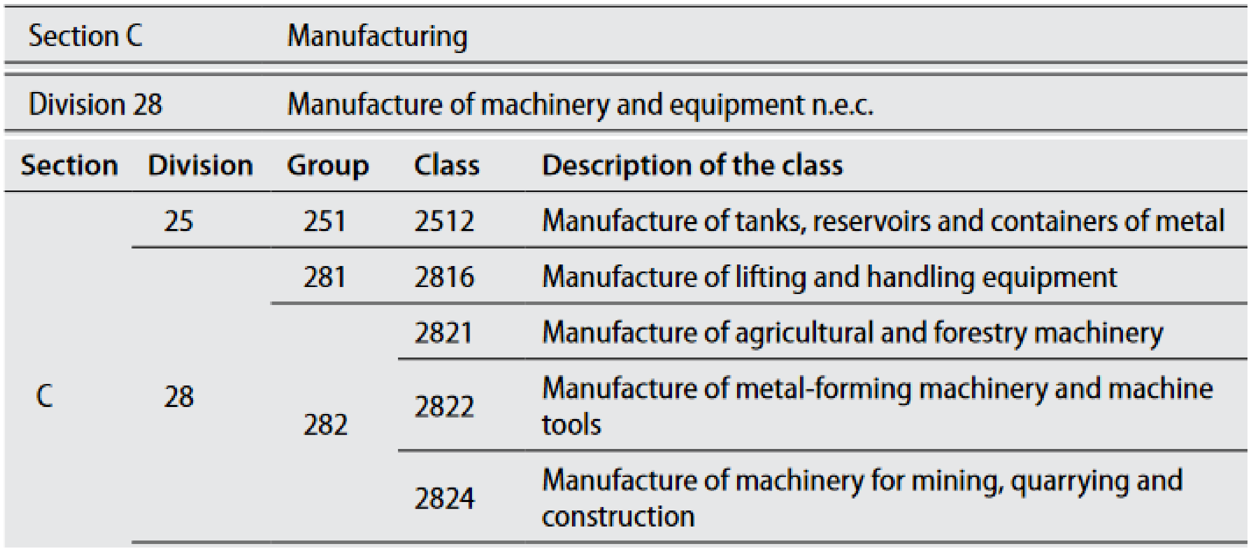
industrycat_isic

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) codes industries by sections, divisions, groups, and classes, in decreasing order of hierarchy.

Figure 19 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 19 - Example of ISIC classification



Section C	Manufacturing
Division 28	Manufacture of machinery and equipment n.e.c.

Section	Division	Group	Class	Description of the class
C	25	251	2512	Manufacture of tanks, reservoirs and containers of metal
	28	281	2816	Manufacture of lifting and handling equipment
		282	2821	Manufacture of agricultural and forestry machinery
			2822	Manufacture of metal-forming machinery and machine tools
			2824	Manufacture of machinery for mining, quarrying and construction

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).

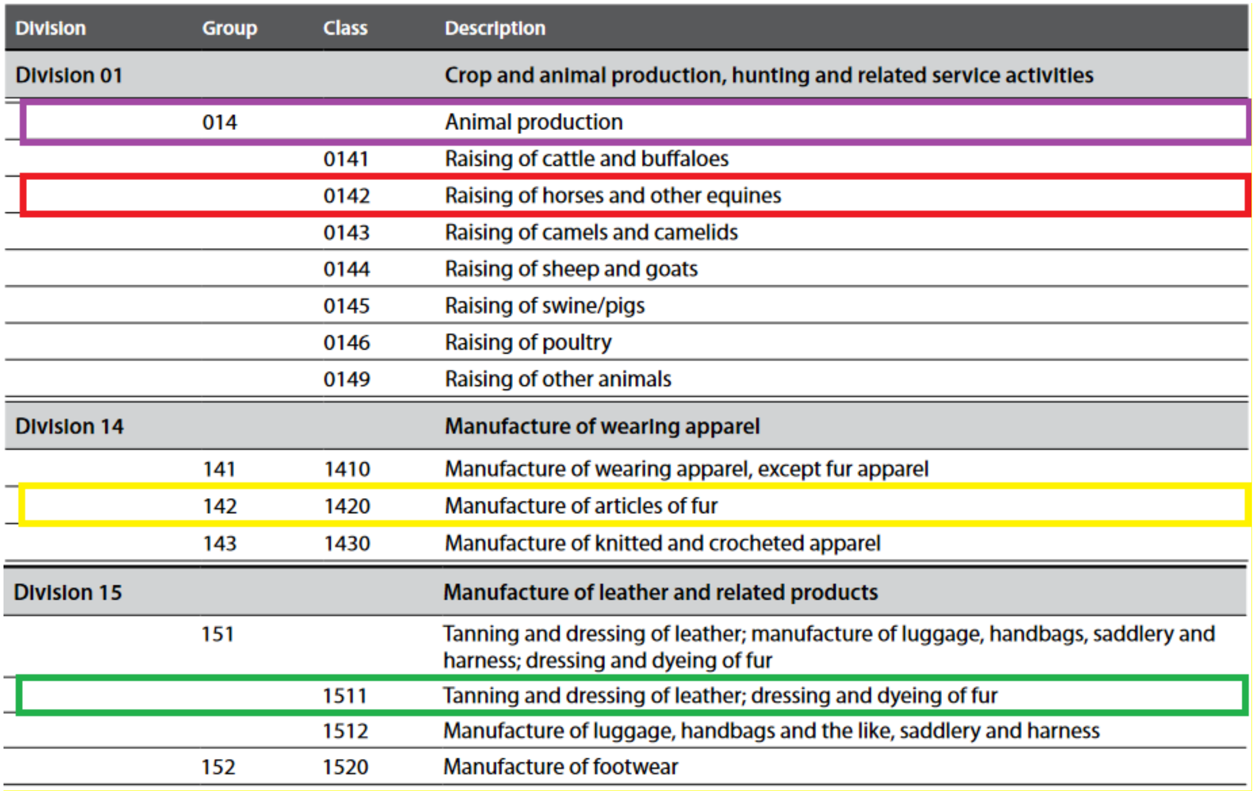
As an example, Figure 20, puts together a few excerpts from ISIC Rev.4. Note that, if we do not codify correctly, group 14 (Animal production) may be misunderstood for division 14 (Manufacture of wearing apparel).

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 20 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 in Figure 20 below) we would code “1511”.

Figure 20 - Examples of different ISIC codes



Division	Group	Class	Description
Division 01			**Crop and animal production, hunting and related service activities**
	014		Animal production
		0141	Raising of cattle and buffaloes
		0142	Raising of horses and other equines
		0143	Raising of camels and camelids
		0144	Raising of sheep and goats
		0145	Raising of swine/pigs
		0146	Raising of poultry
		0149	Raising of other animals
Division 14			**Manufacture of wearing apparel**
	141	1410	Manufacture of wearing apparel, except fur apparel
	142	1420	Manufacture of articles of fur
	143	1430	Manufacture of knitted and crocheted apparel
Division 15			**Manufacture of leather and related products**
	151		Tanning and dressing of leather; manufacture of luggage, handbags, saddlery and harness; dressing and dyeing of fur
		1511	Tanning and dressing of leather; dressing and dyeing of fur
		1512	Manufacture of luggage, handbags and the like, saddlery and harness
	152	1520	Manufacture of footwear

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

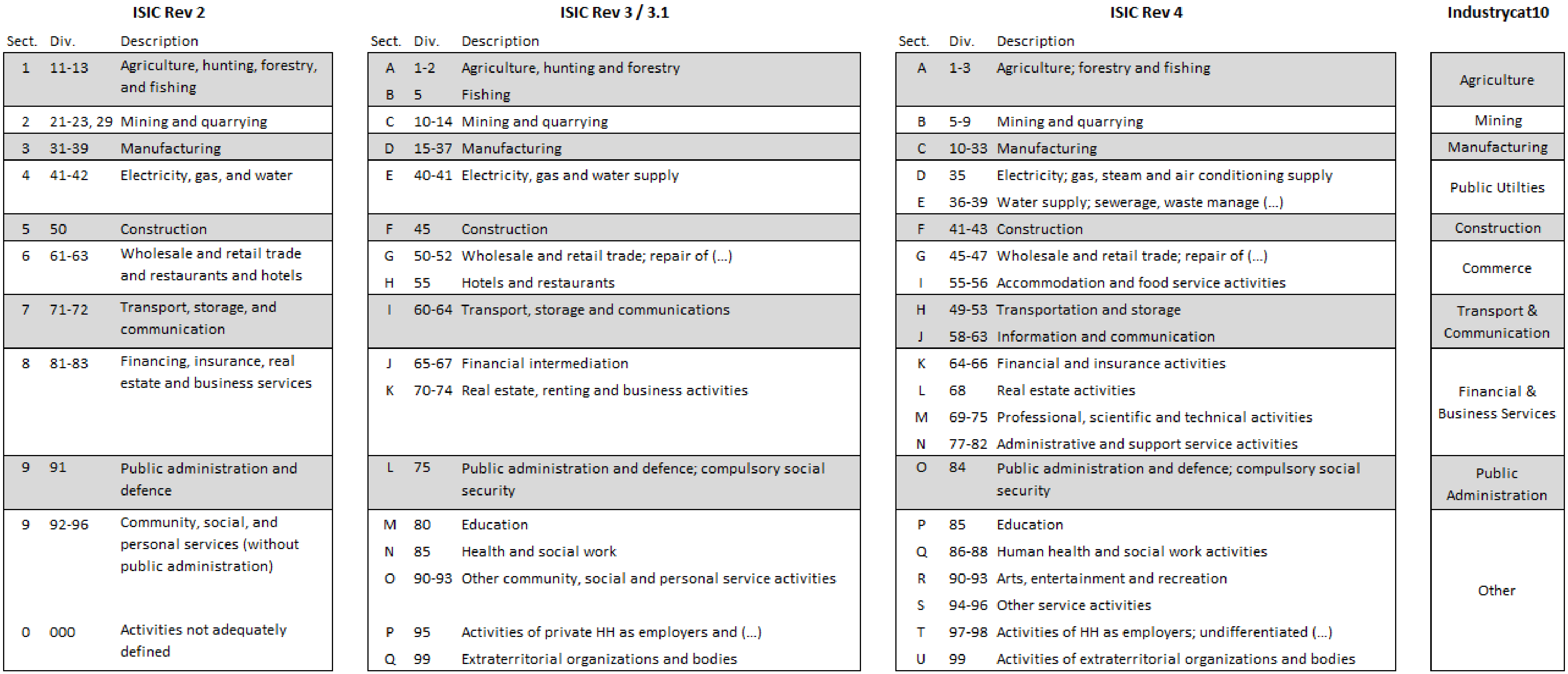
9 = Public Administration

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 Figure 21 shows how to classify the different ISIC revision codes into industrycat10.

Figure 21 - Overview of coding of industrycat based on different ISIC revisions



ISIC Rev 2			ISIC Rev 3 / 3.1			ISIC Rev 4			Industrycat10
Sect.	Div.	Description	Sect.	Div.	Description	Sect.	Div.	Description	
1	11-13	Agriculture, hunting, forestry, and fishing	A	1-2	Agriculture, hunting and forestry	A	1-3	Agriculture; forestry and fishing	Agriculture
			B	5	Fishing				
2	21-23, 29	Mining and quarrying	C	10-14	Mining and quarrying	B	5-9	Mining and quarrying	Mining
3	31-39	Manufacturing	D	15-37	Manufacturing	C	10-33	Manufacturing	Manufacturing
4	41-42	Electricity, gas, and water	E	40-41	Electricity, gas and water supply	D	35	Electricity; gas, steam and air conditioning supply	Public Utilties
						E	36-39	Water supply; sewerage, waste manage (...)	
5	50	Construction	F	45	Construction	F	41-43	Construction	Construction
6	61-63	Wholesale and retail trade and restaurants and hotels	G	50-52	Wholesale and retail trade; repair of (...)	G	45-47	Wholesale and retail trade; repair of (...)	Commerce
			H	55	Hotels and restaurants	I	55-56	Accommodation and food service activities	
7	71-72	Transport, storage, and communication	I	60-64	Transport, storage and communications	H	49-53	Transportation and storage	Transport & Communication
						J	58-63	Information and communication	
8	81-83	Financing, insurance, real estate and business services	J	65-67	Financial intermediation	K	64-66	Financial and insurance activities	Financial & Business Services
			K	70-74	Real estate, renting and business activities	L	68	Real estate activities	
						M	69-75	Professional, scientific and technical activities	
						N	77-82	Administrative and support service activities	
9	91	Public administration and defence	L	75	Public administration and defence; compulsory social security	O	84	Public administration and defence; compulsory social security	Public Administration
9	92-96	Community, social, and personal services (without public administration)	M	80	Education	P	85	Education	Other
			N	85	Health and social work	Q	86-88	Human health and social work activities	
			O	90-93	Other community, social and personal service activities	R	90-93	Arts, entertainment and recreation	
						S	94-96	Other service activities	
0	000	Activities not adequately defined	P	95	Activities of private HH as employers and (...)	T	97-98	Activities of HH as employers; undifferentiated (...)	
			Q	99	Extraterritorial organizations and bodies	U	99	Activities of extraterritorial organizations and bodies	

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.

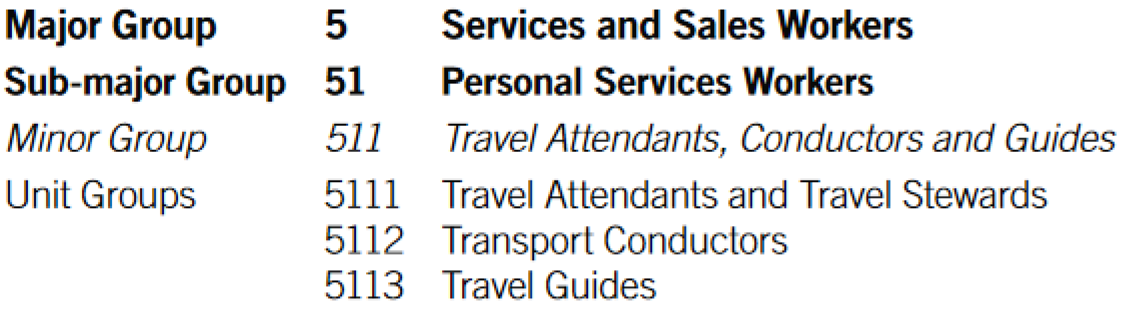
occup_isco

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) 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 22 shows an example of the possible values that can be taken on.

Figure 22 - Example of values for ISCO-08



Major Group	**5**	**Services and Sales Workers**
Sub-major Group	**51**	**Personal Services Workers**
Minor Group	*511*	*Travel Attendants, Conductors and Guides*
Unit Groups	5111	Travel Attendants and Travel Stewards
	5112	Transport Conductors
	5113	Travel Guides

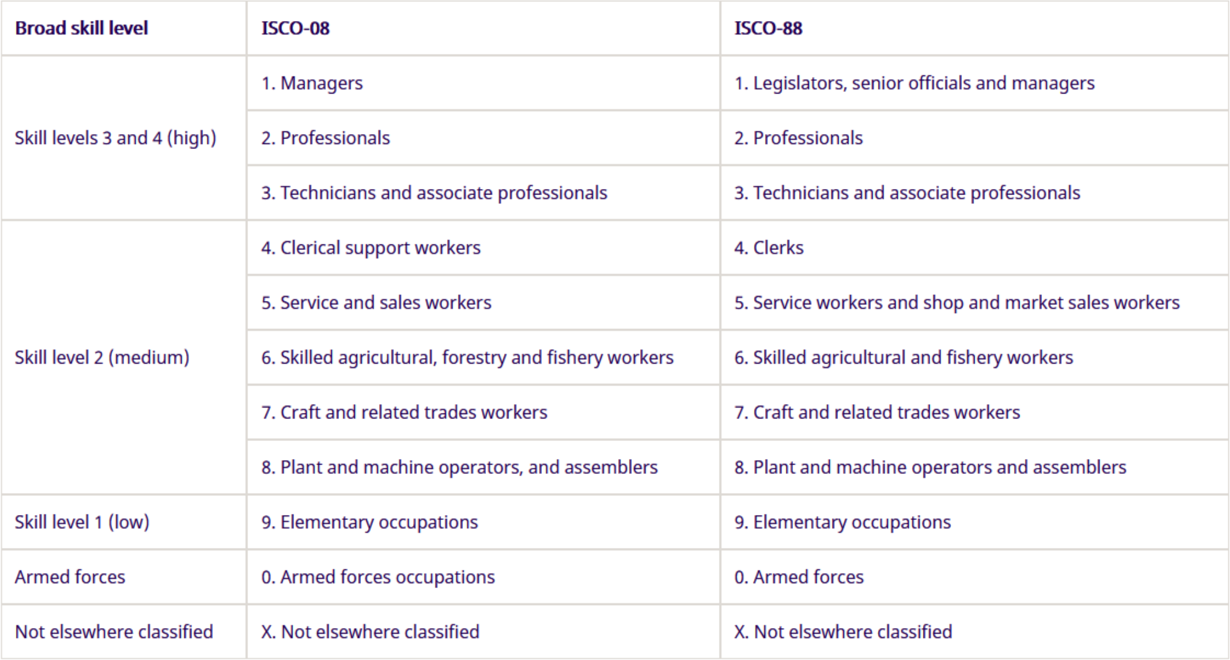
If we only had information at Major Group level, a person working as a Services and Sales Worker ought to be coded as the number 5000. If the information is at Sub-major Group level, it should be codified as the number 5100 for a Personal Services Worker, while a Travel Attendant, a Conductor, or a Guide (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 23.

Figure 23 - ISCO broad skill level classification



Broad skill level	ISCO-08	ISCO-88
Skill levels 3 and 4 (high)	1. Managers	1. Legislators, senior officials and managers
	2. Professionals	2. Professionals
	3. Technicians and associate professionals	3. Technicians and associate professionals
Skill level 2 (medium)	4. Clerical support workers	4. Clerks
	5. Service and sales workers	5. Service workers and shop and market sales workers
	6. Skilled agricultural, forestry and fishery workers	6. Skilled agricultural and fishery workers
	7. Craft and related trades workers	7. Craft and related trades workers
	8. Plant and machine operators, and assemblers	8. Plant and machine operators and assemblers
Skill level 1 (low)	9. Elementary occupations	9. Elementary occupations
Armed forces	0. Armed forces occupations	0. Armed forces
Not elsewhere classified	X. Not elsewhere classified	X. Not elsewhere classified

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 preceding 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 and written – not verbal) 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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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?

4. Validation and quality checks

4.1. How is the harmonization validated?

The harmonization is checked thoroughly and continuously, to ensure the output is the most accurate and reliable. There are three main processes. Chronologically the first are the checks performed during the harmonization. As the code is developed, the harmonizers evaluate the results they are getting on an ongoing basis. Moreover, they try to confirm at early stages some of the more difficult mappings, like the conversion of survey education codes to the categories of the harmonized education variables. This process includes reaching out to World Bank colleagues in country offices and partners at the statistical offices to validate choices.

The second process are the automated quality checks, made up of a standard code that uses the harmonization as input and it for internal and external consistency, as well as checking a series over surveys over time. They are the main source of validation and are discussed further in detail in the remainder of the section.

The third process starts after the harmonization is published by collecting users' feedback. GLD encourages users to review the harmonization and the codes and alert the team of any issues that may still have slipped previous processes. Users can raise issues on GitHub (see How to communicate on any other issue section) or reach out to the GLD Focal Point. The GLD Team has set up code pipelines to try to update harmonizations to respond to user feedback as quickly as possible. Your feedback is an integral part of GLD!

The automated quality checks are two sets of functions, one in `Stata` one in `R` to evaluate the output of the harmonization. The `Stata` code evaluates each survey individually, that is we check the particular "IND_2018_PLFS" harmonization for coherence and consistency, both internally (e.g., is the urban/rural variables only codes urban or rural; do people classified as professionals have, on average, the education level we expect them to have) and externally (is the labor force participation from the survey in line with ILO, WDI?).

To interpret the results, users should differentiate between errors and flags. By errors we mean things in the harmonization that can be evaluated and are seen as wrong. An example would be an observation with value `3` in the `urban` variable, when the only possible values are `0`, `1`, or `NA` if the information is not available.

Flags are issues with the data that hint at an issue. For example, if the variable `urban` has a high share of missing (i.e., `NA`) answers. This is uncommon and thus points to an issue, but it need not be incorrect: the underlying raw data may have an issue and the data are truly not available. For such cases, it is advisable to include the issue in the Country Survey Details, but no further correction of the harmonization code would be warranted.

The quality checks do not specifically name issues as errors or flags. This is for the user to evaluate, based on what the checks do, which is covered in the next two subsections.

4.2. Guide to the GLD single survey quality checks

This is a guide to the single survey quality checks for the Global Labor Database (GLD), one half of the automated quality control process after harmonization (yet before publication). After reading this subsection, users should be able to:

- understand what the quality checks evaluate and how they work

- run the checks by themselves on a newly harmonized file, and

- read and interpret the output

The guide is divided itself in three major sections, one for each learning objective.

Understanding the quality checks

The GLD quality checks intend to make sure that the GLD harmonized output is of sufficient quality to include the evaluated harmonized survey into the database. This is done in five blocks:

- Block 1 checks adherence of the survey data to the GLD formatting requirements (e.g., binary 0/1 variables have no responses outside 0 and 1).

- Block 2 compares key indicators from the harmonization to the same indicators from external sources (e.g., ILO, UN), to ensure external consistency.

- Block 3 assesses whether major variables are missing.

- Block 4 evaluates whether key relationships between pairs of variables align with ex-ante expectations (e.g., individuals occupied as professionals - `occup == 2` - chiefly have post-secondary education).

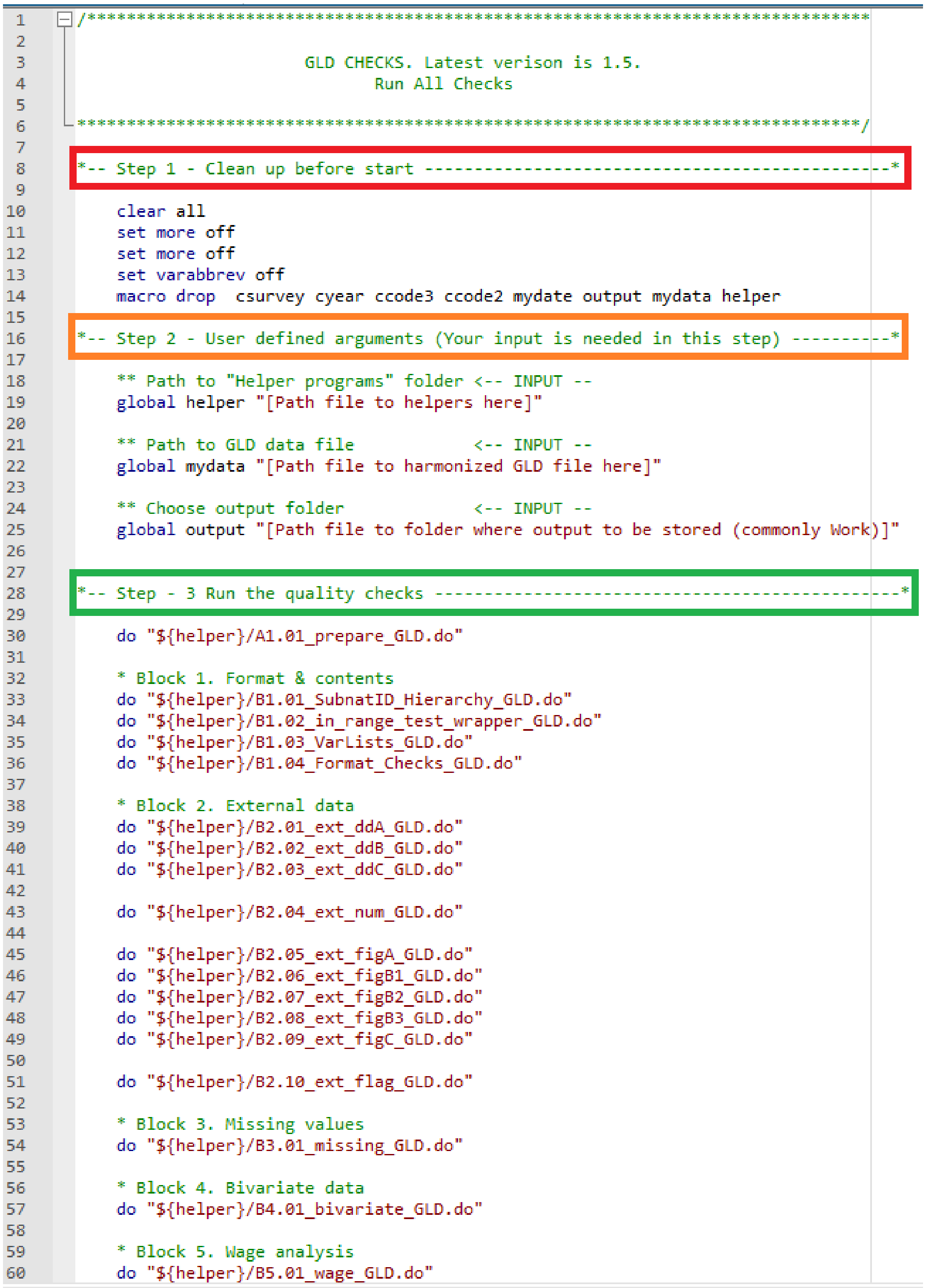
- Block 5 analyses central aspects of wages to ensure wage/earnings information agrees with stylized facts on wage behaviour (e.g., lower skilled workers' wages are, on average, lower than those of high skilled workers).

The next subsections discuss each block in more detail. We start, however, with the overall quality checks template.

4.2.1. Overall quality checks template

The overall quality checks template is the only do-file the user needs to interact with regularly. It defines al relevant arguments and calls the do-files that run blocks 1 through 5. The code is accessible here and shown below in Figure 24:

Figure 24 - GLD single survey quality checks template



```
/*******************************************************************************


                       GLD CHECKS. Latest verison is 1.5.
                              Run All Checks

*******************************************************************************/

*-- Step 1 - Clean up before start ---------------------------------------------*

    clear all
    set more off
    set more off
    set varabbrev off
    macro drop  csurvey cyear ccode3 ccode2 mydate output mydata helper

*-- Step 2 - User defined arguments (Your input is needed in this step) ---------*

    ** Path to "Helper programs" folder <-- INPUT --
    global helper "[Path file to helpers here]"

    ** Path to GLD data file            <-- INPUT --
    global mydata "[Path file to harmonized GLD file here]"

    ** Choose output folder             <-- INPUT --
    global output "[Path file to folder where output to be stored (commonly Work)]"


*-- Step - 3 Run the quality checks ---------------------------------------------*

    do "${helper}/A1.01_prepare_GLD.do"

    * Block 1. Format & contents
    do "${helper}/B1.01_SubnatID_Hierarchy_GLD.do"
    do "${helper}/B1.02_in_range_test_wrapper_GLD.do"
    do "${helper}/B1.03_VarLists_GLD.do"
    do "${helper}/B1.04_Format_Checks_GLD.do"

    * Block 2. External data
    do "${helper}/B2.01_ext_ddA_GLD.do"
    do "${helper}/B2.02_ext_ddB_GLD.do"
    do "${helper}/B2.03_ext_ddC_GLD.do"

    do "${helper}/B2.04_ext_num_GLD.do"

    do "${helper}/B2.05_ext_figA_GLD.do"
    do "${helper}/B2.06_ext_figB1_GLD.do"
    do "${helper}/B2.07_ext_figB2_GLD.do"
    do "${helper}/B2.08_ext_figB3_GLD.do"
    do "${helper}/B2.09_ext_figC_GLD.do"

    do "${helper}/B2.10_ext_flag_GLD.do"

    * Block 3. Missing values
    do "${helper}/B3.01_missing_GLD.do"

    * Block 4. Bivariate data
    do "${helper}/B4.01_bivariate_GLD.do"

    * Block 5. Wage analysis
    do "${helper}/B5.01_wage_GLD.do"
```

The quality checks template proceeds in three steps. Step 1 readies Stata by cleaning up any data that may still be stored in the memory. Step 2 defines the arguments. It is the only section requiring user input. Users need to define three `globals`.

- `helper`: Define the path to the folder that contains all the files that run the quality checks (here, folder Helper_programs_1.5). It is recommended to have this be a central place so it applies to all surveys and can easily be updated if the quality checks are amended.

- `mydata`: Define the path to the harmonized data the user wishes to check.

- `output`: Define the path to the folder the user wants the output to be stored in. It is recommended to make this the `CCC_YYYY_SURV_V0#_M_V0#_A_GLD/Work/Output` folder for consistency so other users may always know where to find the checks output.

Step 3 simply calls the do files from the helper path running the checks. This step no longer requires user involvement. It does, however, run through all the files that make up the checks. What these files cover is the subject of the next section.

4.2.2. Block 1 - GLD format checks

This block (with all files starting with `B1`) is concerned with ensuring that the harmonized output conforms to the GLD data dictionary. The code for the format checks is divided into sections (overall checks, demography checks, education checks, ...) and should be commented enough for users to understand what each step is doing. If you feel more explanations are warranted, please raise and issue detailing the part not understood and we will endeavour to expand on it.

Broadly, the checks first evaluate whether the survey is set up correctly: filenames follow the naming convention and variables are from the data dictionary.

Once this is established, the checks go through variables and, if they are in teh data, investigates whether they are in line with formatting rules. Some examples are:

- Check if vintages of survey (`vermast`/`veralt`) agree with the filename vintages (e.g., if file is `CCC_YYYY_SURV_V01_M_V02_A_GLD` `veralt` needs to be `"02"`).

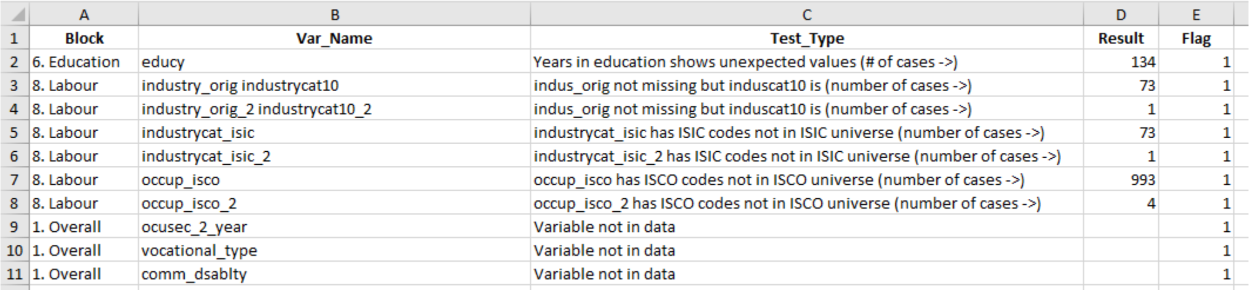
- Check there is a single household head per household.

- Check that categorical variables have answers exclusively within the realm of feasible answers.

- Check that ISCO/ISIC codes are in the universe of possible codes for the pertinent classification (e.g., in ISIC 3 codes starting with `45` may only have either `0` or `1` through `5` as third digit; anything else would be outside the possible universe and flagged).

The format checks create a list of flagged issues. It is stored as a `.dta` file in the `Block1_Format` folder (file `CCC_YYYY_Other Household Survey_Q_Format_Checks.dta`), and as an Excel spreadsheet in the `01_summary` folder (file `B1_format_results.xlsx`). Figure 25 below shows a screenshot of the spreadsheet output:

Figure 25 - Example of the output from the GLD format checks



	A	B	C	D	E
1	**Block**	**Var_Name**	**Test_Type**	**Result**	**Flag**
2	6. Education	educy	Years in education shows unexpected values (# of cases ->)	134	1
3	8. Labour	industry_orig industrycat10	indus_orig not missing but induscat10 is (number of cases ->)	73	1
4	8. Labour	industry_orig_2 industrycat10_2	indus_orig not missing but induscat10 is (number of cases ->)	1	1
5	8. Labour	industrycat_isic	industrycat_isic has ISIC codes not in ISIC universe (number of cases ->)	73	1
6	8. Labour	industrycat_isic_2	industrycat_isic_2 has ISIC codes not in ISIC universe (number of cases ->)	1	1
7	8. Labour	occup_isco	occup_isco has ISCO codes not in ISCO universe (number of cases ->)	993	1
8	8. Labour	occup_isco_2	occup_isco_2 has ISCO codes not in ISCO universe (number of cases ->)	4	1
9	1. Overall	ocusec_2_year	Variable not in data		1
10	1. Overall	vocational_type	Variable not in data		1
11	1. Overall	comm_dsablty	Variable not in data		1

As shown, the output has five columns:

- Column Block describes the data dictionary block the variable causing the flag belongs to (demography, education, ...).

- Column Var_Name denotes the name of the variable flagged.

- Column Test_Type describes the kind of test the variable has undergone.

- Column Result denotes the number or share of cases with the issue (if this can be written).

- Column Flag guides the reader whether the issue is at odds with the data dictionary (`Flag = 1`) or technically possible, yet odd (`Flag = 99`).

4.2.3. Block 2 - GLD external checks

Overview

The GLD external comparison checks are a series of checks that compare GLD data (aggregate moments) with comparable statistics from external sources. It is the most extensive block, covering - currently – 10 code files, all starting with `B2`. In particular, we check the following data from GLD surveys:

Sub-Block 2.1. Demographics (file `B2.01`)

1. Total population (number)
2. Gender split (% of female population)
3. Urban population (% of total)
4. Children (0-14 years old, % of total population)
5. Working Age Population (15-64 years old, % of total population)
6. Seniors (65+ years old, % of total population)

Sub-Block 2.2. Labor Force Variables (file `B2.02`)

7. Labor force size (number)
8. Labor force participation rate (labor force / working age population)
9. Employment (number)
10. Employment to population ratio (employed / working age population)

11. Unemployment rate (unemployed / labor force)
12. Agriculture (% of employment)
13. Industry (% of employment)
14. Services (% of employment)
15. Industry category (% of employment)
16. Status in Employment (% of employment)

Sub-Block 3.3. Wage variables (file `B2.03`)

17. Average hourly wages

Variable aggregation in GLD data

To construct aggregate indicators from the GLD surveys, we weight sums and means by the weight GLD variable. The demographic variables are fully described in the previous section and computed in the do file "01_checks_demo_GLD.do".

The "Labor Force Variables" from block 2 are defined as follows:

- Labor force size is the number of persons aged 15+ that are either employed or unemployed.
- Labor force participation rate is the labor force size divided by the population aged 15+.
- Employment is the number of persons aged 15+ that are employed.
- Employment to population ratio is employment (as defined above) divided by the population aged 15+.
- Unemployment rate is the number of unemployed people over the number of employed plus unemployed people. Note: no age requirement.
- Agriculture is the number of people employed that work in the agricultural or primary sector, over total people employed, times 100. Note: no age requirement
- Industry is the number of people employed that work in the industrial or secondary sector, over total people employed, times 100. Note: no age requirement
- Services is the number of people employed that work in the service or tertiary sector, over total people employed, times 100. Note: no age requirement

Notice that agriculture, industry, and services are three economic activity categories that cover all possible employment sectors.

- Industry category consists of 10 economic activity categories that cover all possible employment sectors, at a more granular level than the categories in the previous point. This variable is actually a set of 10 variables, defined as the number of people in each industry category, divided by the number of people with information in the "industrycat10" variable in GLD.

External comparison sources

Each variable listed in Overview is matched with one or more comparable variables from external sources. The sources used are:

- World Development Indicators (WDI), from the World Bank. This is a collection of development indicators compiled from international sources. The data is downloaded using the Stata module wbopendata.

- ILOSTAT, from the International Labor Organization, accessed through the Stata module dbnomics, developed by the CEPREMAP.
- UNdata, from the United Nations, also accessed through dbnomics.

Note that the external data updates periodically. The GLD checks, will automatically use the latest data available in each of these sources. The Annex contains a complete list of the variables selected from these sources, along with the GLD variables they are matched with.

Output: content overview

The survey checks process will create a folder called `Block2_External`, containing two folders:

- 01_data
- 02_figures

Each file in the 01_data folder can be matched to a figure in 02_figures. It contains the necessary data to generate that figure. Each .dta file consists of the following variables:

- year: the year for each value in the data set.
- value: the relevant point estimate for each source, represented as a dot in the figures.
- ub: the upper bound of the value, set at 1.1*value (10% above) for most variables.
- lb : the lower bound of the value, set at 0.1*value (10% below) for most variables.
- countrycode: the 2 or 3-digit code corresponding to the country of the survey.
- source: string name of the source behind each value. One of the entries in this column will be "GLD", the rest will be the external sources we use to compare.

The figures folder contains 25 pdf files named Fig1.pdf to Fig16.pdf, with 10 versions of Fig15, called Fig15_1.pdf to Fig15_10.pdf. All are graphical representations of the quality check comparisons (more details in next section). Additionally, there are four group figures: 2A collects Figures 1 through 6, 2B1 Figures 7 through 14, 2B2 Figures 15_1 through 15_10, and 2C Figure 16.

Each figure in the folder "02_figures" is the graphical representation of and individual check run on the data from the input GLD survey, consisting of the following elements:

- A title, indicating the variable being checked,
- A y-axis, with the values the variable may take,
- An x-axis, which represents the different data sources: GLD (as the left-most x-axis value) and the one or multiple comparison sources (to its right),
- The data, in red or blue segments:

    - A red segment representing the GLD harmonized survey estimate.
    - A blue segment representing the estimate drawn from an external source, the name of which is denoted in the x-axis tick label.
- A solid line means that the year of the data is the same as the year of the survey while a dashed line means that we are using data from a different year (as data for the given estimate was not available for the survey year)
- Each segment consists of two distinct parts: The first is a central point (dot or triangle) showing the precise estimate for each variable and source. The second are the lines stretching out above and below the central point representing an upper and lower bound of the estimate (commonly 10% above and below the point estimate).

These figures allow us to visually see if the values in GLD are close to the values from external sources. In the Flags Methodology section we describes precisely the circumstances under which we consider the values to be "too far off".

Additionally, the external check process places in the summary folder a PDF file ("B2_external_flags") with all figures that have been flagged. The role of this file is to present a snapshot of the variables whose estimates are clearly different from the external sources we use as comparison, which may indicate a problem in the harmonization or other issues that require our attention.

As well, an overview Excel file (called "B2_external_results.xlsx") is created, detailing the information behind each check. The rows are made up of a first row denoting column names and up to 25 for each of the checks undertaken to the GLD data. If there are fewer than 25 check rows, then some checks have not been performed. This happens when some variables are not available in the original GLD dataset (e.g., no urban/rural data recorded).

The file consists of the following 18 columns:

1. **Year:** Year of the GLD survey, as described in "section 3. Output".
2. **Country:** ISO Alpha-3 country code of the country in the GLD survey.
3. **Survey:** Type of GLD survey (for instance: LFS, QLFS, etc).
4. **Varchecked:** Stands for "Variable checked" and is the variable we are testing in the GLD data.
5. **Varorder:** Stands for "Variable order" and is a numeric variable ranging from 1 to the number of checks for that survey (up to 25). It allows us to sort the checks in the default order (starting with "Total population" and ending with "Average hourly wages").
6. **Val_pe:** Stands for "Value point-estimate". This is the value for the variable in that row computed from GLD.
7. **Val_lb:** Stands for "Value lower-bound". It is the result of multiplying the val_pe by a number 0.95 (0.90 for the variable "average hourly wages"). By construction, this results in a value 5% (or 10%) lower than val_pe.
8. **Val_ub:** Stands for "Value upper-bound". It is the result of multiplying the val_pe by a number 1.05 (1.10 for the variable "average hourly wages"). By construction, this results in a value 5% (or 10%) higher than val_pe.
9. **Com_pe:** Stands for "Comparison point estimate". It is the result of averaging all the values from external sources of the variable we are testing. That is, if we are using two external sources to test

"total population" (eg. ILO and UN), com_pe will be the average of "total population" in each of these two sources. For whichever number of source we are using, com_pe will be the average of the value in the sources we are using.

10. **Com_lb:** Stands for "Comparison lower bound". It is the result of multiplying the com_pe by a number 0.95 (0.90 for the variable "average hourly wages"). By construction, this results in a value 5% (or 10%) lower than com_pe.
11. **Com_ub:** Stands for "Comparison upper-bound". It is the result of multiplying the com_pe by a number 1.05 (1.10 for the variable "average hourly wages"). By construction, this results in a value 5% (or 10%) higher than com_pe.
12. **Com_cases:** Indicates how many actual values we are using to establish the comparison. Often times, two different external sources will report the exact same value, because they take the data from one another. For instance, WDI takes data from sources such as ILO. Therefore, if we look at the labor force size in WDI and in ILO, we will obtain the exact same point. Com_cases keeps track of this, and indicates how many different values we are dealing with, rounded to the 0.01 value.
13. **Diff:** Reports the absolute difference between val_pe and com_pe. In the figures, this represents the distance between the dot in the GLD value and the average of the external sources' dots.
14. **Dist1:** The distance between the val_lb and com_ub. That is, the distance from the lower bound of the GLD value to the upper bound of the (average of) external sources. When this distance is positive, the lower bound of the GLD value is above the upper limit we give to the external sources, ie, our value is way above the external sources.
15. **Dist2:** The distance between com_lb and val_ub. That is, the distance between the lower bound of the external sources and the upper bound of the GLD value. When this distance is positive, it means that the lower bound of the external sources is above the upper limit we give to the GLD value. That is, the external sources are (on average) well above the GLD value. That, our value is way below the external sources.
16. **F1:** An indicator value that takes value of 1 when dist1 is positive (we omit the 0s). That is, it indicates when the GLD value is well above the external sources.
17. **F2:** An indicator value that takes the value of 1 when dist2 is positive (again, we omit the 0s). It indicates when the GLD value is well below the external sources.
18. **Flag:** An indicator value that takes the value 1 when either F1 or F2 are 1, 0 (or missing/blank) otherwise. That is, it indicates when a GLD variable is either well above or below the external sources' estimate. Graphically, it means that the segment of the GLD value does not intersect the imaginary segment generated by the average of the external sources.

Flags Methodology

This section details the flags methodology. Some parts have already been addressed in previous sections. Nonetheless this section explains the methodology in its entirety.

One of the main goals of the checks process is to detect GLD variables that differ substantially from that same variable as reported from external sources.

To conduct this analysis, we need the following elements:

1. A GLD variable,
2. A benchmark to compare the external variable with,

3. A measure of distance between (1) and (2),
4. A threshold above which the distance in (3) is considered too large.

For point (1), we use the GLD variables described in [Overview](#overview).

For point (2), we average the external variables listed in the Annex. Note that for each GLD variable we have one or more variables from external sources we wish to compare it to. By averaging all the variables from external sources, we guarantee we have one single benchmark per variable.

Two further clarifications on the benchmark values:

In the future, we may want to change this to a weighted average, to account for the fact that some sources matter more than others. In particular, we know not all external variables are from the same year. Hence, we may want to give a higher weight to the external sources that are from the same year as the variable.

- We know that some external sources report the same value, because they ultimately come from the same source (e.g., WDI takes data from ILO). Note that at this point, we are not correcting for this: we are giving equal weights at all sources regardless of how "original" their values are.

Finally, it is worth noting that neither of the points mentioned above is guaranteed to affect the results. In fact, the results have proven to be fairly robust, and it is likely that a more elaborate weighting process will not translate in any noticeable differences in the output.

For point (3), to measure the distance between (1) GLD and (2) the average of the external sources, we consider two cases of particular interest:

- For the cases in which the GLD data is above the external data, we calculate the distance between a lower bound of the GLD data and an upper bound of the external data. If this distance is positive, then the GLD is considerably above the external sources. If the distance is 0, it means that the lower bound of the GLD data is exactly the same as the upper bound of the external data. If the distance is negative, it means that the lower bound of the GLD data is below the upper bound of the external data. Thinking in terms of the segments represented in the figures, a negative distance represents an overlap between the GLD segment and the average of the external segments, and a positive distance means no intersection of the segments.
- For the cases in which the GLD data is below the external data, we proceed symmetrically. That is, we compute the difference between lower bound of the average of external values and the upper bound of the GLD data. If this distance is positive, then the lower bound of the external variables is above the upper bound of the GLD variable. That is, the GLD value is very far away from the comparison values. If the distance is 0, then the lower bound of the external variables coincides exactly with the upper bound of the GLD data. If it is negative, then the upper bound of the GLD data is higher than the lower bound of the external data, i.e., the segments intersect.

As explained above, a positive distance between the two points implies a large difference between the GLD values and the external values. Hence, these are the cases that we flag. That is, we will flag a GLD variable from the list in [Overview](##Overview) if any of the two following conditions hold:

- The distance between the lower bound of the GLD data and the upper bound of average of the external data is positive.
- The distance between the lower bound of the average of the external data and the upper bound of the GLD data is positive.

It remains to explain how we chose these upper and lower bounds, which effectively determine the thresholds we use to flag the data (point (4) in the list of required elements). Here, we use a threshold of 5% to compute the upper/lower bounds for all the variables, and 10% for the wage related variables. That is, for most variables, we will flag a GLD variable if the difference between the value it takes and the average of the external sources is equal or higher than 5% de value of the GLD variable plus 5% the value of the external sources.

Graphically, we will flag the variables where the GLD segment and the imaginary segment defined by the average of the external sources do not intersect.

4.2.4. Block 3 - GLD missing checks

Block 3 records the share of responses that are missing for a series of variables. For variables `age`, `male`, and `urban`, the code evaluates if any answers are missing. For variable `lstatus` it is evaluated whether there any responses missing for those above the minimum age to respond to the labour module (variable `minlaborage`). Finally, variables `empstat`, `industrycat4`, `wage_no_compen`, and `unitwage` are checked for missing answers only for those employed (i.e., with `lstatus == 1`).

The code outputs the data in "Block3_Missing/01_data" and places the figure directly in the "01_summary" folder in the user designated output folder.

4.2.5. Block 4 - GLD bivariate checks

The bivariate data checks look at combinations of two or more variables and raise a flag when the expected behavior of the data under these combinations is violated.

There are two groups of bivariate checks. The first group consists of three checks and looks at the conversions of a certain variables. The idea behind these three checks is that we have different versions of a particular variable (say “education” is a categorical variable that can present 4 levels, 5 levels and 7 levels), and we want to make sure that the correspondences are correct. This should be true mechanically, and these checks will confirm that it is the case.

The second group of bivariate checks look at relationships between two variables and raise a flag with the relationship is different from what we expect. Consider, for instance the variables education and age. Common sense tells us that we should expect the number of children under 12 years old with postgraduate education to be very low. Hence, there is a check that computes this share, and flags it if it is above certain threshold. If many children appeared in our data with post-secondary education, we would need to make sure that the variables are correctly harmonized. Therefore, bivariate checks can be useful to identify potential errors in the harmonization process.

In the next section, we list 8 groups of variables. The first three sections explain the *cross-categories* consistency checks while the latter five sections explain the *bivariate* checks conducted on pairs of variables.

Education: cross-categories

The GLD data includes three different education variables: educat7, educat5 and educat4. These are categorical variables with different number of levels: educat4 classifies education in 4 groups, educat5 does so in 5 groups, and educat7 in 7 groups.

We expect these three variables to be coherent. That is, if an individual has completed primary education, all three variables should reflect this information. Hence, we construct a table with the equivalences between these variables and raise a flag when the data deviates from them.

Industry: cross-categories

This check follows the same structure as the "Education: cross-categories" check, with the variables industrycat_isic, industrycat10 and industrycat4.

The appendix contains a table with the correspondences that the code checks for. Notice that these follow the Version 4 of ISIC. Whenever the GLD data follows a different version, or the version is not indicated in the GLD database, this check is skipped.

Occupation: cross categories

This check looks at the conversion of the occup_isco, occup_skill and occup variables.

The correspondence between these variables follows the ISCO version 1988, listed in the appendix.

Whenever the data presents a different conversion, a flag is raised. If the data follows a different ISCO version, or if the version is not known, the check is skipped.

Labor force versus employment

This check looks at the joint distribution of the variables lstatus and empstat. By definition, when lstatus equals "Employed", empstat should contain information. If lstatus is either "Unemployed" or "Not in LF", then empstat should be missing. Whenever this rule is not followed in the data, a flag is raised.

Labor force status versus age

This check looks at the joint distribution of labor force status (lstatus) and age. To make the check easier, we recode the variable age into a categorical variable with 5 groups, defined as follows:

- Child, if an individual is between 0 and 14 years old;
- Youth, if an individual is between 15 and 24 years old;
- Adult, if an individual is between 25 and 54 years old;
- Senior, if an individual is between 55 and 70 years old;
- Retiree, if an individual is between 71 and 120 years old.

We compute the join distribution of age and lstatus, and raise a flag whenever the data violates one of the following 6 conditions:

1. 90%+ of children should be out of the labor force or have lstatus missing;
2. 50%+ of youth should be out of the labor force;

3. 50%+ of adults should be in the labor force;
4. Seniors must be in the labor force at a lower rate than adults;
5. 80%+ of retirees should be out of the labor force
6. All lstatus groups should be positive (>0) for working-age people (youth, adult, and senior).

Education versus age

This check looks at the joint distribution of education (as defined by educat4) and age (as re-defined in section 2.5), and raise a flag whenever one of following 3 conditions are not met in the GLD data:

1. 90%+ of children should have no education or primary only;
2. None of the primary, secondary and post-secondary categories should be 0 for people 15+;
3. For all age groups, no single education level should be 100%.

Industry versus occupation

This check considers the join distribution of industrycat4 and occup, and raises a flag when any of the following 15 conditions is violated:

1. Managers should not be prevalent in agriculture (share under 20%);
2. Professionals should be low in agriculture (under 10%);
3. Professionals should be mainly in services (over 60%);
4. Technicians should be low in agriculture (under 10%);
5. Clerks should be low in agriculture (under 10%);
6. Clerks should be mainly in services (over 60%);
7. Service and market should be low in agriculture (under 10%);
8. Service and market should be almost only in services (over 80%);
9. Skilled agricultural should be almost only in agriculture (over 80%);
10. Craft workers should be low in agriculture (under 10%);
11. Craft workers should be mainly in industry (over 60%);
12. Machine operators should be low in agriculture (under 10%);
13. In all ocuppations, “other” should be low (under 25%);
14. In any ocuppation, no industry should be exclusive (none is 100%);
15. In any occupation, not all industries should not be missing.

Occupation versus education

This check looks at the joint distribution of occupation and education, as measured by occup and educat4.

A flag is raised whenever one of the following 6 conditions is not met:

1. Professionals with no education should be low (under 10%);
2. Professionals should mainly have post-secondary education (over 60%);
3. Technicians with no education should be low (under 10%);
4. Clerks with no education should be low (under 10%);
5. Machine operators with post-secondary education should not be prevalent (under 20%);
6. Elementary occupations with post-secondary education should be low (under 10%).

Output produced

Each of the 8 checks described above produces a file with the key data used to conduct a check (a list of correspondences or a joint bivariate distribution), as well as a sequence of 0s and 1s indicating if the conditions are met. If any of the required conditions is not met, a string variable explaining the reason behind the flag is created. Additionally, a file is created listing all flagged issues. All files are stored under "Block4_Bivariate/01_data".

The code stores in the "01_summary" folder, an extra file called "B4_bivariave_results.xlsx" “bivariave_results.xlsx”. This file contains a list of cases in which a flag has been raised and will be empty when the data does not raise any flags.

4.2.6. Block 5- GLD wage checks

The goal of this section is to look at the GLD wage data (wage_no_compen variable for those with empstat == 1), and to use this information to detect potential harmonization errors.

We compute 4 series of wages: by education, by occupation, by industry and by age, and flag a series whenever the wage behaves differently than expected. Note than in all cases, we convert the wages to hourly wages.

Wage series & checks

Wage by education

To compute wage by education, we use the education variable educat4, a categorical variable that takes 4 values. We expect wages to be increasing in education, and we flag the series when this is not the case.

Wage by occupation

To compute wage by occupation, we use the education variable occup_skill, a categorical variable that takes 3 values. We expect wages to be increasing in skill level, and we flag the series when this is not the case.

Wage by industry

To compute wage by industry, we use the education variable industrycat4 skill, a categorical variable that takes 4 values, of which we consider 3: agriculture, industry and services. We expect wages to be highest in services, followed by industry and finally agriculture. We flag the series when this is not the case.

Wage by age

To compute wage by age, we use a re-coded version of the age variable, which classifies population in 5 groups: child (0 to 14 years old), young (15 to 24 years old), adult (25 to 54 years old), senior (55 to 70 years old) and retiree (71 to 120 years old). We expect wages to increase until adulthood, and we expect wages of retirees to be lower than wages of senior people. We remain agnostic about how wages for adults and for seniors compare. We flag the series when these rules are violated.

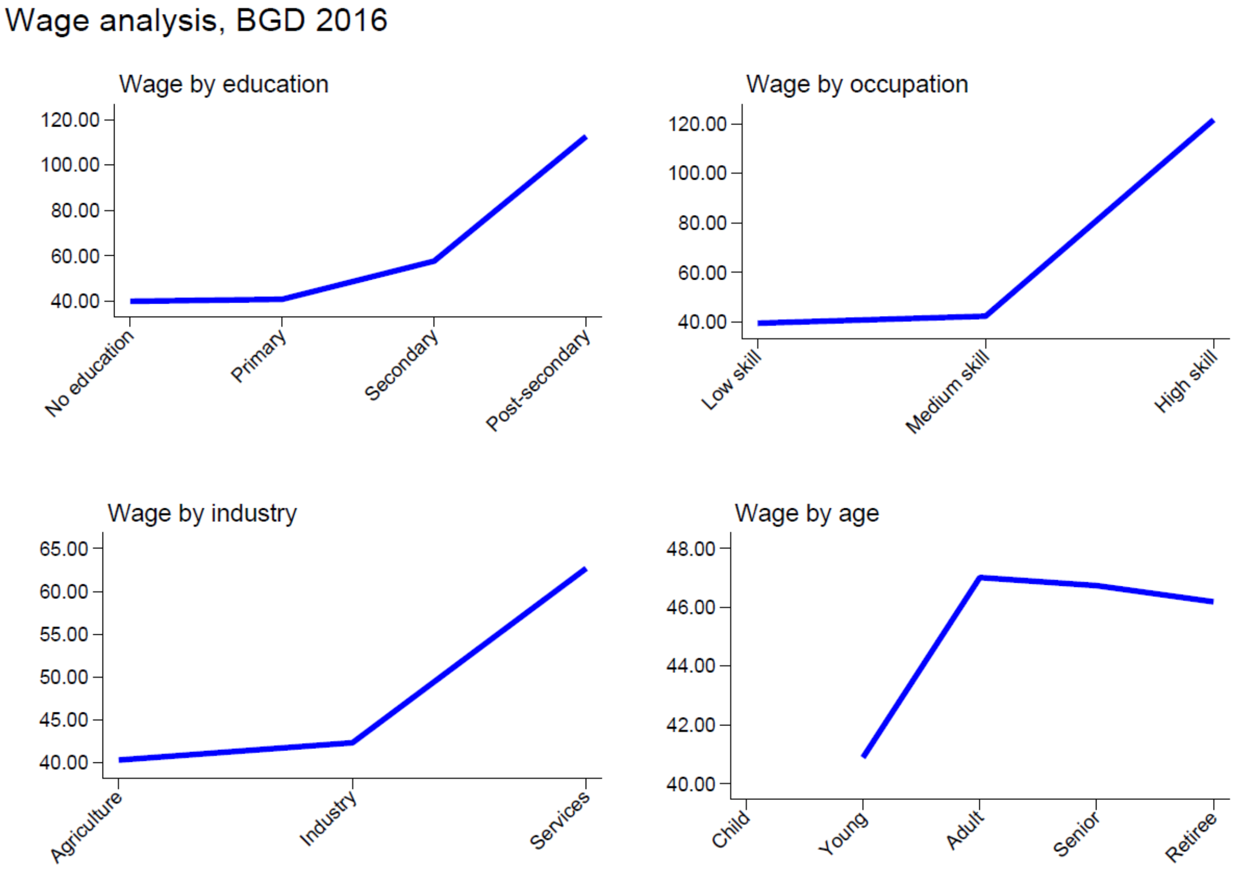
Output produced

Regardless of the results of the checks, we produce the following two pieces out output:

*	A small database for each wage series, with a binary variable indicating if any condition is violated. These are stored in "Block5_Wage/01_data".

* A pdf image consisting on four figures, one for each series. This file can be found in "Block5_Wage/02_figures" and "01_summary". The graphs are colour coded. When the series is depicted in blue it conforms with our ex-ante expectations. If the line is drawn red, the series raises at least one flag. An example of the output produced, for the case of the 2016 Bangladesh LFS is shown below in Figure 26.

Figure 26 - Example of the wage analysis checks.



4.2.7. Annex to Block 2 – GLD external checks

The below is the list of exact indicators read through WDI or dbnomics and used to create external comparators.

1. Total population
   1.1. `SP.POP.TOTL`, Total population, WDI
   1.2. `A.N.”CC”.W0.S1.S1._Z.POP._Z._Z._Z.PS._Z.N`, Total population, NA Main Aggregates, UNData
      - Unit of measure: “PS” (persons)
   1.3. `POP_2POP_GEO_NB`, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO
      - Classif1: "`GEO_COV_NAT`" (national geographic coverage)

1.4. POP_2POP_SEX_AGE_NB, Population by sex and age -- UN estimates and projections, Nov. 2021 (thousands), ILO
   - Sex: SEX_T (total)
   - Classif1: "AGE_10YRBANDS_TOTAL" (total age, classified in 10 year bands)

2. Gender split (% of female population)

2.1. SP.POP.TOTL.FE.ZS, Population, female (% of total population), WDI

2.2. POP_2POP_SEX_AGE_NB, Population by sex and age -- UN estimates and projections, Nov. 2021 (thousands), ILO
   - Sex: “SEX_F” & “SEX_T” (Female & Total. Female for the numerator, total for the denominator)
   - Classif1: AGE_AGGREGATE_TOTAL (total age, classified in aggregates)

2.3. POP_2POP_GEO_NB, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO
   - Sex: “SEX_F” & “SEX_T” (Female & Total. Female for the numerator, total for the denominator)
   - Classif1: “AGE_AGGREGATE_TOTAL” (total age, classified in aggregates)
   - Classif2: “GEO_COV_NAT” (national geographic coverage)

3. Urban population (% of total)

3.1. SP.URB.TOTL.IN.ZS, Urban population (% of total population), WDI

3.2. POP_2POP_GEO_NB, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO
   - Classif1: “GEO_COV_NAT” & “GEO_COV_URB” (national & urban geographic coverage)

4. Children (0-14 years old, % of total population)

4.1. SP.POP.0014.TO.ZS, Population ages 0-14 (% of total population), WDI

4.2. POP_2POP_SEX_AGE_NB, Population by sex and age -- UN estimates and projections, Nov. 2021 (thousands), ILO
   - Sex: SEX_T (total)
   - Classif1: “AGE_5YRBANDS_TOTAL”. In particular, we select age groups “AGE_5YRBANDS_Y00-04”, “AGE_5YRBANDS_Y05-09” and “AGE_5YRBANDS_Y10” to compute the numerator.

4.3. POP_2POP_GEO_NB, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO
   - Sex: SEX_T (total)
   - Classif1: "AGE_5YRBANDS_TOTAL". In particular, we select age groups “AGE_5YRBANDS_Y00-04”, “AGE_5YRBANDS_Y05-09” and “AGE_5YRBANDS_Y10” to compute the numerator.
   - Classif2: "GEO_COV_NAT"

5. Adults (15-64 years old, % of total population)

5.1. `SP.POP.1564.TO.ZS`, Population ages 15-64 (% of total population), WDI

5.2. `POP_2POP_SEX_AGE_NB`, Population by sex and age -- UN estimates and projections, Nov. 2021 (thousands), ILO

- Sex: `SEX_T` (Total)
- Classif1: "`AGE_5YRBANDS_TOTAL`". In particular, we select age groups "`AGE_5YRBANDS_Y15-19`" , "`AGE_5YRBANDS_Y20-24`", "`AGE_5YRBANDS_Y25-29`", "`AGE_5YRBANDS_Y30-34`", "`AGE_5YRBANDS_Y35-39`”, "`AGE_5YRBANDS_Y40-44`", "`AGE_5YRBANDS_Y45-49`”, "`AGE_5YRBANDS_Y50-54`", "`AGE_5YRBANDS_Y55-59`" and "`AGE_5YRBANDS_Y60-64`" to compute the numerator.

5.3. `POP_2POP_GEO_NB`\`, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO

- Sex: `SEX_T` (total)
- Classif1: "`AGE_5YRBANDS_TOTAL`". In particular, we select age groups "`AGE_5YRBANDS_Y15-19`" , "`AGE_5YRBANDS_Y20-24`", "`AGE_5YRBANDS_Y25-29`", "`AGE_5YRBANDS_Y30-34`", "`AGE_5YRBANDS_Y35-39`”, "`AGE_5YRBANDS_Y40-44`", "`AGE_5YRBANDS_Y45-49`”, "`AGE_5YRBANDS_Y50-54`", "`AGE_5YRBANDS_Y55-59`" and "`AGE_5YRBANDS_Y60-64`" to compute the numerator.
- Classif2: "`GEO_COV_NAT`" (national geographic coverage)

6. Seniors (65+ years old, % of total population)

6.1. `SP.POP.65UP.TO.ZS`, Population ages 65 and above (% of total population), WDI

6.2. `POP_2POP_SEX_AGE_NB`, Population by sex and age -- UN estimates and projections, Nov. 2021 (thousands), ILO

- Sex: `SEX_T` (total)
- Classif1: "`AGE_5YRBANDS_TOTAL`". In particular, we select age group "`AGE_5YRBANDS_YGE65`" to compute the numerator.

6.3. `POP_2POP_GEO_NB`, Population by rural / urban areas -- UN estimates and projections, Nov. 2020 (thousands), ILO

- Sex: `SEX_T` (total)
- Classif1: "`AGE_5YRBANDS_TOTAL`". In particular, we select age group "`AGE_5YRBANDS_YGE65`" to compute the numerator.
- Classif2: "`GEO_COV_NAT`" (national geographic coverage)

7. Labor force size (number)

7.1. `SL.TLF.TOTL.IN`, Labor force, total, WDI

7.2. `EAP_TEAP_SEX_AGE_NB`, Labour force by sex and age (thousands), ILO

- Sex: `SEX_T` (total)

- Classif1: "AGE_10YRBANDS_TOTAL"

7.3. EAP_2EAP_SEX_AGE_NB, Labour force by sex and age -- ILO modelled estimates, Nov. 2021 (thousands), ILO

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

8. Labor force participation rate (labor force/ total population)

8.1. SL.TLF.CACT.ZS, Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate), WDI

8.2. SL.TLF.CACT.NE.ZS, Labor force participation rate, total (% of total population ages 15+) (national estimate), WDI

8.3. SL.TLF.CACT.NE.ZS, Labor force participation rate, total (% of total population ages 15+) (national estimate), WDI

8.4. EAP_DWAP_SEX_AGE_RT, Labour force participation rate by sex and age (%), ILO

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

8.5. EAP_2WAP_SEX_AGE_RT, Labour force participation rate by sex and age – ILO modelled estimates, Nov. 2021 (%), ILO

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

9. Employment (number)

9.1. EMP_TEMP_SEX_AGE_NB, Employment by sex and age (thousands), ILO

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

9.2. EMP_2EMP_SEX_AGE_NB, Employment by sex and age -- ILO modelled estimates, Nov. 2021 (thousands)

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

10. Employment to population ratio (employed/ total population)

10.1. SL.EMP.TOTL.SP.ZS, Employment to population ratio, 15+, total (%) (modeled ILO estimate), WDI

10.2. SL.EMP.TOTL.SP.NE.ZS, Employment to population ratio, 15+, total (%) (national estimate), WDI

10.3. EMP_DWAP_SEX_AGE_RT, Employment-to-population ratio by sex and age (%), ILO

- Sex: SEX_T (total)
- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

10.4. EMP_2WAP_SEX_AGE_RT, Employment-to-population ratio by sex and age -- ILO modelled estimates, Nov. 2021 (%), ILO

- Sex: SEX_T (total)

- Classif1: "AGE_YTHADULT_YGE15" (Age +15)

11. Unemployment rate (unemployed/ labor force)
    - 11.1. SL.UEM.TOTL.ZS, Unemployment, total (% of total labor force) (modeled ILO estimate), WDI
    - 11.2. SL.UEM.TOTL.NE.ZS, Unemployment, total (% of total labor force) (national estimate), WDI
    - 11.3. UNE_DEAP_SEX_AGE_RT Unemployment rate by sex and age (%), ILO
        - Sex: SEX_T (total)
        - Classif1: "AGE_AGGREGATE_TOTAL" (Total age)
    - 11.4. UNE_2EAP_SEX_AGE_RT, Unemployment rate by sex and age -- ILO modelled estimates, Nov. 2021 (%), ILO
        - Sex: SEX_T (total)
        - Classif1: "AGE_YTHADULT_YGE15" (Age +15)

12. Agriculture (% of employment)
    - 12.1. SL.AGR.EMPL.ZS, Employment in agriculture (% of total employment) (modeled ILO estimate), WDI
    - 12.2. EMP_TEMP_SEX_ECO_NB, Employment by sex and economic activity (thousands), ILO
        - Sex: SEX_T (total)
        - Classif1: "ECO_SECTOR_AGR" and "ECO_SECTOR_TOTAL" (agriculture and total)
    - 12.3. EMP_2EMP_SEX_ECO_NB, Employment by sex and economic activity -- ILO modelled estimates, Nov. 2020 (thousands), ILO
        - Sex: SEX_T (total)
        - Classif1: "ECO_SECTOR_AGR" and "ECO_SECTOR_TOTAL" (agriculture and total)

13. Industry (% of employment)
    - 13.1. SL.IND.EMPL.ZS, Employment in industry (% of total employment) (modeled ILO estimate), WDI
    - 13.2. EMP_TEMP_SEX_ECO_NB, Employment by sex and economic activity (thousands), ILO
        - Sex: SEX_T (total)
        - Classif1: "ECO_SECTOR_IND" and "ECO_SECTOR_TOTAL" (industry and total)
    - 13.3. EMP_2EMP_SEX_ECO_NB, Employment by sex and economic activity -- ILO modelled estimates, Nov. 2020 (thousands), ILO
        - Sex: SEX_T (total)
        - Classif1: "ECO_SECTOR_IND" and "ECO_SECTOR_TOTAL" (industry and total)

14. Services (% of employment)
    - 14.1. SL.SRV.EMPL.ZS, Employment in services (% of total employment) (modeled ILO estimate), WDI
    - 14.2. EMP_TEMP_SEX_ECO_NB, Employment by sex and economic activity (thousands), ILO
        - Sex: SEX_T (total)

- Classif1: "ECO_SECTOR_SER" and "ECO_SECTOR_TOTAL" (services and total)

14.3. EMP_2EMP_SEX_ECO_NB, Employment by sex and economic activity -- ILO modelled estimates, Nov. 2020 (thousands), ILO
- Sex: SEX_T (total)
- Classif1: "ECO_SECTOR_SER" and "ECO_SECTOR_TOTAL" (services and total)

15. Industry category (% of employment)

15.1. EMP_TEMP_SEX_ECO_NB, Employment by sex and economic activity (thousands), ILO
- Sex: SEX_T (total)
- Classif1: ECO_ISIC4_TOTAL & ECO_ISIC4_A - ECO_ISIC4_U or ECO_ISIC3_TOTAL and ECO_ISIC3_A - ECO_ISIC3_X

15.2. EMP_2EMP_SEX_ECO_NB`, Employment by sex and economic activity -- ILO modelled estimates, Nov. 2020 (thousands), ILO
- Sex: SEX_T (total)
- Classif1: ECO_DETAILS_TOTAL & ECO_ DETAILS _A - ECO_ DETAILS_RSTU

16. Status in Employment (% of employment)

16.1. EMP_2EMP_AGE_STE_NB, Employment by age and status in employment (thousands), ILO
- Classf1: AGE_YTHADULT_YGE15
- frequency: A

17. Hourly wages

17.1. EAR_HEES_SEX_OCU_NB, Mean nominal hourly earnings of employees by sex and occupation (local currency), ILO
- Sex: SEX_T (total)
- Classif1 == "OCU_ISCO88_TOTAL" (Total occupations)

17.2. NY.GDP.PCAP.CN, GDP per capita (current LCU), ILO
- Times 2/3 (labor share), divided by 2080 (number of full time working-hours in a year, 40h/week*52weeks = 2080 h)

4.3. Guide to the GLD survey series quality checks

4.3.1. Overview

The series checks aim to help users inspect graphically a set of surveys for a country over time to ensure they are coherent and consistent.

The only file the user needs to use is the process_time_series.R file, consisting of seven steps, that are detailed here in the following.

4.3.2. Step 1 - Define user variables

Step 1 is the only section requiring user input. The user ought to input the country ISO alpha-3 code as well as the variables that they want to inspect, provided they are categorical (as are the variables listed below). Note that the wage information is automatically included and thus needs not to be listed.

Subsequently the user defines three paths, the first (`path_in`) points to the folder that contains all the surveys for that country. That is, the path to folder that contains the top level survey folders (i.e., folders of the "CCC_YYYY_[Survey_Name]" level). The second is to the folder where the user wants to store the graphs created for their inspection. The last path points to where the functions are stored (i.e., where all the files listed in this GLD directory starting with "function" are stored on the user's system).

```
# Enter country ISO Alpha 3 code
country <- "[CCC]"

# Enter variables that ought to be analysed (or leave standard
variables)
vars_to_study <- c("empstat", "educat7", "educat4", "industrycat10",
"industrycat4", "occup", "lstatus")
# Note that wage will be included by default, no need to include here

# Define the path to the folder holding the series
path_in <- "[For example: Z:/GLD-Harmonization/123456_AZ/ZAF]"

# Define the path to the folder where the graphs ought to be stored in
path_out <- "[Path to output folder]"

# Define the path to the folder where this code and the other functions
of the survey series checks are stored
dir_w_functions <- "[Path to folder with functions]"
```

4.3.3. Step 2 - Call libraries, functions

Step 2 defines the packages that are needed, ensures they are installed and loads them. It then runs the GLD functions for the survey series checks.

```
# List packages that we need
pkgs <- c("Hmisc", "tidyverse")

# Check they are installed
pkgs_2_install <- pkgs[!(pkgs %in% installed.packages())]

# Install if not on system
```

```
for (pkg in pkgs_2_install) {
  install.packages(pkg)
}

# Load packages
for (pkg in pkgs) {
  library(pkg,character.only=TRUE)
}

purrr::walk(dir(dir_w_functions, pattern = "^function", full.names =
T), ~source(.x))
```

4.3.4. Step 3 - Load DF

Step 3 uses the function defined in function_load_df.R. It uses as input the path defined at the start. The function will enter the path and extract the latest harmonized dataset for each folder. That is, if the BRA_2018_PNADC folder has four harmonized versions (from V01_A to V04_A), the function will select the latest (here V04_A). It then will extract the necessary variables (e.g., age, weight, unitwage, ...) plus the variables defined at the start to inspect and append all data files into a single dataset. If the option wap_only is set to TRUE (as shown below) the data will be restricted to the working age population.

```
df <- load_df(path_in = path_in,
              wap_only = TRUE,
              vars_to_study = vars_to_study)
```

4.3.5. Step 4 - Make time series data frames

Step 4 makes a list of aggregated data for each variable in the vector vars_to_study provided by the user in Step 1 using the function defined in function_make_cat_ts.R. For each variable the shares per year and variable category (excluding NAs), so that the shares sum to 1 for each year.

If the option employed is set to TRUE (as shown below), the calculations apply only to those who are employed (i.e., for those with lstatus == 1). Note that the function automatically will not apply this reduction to the employed sub-sample if the variable passed is lstatus (as otherwise there would be no information value).

```
summary_df_list <-
  purrr::map(
    vars_to_study,
    ~make_cat_ts(df = df,
                 var = .x,
                 employed = TRUE))
```

4.3.6. Step 5 - Make wage time series data frame

Step 5 uses the function defined in function_make_wage_ts.R to calculate for the wage employed (empstat == 1) the mean and median hourly wages as well as the 10th, 25th, 75th, and 90th percentile for each year. It also stores the sample size of respondents for which there are answers for all variables used to calculate the hourly wage (unitwage, whours, and wage_no_compen)

```
wage_df <-
  make_wage_ts(
    df = df,
    country = country)
```

4.3.7. Step 6 - Make time series plots

Step 6 creates a list of plots. Firstly, the list of shares over the categorical variables from Step 4 is run to create a graph for each, using the function defined in function_plot_cat_ts.R. Then, an additional plot for the wage information created in Step 5 is added using the function defined in function_plot_wage_ts.R.

```
summary_plots <- purrr::map(summary_df_list, ~plot_cat_ts(.x))

# Add wage plot as the n_th plot
n_th <- length(vars_to_study) + 1
summary_plots[[n_th]] <- plot_wage_ts(wage_df = wage_df)
```

4.37. Step 7 - Save all summary plots

Step 7, the last step, simply takes the list of plots from Step 6 and stores them in the folder defined by path_out given by the user in Step 1 using the function defined in function_save_ts_plots.R.

```
purrr::walk(summary_plots,
            ~save_ts_plots(.x,
                           out_folder = path_out,
                           country = country))
```

`

5. Using the GLD

5.1. How to use the harmonized dta file

How to use a *single* harmonized dta file

A single file can be accessed, as discussed above, via the GLD server, the Microdata Library, or datalibweb. Each file represents the answers of individuals to the standard variables contained in the GLD dictionary. It can be directly used to calculate summary statistics (over the appropriate subgroups) and regressions.

The variables of the first block contain all relevant variables (psu, ssu, strata, …) to correctly set the survey setting with the exception of the finite population correction factor, e.g., the total number of primary sampling units from which the PSUs where selected from. This information is not available to us and users should either try to approximate the number or use caution when reading results that implicitly assume an infinite population.

How to access several harmonized dta files

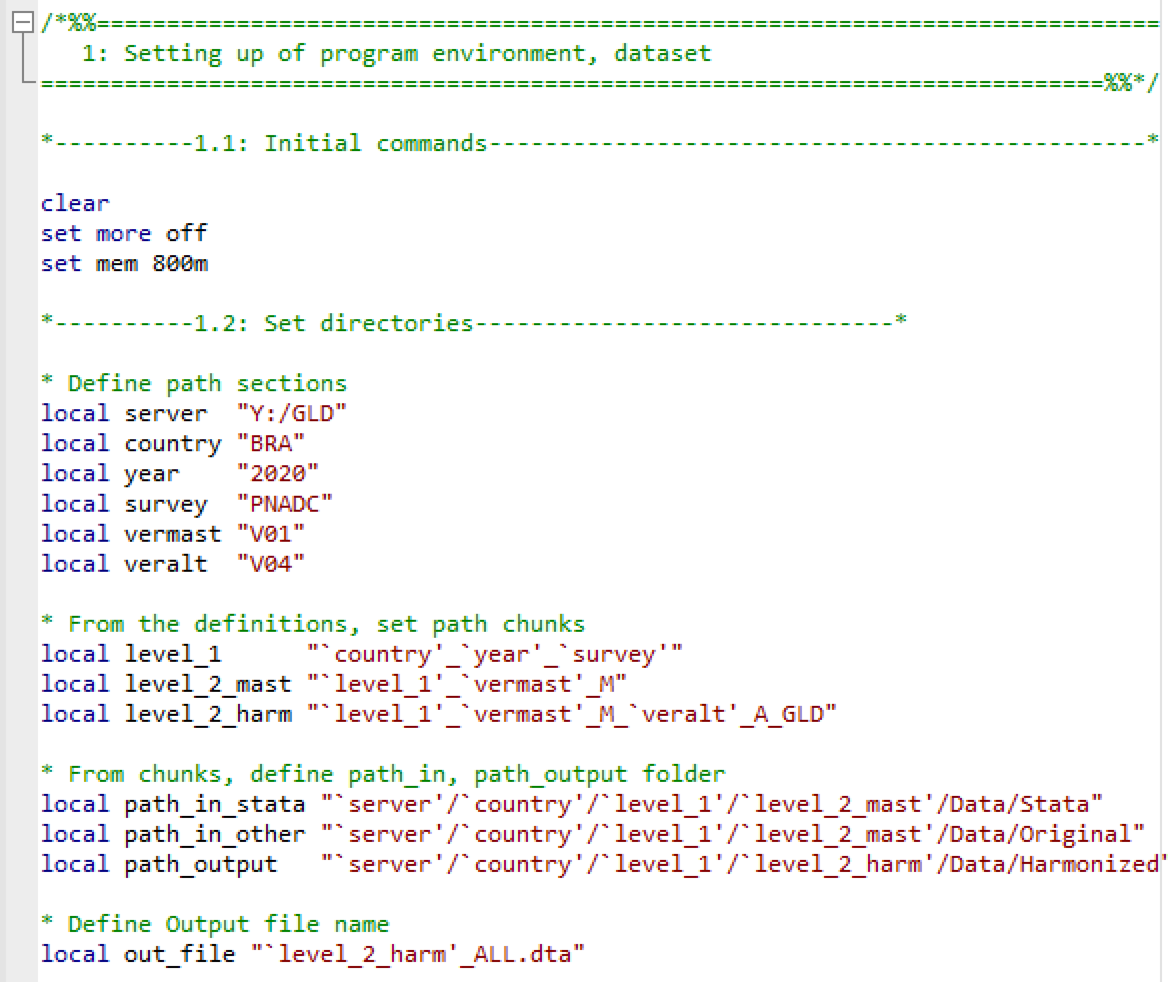
Users may access several surveys at a time and append them but should note that there are occasions when then format of a variable (namely the "_orig" variables). The GLD has prepared a Stata tool for users with access to the server to load the latest files from any given country. All explanations from how to install it to how to use it are available here on the GLD GitHub repository.

5.2. How to use the harmonization code?

The harmonization code is designed for users to exploit it by amending and adding variables. To amend or edit there are three sections consider. Firstly, amending the filenames and paths, secondly, amending or adding the variable(s) of interest, and thirdly, amending the "Final steps Section" that cleans the dataset.

For the first part, users need to evaluate the paths and overwrite the ones laid out in subsection `1.2` (see example below in Figure 27 of Step 1 of the harmonization code for the Brazilian 2020 PNADC) to ensure the files are read from and stored in the folders they wish to use.

Figure 27 - Step 1 of the harmonization code

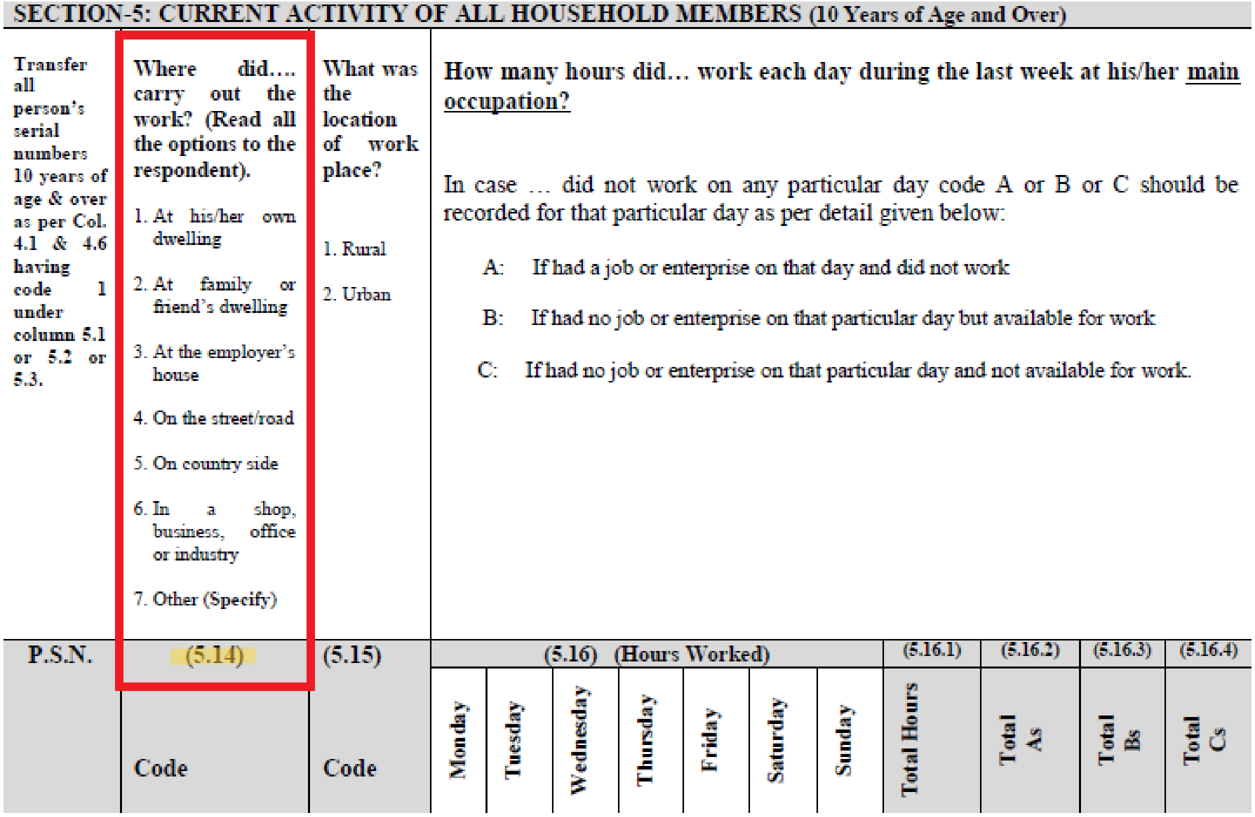


```
/*%%============================================================================
   1: Setting up of program environment, dataset
============================================================================%%*/

*----------1.1: Initial commands-----------------------------------------------*

clear
set more off
set mem 800m

*----------1.2: Set directories------------------------------*

* Define path sections
local server  "Y:/GLD"
local country "BRA"
local year    "2020"
local survey  "PNADC"
local vermast "V01"
local veralt  "V04"

* From the definitions, set path chunks
local level_1      "`country'_`year'_`survey'"
local level_2_mast "`level_1'_`vermast'_M"
local level_2_harm "`level_1'_`vermast'_M_`veralt'_A_GLD"

* From chunks, define path_in, path_output folder
local path_in_stata "`server'/`country'/`level_1'/`level_2_mast'/Data/Stata"
local path_in_other "`server'/`country'/`level_1'/`level_2_mast'/Data/Original"
local path_output   "`server'/`country'/`level_1'/`level_2_harm'/Data/Harmonized"

* Define Output file name
local out_file "`level_2_harm'_ALL.dta"
```

When adding or amending variables, users could add the variable at any point after assembly (i.e., after section 1) and before the final steps. The GLD team recommends adding variables at the end of the section or subsection to which the variable belongs to. That is, if the user is adding an education variable, to do so after the last education variable. If the variable concerns the secondary job over the 12-month recall to add it after the last variable for that section.

For example, the Pakistani 2020 LFS includes question 5.14 (Figure 28 below) that informs about the location where an activity is carried out.

Figure 28 - Excerpt of the 2020 PAK LFS questionnaire

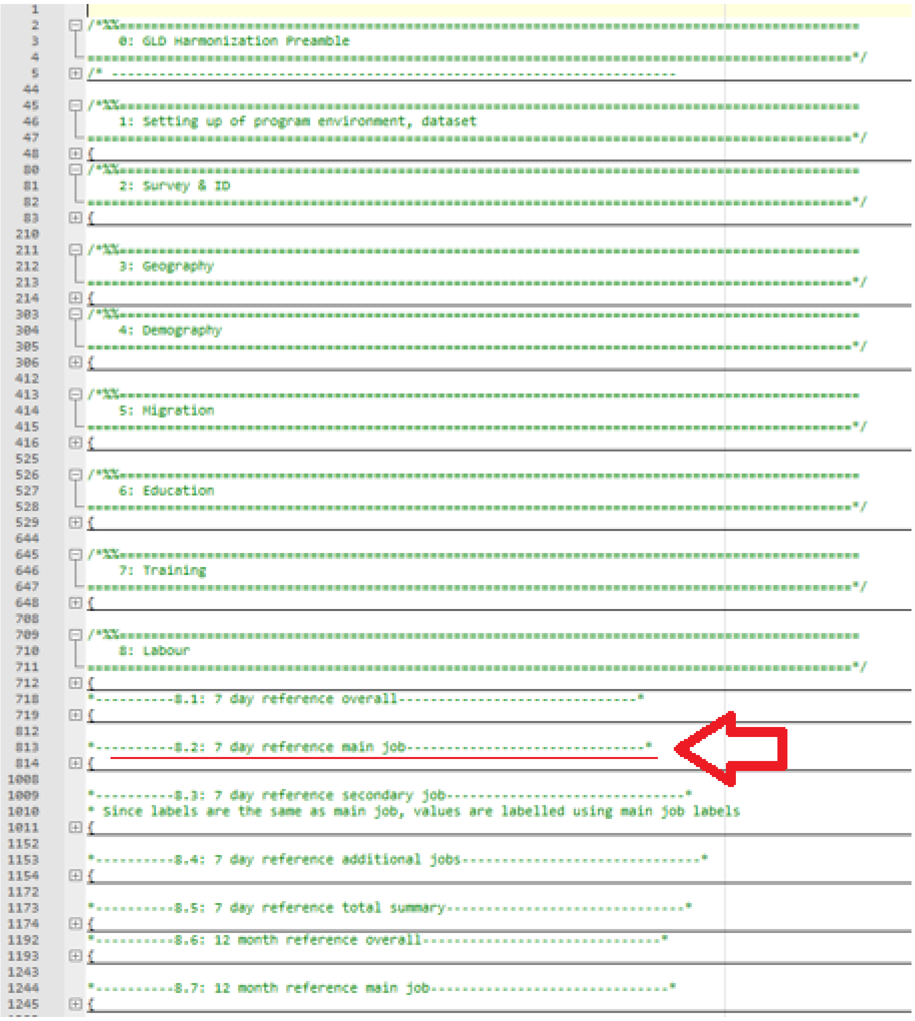


SECTION-5: CURRENT ACTIVITY OF ALL HOUSEHOLD MEMBERS (10 Years of Age and Over)

Transfer all person's serial numbers 10 years of age & over as per Col. 4.1 & 4.6 having code 1 under column 5.1 or 5.2 or 5.3.	**Where did…. carry out the work? (Read all the options to the respondent).** 1. At his/her own dwelling 2. At family or friend's dwelling 3. At the employer's house 4. On the street/road 5. On country side 6. In a shop, business, office or industry 7. Other (Specify)	**What was the location of work place?** 1. Rural 2. Urban	**How many hours did… work each day during the last week at his/her main occupation?** In case … did not work on any particular day code A or B or C should be recorded for that particular day as per detail given below: A: If had a job or enterprise on that day and did not work B: If had no job or enterprise on that particular day but available for work C: If had no job or enterprise on that particular day and not available for work.										
P.S.N.	**(5.14)**	**(5.15)**	**(5.16) (Hours Worked)**							**(5.16.1)**	**(5.16.2)**	**(5.16.3)**	**(5.16.4)**
	Code	**Code**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**	**Total Hours**	**Total As**	**Total Bs**	**Total Cs**

As this question pertains to the main job over the 7-day recall period it should be placed at the end of the relevant section, namely section 8.2 (see Figure 29 below).

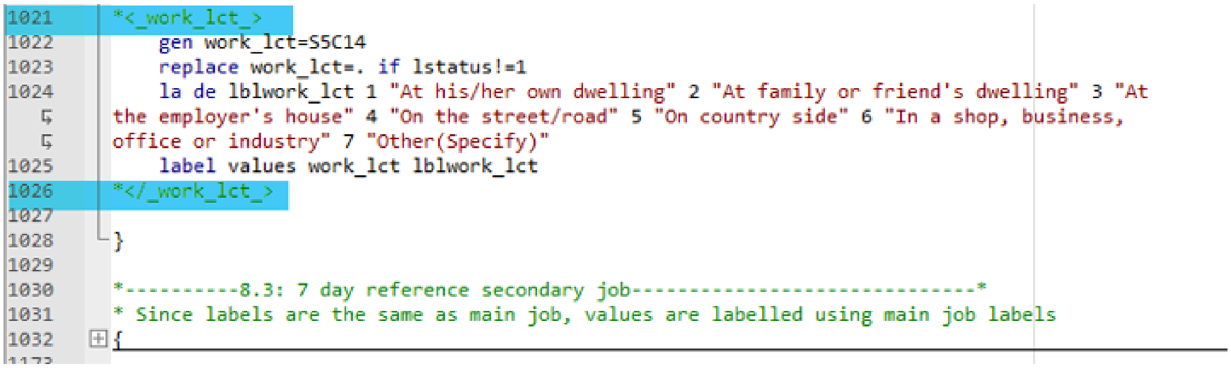
Figure 29 - Indication of the recommended location for including an additional 7 day labor recall variable



```
/*%%=============================================================================
	0: GLD Harmonization Preamble
================================================================================*/
/* ----------------------------------------------------------------------

/*%%=============================================================================
	1: Setting up of program environment, dataset
================================================================================*/
{
/*%%=============================================================================
	2: Survey & ID
================================================================================*/
{

/*%%=============================================================================
	3: Geography
================================================================================*/
{
/*%%=============================================================================
	4: Demography
================================================================================*/
{

/*%%=============================================================================
	5: Migration
================================================================================*/
{

/*%%=============================================================================
	6: Education
================================================================================*/
{

/*%%=============================================================================
	7: Training
================================================================================*/
{

/*%%=============================================================================
	8: Labour
================================================================================*/
{
*----------8.1: 7 day reference overall------------------------------*
{

*----------8.2: 7 day reference main job------------------------------*
{

*----------8.3: 7 day reference secondary job------------------------------*
* Since labels are the same as main job, values are labelled using main job labels
{

*----------8.4: 7 day reference additional jobs------------------------------*
{

*----------8.5: 7 day reference total summary------------------------------*
{
*----------8.6: 12 month reference overall------------------------------*
{

*----------8.7: 12 month reference main job------------------------------*
{
```

The next step in this process is then to add the variable. Users are free to do this as they wish, but the GLD team recommends keeping with the GLD format of wrapping variable code in tags, as shown below (Figure 30).

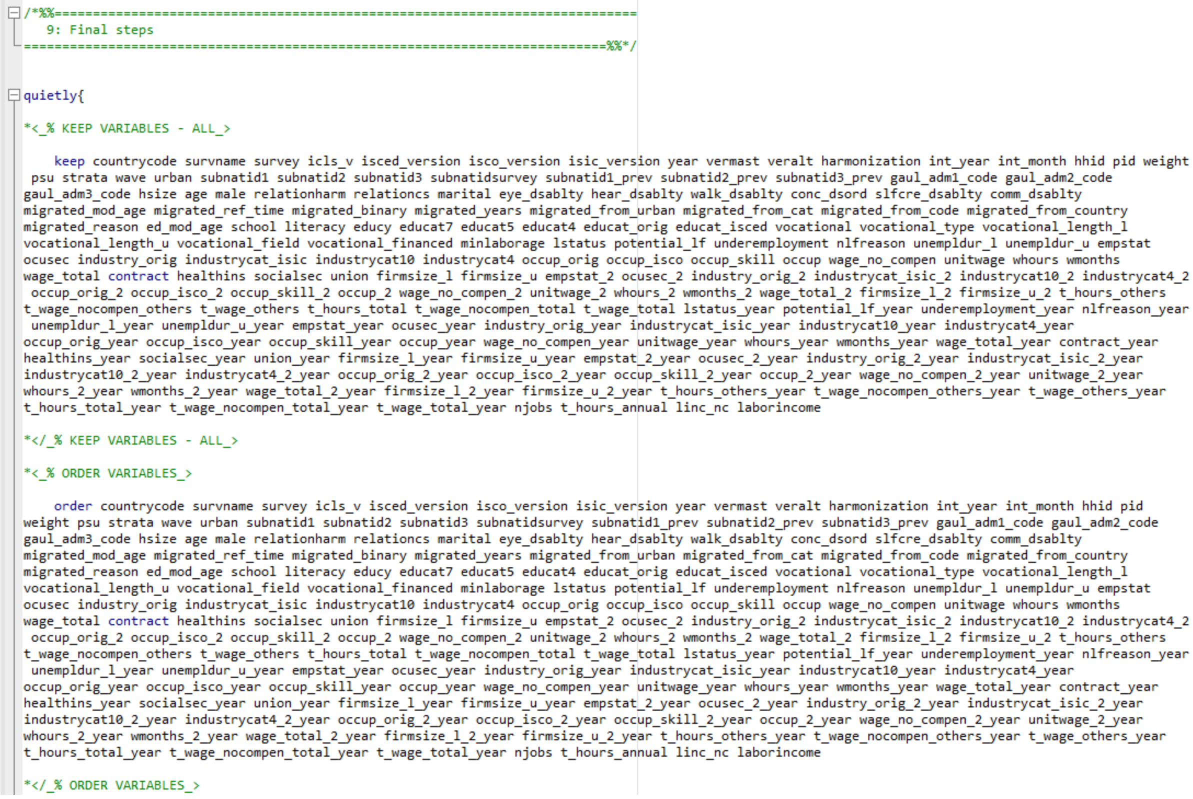
Figure 30 - Example of the code for an additional, non-standard variable



```
*<_work_lct_>
	gen work_lct=S5C14
	replace work_lct=. if lstatus!=1
	la de lblwork_lct 1 "At his/her own dwelling" 2 "At family or friend's dwelling" 3 "At
the employer's house" 4 "On the street/road" 5 "On country side" 6 "In a shop, business,
office or industry" 7 "Other(Specify)"
	label values work_lct lblwork_lct
*</_work_lct_>

}

*----------8.3: 7 day reference secondary job------------------------------*
* Since labels are the same as main job, values are labelled using main job labels
{
```

Once the variable(s) ha(s/ve) been added it needs to be included in the final section. This is because the final section not only drops variables that have missing answers for all observations but also reduces the data only to the variables in the standard dictionary and orders them accordingly. Since, by definition, any additional variable is not part of the standard dictionary it will not be part of the keeping and ordering sections (Figure 31 below). Users need to include it manually.

Figure 31 - Step 9 of the harmonization code



```
/*%%=============================================================================
   9: Final steps
=============================================================================%%*/


quietly{

*<_% KEEP VARIABLES - ALL_>

    keep countrycode survname survey icls_v isced_version isco_version isic_version year vermast veralt harmonization int_year int_month hhid pid weight
 psu strata wave urban subnatid1 subnatid2 subnatid3 subnatidsurvey subnatid1_prev subnatid2_prev subnatid3_prev gaul_adm1_code gaul_adm2_code
gaul_adm3_code hsize age male relationharm relationcs marital eye_dsablty hear_dsablty walk_dsablty conc_dsord slfcre_dsablty comm_dsablty
migrated_mod_age migrated_ref_time migrated_binary migrated_years migrated_from_urban migrated_from_cat migrated_from_code migrated_from_country
migrated_reason ed_mod_age school literacy educy educat7 educat5 educat4 educat_orig educat_isced vocational vocational_type vocational_length_l
vocational_length_u vocational_field vocational_financed minlaborage lstatus potential_lf underemployment nlfreason unempldur_l unempldur_u empstat
ocusec industry_orig industrycat_isic industrycat10 industrycat4 occup_orig occup_isco occup_skill occup wage_no_compen unitwage whours wmonths
wage_total contract healthins socialsec union firmsize_l firmsize_u empstat_2 ocusec_2 industry_orig_2 industrycat_isic_2 industrycat10_2 industrycat4_2
 occup_orig_2 occup_isco_2 occup_skill_2 occup_2 wage_no_compen_2 unitwage_2 whours_2 wmonths_2 wage_total_2 firmsize_l_2 firmsize_u_2 t_hours_others
t_wage_nocompen_others t_wage_others t_hours_total t_wage_nocompen_total t_wage_total lstatus_year potential_lf_year underemployment_year nlfreason_year
 unempldur_l_year unempldur_u_year empstat_year ocusec_year industry_orig_year industrycat_isic_year industrycat10_year industrycat4_year
occup_orig_year occup_isco_year occup_skill_year occup_year wage_no_compen_year unitwage_year whours_year wmonths_year wage_total_year contract_year
healthins_year socialsec_year union_year firmsize_l_year firmsize_u_year empstat_2_year ocusec_2_year industry_orig_2_year industrycat_isic_2_year
industrycat10_2_year industrycat4_2_year occup_orig_2_year occup_isco_2_year occup_skill_2_year occup_2_year wage_no_compen_2_year unitwage_2_year
whours_2_year wmonths_2_year wage_total_2_year firmsize_l_2_year firmsize_u_2_year t_hours_others_year t_wage_nocompen_others_year t_wage_others_year
t_hours_total_year t_wage_nocompen_total_year t_wage_total_year njobs t_hours_annual linc_nc laborincome

*</_% KEEP VARIABLES - ALL_>

*<_% ORDER VARIABLES_>

    order countrycode survname survey icls_v isced_version isco_version isic_version year vermast veralt harmonization int_year int_month hhid pid
weight psu strata wave urban subnatid1 subnatid2 subnatid3 subnatidsurvey subnatid1_prev subnatid2_prev subnatid3_prev gaul_adm1_code gaul_adm2_code
gaul_adm3_code hsize age male relationharm relationcs marital eye_dsablty hear_dsablty walk_dsablty conc_dsord slfcre_dsablty comm_dsablty
migrated_mod_age migrated_ref_time migrated_binary migrated_years migrated_from_urban migrated_from_cat migrated_from_code migrated_from_country
migrated_reason ed_mod_age school literacy educy educat7 educat5 educat4 educat_orig educat_isced vocational vocational_type vocational_length_l
vocational_length_u vocational_field vocational_financed minlaborage lstatus potential_lf underemployment nlfreason unempldur_l unempldur_u empstat
ocusec industry_orig industrycat_isic industrycat10 industrycat4 occup_orig occup_isco occup_skill occup wage_no_compen unitwage whours wmonths
wage_total contract healthins socialsec union firmsize_l firmsize_u empstat_2 ocusec_2 industry_orig_2 industrycat_isic_2 industrycat10_2 industrycat4_2
 occup_orig_2 occup_isco_2 occup_skill_2 occup_2 wage_no_compen_2 unitwage_2 whours_2 wmonths_2 wage_total_2 firmsize_l_2 firmsize_u_2 t_hours_others
t_wage_nocompen_others t_wage_others t_hours_total t_wage_nocompen_total t_wage_total lstatus_year potential_lf_year underemployment_year nlfreason_year
 unempldur_l_year unempldur_u_year empstat_year ocusec_year industry_orig_year industrycat_isic_year industrycat10_year industrycat4_year
occup_orig_year occup_isco_year occup_skill_year occup_year wage_no_compen_year unitwage_year whours_year wmonths_year wage_total_year contract_year
healthins_year socialsec_year union_year firmsize_l_year firmsize_u_year empstat_2_year ocusec_2_year industry_orig_2_year industrycat_isic_2_year
industrycat10_2_year industrycat4_2_year occup_orig_2_year occup_isco_2_year occup_skill_2_year occup_2_year wage_no_compen_2_year unitwage_2_year
whours_2_year wmonths_2_year wage_total_2_year firmsize_l_2_year firmsize_u_2_year t_hours_others_year t_wage_nocompen_others_year t_wage_others_year
t_hours_total_year t_wage_nocompen_total_year t_wage_total_year njobs t_hours_annual linc_nc laborincome

*</_% ORDER VARIABLES_>
```

Once these three steps have been completed, the user will have created a customized output that expands on the GLD harmonization without having to go through all the previous steps to create a microdata file from the raw data that can be exploited. The harmonization is no longer a “take it or leave it” product to become sandbox with a sandcastle you can easily transform.

5.3. How to cite the GLD

Citing the GLD is not as straightforward as citing an academic journal article. Imagine a GLD user who has analyzed, for example, the Brazilian PNADC, and used the harmonization code to build their data file but also added other elements. Moreover, the user read the Country Survey Details and learn something about the survey they may have learned anyway, yet much more quickly because of the CSD. Under a table of summary statistics, should this user reference the GLD or the IBGE, the Brazilian statistics office, who run and publish PNADC?

The GLD team thinks in this case, the IBGE has done the bulk of the work and ought to be named first. We propose to include in the citation that data was mediated by the GLD, that data came *via* the GLD. The below Table 5 is a mock example of the table with such a citation.

Table 5 - Example mock summary statistics table

Year	Country	# of concept A	% of concept B
2014	BRA	29,453	23.4
2015	BRA	29,209	24.5
…	…	…	…
2014	TZA	18,463	9.34
2015	TZA	18,796	8.57

Source: BRA PNADC, IBGE; …; TZA LFS, NBS via World Bank Global Labor Database

In the bibliography we propose the following citation:

> World Bank Jobs Group. (n.d.). Global Labor Database. World Bank. Retrieved (Date of retrieval), from https://github.com/worldbank/gld.

The reader should be advised that the Date of retrieval of the data should correspond not to the last day the website was viewed, but to the day the data was last downloaded. Suppose a user downloaded the data by the 10th of Month of Year and on the 11th of that Month, data for the CCCC_YYYY_LFS was updated from CCCC_YYYY_LFS_V01_M_V0***2***_GLD to CCCC_YYYY_LFS_V01_M_V0***3***_GLD. The user’s data still contains V02. Then, even though the website was last accessed on the 13th of Month, the effective date for the purposes of the reference should remain the 10th of Month.

6. Contributing to GLD’s quality and expansion

This section covers a description of how to collaborate with the GLD team, either improving the current offering or expanding it.

6.1. General rules to collaborate with GLD

The GLD team is open to any contribution or collaboration and appreciates your help. From pointing out a typo in a harmonization (Rio de Jane**i**ro, not Rio de Janero) to providing us with a fully worked out harmonization your team had worked on anyway.

The only rules are, firstly, to be mindful of each other’s timelines and workstreams so that we set out the right expectations about when a product can be delivered, and, secondly, to adhere – if you interact on our GitHub repository – by our Community Code of Conduct.

6.2. How to share new raw data with GLD

If you have a new survey not included in GLD that you would like us to collaborate on, please reach out to the GLD Focal Point.

We kindly request to make us aware of all data access restrictions on the data you would like to share. For data not in the public domain, we highly recommend the data be made available to all staff as a *development data* for *official use*. The former concept means the data does not contain personal information like names or telephone numbers (age and sex are acceptable), the latter that the data can be shared with all World Bank colleagues without restriction.

If the data is provided by the producer (commonly the NSO) free of charge, you may use the World Bank’s Data Acquisition Template (here under the heading Templates for Data Acquisition. If the data is purchased from the provider, the process depends on the amount. Below a certain threshold it can by acquired by your business unit, above a threshold the acquisition needs to go through Central Procurement. For more details and general support, please visit the (intranet) site of Central Procurement.

For restricted surveys, that is, once that cannot be shared even within the World Bank, we can keep the survey stored on the GLD server, which is only accessible to the core team and share it on a case-by-case basis. Moreover, via the datalibweb distribution system we can not only share it, equally, case-by-case, but put you as the Task Leader for the data so request and the power of approval remain with you.

6.3. How to share a harmonization for a survey not covered by GLD

If you wish to share a harmonization for a survey not covered by GLD, first of all: thank you! That is the data equivalent of finding a banknote on the street. As such, we will try to accommodate you as best as we can.

The easiest way is to reach out us is via an email to the GLD Focal Point. We can grant you temporary access to the core team GLD server where data can be exchanged. For users familiar with GitHub (or those who want to be), you may also “clone” or copy our repository, add your harmonization (following the GLD folder structure) and make a pull request. For more information, see details here on cloning as well as here on making a pull request.

6.4. How to collaborate with GLD on a new harmonization

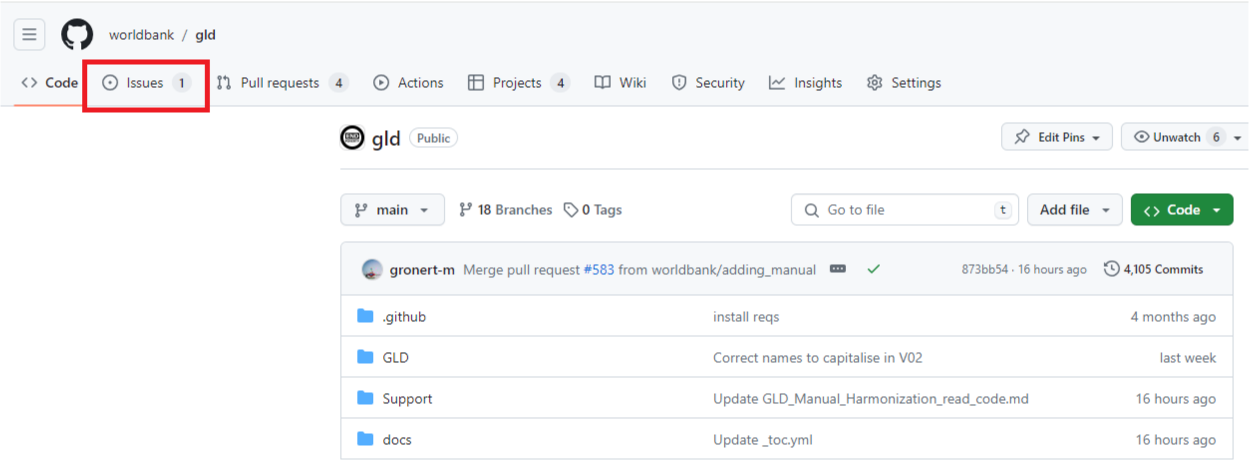
The GLD team is happy to collaborate with you on a new harmonization, divide tasks and reduce duplication. You can reach us via an email to the GLD Focal Point or by raising a blank issue on the GitHub repository. For more details on how to do this, please see the last sub-section of this section.

6.5. How to correct/ expand an existing harmonization

The GLD team is happy to rectify errors in the harmonization or add a variable we had not included previously. Despite our best efforts and extensive validation efforts some things fall through the cracks. Only extensive use of the data and your collaboration can add value at this stage. For either issue you can email the GLD Focal Point directly.

Additionally, you can raise an issue about this on our GitHub repository. This provides a standardized form to give feedback we may be able to incorporate more quickly. To do so, you may navigate to github.com/worldbank/gld/issues or click on the issues tab on the GLD repository (see red square in Figure 32 below).

Figure 32 - Finding the Issues tab on the GLD GitHub repository

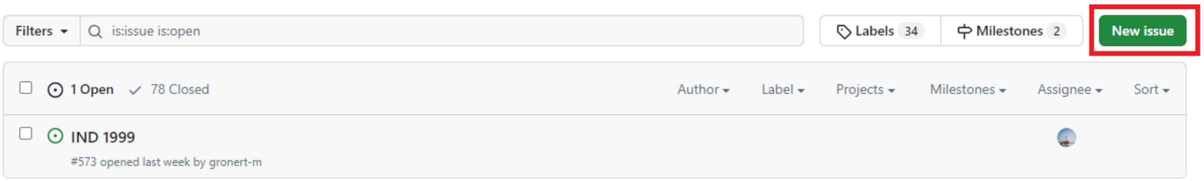


gronert-m Merge pull request #583 from worldbank/adding_manual ✓ 873bb54 · 16 hours ago 4,105 Commits

.github	install reqs	4 months ago
GLD	Correct names to capitalise in V02	last week
Support	Update GLD_Manual_Harmonization_read_code.md	16 hours ago
docs	Update _toc.yml	16 hours ago

Users then will see the list of open issues and can, by clicking on the button on the top right, open a new issue (see red square in Figure 33 below).

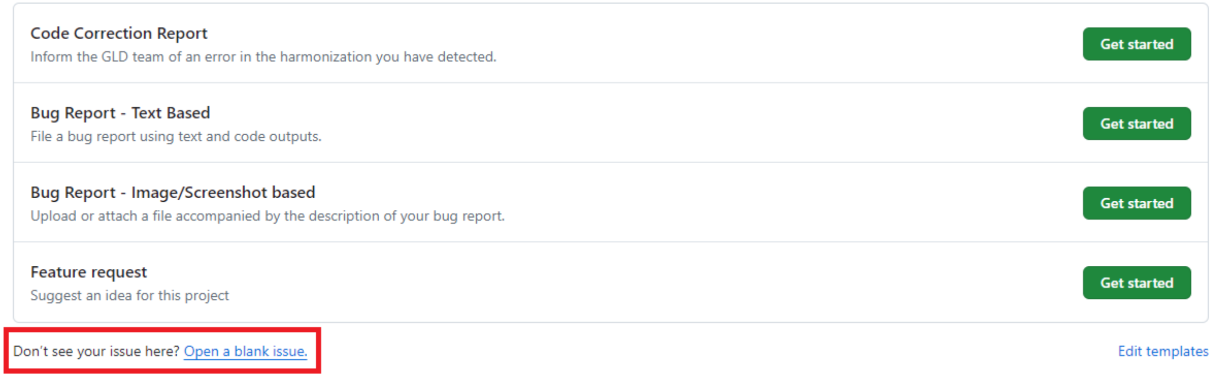
Figure 33 - Example of how to raise a new issue on the GLD GitHub repository



□ IND 1999
#573 opened last week by gronert-m

Now users have five options at their disposal (see image below). The four options listed, each with a green button stating "Get started" to initiate the issue, plus the option at the bottom (see red square in Figure 34 below) to open a blank issue, that is one that has no prior formatting.

Figure 34 - Types of issues to raise on the GLD GitHub repository



Code Correction Report
Inform the GLD team of an error in the harmonization you have detected.

Bug Report - Text Based
File a bug report using text and code outputs.

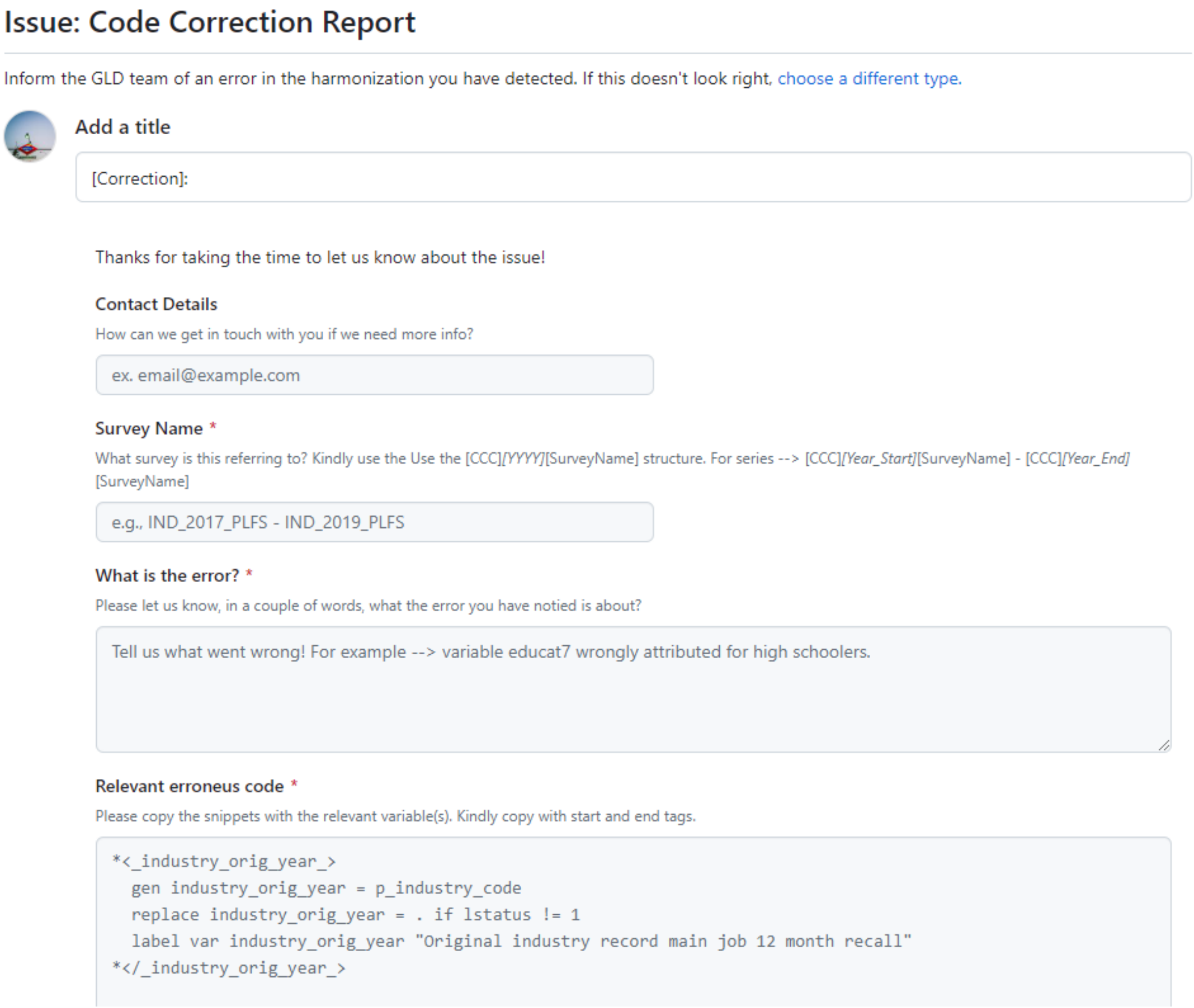
Bug Report - Image/Screenshot based
Upload or attach a file accompanied by the description of your bug report.

Feature request
Suggest an idea for this project

Don't see your issue here? Open a blank issue.

For reporting an error in the code, we recommend choosing the "Code Correction Report" option (the first). This will lead to the page shown below, where the user is provided with some boxes to make sure the information can be used directly in our data updating process, namely, to enter which surveys this applies to, a description of the error, the erroneous code and (not shown in Figure 35 below) a box for a proposed corrected code.

Figure 35 - GLD code correction issue template



Issue: Code Correction Report

Inform the GLD team of an error in the harmonization you have detected. If this doesn't look right, choose a different type.

Add a title

[Correction]:

Thanks for taking the time to let us know about the issue!

Contact Details

How can we get in touch with you if we need more info?

ex. email@example.com

Survey Name *

What survey is this referring to? Kindly use the Use the [CCC]*[YYYY]*[SurveyName] structure. For series --> [CCC]*[Year_Start]*[SurveyName] - [CCC]*[Year_End]*[SurveyName]

e.g., IND_2017_PLFS - IND_2019_PLFS

What is the error? *

Please let us know, in a couple of words, what the error you have notied is about?

Tell us what went wrong! For example --> variable educat7 wrongly attributed for high schoolers.

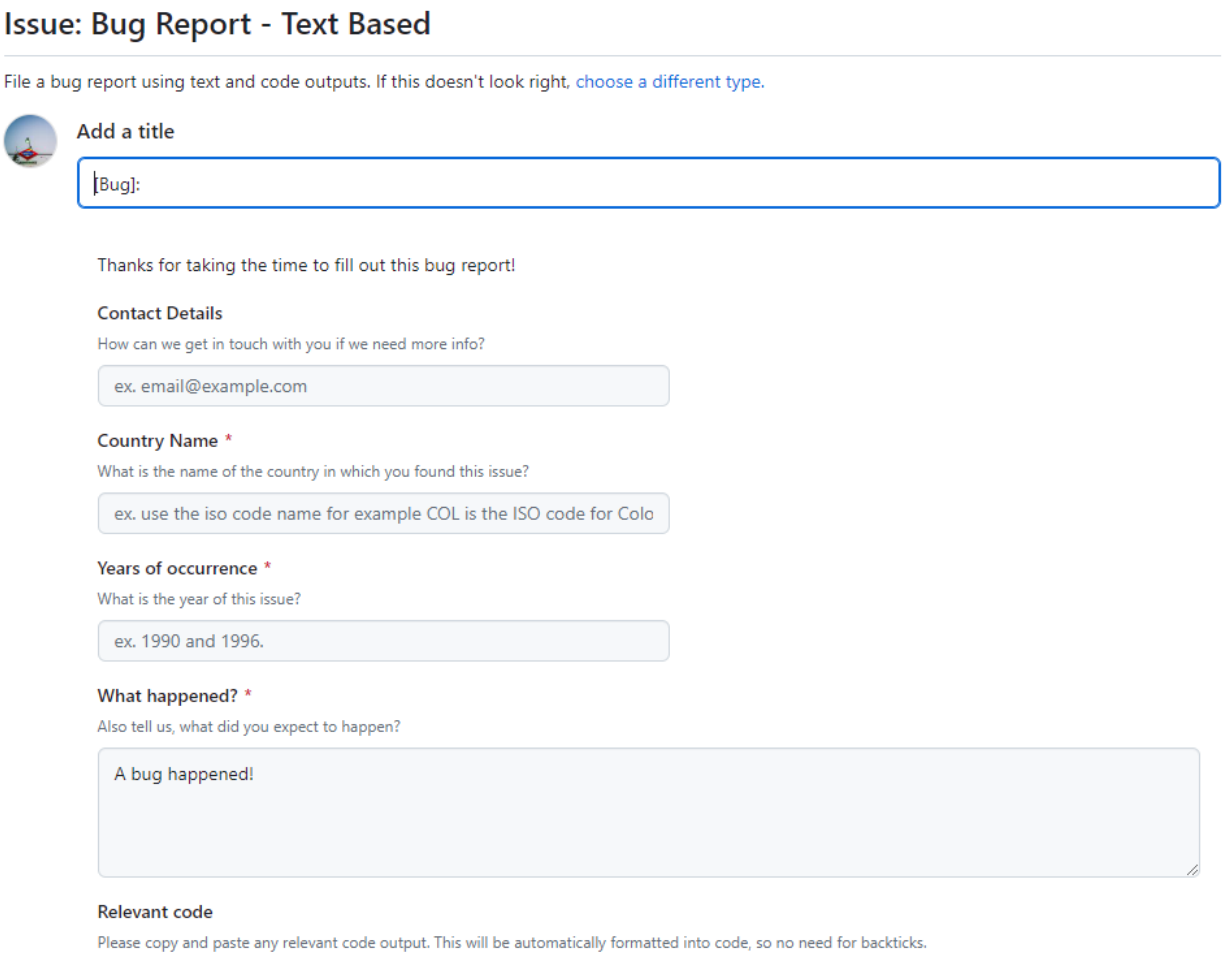
Relevant erroneus code *

Please copy the snippets with the relevant variable(s). Kindly copy with start and end tags.

```
*<_industry_orig_year_>
  gen industry_orig_year = p_industry_code
  replace industry_orig_year = . if lstatus != 1
  label var industry_orig_year "Original industry record main job 12 month recall"
*</_industry_orig_year_>
```

Users can also alert us of a bug in the code via issues. There are two prepared formats for this. The first is the text-based alert (shown below in Figure 36). The user is requested to give information about the survey, describe the issue and, if possible, add the code that is causing the issue.

Figure 36 - GLD text-based bug report template



Issue: Bug Report - Text Based

File a bug report using text and code outputs. If this doesn't look right, choose a different type.

Add a title

[Bug]:

Thanks for taking the time to fill out this bug report!

Contact Details

How can we get in touch with you if we need more info?

ex. email@example.com

Country Name *

What is the name of the country in which you found this issue?

ex. use the iso code name for example COL is the ISO code for Colo

Years of occurrence *

What is the year of this issue?

ex. 1990 and 1996.

What happened? *

Also tell us, what did you expect to happen?

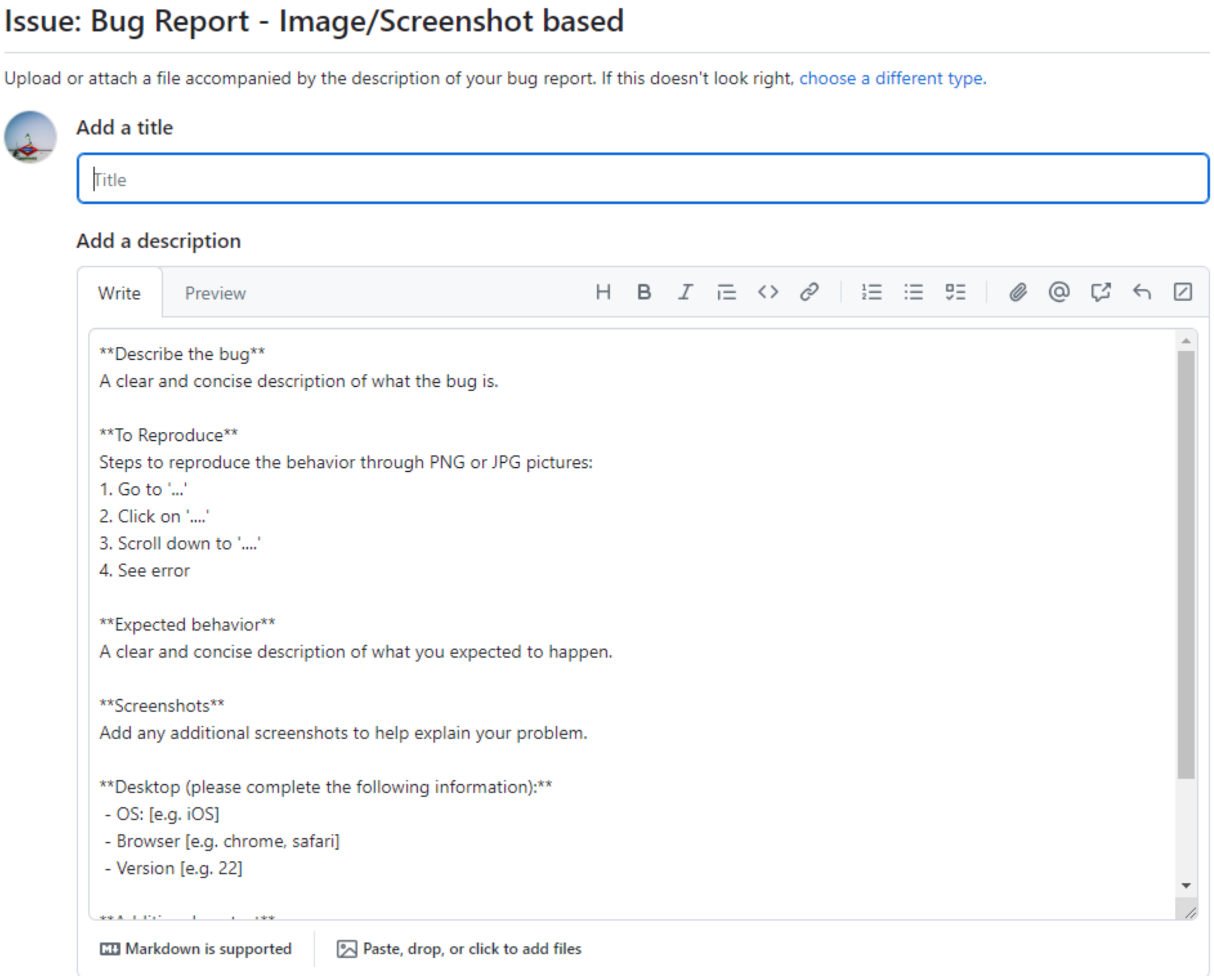
A bug happened!

Relevant code

Please copy and paste any relevant code output. This will be automatically formatted into code, so no need for backticks.

The alternative is the image/screenshot based issue report (third option on the issues options list). The text box is pre-filled with indications of what kind of information would be most useful to the GLD team (see Figure 37 below).

Figure 37 - GLD image-based bug report template



Issue: Bug Report - Image/Screenshot based

Upload or attach a file accompanied by the description of your bug report. If this doesn't look right, choose a different type.

Add a title

Title

Add a description

```
**Describe the bug**
A clear and concise description of what the bug is.

**To Reproduce**
Steps to reproduce the behavior through PNG or JPG pictures:
1. Go to '...'
2. Click on '....'
3. Scroll down to '....'
4. See error

**Expected behavior**
A clear and concise description of what you expected to happen.

**Screenshots**
Add any additional screenshots to help explain your problem.

**Desktop (please complete the following information):**
 - OS: [e.g. iOS]
 - Browser [e.g. chrome, safari]
 - Version [e.g. 22]
```

6.6. How to correct and expand the Country Survey Details

There is no specific method to expand or correct the Country Survey Details (CSD). Users are asked to either reach out to the GLD Focal Point or raise a blank issue on GitHub detailing the problems with the CSD. How to do this is described in the last sub-section of this page.

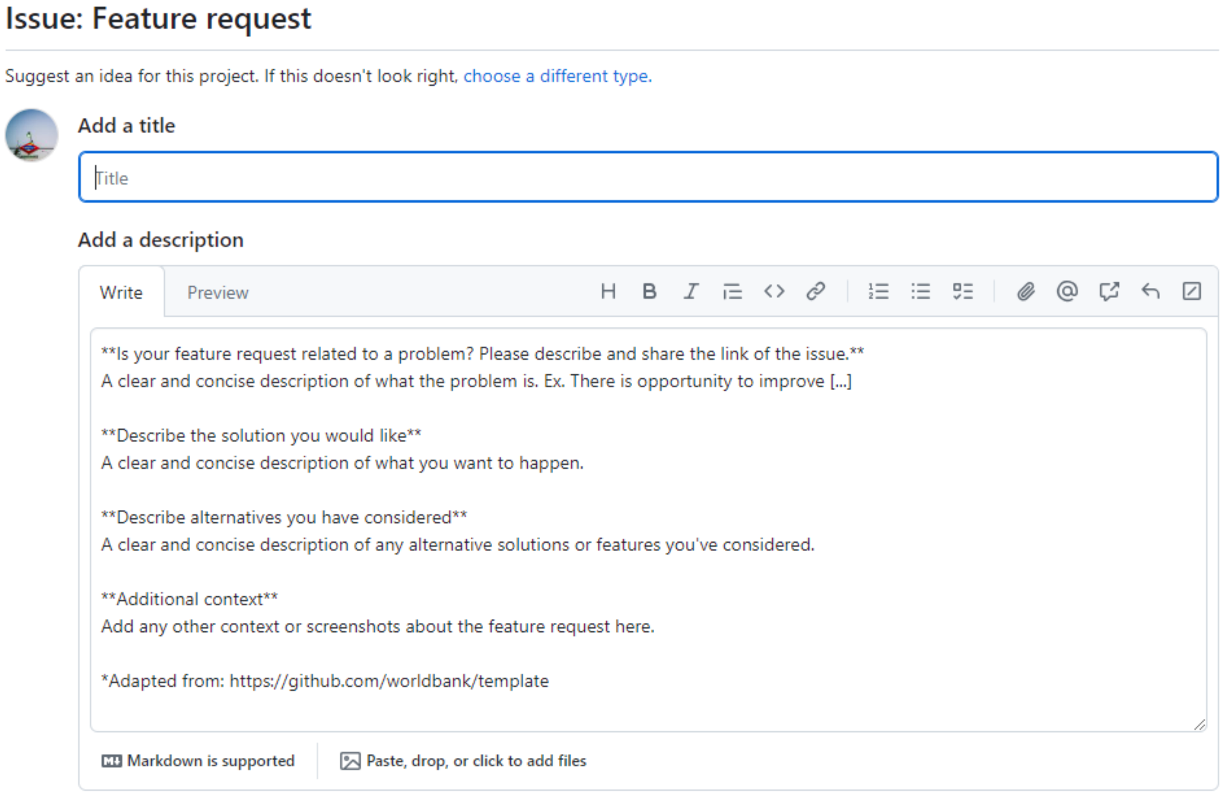
Alternatively, users can make a pull request from a clone of the GLD repository where they have corrected or amended the text, so that we may review their request and integrate it. You can find details here on cloning as well as here on making a pull request.

6.7. How to correct and expand the GLD tools

The GLD team is happy to receive updates on the tools we have created as well as on request for new tools. For corrections users can reach out to the GLD Focal Point or raise a blank issue on GitHub (see next sub-section on blank issues). They may also, as detailed above, update information on a copy of the repository and propose to merge it into the GLD via a pull request.

For requesting new tools, in addition to emails and blank issues, there is also the “Feature Request” issue (Figure 38 below). The Description box guides users on what kind of information would be most helpful for us to create a tool that solves issues you may be encountering when using the GLD.

Figure 38 - GLD feature request template



Issue: Feature request

Suggest an idea for this project. If this doesn't look right, choose a different type.

Add a title

Title

Add a description

Is your feature request related to a problem? Please describe and share the link of the issue.
A clear and concise description of what the problem is. Ex. There is opportunity to improve [...]

Describe the solution you would like
A clear and concise description of what you want to happen.

Describe alternatives you have considered
A clear and concise description of any alternative solutions or features you've considered.

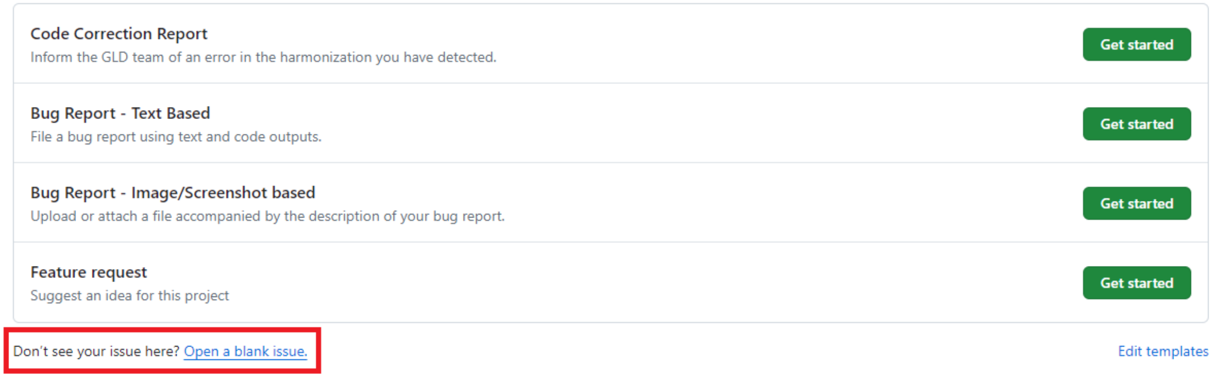
Additional context
Add any other context or screenshots about the feature request here.

*Adapted from: https://github.com/worldbank/template

6.8. How to communicate on any other issue

The GLD team encourages you to reach out on any issue. To inform us you can either reach out via email or create a blank issue on GitHub. A blank issue can be created, after clicking on the "Issues" tab in the repository, by selecting the "Open a blank issue" option at the end of the list of issue templates (red box in Figure 39 below).

Figure 39 - Example of how to raise a blank issue on the GLD GitHub repository



Code Correction Report
Inform the GLD team of an error in the harmonization you have detected.

Bug Report - Text Based
File a bug report using text and code outputs.

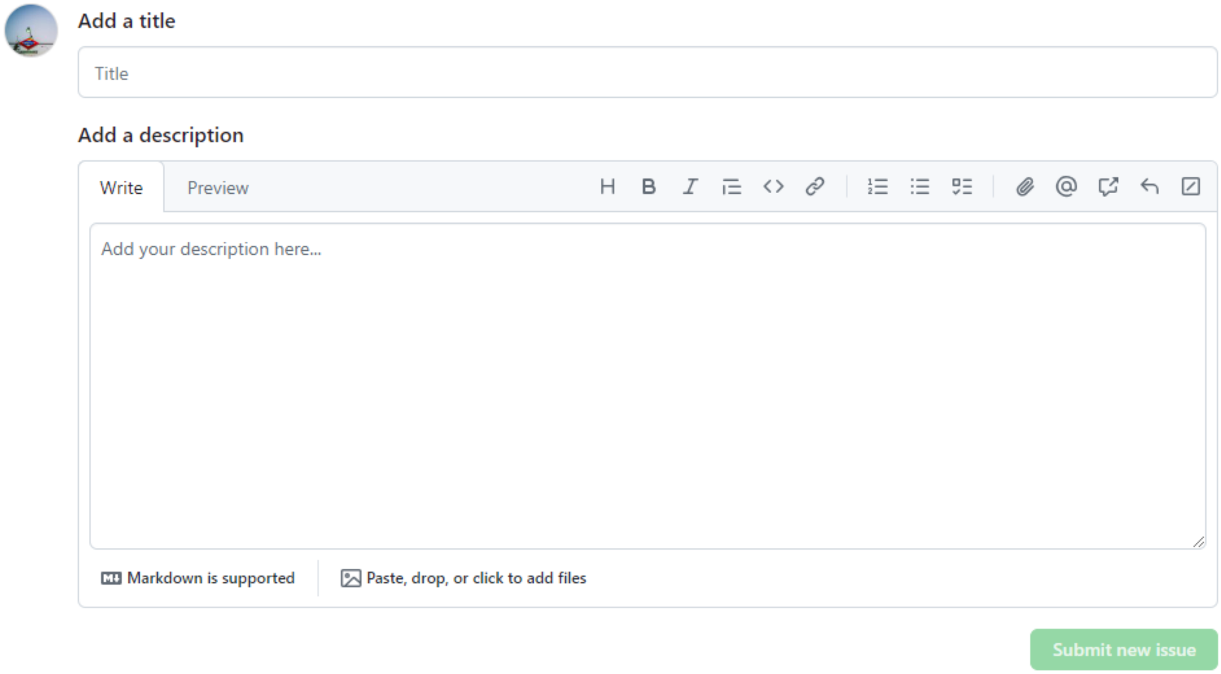
Bug Report - Image/Screenshot based
Upload or attach a file accompanied by the description of your bug report.

Feature request
Suggest an idea for this project

Don't see your issue here? Open a blank issue.

Here the user is only requested to enter a title to their issue and they have a blank box at their disposal to detail any issue they wish to communicate to us (see Figure 40 below).

Figure 40 - GLD blank issue template



Add a title
Title
Add a description
Write
Preview
Add your description here...
Markdown is supported
Paste, drop, or click to add files
Submit new issue