

README

Project: *Rationalizing Informality: Self-Employment, Job Quality and Growth (P180914)*

Date: 11/05/2026

1. Overview

Overview

This repository contains the **code required to reproduce the figures and tables** of the *Informality Report*.

The workflow is organized by **chapters**, and a master do-file runs all the individual scripts to generate the outputs.

Figures and tables sourced from external references

The following figures and tables included in the report are **reproduced or cited from external studies and data sources**. As they are not originally generated within this project's codebase, they are **not included in the reproducibility package**. For access to these materials, please refer to the **original publications and data sources cited in the report**. These figures can be consulted directly in their respective references.

Exhibit	Source
Fig 1.8	Fietz, K., C. Joubert, H. Nopo et al. 2025. (In)Formalizing Jobs in Latin America and the Caribbean: Taxes, Benefits, and Labor Market Incentives. World Bank. https://openknowledge.worldbank.org/server/api/core/bitstreams/0950285f-07d4-4164-a3e7-8332632e2195/content . Page 125.
Fig 2.1a	Evans, D. S., and L. S. Leighton. 1989. "Some Empirical Aspects of Entrepreneurship." American Economic Review 79 (3): 519–35. https://www.edegan.com/pdfs/Evans%20Leighton%20(1989)%20-%20Some%20Empirical%20Aspects%20Of%20Entrepreneurship.pdf . Page 7.
Fig 2.1c	Fajnzylber, P., W. Maloney, and G. Montes-Rojas. 2006. "Microenterprise Dynamics in Developing Countries: How Similar Are They to Those in the Industrialized World? Evidence from Mexico." World Bank Economic Review 20 (3): 389–419. https://academic.oup.com/wber/article-abstract/20/3/389/1664195?_cf_chl_f_tk=E946JvnCY4cBY3bRbCY0PijCzyaxVBYplswgui07Rs-1782901898-1.0.1.1-NvzPUbjCclo3HjcZmRglkLx6FkDGqtE2ZBlpaAsW2Zk . Page 14.
Table 2.1	Beylis, G., E. Cerutti, and W. Maloney. Forthcoming. "Formality as an Input in Production: Micro-Entrepreneurship and Informality in Chile, Colombia, and Mexico." Background paper for the study. Forthcoming.
Table 2.2 (US)	2018 Report on Nonemployer Firms: Findings from the 2017 Small Business Credit Survey. Page 12 (https://www.fedsmallbusiness.org/reports/survey/2018/2018-report-on-nonemployer-firms).
Fig 2.8	Maloney, W. F., G. Vuletin, P. Garriga, and R. Morales. 2025. Transformational Entrepreneurship for Jobs and Growth. Latin America and the Caribbean Economic Review (October). World Bank. http://hdl.handle.net/10986/43741 .

	https://openknowledge.worldbank.org/server/api/core/bitstreams/5adee9c4-8539-4292-86fb-7773573fe9a5/content . Page 51.
Fig 2.9	Beylis, G., W. Maloney, and A. Zambrano. 2026. "Self-Employment and Development: Job Amenities, (Mis)allocation, and Informality." World Bank. Forthcoming.
Fig 2.10	Levenson, A. R., and W. F. Maloney. 1998. "The Informal Sector, Firm Dynamics, and Industrial Participation." Policy Research Working Paper 1988, World Bank Group. Link: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5180 Page: The figure is based on the results shown in Table 4 of the document, located on page 36. Note that Panel A corresponds to the "Logarithmic Revenue" column of the table, while Panel B corresponds to the "Years in Business" column.
Map 3.1	Ianchovichina, E. 2024. La evolución geográfica de la productividad y el empleo: Ideas para lograr un crecimiento inclusivo a través de una perspectiva territorial en América Latina y el Caribe. Estudios del Banco Mundial sobre América Latina y el Caribe, World Bank. https://www.bancomundial.org/es/region/lac/publication/the-evolving-geography-of-productivity-and-employment . Page 30 (Map PG.10).
Fig 3.22	Feinmann, J., A. P. Franco, P. Garriga, N. Gonzalez-Prieto, R. Hsu Rocha, and M. Lauletta. Forthcoming. "Payments under the Table in Latin America: Evidence from Argentina, Brazil, Colombia, Chile, Mexico, and Peru." Background paper for this volume, unpublished.
Fig 4.1a	Beylis, G., W. Maloney, and A. Zambrano. 2026. "Self-Employment and Development: Job Amenities, (Mis)allocation, and Informality." World Bank. Forthcoming.
Fig 4.1b	Beylis, G., W. Maloney, and A. Zambrano. 2026. "Self-Employment and Development: Job Amenities, (Mis)allocation, and Informality." World Bank. Forthcoming.

2. Repository structure

Root:

1. Analysis/
 - a. Codes/ *(All do-files, organized by chapter)*
 - i. Intro
 - ii. CH1
 - iii. CH2
 - iv. CH3
 - v. Master.do (master dofile to run everything)
2. Data/ *(Input datasets, multiple sources)*
3. Outputs/
 - a. Figures *(All exported figures)*
 - i. Intro
 - ii. CH1
 - iii. CH2
 - iv. CH3
 - b. Tables *(All exported tables)*
 - i. Intro
 - ii. CH1
 - iii. CH2

iv. CH3

v.

- **How to Run the Project**

1. Open Stata
2. Set the correct root directory in master.do:

"global root "your/local/path/to/project"

3. Run:

do master.do

This will sequentially execute all do-files and export figures and tables into the corresponding folders.

- **Workflow Description**

- Each figure or table is generated by a separate do-file. (/Analysis/Code...)
- Do-files are grouped by chapter
- The master.do file calls each dofile in the correct order
- Outputs are exported automatically
- Each do-file contains its own data loading and transformations
- The master do-file acts only as an execution pipeline
- The code is designed to be run through the master do-file, which defines all global paths and required settings. Users can also run individual do-files independently; however, in such cases, they must replicate the global path definitions from the master do-file within each individual file to ensure proper execution.

4. Data Sources

The datasets included in the package correspond to intermediate harmonized analytical files constructed from multiple household surveys and administrative sources.

Due to restrictions associated with redistribution agreements and the extensive harmonization process conducted over several years, the full raw-to-final construction pipeline is not fully automated in a single reproducible script.

However, all figures and tables in the report can be reproduced directly from the analytical datasets included in the package and the accompanying do-files. The reproducibility package combines publicly available datasets and restricted-access microdata. Restricted datasets cannot be redistributed due to licensing agreements and confidentiality restrictions imposed by the original data providers. All analysis files included in this package were constructed from the original sources listed below.

Note: The URLs provided in this README correspond to the original raw data sources, when publicly available. Several datasets used in the analysis are harmonized and processed versions

constructed by the authors from the original microdata. These harmonized analytical datasets are not publicly distributed due to data-use agreements and confidentiality restrictions imposed by the original data providers.

The analysis relies on multiple data sources, including:

a. Data name: Adult_pop_WDI_data.xlsx

Source: World Development Indicators (WDI), World Bank.

Access date: May 2026

Access type: Public data.

Access instructions: Available through the World Bank Data Catalog and WDI database.

URL: <https://databank.worldbank.org/source/world-development-indicators>

Notes: Contains adult population indicators used in descriptive analysis and figure construction. Data corresponds to the year 2022. The variables used from the source dataset are country identifier (country_name) and adult population. No harmonization of the original dataset was performed. The only transformation applied was the construction of a regional classification variable.

Regions were defined as follows: Latin America and the Caribbean (LAC): Argentina, Brazil, Chile, Colombia, Guatemala, Panama, Peru, Puerto Rico, and Uruguay. OECD non-LAC: Austria, Canada, France, Germany, Greece, Ireland, Israel, Italy, Japan, Luxembourg, Netherlands, Poland, Slovenia, South Korea, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States. Sub-Saharan Africa: Angola, Madagascar, and Sudan. North Africa: Egypt and Morocco. Middle East: Cyprus, Iran, Lebanon, Qatar, Saudi Arabia, and United Arab Emirates. Eastern Europe and Eurasia: Bulgaria, Croatia, Russia, and Slovakia. East Asia: China and Taiwan. Southeast Asia: Indonesia and Thailand. South Asia: India.

The table also incorporates information from the restricted dataset “GEM_2018_Data.dta”. Due to data access restrictions, this dataset cannot be redistributed as part of the replication package. No additional harmonization was performed on the non-restricted source data.

b. Data name: Implicit tax data.xlsx

Source: Beylis, G., A. Garin, N. Gonzalez-Prieto, and D. Koustas. Forthcoming. “Valuing Alternative Work Arrangements: Evidence from 7 Countries with High Rates of Informality.” Policy Research Working Paper, World Bank; Non-salary labor costs of Universidad de los ANDES

Access date: January 2026

Access type: Restricted/ Constructed analytical dataset.

Note: This dataset combines information on workers' willingness to pay (WTP) for social security benefits with statutory social security contribution rates across countries. The WTP estimates are derived from the study by Beylis, Garin et al. (2026). Information on employer non-wage labor costs and employee social security contributions was compiled directly from the social security and labor legislation of each country included in the analysis: Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, and Peru. The dataset was constructed by combining these sources at the country level. The resulting variables capture the statutory contribution burden borne by employers and employees, together with the corresponding WTP estimates used in the analysis.

This dataset cannot be included in the reproducibility package because it incorporates information derived from external sources that are not authorized for redistribution. The WTP estimates originate from Beylis, Garin et al. (2026). Researchers interested in accessing these data should consult the authors of the study or the data repository associated with the publication, subject to the corresponding access conditions. Information on employer and employee social security contribution rates was compiled from publicly available national legislation in Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, and Peru. These statutory parameters can be reconstructed by consulting the relevant legal and regulatory sources for each country.

c. Data name: Data_LAMRIG_to_IMF.xls

Source: Authors' calculations based on IMF and labor market rigidity indicators.

Access date: May 2026

Access type: Restricted/Constructed data.

Access instructions: Constructed by the authors from publicly available sources.

Notes: Intermediate file used for harmonization and comparison exercises. The dataset contains country-level observations by year and period, including the labor market regulation indicator (LAMRIG). The LAMRIG indicator is derived from the dataset underlying: Campos, N. F., & Nugent, J. B. (2012). *The dynamics of the regulation of labor in developing and developed countries since 1960*. IZA Discussion Paper No. 6881, Institute for the Study of Labor (IZA). The file used in this project was obtained from the data associated with this study and contains the country-period observations employed in the analysis. No code used to construct the original LAMRIG dataset is available as part of this replication package. Researchers interested in reproducing or extending the construction of the indicator should consult the original study and contact the authors to request access to the underlying data and documentation. The dataset is not included in the reproducibility package. It was obtained from the data associated with Campos and Nugent (2012). Researchers interested in accessing the data should consult the original publication and contact the authors regarding data availability and access conditions.

d. Data name: Data_all_variables.dta

Source: Authors' calculations based on ILO, World Bank and WHO

Access date: March 2026

Access type: Restricted/Constructed analytical dataset.

Access instructions: Constructed by the authors from the sources listed above.

Notes: Combined dataset used for cross-country analysis and figure production. The dataset combines information from three international sources:

* **ILO (International Labour Organization):** informality rate.

* **World Bank (WB):** GDP (PPP-adjusted) and number of employed workers.

* **World Health Organization (WHO):** indicators of health coverage

The dataset was constructed by harmonizing country-year observations across sources using standardized country identifiers. Country names were standardized and matched to ISO country codes before merging the datasets. The resulting database contains annual observations through 2023 and includes a regional classification following World Bank geographic regions.

Missing observations were interpolated within countries when sufficient time-series information was available. In addition, GDP per worker was computed as the ratio of GDP to the number of employed workers. For selected analyses, indexed measures were constructed using year 2000 as the base year (2000 = 100), including GDP per worker and health coverage indicators.

The harmonization and construction procedures were implemented through project-specific data processing scripts that are not distributed as part of the replication package. The final constructed dataset and the harmonization files are not included in the reproducibility package. The underlying data sources are publicly available through the ILO, World Bank, and WHO data portals. Users can access these sources directly from the respective institutional databases under their standard access conditions.

URL: <https://www.who.int/data/data-collection-tools/health-service-data>;
<https://data.worldbank.org/indicator>; <https://ilostat.ilo.org/es/topics/informality/>

e. Data name: ENAMIN, EME VII and EMICRON.dta

Source: Encuesta Nacional de Micronegocios (ENAMIN), Mexico; EME VII, Chile; and EMICRON, Colombia.

Access date: April 2026

Access type: Restricted /harmonized microdata.

Notes: Harmonized microenterprise survey data used in comparative analysis. The figure uses microdata from three surveys:

* Chile: Encuesta de Microemprendimiento (EME), VI wave (2022)

* Colombia: National Microenterprise Survey (EMICRON, 2024)

* Mexico: Encuesta Nacional de Micronegocios (ENAMIN, 2012)

The analysis uses firm-level information on firm age, federal registration status, and firm revenue. Revenue is available in the original surveys and is used to construct a logarithmic

transformation, which is the variable employed in the estimation.

A probit model is estimated to obtain predicted probabilities of federal registration. The specification includes firm age group indicators and controls for log revenue. Predicted margins by firm age group are computed holding log revenue constant, and 95% confidence intervals are reported.

Variables were harmonized across surveys to ensure consistency in definitions and comparability across countries. The harmonized dataset used for the analysis is not included in the reproducibility package. The underlying microdata come from official national surveys (EME, EMICRON, and ENAMIN), which are subject to the access conditions of the respective statistical agencies. Researchers interested in accessing the data should refer to the corresponding national data providers.

URL: <https://www.inegi.org.mx/programas/enamin/2012/>;
<https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/micronegocios>;
<https://www.economia.gob.cl/2023/06/16/septima-encuesta-de-microemprendimiento-eme-7.htm>

f. Data name: ENAMINFINAL2012.dta

Source: Encuesta Nacional de Micronegocios (ENAMIN) 2012, Instituto Nacional de Estadística y Geografía (INEGI), Mexico.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Original survey available through INEGI. Harmonized analytical version is not publicly redistributed.

Notes: Processed ENAMIN 2012 microdata used in the analysis. The dataset corresponds to Mexico's National Microenterprise Survey (ENAMIN) 2012. The analysis uses individual-level information on sex and the main reason for leaving a previous job (Question 23 of the survey: "¿Cuál fue el motivo principal por el que dejó ese trabajo?").

The construction of the analytical dataset involves merging several survey modules (household roster, income/expenditure, and individual-level employment information) using a unique household-respondent identifier. The data are first standardized at the person level and reshaped to ensure consistent individual identifiers across modules.

The final dataset is built by linking household and individual modules, retaining matched observations only. Survey variables are renamed to ensure internal consistency across files, and identifiers are converted to numeric format for analysis. The resulting dataset is saved as the final analytical file used in the estimations.

No additional harmonization files are included in the reproducibility package; only the final constructed dataset is provided. The ENAMIN 2012 microdata were accessed through the National Institute of Statistics and Geography (INEGI, Mexico). The data are subject to the

access conditions established by INEGI and can be obtained directly from its official data portal.

The replication package does not redistribute the microdata; only the constructed analytical dataset is used in the analysis.

URL: <https://www.inegi.org.mx/programas/enamin/2012/>

g. Data name: ENOE and ENAMIN.dta

Source: Encuesta Nacional de Ocupación y Empleo (ENOE) and Encuesta Nacional de Micronegocios (ENAMIN), Mexico.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to the conditions established by INEGI.

Notes: Combined analytical dataset merging labor force and microenterprise surveys. The dataset corresponds to Mexico, using ENAMIN 2012 microdata combined with ENOE (National Survey of Occupation and Employment) data.

The main variables used in the analysis are income deciles, current income, and previous salary, together with an indicator identifying self-employed individuals in the ENAMIN survey. The outcomes are expressed as ratios of required income relative to current income and previous salary, respectively.

The construction of the dataset involves linking ENAMIN respondents to ENOE records using common household and individual identifiers. This matching procedure allows the creation of a structure combining self-employment income information (ENAMIN) with labor market history and earnings information (ENOE). Only successfully matched observations are retained for analysis.

The dataset used for estimation is constructed by merging ENAMIN with ENOE, followed by sample restriction to matched individuals and aggregation for analysis. No intermediate construction code is included in the reproducibility package; only the final analytical dataset is provided. The constructed dataset combining ENAMIN and ENOE data is not included in the reproducibility package. The underlying microdata are obtained from INEGI (Mexico) through its official data access procedures. The replication package does not redistribute either the raw microdata or the matched dataset.

URL:

<https://www.inegi.org.mx/programas/enamin/>; <https://www.inegi.org.mx/programas/enoe/15ymas/>

h. Data name: entrepreneurship_data.dta

Source: Authors' calculations based on WDI, OECD, and household survey data.

Access date: May 2026

Access type: Constructed analytical dataset.

Access instructions: Constructed by the authors from publicly available and restricted-access sources.

Notes: Dataset used to construct entrepreneurship and self-employment figures. The analysis

uses country-level data for 2020 from three sources. For Latin American and Caribbean countries, the self-employment rate was obtained from SEDLAC, a database that harmonizes national household surveys following World Bank methodological guidelines. Both the harmonized database and the underlying harmonization procedures are restricted-access and cannot be shared as part of the replication package. For countries outside Latin America and the Caribbean, self-employment rates were obtained from the OECD Self-Employment Rate database. GDP data were obtained from the World Bank's World Development Indicators database. The main variables used in the analysis are the percentage of self-employed individuals in total employment and GDP. The final analytical dataset was constructed by combining the self-employment indicators from SEDLAC (for Latin American and Caribbean countries) and OECD (for all other countries) with GDP data from the World Bank at the country-year level for 2020. No additional harmonization of the original SEDLAC data was performed by the authors. The SEDLAC harmonized dataset cannot be included in the reproducibility package because both the harmonized database and the underlying harmonization procedures are restricted-access. The data were accessed through SEDLAC, a database jointly maintained by the World Bank and CEDLAS. Researchers interested in accessing these data should contact the SEDLAC team and comply with the applicable access requirements established by the data providers. In contrast, the OECD self-employment data used in the analysis are publicly available through the OECD Self-Employment Rate database and can be directly downloaded from the OECD website. Likewise, the GDP data are publicly available through the World Bank's World Development Indicators database and can be accessed and downloaded from the World Bank Data portal. Therefore, only the OECD and World Bank data can be independently retrieved from public sources, while access to the SEDLAC harmonized data is subject to the restrictions established by the database administrators.

URL: <https://www.oecd.org/en/data.html>; <https://databank.worldbank.org/source/world-development-indicators>

i. Data name: gdp_2021_per_capita_ppp.xlsx

Source: World Development Indicators (WDI), World Bank.

Access date: April 2026

Access type: Public data.

Access instructions: Available through the World Bank Data Catalog.

Notes: GDP per capita PPP indicators used in cross-country comparisons.

Indicators used: GDP, PPP (constant 2021 international \$), GDP per capita, PPP (current international \$), GDP per capita, PPP (constant 2021 international \$)

URL: <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD>

j. Data name: gdp_quarterly_long_LAC.dta

Source: Authors' calculations based on national accounts and macroeconomic data sources.

Access date: April 2026

Access type: Constructed data.

Access instructions: Constructed by the authors from publicly available macroeconomic datasets.

Notes: Quarterly GDP dataset used for macroeconomic comparisons. The analysis uses quarterly GDP data for Latin American and Caribbean countries covering the period from 2002 onward. The primary variable is real GDP measured at constant prices in national currency units, obtained from World Bank sources and organized into a country-quarter panel dataset. In addition to the original quarterly GDP series, the dataset includes quarterly and annual GDP growth rates, which were calculated directly from the quarterly GDP variable. The construction process consisted of extracting the quarterly GDP series for each country and deriving the corresponding growth measures before combining all information into a harmonized database covering the study period. Although the specific series used in the analysis were obtained through World Bank sources, equivalent GDP data at constant prices in national currency units are publicly available from other repositories, including CEPALSTAT (<https://statistics.cepal.org/portal/cepalstat/dashboard.html>). Researchers seeking to reproduce the analysis may access these publicly available sources and retrieve the corresponding quarterly GDP series for the countries and periods covered in the study.

k. Data name: GDPpc_WDI_data_for_ISSP.xlsx

Source: World Development Indicators (WDI), World Bank.

Access date: May 2026

Access type: Public data.

Access instructions: Available through the World Bank Data Catalog.

Notes: GDP per capita indicators prepared for ISSP comparisons. The dataset uses country-level GDP per capita based on purchasing power parity (PPP), measured at constant 2021 international dollars. The data correspond to the year 2005 and were obtained from the World Bank World Development Indicators (WDI), specifically the indicator NY.GDP.PCAP.PP.KD. The variable is used in levels and includes no additional transformations beyond the direct extraction of the indicator from the WDI database. The data are publicly available and can be accessed through the World Bank Data portal.

URL: <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD>

l. Data name: GEIH_2022

Source: Gran Encuesta Integrada de Hogares (GEIH), Departamento Administrativo Nacional de Estadística (DANE), Colombia.

Access date: January 2026

Access type: Public/harmonized microdata.

Access instructions: The .dta files contained in the GEIH 2022 folder correspond to the original monthly microdata from the Gran Encuesta Integrada de Hogares (GEIH) for 2022. Access to this data is available via the link provided below.

The included do-files, used to generate the figures from these databases, contains the procedures for harmonizing the monthly microdata, as well as the transformations necessary to generate the variables and results used in the analysis.

Notes: Household survey microdata used for labor market analysis. The data in the GEIH 2022 folder correspond to monthly microdata from the Gran Encuesta Integrada de Hogares (GEIH) for Colombia. The underlying microdata are publicly available and provided by the National Administrative Department of Statistics (DANE).

The construction process assumes that the different survey modules for each month have already been obtained and integrated at the monthly level. Starting from these monthly datasets, the do-file used to generate the figures and tables performs the following steps:

- Harmonization of variables within each monthly dataset once all survey modules have been merged at the individual level.
- Standardization and recoding of key labor market variables to ensure consistency across months.
- Appending of all monthly datasets for 2023 into a single annual dataset.
- Construction of additional variables required for the empirical analysis.
- Use of the resulting harmonized annual dataset for the production of figures and tables.

No external data sources are used beyond the GEIH microdata provided by DANE. As the harmonised version cannot be included in the reproducibility package, please share information on how potential replicators might access it.

The harmonized dataset cannot be included in the reproducibility package due to size and redistribution restrictions. However, full replication is possible using publicly available data.

The original GEIH 2023 microdata can be accessed through the DANE (Departamento Administrativo Nacional de Estadística) data portal. Researchers can obtain the monthly survey files, merge the relevant modules at the monthly level, and then run the provided do-file, which performs the harmonization, variable construction, and monthly append to generate the final analytical dataset.

URL: <https://microdatos.dane.gov.co/index.php/catalog/782>

m. Data name: GEM_2018_Data.dta

Source: Global Entrepreneurship Monitor (GEM) 2018.

Access date: May 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to GEM data access conditions.

Notes: Entrepreneurship indicators used for comparative analysis. The dataset uses country-level information from the Global Entrepreneurship Monitor (GEM) for the year 2018. The variables extracted correspond to standard GEM indicators on entrepreneurship, with no structural modifications to the underlying data. Any changes applied to the original variables are limited to renaming or relabeling for consistency within the analytical dataset, and these

adjustments are fully documented in the corresponding replication code (do-files) used to generate the figures and tables. The construction process therefore consists of extracting the relevant GEM indicators for 2018 and preparing them for analysis without altering the underlying measurement or structure of the original dataset. The dataset cannot be included in the reproducibility package due to restrictions on redistributing the compiled dataset and any derived harmonization files. However, the underlying data are publicly available through the Global Entrepreneurship Monitor (GEM) data portal (<https://www.gemconsortium.org/data>). Researchers can access the original GEM indicators directly from this source and reconstruct the dataset using the same 2018 country-level variables employed in the analysis. No proprietary transformation of the data is required beyond the variable selection and labeling steps, which are fully documented in the replication code.

URL: <https://www.gemconsortium.org/data>

n. Data name: ISSP_2005_Data.dta

Source: International Social Survey Programme (ISSP) 2005.

Access date: May 2026

Access type: Public/restricted depending on registration status.

Access instructions: Available through the ISSP data archive subject to registration requirements.

Notes: Survey data used for attitudinal comparisons. The dataset is based on the 2005 wave of the International Social Survey Programme (ISSP), available through GESIS (ZA4350). The data used in the analysis are fully derived from this source, and no additional transformations are applied beyond those explicitly documented in the replication code used to generate the figures. The construction process consists of selecting the relevant variables from the ISSP 2005 dataset and preparing them for analysis, with all variable definitions, recoding, and any renaming fully documented in the corresponding do-files. No further modifications to the original data structure are performed. The ISSP 2005 dataset cannot be included in the reproducibility package due to data redistribution restrictions. However, the data are publicly available and can be accessed through GESIS at https://search.gesis.org/research_data/ZA4350. Researchers can obtain the original ISSP 2005 microdata from this source and replicate the analysis using the same variable definitions and sample selection criteria as specified in the replication code. All transformations, including any recoding or variable renaming, are fully documented in the provided do-files, and no additional proprietary processing was applied beyond these steps.

o. Data name: LAC_allyears.dta

Source: Harmonized Latin American household survey microdata compiled from SEDLAC and national statistical offices.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to the conditions established by SEDLAC and national statistical agencies.

Notes: Multi-country pooled household survey dataset. The dataset is based on SEDLAC harmonized household survey microdata for Latin American and Caribbean countries. It covers

all available survey years from 2005 to 2023 and includes repeated cross-sections at the individual level. The analytical sample is restricted to individuals aged 18 to 65 who are employed and have non-missing information on education, sector of activity, firm size, and pension contribution status.

The SEDLAC database is constructed by the World Bank and CEDLAS through harmonization of national household surveys; both the underlying microdata and the harmonized outputs are restricted-access and cannot be redistributed. The variables used in the analysis are extracted directly from the harmonized SEDLAC dataset and are fully specified in the corresponding replication do-files associated with each figure or table. Any additional variable transformations or constructed variables are implemented in the replication code and do not modify the underlying structure of the SEDLAC data. This dataset cannot be included in the reproducibility package due to restrictions on both the SEDLAC harmonized microdata and the associated harmonization procedures. Access to SEDLAC is granted through the World Bank and CEDLAS under specific data use conditions. Researchers interested in using the data must request access through the SEDLAC platform and comply with the corresponding data agreements.

As an alternative, users may access the original national household survey microdata through the respective national statistical offices of Latin American and Caribbean countries, subject to each agency's data access policies. The analytical dataset used in this study is fully reproducible in principle from the underlying surveys using the harmonization framework, but the specific SEDLAC harmonized version employed here is not publicly distributable.

p. Data name: LAC_allyears_var.dta

Source: Authors' harmonization of Latin American household survey microdata from SEDLAC and national statistical offices.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Original surveys accessible through SEDLAC and national statistical agencies. Harmonized analytical version is not publicly redistributed.

Notes: Main analytical dataset used throughout the report. The dataset is based on SEDLAC harmonized household survey microdata for Latin American and Caribbean countries. It covers all available survey years from 2005 to 2023 and includes repeated cross-sections at the individual level. The analytical sample is restricted to individuals aged 18 to 65 who are employed and have non-missing information on education, sector of activity, firm size, and pension contribution status.

The SEDLAC database is constructed by the World Bank and CEDLAS through harmonization of national household surveys; both the underlying microdata and the harmonized outputs are restricted-access and cannot be redistributed. The variables used in the analysis are extracted directly from the harmonized SEDLAC dataset and are fully specified in the corresponding replication do-files associated with each figure or table. Any additional variable transformations or constructed variables are implemented in the replication code and do not modify the underlying structure of the SEDLAC data.

This dataset cannot be included in the reproducibility package due to restrictions on both the SEDLAC harmonized microdata and the associated harmonization procedures. Access to SEDLAC is granted through the World Bank and CEDLAS under specific data use conditions. Researchers interested in using the data must request access through the SEDLAC platform and comply with the corresponding data agreements.

As an alternative, users may access the original national household survey microdata through the respective national statistical offices of Latin American and Caribbean countries, subject to each agency's data access policies. The analytical dataset used in this study is fully reproducible in principle from the underlying surveys using the harmonization framework, but the specific SEDLAC harmonized version employed here is not publicly distributable.

q. Data name: LAC_allyears_var_family_2023.dta

Source: Authors' harmonization of household-level survey data from Latin American national household surveys.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to the conditions of the original data providers.

Notes: Household-level analytical dataset. The dataset is based on SEDLAC harmonized household survey microdata for Latin American and Caribbean countries and uses all available survey years from 2005 to 2023, with the analytical sample restricted to the year 2023. The universe of analysis includes individuals aged 18 to 65 who are employed and have non-missing information on education, sector of activity, firm size, and pension contribution status.

The construction of the dataset involves additional transformations applied to the SEDLAC harmonized variables in order to define household composition and labor market status. Household structure is constructed using information on age and household relationship variables to identify household members' roles (e.g., head, spouse, children, and other members) and to classify household composition categories. Labor market status is further refined by combining employment status variables and pension contribution information to distinguish between self-employed, formal salaried, informal salaried, unpaid workers, inactive individuals, unemployed, and retired individuals. All variable definitions, recoding, and additional transformations are fully documented in the replication do-file associated with the corresponding tables and figures. No modifications are made to the underlying SEDLAC harmonized data beyond these reproducible coding steps. This dataset cannot be included in the reproducibility package due to restrictions on the SEDLAC harmonized microdata and the derived harmonization procedures. Access to SEDLAC is provided through the World Bank and CEDLAS under restricted-use conditions and subject to data use agreements.

The analytical dataset used in this study is fully reproducible from the underlying SEDLAC harmonized microdata by applying the coding procedures described in the replication do-file.

However, the constructed dataset itself is not redistributable. Researchers interested in replication must obtain access to SEDLAC and reconstruct the variables following the same sample restrictions and transformation rules described in the code.

r. Data name: LAC_lastyear_var.dta

Source: Authors' harmonization of the latest available Latin American household survey data.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to authorization from original statistical agencies.

Notes: Latest-year harmonized survey dataset. The dataset is based on SEDLAC harmonized household survey microdata for Latin American and Caribbean countries and uses the same variables and construction procedures as the full-panel version of the dataset. The difference is that this dataset is restricted to the most recent available year (2023). The analytical sample includes individuals aged 18 to 65 who are employed and have non-missing information on education, sector of activity, firm size, and pension contribution status.

The construction process follows the same coding and variable definitions as the full SEDLAC-based dataset. Household composition variables, labor market status categories, and all other constructed indicators are derived using the same procedures documented in the replication do-file. No additional transformations are introduced beyond restricting the sample to the 2023 cross-section. This dataset cannot be included in the reproducibility package due to restrictions on the SEDLAC harmonized microdata and derived variables. Access to SEDLAC is provided by the World Bank and CEDLAS under restricted-use agreements.

The dataset is fully reproducible from the underlying SEDLAC harmonized microdata by applying the same code used in the full-panel version and restricting the sample to the 2023 wave. However, the constructed dataset itself is not redistributable.

s. Data name: LAC_lastyear2_var.dta

Source: Authors' harmonization of the latest available Latin American household survey data.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to authorization from original statistical agencies.

Notes: Supplementary latest-year harmonized survey dataset. The dataset is based on SEDLAC harmonized household survey microdata for Latin American and Caribbean countries and corresponds to the most recent available year (2023). The dataset follows the same variable definitions, sample restrictions, and construction procedures as the baseline SEDLAC dataset, including individuals aged 18 to 65 who are employed and have non-missing information on education, sector of activity, firm size, and pension contribution status.

An updated version of the data is used for Honduras, as the original SEDLAC coverage for this

country was only available up to 2019. The dataset incorporates updated 2023 information for Honduras, ensuring consistency with the rest of the sample. All variable construction procedures, including household composition and labor market classification, follow the same replication code as the main dataset. No additional modifications are introduced beyond the country-specific data update and the standard coding procedures documented in the replication do-file.

This dataset cannot be included in the reproducibility package due to restrictions on the SEDLAC harmonized microdata and derived variables. Access to SEDLAC is provided through the World Bank and CEDLAS under restricted-use agreements.

The dataset is fully reproducible from the underlying SEDLAC harmonized microdata by applying the same code used in the main specification, with the exception of Honduras, which requires updated data for 2023. For Honduras, a proxy definition of formal employment is used due to missing information in the original SEDLAC harmonized dataset. This proxy is based on education level, employment type, and firm size, as follows: non-salaried workers without tertiary education; non-salaried workers with tertiary education employed in small private firms; salaried workers without education; and salaried workers with some education employed in small private firms.

However, in the final specification Honduras is excluded from the analysis. Results are therefore not affected when using the alternative dataset, provided that Honduras is consistently excluded. The constructed dataset itself is not redistributable.

t. Data name: MEX_ECV_v&iv.dta

Source: ENOE (Mexico) survey microdata harmonized by the authors.

Access date: April 2026

Access type: Restricted/harmonized microdata.

Access instructions: Access subject to original data provider restrictions.

Notes: Analytical dataset used in Mexico-specific analysis. The dataset is based on the Mexican National Survey of Occupation and Employment (ENOE) for the year 2024. The analysis uses the extended questionnaire and focuses on individuals aged 18 to 65 who are employed. The dataset combines information on job separation histories and employment transitions, with additional variables used to classify reasons for job loss and transitions across employment states.

The construction process relies on the ENOE questionnaire modules on labor market transitions and job separation (in particular questions P9, P9A, P9B, P9C, and P9D), which identify whether individuals have ever lost a job or business and the reasons for such separation. These responses are used to construct a classification of voluntary and involuntary job exits. Involuntary separations include events such as layoffs, firm closures, contract non-renewal, lack of work, illness, accidents, or other exogenous shocks. Voluntary separations include resignations driven by preferences over wages, working conditions, career changes, education, or family-related decisions. Individuals with missing information on all follow-up separation questions are not classified.

In addition, labor market transitions between employment states (formal salaried, informal salaried, and self-employment) are constructed using ENOE information on access to medical care in the previous job (P9K), which is used to proxy formal employment status. Current employment status is aligned with SEDLAC-based classification using information on labor status and social security contributions. The final dataset distinguishes transitions between formal salaried employment, informal salaried employment, and self-employment. This dataset cannot be included in the reproducibility package due to restrictions on redistribution of the processed ENOE-based analytical dataset and derived classification variables. However, the underlying ENOE microdata are publicly available through INEGI.

Researchers can access the original survey data and replicate the construction by using the ENOE 2024 extended questionnaire microdata and applying the same classification rules for voluntary and involuntary job separation using variables P9, P9A, P9B, P9C, P9D, and P9K. The construction of labor market transitions requires combining these variables with standard employment status definitions.

The analytical dataset is derived entirely from publicly available microdata, but the processed version used in the analysis is not redistributable.

URL: <https://www.inegi.org.mx/programas/enoe/15ymas/>

u. Data name: OECD_self_employment.xlsx

Source: OECD Statistics Database.

Access date: May 2026

Access type: Public data.

Access instructions: Available through OECD statistics portal.

Notes: OECD self-employment indicators used in cross-country comparisons.

URL: <https://www.oecd.org/en/data/indicators/self-employment-rate.html>

v. Data name: Panels_LAC_2version.dta

Source: Authors' construction based on harmonized Latin American household survey panel data.

Access date: April 2026

Access type: Restricted/constructed microdata.

Access instructions: Derived from restricted household surveys.

Notes: Panel dataset used for transition analysis. The longitudinal analysis is based on labor market panels harmonized using the do-files published by Donovan et al. in their project Labor Market Dynamics and Development. The final datasets are constructed from raw microdata obtained from official national statistical agencies and correspond to nationally representative labor force and household surveys.

The original data sources vary by country and cover the following periods:

Argentina (EPH): 2003–2023
Bolivia (ECE): 2015–2022
Brazil (PNADC: 2012–2023; PME: 2003–2012)
Chile (ENE): 2010–2024
Costa Rica (ECE): 2010–2024
Dominican Republic (ENCFT): 2016–2024
Ecuador (ENEMDU): 2010–2019
Mexico (ENOE): 2005–2023
Peru (EPE): 2003–2019
Paraguay (EPHC): 2010–2017

The do-files follow a standardized structure and generate two main outputs: a cross-sectional dataset and a wide-format panel dataset. These datasets include a common set of harmonized labor market variables such as country, year, quarter, age, education, gender, employment. The longitudinal analysis is based on labor market panels harmonized using the do-files published by Donovan et al. in their project Labor Market Dynamics and Development. The final datasets are constructed from raw microdata obtained from official national statistical agencies and correspond to nationally representative labor force and household surveys.

The original data sources vary by country and cover the following periods:

Argentina (EPH): 2003–2023
Bolivia (ECE): 2015–2022
Brazil (PNADC: 2012–2023; PME: 2003–2012)
Chile (ENE): 2010–2024
Costa Rica (ECE): 2010–2024
Dominican Republic (ENCFT): 2016–2024
Ecuador (ENEMDU): 2010–2019
Mexico (ENOE): 2005–2023
Peru (EPE): 2003–2019
Paraguay (EPHC): 2010–2017

The do-files follow a standardized structure and generate two main outputs: a cross-sectional dataset and a wide-format panel dataset. These datasets include a common set of harmonized labor market variables such as country, year, quarter, age, education, gender, employment status, hours worked, earnings, sector, occupation (ISCO codes), informality indicators, and related labor market characteristics.

Key variables include, among others: country, year, qtr, age, edu, sex, emp, hours, earnings, sector, occisco, formal, search_length, employer, weight, as well as additional labor, household, and job-related variables.

The data construction process consisted of (i) downloading raw microdata directly from each country's official statistical agency, (ii) adapting and running the original Donovan et al. do-files to incorporate the most recent available survey waves, and (iii) extending the code to construct informality indicators following closely the definitions used in SEDLAC, in order to ensure consistency and cross-country comparability.

This dataset cannot be directly included in the reproducibility package. However, the full construction process is reproducible. Users can independently access the original microdata through the official websites of each country's statistical agency, download

the relevant survey waves, and execute the adapted version of the Donovan et al. do-files used in this project.

The project is built upon publicly available replication do-files provided by Donovan et al., which fully document the structure of the harmonized datasets and allow for reconstruction of the panels from the raw data. Our contribution consists of updating the sample period to include the most recent available survey waves and extending the code to construct additional variables, including harmonized informality indicator

w. Data name: PNAD_person.dta

Source: Pesquisa Nacional por Amostra de Domicílios (PNAD), Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil.

Access date: January 2026

Access type: Restricted/harmonized microdata.

Access instructions: Original survey microdata available through IBGE. Harmonized analytical version used in this project is not publicly redistributed.

Notes: Individual-level household survey microdata. The dataset is based on the 1990 Brazilian National Household Survey (PNAD), using microdata from the individual (person-level) file. The original data were obtained from the Princeton University Brazil Data Center archive in January 2026. At the time of access, the raw survey files and accompanying documentation (README) were available through the repository; however, the dataset is no longer publicly accessible through that link.

The original PNAD release included two separate files: a household file and a person-level file. Only the person-level dataset was used in this study.

The raw data were provided in .dat format and processed according to the instructions in the original README file. The preprocessing involved cleaning and structuring the dataset by mapping variables directly from the fixed-width format using the positions specified in the documentation.

The main variables used in the analysis were constructed based on the following README-defined fields:

le_gustaria_empleo_formal (columns 285–285) — important variable

razones_no_le_gustaria (columns 286–286) — important variable

condicion_laboral (columns 280–281) — important variable

The survey sampling weight used corresponds to the 1991 PNAD weight provided in the original dataset.

All variable construction follows directly from the original codebook/README, with no additional

modifications beyond standard data cleaning and formatting required to convert the fixed-width .dat file into an analyzable dataset. This dataset cannot be included in the reproducibility package, and neither the raw microdata nor the harmonized version can be shared due to confidentiality and data access restrictions.

The original microdata were obtained from the Princeton University Brazil Data Center archive in January 2026. At that time, the raw PNAD 1990 files and accompanying documentation were accessible through that repository. Although the archive is no longer publicly available, the underlying PNAD microdata may still be obtainable through alternative official or archival channels upon request.

However, the harmonized dataset used in this project is not distributable. Replication would therefore require independent access to the original PNAD microdata and full reconstruction of the dataset following the processing steps described in the documentation.

All data processing procedures are fully documented and rely on the original README file provided with the PNAD 1990 release. This includes the mapping of variables from the fixed-width .dat file using the specified column positions and standard cleaning procedures. Once access to the original microdata is obtained, the dataset can be reconstructed by following these documented steps.

URL: <https://opr.princeton.edu/brazils-national-household-sample-survey>. The original URL previously used to access PNAD harmonized files is no longer active. Data were obtained from the Princeton University Brazil Data Center archive in January 2026.

x. Data name: Processed_data_final_created_weights.dta

Source: Beylis, G., A. Garin, N. Gonzalez-Prieto, and D. Koustas. Forthcoming. “Valuing Alternative Work Arrangements: Evidence from 7 Countries with High Rates of Informality.” Policy Research Working Paper, World Bank.

Access date: April 2026

Access type: Restricted data.

Access instructions: Access subject to authorization from the original authors.

Notes: Processed analytical dataset used for willingness-to-pay analysis. Processed analytical dataset used for willingness-to-pay analysis. The dataset used in this study is a processed analytical dataset constructed for the willingness-to-pay (WTP) analysis. It is based on survey data from a discrete choice experiment (DCE) implemented across seven countries as part of the study by Beylis, Garin, González-Prieto, and Koustas (2026), “Valuing Alternative Work Arrangements: Evidence from 7 Countries with High Rates of Informality”, World Bank Policy Research Working Paper.

The dataset covers the following countries: Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, and Peru. Both the raw survey data and the harmonized dataset were originally constructed and provided by the authors of the study.

For the purposes of this project, the analytical dataset was used as provided, with no further harmonization of variables beyond standard preparation required for estimation. Any additional processing was limited to generating analysis-specific variables required for the willingness-to-

pay estimation framework and ensuring consistency in sample selection across countries. No modifications were made to the underlying survey structure or core harmonized variables developed by the original authors.

This dataset cannot be included in the reproducibility package, and access to both the raw and harmonized data is subject to authorization by the original authors of the study.

The data were provided directly by the authors of Beylis et al. (2026) under the conditions specified in their data use agreement. As such, redistribution is not permitted.

Researchers interested in accessing the dataset for replication purposes should contact the authors of the study directly. The corresponding contact information can be found through their institutional affiliations listed in the publication or on their official professional webpages.

y. Data name: pwt1001.dta

Source: Penn World Table version 10.01.

Access date: May 2026

Access type: Public data.

Access instructions: Available through the Penn World Table database.

Notes: Macroeconomic indicators used in cross-country analysis.

URL: <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt1001?lang=en>

z. Data name: SE_type_rates_temp.dta

Source: Authors' calculations based on harmonized household survey microdata.

Access date: May 2026

Access type: Constructed analytical dataset.

Access instructions: Constructed by the authors from restricted-access household surveys.

Notes: Intermediate dataset used to generate self-employment rate figures.

ab. Data name: SE_type_rates_temp_last_year.dta

Source: Authors' calculations based on harmonized household survey microdata.

Access date: May 2026

Access type: Constructed analytical dataset.

Access instructions: Constructed by the authors from restricted-access household surveys.

Notes: Intermediate dataset using the latest available year.

ac. Data name: Self_Employment_GDP_WDI_1999_2025.xlsx

Source: World Development Indicators (WDI), World Bank, combined with authors' calculations.

Access date: April 2026

Access type: Public/constructed data.

Access instructions: WDI indicators available through the World Bank Data Catalog.

Notes: Dataset combining GDP and self-employment indicators. The dataset does not include original statistical estimations by the authors. All variables are constructed using publicly available data sources, with no independent estimation of underlying labor market indicators.

The only element that may be interpreted as an “author calculation” in the README refers to the construction of derived indicators based on pre-existing published series. In particular, the self-employment variable is based on the World Bank indicator “Self-employed, total (% of total employment) (modeled ILO estimate)”, which is produced by the International Labour Organization (ILO) through model-based estimation procedures.

URL: <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD>

ad. Data name: Transitions_BWUN_JJ.dta

Source: Donovan, K., W. J. Lu, and T. Schoellman. 2023. “Labor Market Dynamics and Development.” *Quarterly Journal of Economics* 138 (4): 2287–325.

Access date: March 2026

Access type: Public.

Access instructions: The database is available for download at the URL provided below.

Notes: Dataset used for labor market transition analysis.

URL: <https://doi.org/10.7910/DVN/RXWKTV>

ae. Data name: GEIH_2023.dta

Source: Gran Encuesta Integrada de Hogares (GEIH), Departamento Administrativo Nacional de Estadística (DANE), Colombia.

Access date: January 2026

Access type: Public/harmonized microdata.

Access instructions: Original survey available through DANE. Harmonized analytical version used in this project is not publicly redistributed.

Notes: Household survey microdata used for labor market analysis. The dataset is constructed using microdata from the Colombian Household Survey (GEIH) for 2023, accessed through SEDLAC harmonized outputs. The construction process follows the SEDLAC harmonization framework; however, the specific harmonization procedures and derived dataset used in this project cannot be shared as part of the reproducibility package. The analytical sample is restricted to employed individuals aged 18 to 65 with non-missing educational attainment. The main variables are derived from the original GEIH microdata, including variables P1805 and P1879, which are used to identify self-employed workers' preferences for formal employment and the reasons for self-employment, respectively. Additional variables include age and educational attainment. The final dataset is constructed by filtering the sample according to the

inclusion criteria and combining the selected variables at the individual level for 2023. The harmonized SEDLAC-based dataset cannot be included in the reproducibility package due to restrictions on the harmonized outputs and the associated harmonization procedures. However, the underlying microdata come from the Colombian Household Survey (GEIH), which are publicly available through DANE (Departamento Administrativo Nacional de Estadística). Researchers can access the GEIH microdata directly from the DANE website and reconstruct the variables used in this analysis using the original survey variables (including P1805 and P1879), together with age and educational attainment variables, applying the same sample restrictions (individuals aged 18–65 who are employed and report non-missing education). The SEDLAC harmonized version is not redistributable, but it is based entirely on these publicly available GEIH microdata.

URL: <https://microdatos.dane.gov.co/index.php/catalog/782>

af. Data name: EMICRON_2023.dta

Source: Encuesta de Micronegocios (EMICRON), Departamento Administrativo Nacional de Estadística (DANE), Colombia.

Access date: March 2026

Access type: Public.

Access instructions: The database is available for download at the URL provided below.

URL: <https://microdatos.dane.gov.co/index.php/catalog/832/get-microdata>

Note: Source: Encuesta de Micronegocios (EMICRON), Departamento Administrativo Nacional de Estadística (DANE), Colombia. Access date: March 2026.

The dataset is constructed using microdata from the 2023 Colombian Microbusiness Survey (EMICRON), accessible through the official website of the National Administrative Department of Statistics (DANE). The construction process begins with the inclusion of the different survey modules; however, the specific harmonization procedures and the derived dataset used in this project cannot be shared as part of the reproducibility package.

The analytical sample is restricted to employed individuals. The main variables are derived from the original EMICRON microdata, including variables P1765 and P1567, which are used to identify whether entrepreneurs have access to financial credit and, for those who do not, the reasons why. Additional variables include whether the microbusiness is formal or informal. The final dataset is constructed by filtering the sample according to the inclusion criteria and combining the selected variables at the individual level for 2023. The harmonized dataset cannot be included in the reproducibility package due to restrictions on harmonized outcomes and associated harmonization procedures. However, the underlying microdata comes from the Microbusiness Survey (EMICRON), which is publicly available through DANE. Researchers can access the EMICRON microdata directly from the DANE website and reconstruct the variables used in this analysis using the original survey variables (including P1765 and P1567), along with the formality variable for disaggregation. The harmonized version is not redistributable but is based entirely on this publicly available EMICRON microdata

Notes:

- The datasets stored in the */Data* folder are already cleaned and harmonized. These files are ready to be used for replication and analysis, and no additional data processing is required prior to running the code.
- Some figures and tables rely on intermediate datasets constructed within individual do-files. Each script loads and processes the required data independently to ensure modular reproducibility.

5. Software Requirements

☐ **Stata:** Version MP 18.5 for Windows (64-bit x86-64) or more advanced versions.

☐ **Required packages:**

The main user-written commands are automatically installed within the master do-files. Therefore, users do not need to manually install these packages before running the code.

Appendix

1. Mapping: Figures and Tables

- Introduction

Output	Do-file
Fig I.1	Intro/FigI.1 Informality rate by Region.do
Fig I.2	Intro/FigI.2 The share of informal workers micro entrepreneurship.do
Table I.1	Intro/TableI.1 Labor characteristics by labor status.do
Fig I.3	Intro/FigI.3 Life-cycle patterns for different kinds of work vary.do
Fig I.4	Intro/FigI.4 Contribution SS by firm size.do
Fig I.5	Intro/FigI.5 Rise of non contributory social protection systems.do

- Chapter 1

Output	Do-file
Fig 1.1.A	CH1/Fig 1.1.A Self-employment across the development.do
Fig 1.1.B	CH1/Fig 1.1.B Self-employment across the development.do
Figure 1.2	CH1/ Fig1.2 Self-employment rate by GDP per capita.do"
Table 1.1.A	CH1/ Table1.1.A Motivations for Self-Employment.do
Table 1.1.B	CH1/Table1.1.B Motivations for Self-Employment.do
Fig 1.3.A	CH1/Fig1.3.A Preference for SE among salaried workers by level of development.do
Fig 1.3.B	CH1/Fig1.3.B Prevalence of SE.do
Fig 1.4	CH1/ Fig1.4 Share of informal SE searching for a formal job.do
Table 1.2	CH1/ Table1.2 Comparison of respondents motivations for starting a microenterprise.do
Fig 1.5	CH1/ Fig1.5 Share of opportunity microentrepreneurs.do

Output	Do-file
Fig 1.6	CH1/ Fig1.6 Share of necessity unskilled microentrepreneurs.do
Fig 1.7	CH1/ Fig1.7 Work-family interference perceptions by level of development.do
Fig 1.9	CH1/ Fig1.9 Labor rigidity index by GDP per capita.do
Fig 1.10	CH1/Fig1.10 SE and MW-P50 labor income.do
Fig 1.11	CH1/Fig1.11 Compensation Required for Formalization.do
Fig 1.12.A	CH1/Fig1.12.A WTP for schedule flexibility for men and women.do
Fig 1.12.B	CH1/ Fig1.12.B WTP for autonomy in performing tasks for men and women.do
Fig 1.13.A	CH1/ Fig1.13.A WTP schedule by country.do
Fig 1.13.B	CH1/ Fig1.13.B WTP independence by country.do
Fig 1.14	CH1/ Fig1.14 SE-FS hourly wage gap by gender and education.do

- **Chapter 2**

Output	Do-file
Fig 2.1.B	CH2/ Fig2.1.B Entry and exit Rates in Colombia.do
Fig 2.2	CH2/ Fig2.2 Transition FS-SE and RealGDP.do
Fig 2.3	CH2/ Fig2.3 Relative work finding rates U-S.do
Fig 2.4	CH2/Fig2.4 Relative work finding rates.do
Fig 2.5.A	CH2/ Fig2.5.A Preference for SE among self-employed by age cohort.do
Fig 2.5.B	CH2/ Fig2.5.B Preference for SE among employees by age cohort.do
Fig 2.6.A	CH2/ Fig2.6.A Self-Employed Who Say They Would Prefer to Be Formal Salaried.do
Fig 2.6.B	CH2/Fig2.6.B Fig2.6.B Self-Employed Who Say They Would Prefer to Be Formal Salaried.do

Output	Do-file
Fig 2.6.C	CH2/ Fig2.6.C Self-Employed Who Say They Would Prefer to Be Formal Salaried.do
Fig 2.7	CH2/Fig2.7 WTP ratio by age group.do
Fig 2.11	CH2/ Fig2.11 Probability of Federal Registration.do
Table 2.2	CH2/ Table2.2. Microfirms no credit US, MX, COL.do

- **Chapter 3**

Output	Do-file
Fig 3.1	CH3/ Fig3.1 IS vs FS characteristics.do
Fig 3.2	CH3/ Fig3.2 Correlation between IS and workers in micro firms.do
Fig 3.3	CH3/ Fig3.3 Transitions voluntary and involuntary Mexico.do
Fig 3.4	CH3/ Fig3.4 Share of workers who say they are not satisfied with their job and searching.do
Table 3.1	CH3/ Table3.1 IS workers in Brazil not wanting to be formal.do
Table 3.2	CH3/ Table3.2 Characterization by laboral status.do
Fig 3.5	CH3/ Fig3.5 Benefits in contracts of FS and IS workers.do
Fig 3.6	CH3/ Fig3.6 Implicit costs.do
Fig 3.7	CH3/Fig3.7 WTP for formal job by status laboral.do
Fig 3.8	CH3/ Fig3.8 Cost of formality according to workers.do
Fig 3.9	CH3/ Fig3.9 Associative matching between the occupation of the head of household and the spouse.do
Table 3.3	CH3/ Table3.3 Informal workers with various types of health coverage or benefits.do
Fig 3.10	CH3/ Fig3.10 Formality marriage kids general.do
Fig 3.11	CH3/ Fig3.11 IS Vs MWP50.do

Output	Do-file
Fig 3.12	CH3/ Fig3.12 IS-FS hourly wage gap by gender and education.do
Fig 3.13	CH3/ Fig3.13 Employment status laboral by age.do
Fig 3.14	CH3/ Fig3.14 Tenure by group age and status laboral.do
Fig 3.15	CH3/ Fig3.15 Transition_ItoF_ItoSE_general_stock.do
Fig 3.16	CH3/ Fig3.16 Transition_ItoF_ItoSE_education.do
Table 3.4	CH3/ Table3.4 IS prime-age workers by firm size.do
Fig 3.17	CH3/ Fig3.17 IS ratio (preprime-prime) Vs MWP50.do
Fig 3.18	CH3/ Fig3.18 Unpaid workers living with a SE.do
Fig 3.19	CH3/ Fig3.19 Unpaid workers by relationship.do
Fig 3.20	CH3/ Fig3.20 Correlation between unpaid workers and dependent workers in micro firms.do
Fig 3.21	CH3/ Fig3.21 IS for firms with 50 plus workers by country.do