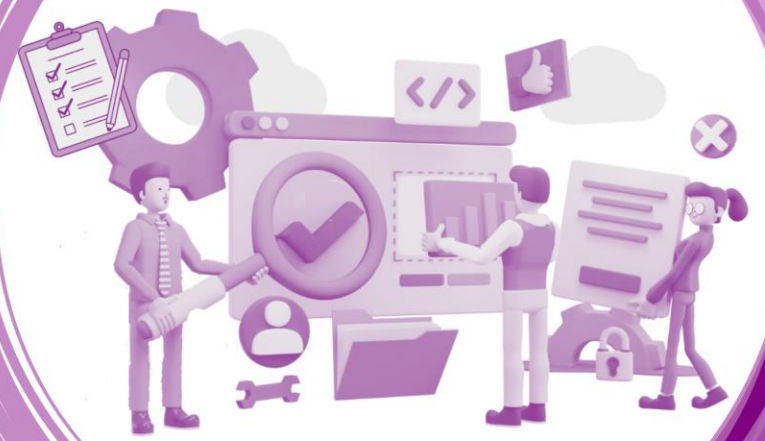
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
Academy • Development Data

Reproducible Research Repository

World Bank Reproducibility Initiative

How to Create a Successful Reproducibility Package



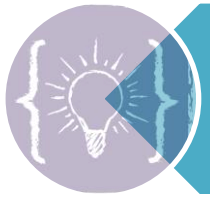
20 June 2025

Reproducible Research Initiative

Reproducibility packages encouraged for all World Bank publications

Verified and published to the Bank's Reproducible Research Repository

Reproducible publications marked with a 'Reproducible Research' seal



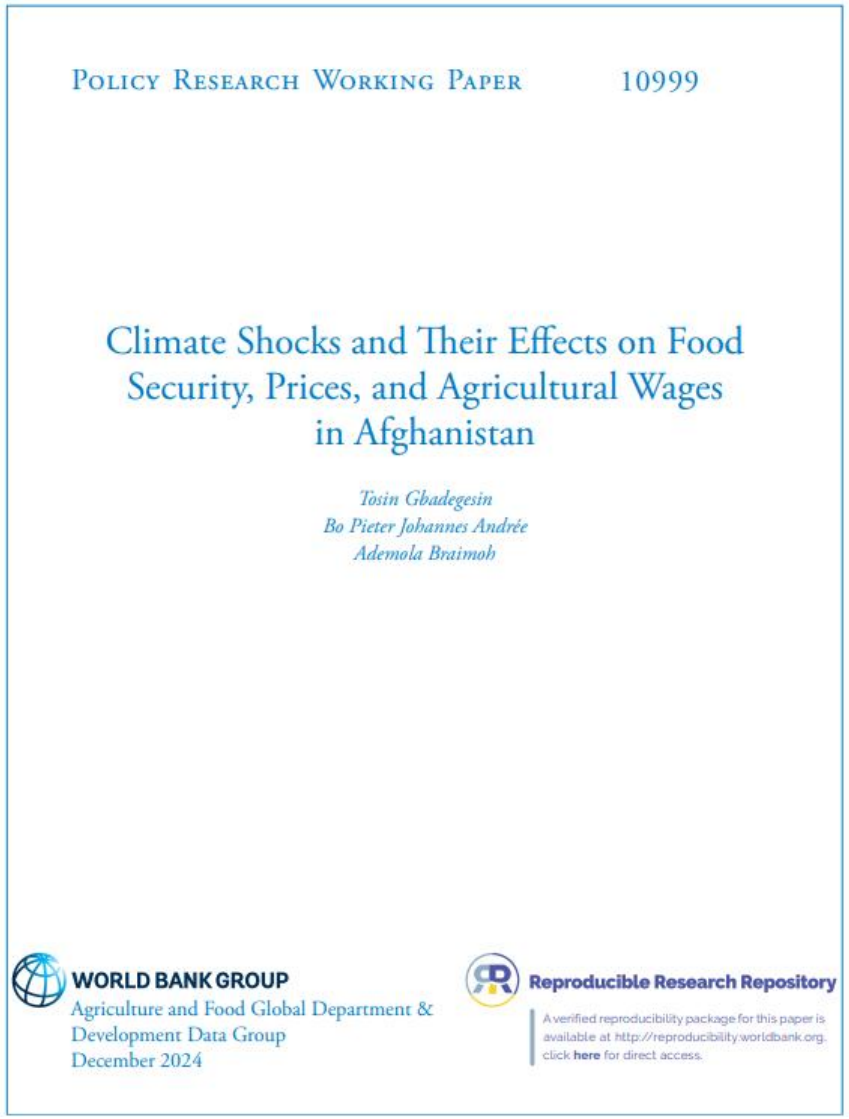
DIME Analytics conducts verifications, builds tools, and leads trainings

Reproducible Research Workflow

Public packages include **code and documentation**. Data included *only if* redistribution is permitted. Options for temporary embargos on data. All packages include a detailed data availability statement, specifying data sources and whether/how raw data files can be accessed by a third party.

http://reproducibility.worldbank.org

Reproducible Research Repository



Reproducibility package for Climate Shocks and Their Effects on Food Security, Prices, and Agricultural Wages in Afghanistan

2025 GET REPRODUCIBILITY PACKAGE

Reference ID	RR_AFG_2024_241
DOI	https://doi.org/10.60572/frgy-9910
Author(s)	Tosin Gbadegesin, Bo Pieter Johannes Andrée
Collections	World Bank Policy Research Working Papers
Metadata	JSON

CITATION

Gbadegesin, T., & Andrée, B. P. J. (2025). *Reproducibility package for Climate Shocks and Their Effects on Food Security, Prices, and Agricultural Wages in Afghanistan*. World Bank. https://doi.org/10.60572/FRGY-9910

Export citation: RIS | BibTeX | Plain text

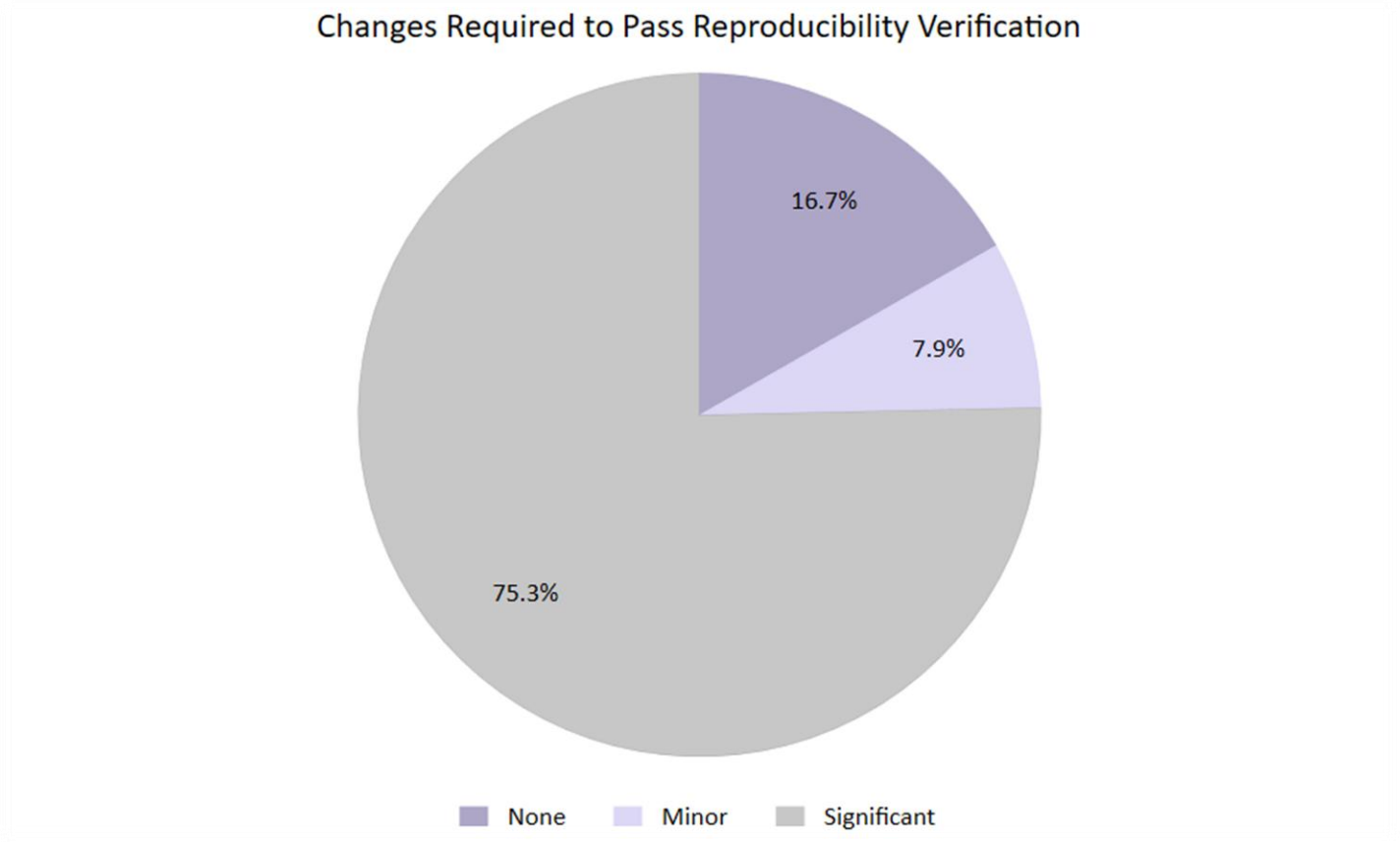
Reproducibility Verification

Reproducibility package is complete and functional

Reproducibility team can reproduce the exact findings in the publication **using the data, code, and documentation provided by the author**

Not a review of quality of code, data, or methods

Getting the same results from the same data and code … that should be easy, right?



1 in 6 packages are reproducible as submitted

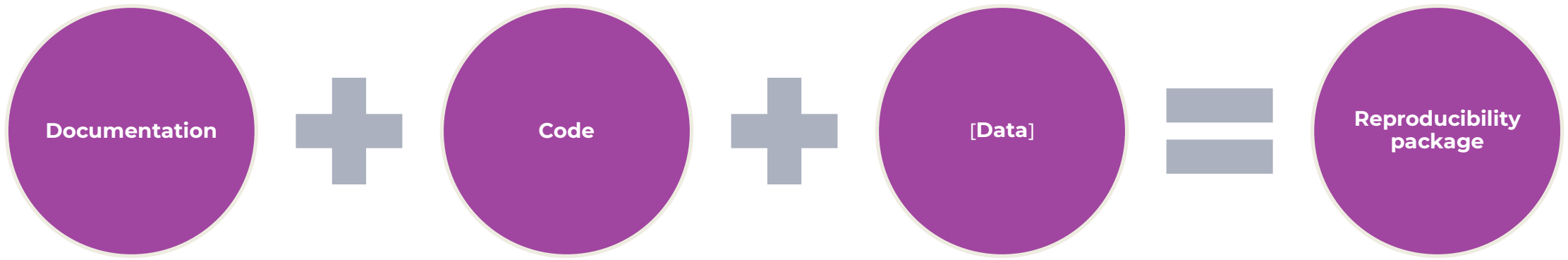
Roadmap

How to build a reproducibility package

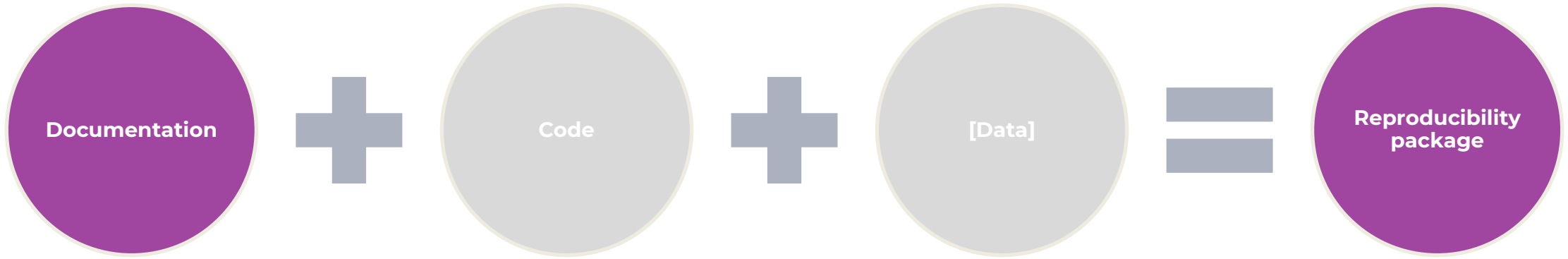
Common reasons reproducibility packages fail

Special considerations for Complex Reports

What is a reproducibility package?



Reproducibility Package Components



Documentation: README

Objective

- Provide enough detail for a 3rd party to understand how to interact with the package to reproduce the results in the paper (and any limitations to doing so)

Components

- Instructions and explanation of package structure (e.g. files organized by chapter)
- Data availability statement (lists **every** source of data used in the paper/Flagship, links when possible)
- List of exhibits mapping to sources

Tools

- Template README for the Reproducible Research Repository includes minimum requirements

Example README: overview (Stata)

Contents of the Package

The replication package includes the following key components:

- **Master Files:** Located in the `master files` folder.
  - `f1_user_profile.do`: Allows the user to set the working directory. This step is crucial for running the other master files.
  - `f2_inputs.do`: Runs scripts to assemble the EQOSOGI 2024 dataset, integrating external indicators from Worldwide Governance Indicators, World Development Indicators, and Women Business and the Law.
  - `f3_outputs.do`: Generates the report's tables and figures in a clear, organized manner.
- **Scripts:** Located in the `scripts` folder.
  - `EQOSOGI_database_ensambling.do`
  - `EQOSOGI_external_indicators.do`
- **Inputs:** All inputs required to run the master files and scripts are included in this package.
  - Coding sheets for the 64 EQOSOGI 2024 countries.
  - Intermediate databases:
    - `EQOSOGI 2_List of Countries_FY24 Classification.xlsx`
    - `country_risk_ratings_Allianz_2023.xlsx`
    - `ssgd_v_2_0.dta`
    - `WBL1971-2023.dta`
- **Outputs:** Once the `f2_inputs.do` file is run, the EQOSOGI 2024 dataset will be generated and stored in the `outputs` folder.
- **Reproducibility in Excel:** Contains all elements of the EQOSOGI 2024 report that can be replicated using Excel. This folder includes the following files:
  - `EQOSOGI_ch1_outputs.xlsx`
  - `EQOSOGI_ch2_outputs.xlsx`
  - `EQOSOGI_ch3_outputs.xlsx`
  - `EQOSOGI_online_appendix_outputs.xlsx`

Replication Instructions

To replicate the analysis, follow these steps:

1. Open the `f1_user_profile.do` file and set the working directory to the path where the replication package is extracted.
2. Run the `f2_inputs.do` file to process the data and assemble the EQOSOGI 2024 dataset.
3. Execute the `f3_outputs.do` file to produce the report's tables and figures. The outputs are organized clearly within this script.

Important: The master files must be run in the specified order: `f1_user_profile.do` first, followed by `f2_inputs.do`, and finally `f3_outputs.do`. After completing step 3, examine the contents of the `Reproducibility in Excel` folder to review the elements of the report that have been replicated.

https://reproducibility.worldbank.org/index.php/catalog/182

Example README: overview (R)

Reproduction Instructions

1. Get access to the data not included in the reproducibility package. See the tab le below for details.
2. Add the global path in the dofiles:
    a. Introduction/Figurel.1.do
    b. Chapter 2/Replication Chapter 2/Macros.do
    c. Spotlight 2/main.do
3. Add the main path in the R scripts:
    a. Chapter 2/Replication Chapter 2/Code/Graph Creation/ Bubble Graph.R
    b. Chapter 2/Replication Chapter 2/Code/Graph Creation/Mics_visualization.R
    c. Chapter 2/Replication Chapter 2/Code/Graph Creation/ Old_new_transition_graph.R
4. Run the do-files and R scripts in this order:
    a. Introduction/Figurel.1.do
    b. Chapter 2/Replication Chapter 2/Master.do
    c. Chapter 2/Replication Chapter 2/Code/Graph Creation/ Bubble Graph.R
    d. Chapter 2/Replication Chapter 2/Code/Graph Creation/Mics_visualization.R
    e. Chapter 2/Replication Chapter 2/Code/Graph Creation/ Old_new_transition_graph.R
    f. Spotlight 2/main.do
5. Map 2.1a, 2.1b, and 2.1c should be reproduced in QGIS with code results.

https://reproducibility.worldbank.org/index.php/catalog/215

Example README: List of Exhibits

Exhibit name	Output filename	Script	Note
Table 1	balancetable.csv	02_analysis.do (line 23)	Found in Outputs/tables/main
Figure 1	regresults.png	02_analysis.do (line 40)	Found in Outputs/figures/annex
Table A1	Appendix.xlsx; Sheet “Table A1”	03_analysis_appendix.do (line 32)	This file was created in Excel using calculations based on data exported from the Stata code. **Please note** that deleting this file will remove the embedded formulas, making it impossible to replicate the results. The file is automatically updated each time the code is run.

If not all data is publicly accessible, then the list of tables should additionally indicate which exhibits can be reproduced with the public material.

Documentation: Data Availability Statement

- DAS describes **every dataset** used for the analysis, regardless of size / scope / accessibility
  - **All** data should have a documented source

- Data generated by the team (or other Bank staff) must be in a permanent archive (not OneDrive) and be documented (e.g. codebook, questionnaire, data generation details)

Documentation: Data Availability Statement

If using data generated by others:

- **Provide data citation / DOI, or:**
  - **Filename**: Exact file name shown on the source website
  - **Source ***: Name of the source / link to website
  - **URL**: Exact downloadable URL of the data used, if available
  - **Access date:** Month/year when authors accessed the data.
- **If known, include:**
  - **License**: who can access the data? Can it be republished?
- **If you extracted a subset of the data (especially if using 'point-and-click'):**
  - **Details (e.g. time period, units or variable names extracted → WDI)**

If using data generated by the team:

- **Document the data**
  - in Microdata Library or Development Data Hub (depending on data type)
- **Choose appropriate terms of use**
  - internal / external, fully open, by request only, etc
  - **Provide data citation or DOI**

*If the dataset is restricted—for example, if it was purchased or provided by a government agency—please include details on how it was obtained. This should include the contact person or institution, a URL if available, and any steps required to request access.

Data Availability and Statement about rights

Make sure to list all the datasets used and categorize them as follows:

- [ ] All data are publicly available.
- [ ] Some data cannot be made publicly available.
- [ ] No data can be made publicly available.

Statement about Rights

- [ ] I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- [ ] I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

Example: Data Availability Statement

Data availability

All data used in the study are publicly available.

Data sources

There are three sources of data:

1. **World Values Survey Wave 7, version 5.**
    a. **Use:** This was used for Figures 5-8 and Summary Figure C.
    b. **Citation, as instructed online**:
        Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi:10.14281/18241.24
    c. **Filenames:** The paper used "WVS_Cross-National_Wave_7_Stata_v5_0.dta" which was the latest available at the time. For the purposes of replication, a smaller dataset including only the necessary variables from version 5 is provided in the replication folder, as well as the code which generated that dataset. The smaller dataset is called "WVS_key_variables.dta".
    d. **Source:** The WVS data can be accessed at the World Values Survey website. Data, instructions for access, and questionnaires are publicly available here:
    e. **URL:** https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp
    f. **Access year:** Version 5 datafile downloaded in December, 2022.

https://reproducibility.worldbank.org/index.php/catalog/196

Note on README for flagships

Flagships normally have more complex structures and more data than working papers. The README **must** reflect that. Pay special attention to:

- ✅ Clearly define **all source datasets** (access date, subset, license)
- 📦 Archive and document the **exact version** (ensure team consistency)
- 📊 Track figures from other sources (**cite and document**)

Best practices example here

Data

Each data folder in each chapter consists of a "raw" and a "processed" folder. There are two versions of this replication package. The "raw" data folders can be accessed only World Bank internal users, while the data stored in "processed" has been cleaned and aggregated so that it is accessible to external users. All figures can be replicated with the "processed" data only.

In the following, all datasets and sources are listed. Data files in processed folders that are created from files in the raw data folders are not listed separately.

1. Summary

a. raw

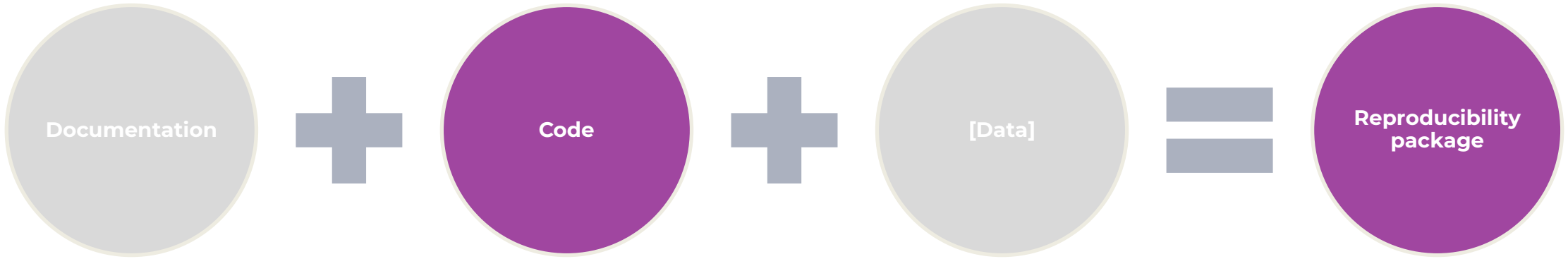
- \classifications\CLASS.dta
  - Source: https://github.com/GPID-WB/Class/tree/master/OutputData
  - Access year: 2024
  - Availability: All data are publicly available
  - Description: Contains country classifications
  - Files created in processed folder: countrycodes_yearly.dta, countrycodes_curent.dta
  - Note: No modifications have been made. Whole dataset has been downloaded.
- \emissions\edgar\CSC-GHG_emissions-2024.03.29b.xlsx
  - Source: https://scorecard.worldbank.org/en/scorecard/our-vision#planet
  - Detailed documentation and instructions on how the dataset was created by the World Bank here: https://scorecard.worldbank.org/content/dam/sites/scorecard/doc/methodology/VISION-GHG-Emissions-Clean.pdf under "data source(s)".
  - Access year: 2024
  - Availability: Available in this package
  - Description: Greenhouse gas emission data, compiled from EDGAR and Grassi et al. (2023)

Documentation: Resources

- **Template README**
  - Word format
  - Markdown format
  - README generator (beta version)

- **Step-by-step guidance for flagships**
- **Excel guidelines**

Reproducibility Package Components



Code Files: Guidance

Starting point: Original data

- **All code files necessary to go *from original data sources to results in the final manuscript***
- Best practice: modular scripts, separated by tasks (e.g., one script for data cleaning, one for analysis)

One ‘main’ script that runs all others

- Enables ‘one-click reproducibility’
- If no main script, the README must include detailed instructions for running code, including order in which code files must be run

Dependencies

- Include all the packages not in base software (e.g., .ado files, r packages, python libraries)

Code Files in Flagships

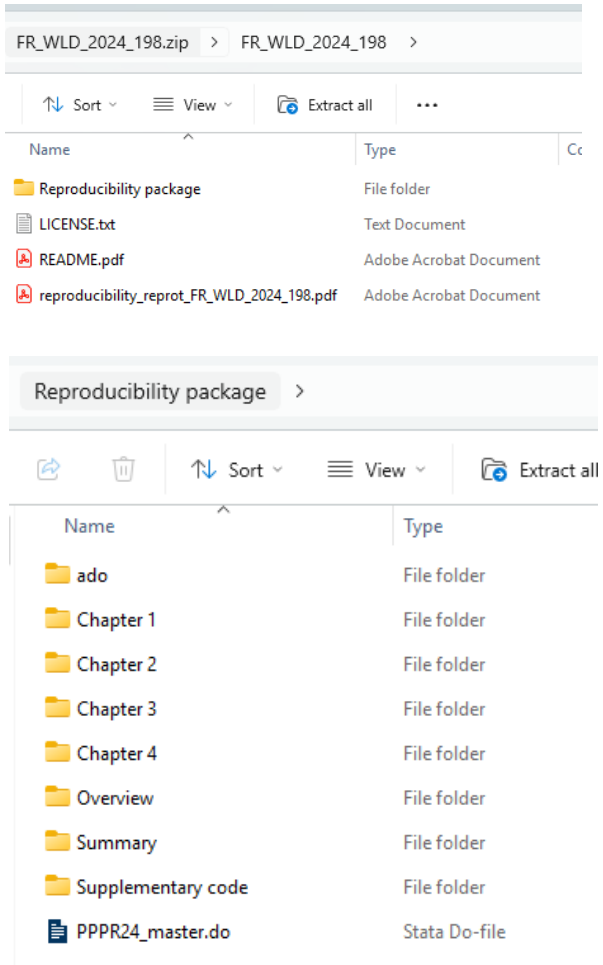
Why it matters:

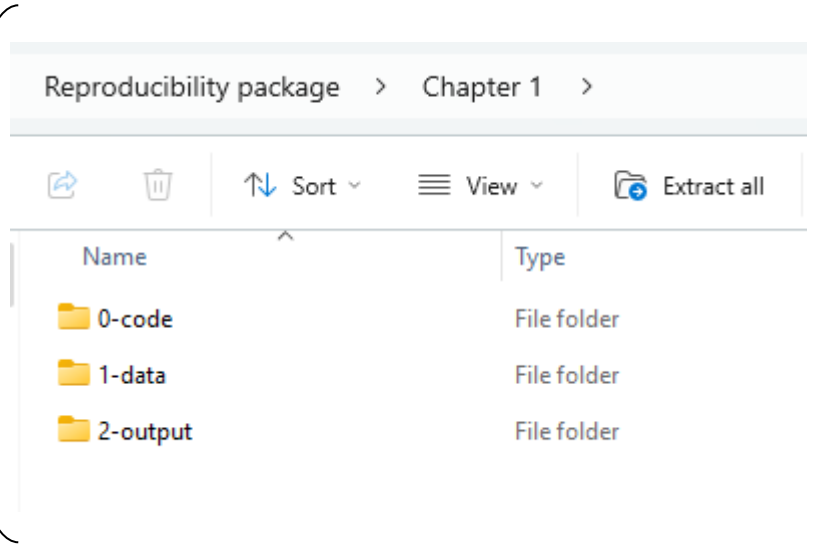
- Flagships often involve multiple teams and chapters → planning a shared file structure from the beginning is essential
- A clear folder structure ensures smooth collaboration and reproducibility

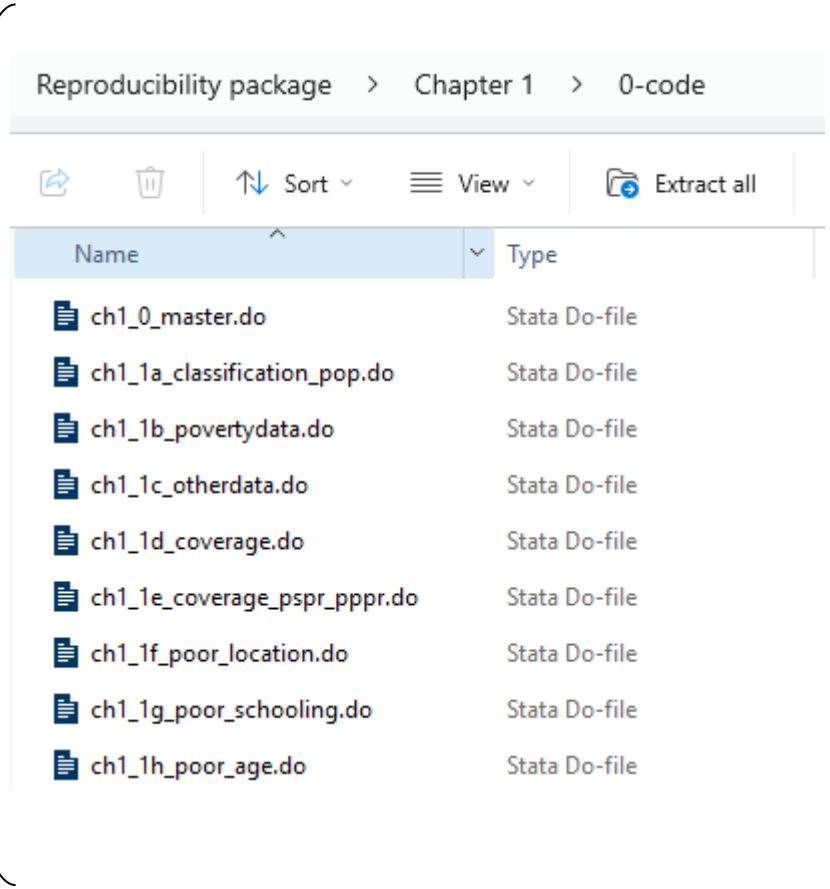
Tips:

- 📁 Define the code and folder structure **at the start** of the project
- 🤝 Agree on the structure **with all team members**
- 🧩 Organize **by chapter** or module (e.g., Chapter 1, Chapter 2…)
- 🧵 Use **a master script** per chapter, plus **one main master script** to run all

Example: Code File Organization







https://reproducibility.worldbank.org/index.php/catalog/250

Code Files: Resources

- **Template Main Files**
  - R Main File: main.R
  - Stata Main File: main.do
- **Instructions for isolating your environment**
  - Python Environment Setup: Python Instructions
  - R Environment Setup: R Instructions
  - Stata Environment Setup: Stata Instructions
- **All resources available through the RRR GitHub**
  - https://worldbank.github.io/wb-reproducible-research-repository/resources.html

Reproducibility Package Components

Data

For verification

- All data that can be shared to WB staff for limited purpose of verification

Public package

- Only data that can legally be redistributed
- Option to temporarily embargo data for Working Papers, if seeking academic publication

Data Statement on the Reproducible Research Repository

- Provides details of each data source and any access restrictions
- Verification report notes differences between public package and package used for review

Reproducibility verification is possible for packages with data access restrictions

Data access for verification only

- Reproducibility team requests temporary data access from source
- And/or reproducibility team signs non-disclosure agreement

Virtual verification

- Author runs package in prepared environment, reviewer observes, author shares outputs

Synthetic data

- Authors share data that mimics data structure; reviewer verifies package runs but not results
- Can complement virtual verification [review verifies actual data; public package contains synthetic data]

Partial verification

- Authors share a subset of the data and a modified package that runs on that subset

Example Data Statement: Restricted data

DATASETS

Statistics on Income and Living Conditions (SILC) of the European Union

Name

Statistics on Income and Living Conditions (SILC) of the European Union

Note

EU statistics on income and living conditions (EU-SILC) for Romania. Data file: "ROU_SILC_2021.dta".

Access policy

The EU-SILC dataset can be requested from Eurostat under a data usage agreement. For more information on how to get access to the EU-LFS, see the data URL.

Data URL

https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

https://reproducibility.worldbank.org/index.php/catalog/151

Example Verification Report: Restricted data

- Data access restrictions did not allow the replicators to access the full raw data to conduct the verification. Consequently, we conducted a virtual reproducibility verification session on May 10, 2024. During this session, the authors executed the code and saved the output files and a log with the results. These were subsequently provided to the replicators for further verification of this paper. This report presents the findings of having compared these results with the version of the manuscript provided by the authors.
- The code takes approximately 6 hours to run.
- All of the exhibits were reproduced.
- **Reproducibility Summary**:
  - **Data**: All data is confidential and has not been included in the reproducibility package. For more details, please refer to the README file.
  - **Code**: All code files (from cleaning to analysis) are included in the reproducibility package.
  - **Outputs**: Some outputs are not directly generated by code but instructions for producing these outputs are included in the README file.
  - **Reproducibility verification**: Reviewers verified the package via virtual verification and did not have access to the data.

https://reproducibility.worldbank.org/index.php/catalog/151/

Note on data for flagships

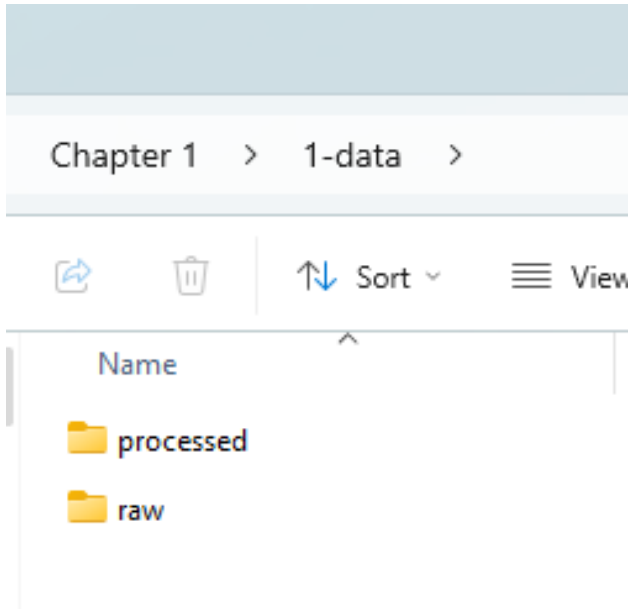
Flagships often involve multiple datasets. **Clarity on which sources and versions of data to use are essential**.

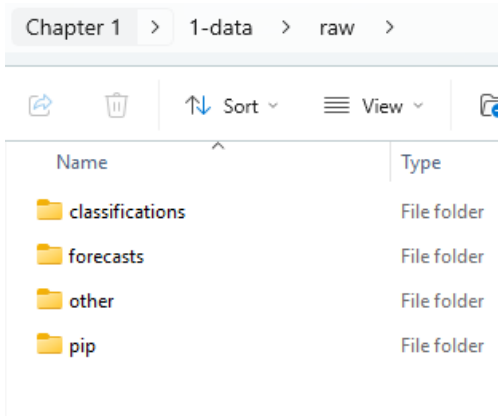
Key practices:

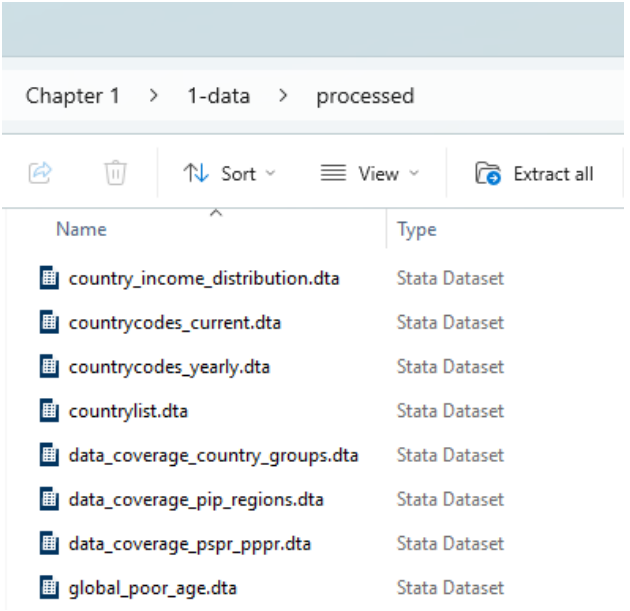
- 🗃️ **Document all changes** made to source data
  → Best if done by code
- 📁 **Clarify raw vs. processed**
  → Keep raw data untouched, store processed versions separately
- 📌 **Track and archive versions** of all data used in the analysis
- 🧹 **Clean the package** before submission
  → Remove unused datasets and keep only what's needed for the report

💡 **Tip:**
Start documenting your datasets in the README **from the beginning.**
→ Saves time later and avoids guesswork at submission

Example on data for flagships

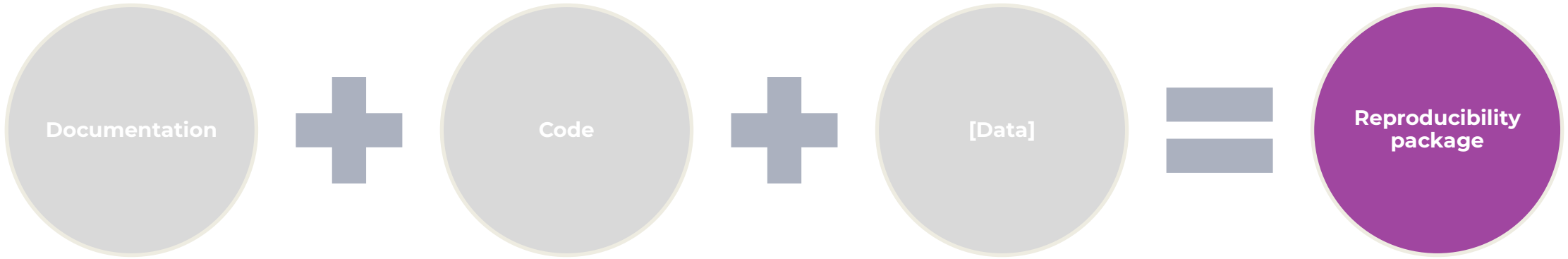






https://reproducibility.worldbank.org/index.php/catalog/250

Reproducibility Package Components

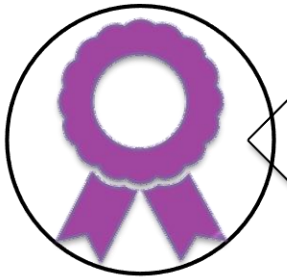


Final Reproducibility Package

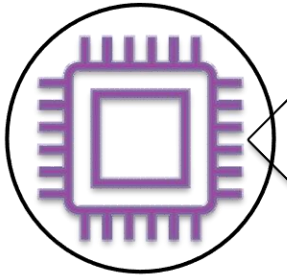
In addition to documentation, code and data, final published reproducibility package includes:

- Reproducibility verification report
- Catalogue entry on the Reproducible Research Repository with detailed metadata, DOI and citation

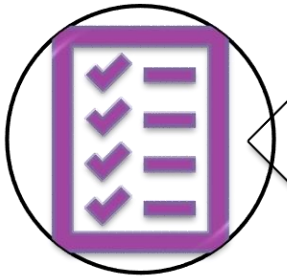
Reproducibility Report



Certifies the package is functional, stable, and exactly matches the paper



Details technical environment for reproduction (OS, processor, memory available, software version)



Lists reproducibility status of every exhibit in the paper

Reproducibility Report - Standard case

Contents in this review:

1. Main findings
2. List of exhibits and reproducibility status
3. Reproduction Environment

Main findings

- Every exhibit has been reproduced accurately.
- The code was successfully executed on a new computer after:
  1. Setting the file paths in the script *o_master_ETR_GMT.R* and in the do-files *o_master_cleaning.do*, *2_4_GMT_output.do*, and *Code_FA2.do*.
- The output demonstrates consistent stability across multiple runs. Specifically, executing the code two times consecutively yielded identical results.
- The code takes approximately 2 hours to run.
- We conducted our reproducibility analysis based on the paper shared by the authors on February 12, 2025 using the project's private GitHub repository.
- **Reproducibility Summary**:
  - **Data**: Some data is restricted and has not been included in the reproducibility package. For more details, please refer to the README file.
  - **Code**: All code files (from cleaning to analysis) are included in the reproducibility package.
  - **Outputs**: All outputs are generated by code included in the reproducibility package.
  - **Reproducibility verification**: Reviewers used data provided directly by the authors to conduct the reproducibility verification, and this is not included in the public reproducibility package. The reviewers did not verify if publicly available data matches the data provided by the authors.
  - **Dependencies environment**: The reviewers created a new environment using the latest versions of dependencies available at the moment of the review.

https://reproducibility.worldbank.org/index.php/catalog/284

Reproducibility Report -Standard case

List of exhibits and reproducibility status

Results in the Main Section of the Paper

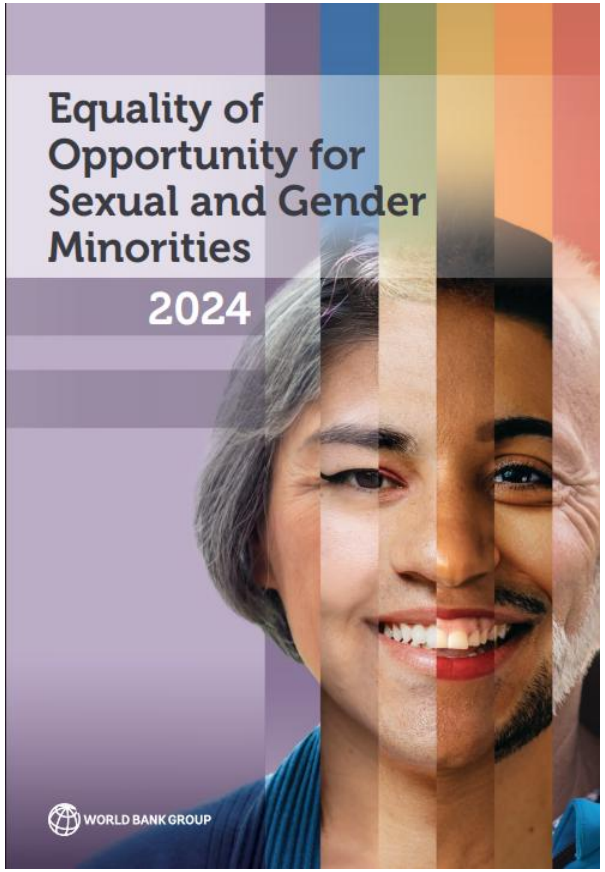
- **Figure 1** Does not apply.
- **Figure 2** Reproduced. The exhibit was compared against code output attendance_pdf.png.
- **Figure 3** Reproduced. The exhibit was compared against code outputs: cdf_phq8_iptg_r1.png, cdf_ghq12_iptg_r1.png, cdf_phq8_r2.png, cdf_ghq12_r2.png, and cdf_phq8_r3.png.
- **Table 1** Does not apply.
- **Table 2** Does not apply.
- **Table 3** Reproduced. The exhibit was compared against the code output in balance.tex.
- **Table 4** Reproduced. The exhibit was compared against code output attrition.tex.
- **Table 5** Reproduced.The exhibit was compared against the code output in compliance.tex.
- **Table 6** Reproduced.The exhibit was compared against the code output in results-mh-both1.tex, results-mh-both2.tex, and results-mh-both3.tex.
- **Table 7** Reproduced.The exhibit was compared against the code output in results-mh-both2-cash.tex and results-mh-both3-cash.tex.
- **Table 8** Reproduced.The exhibit was compared against the code output in results-hc-sec-ther-r1only.tex.
- **Table 9** Reproduced.The exhibit was compared against the code output in results-hc-both2.tex and results-hc-both3.tex.
- **Table 10** Reproduced.The exhibit was compared against the code output in results-hc-both2-cash.tex and results-hc-both3-cash.tex
- **Table 11** Reproduced.The exhibit was compared against the code output in results-covid2.tex and results-covid3.tex

Results in the Annex

- **Table A1** Reproduced. The exhibit was compared against code output attrition_blcovars.tex.
- **Table A2** Reproduced. The exhibit was compared against code output attendance_mean_outcomes1.tex.
- **Table A3** Reproduced. The exhibit was compared against code output results-mh-both1-adj.tex, results-mh-both2-adj.tex, and results-mh-both3-adj.tex.
- **Table A4** Reproduced. The exhibit was compared against code output results-mh-late1.tex, results-mh-late2.tex, and results-mh-late3.tex.
- **Table A5** Reproduced. The exhibit was compared against code output results-mh-mdd1.tex, results-mh-mdd2.tex, and results-mh-mdd3.tex.

https://reproducibility.worldbank.org/index.php/catalog/166

RRR catalogue entry



Reproducible Research Repository
https://reproducibility.worldbank.org

A reproducibility package is available for this book in the Reproducible Research Repository at https://reproducibility.worldbank.org/index.php/catalog/182.

Reproducibility package for Equality of Opportunity for Sexual and Gender Minorities 2024

2024 GET REPRODUCIBILITY PACKAGE

Reference ID	FR_WLD_2024_185
DOI	https://doi.org/10.60572/6tm1-cm51
Author(s)	Paola Ballon, Omar Alburqueque
Collections	Flagships and Reports
Metadata	JSON

Citation

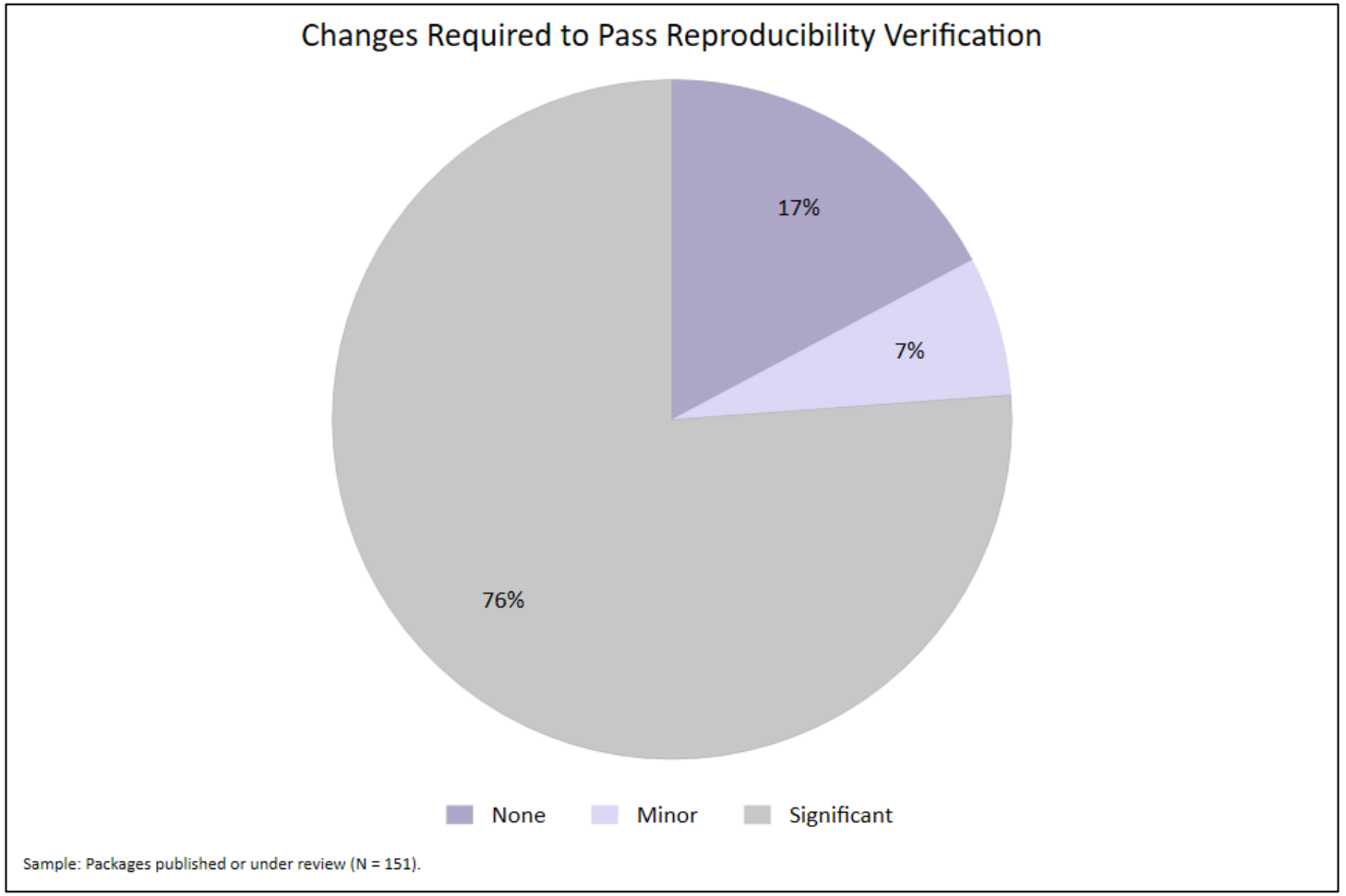
CITATION

Ballon, P., & Alburqueque, O. (2024). *Reproducibility package for Equality of Opportunity for Sexual and Gender Minorities 2024*. World Bank. https://doi.org/10.60572/6TM1-CM51

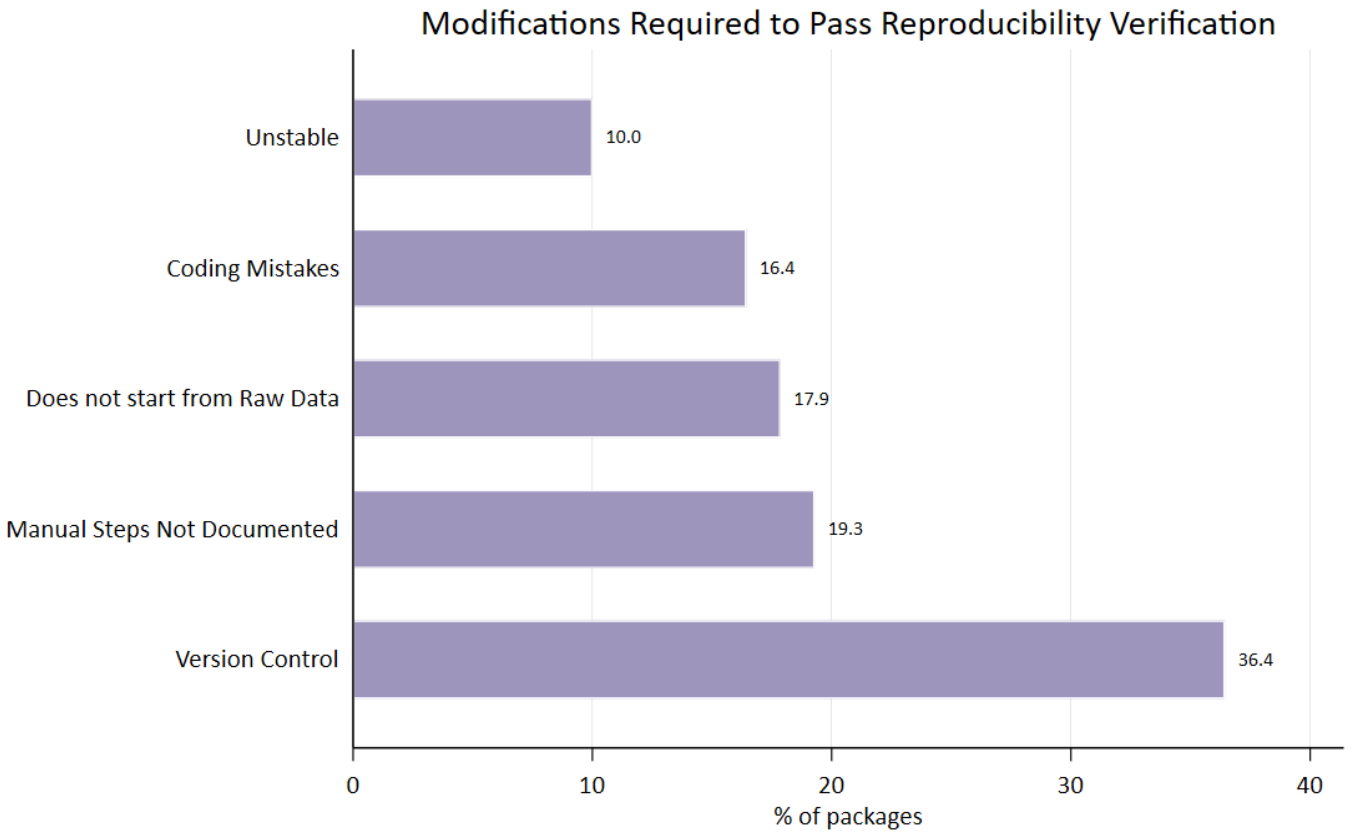
Export citation: RIS | BibTeX | Plain text

Common reasons reproducibility packages fail

Most packages do not reproduce as submitted



Common reasons packages fail to reproduce



Sample: Packages published or under review (N = 140).
Certain packages are categorized under more than one category.

Reproducibility Failure: Version control

- **Problem:** Outputs from code do not match exhibits in report
  - Usually a copy-paste problem (paper not fully updated)
  - We request raw outputs in the package you submit (for verification only)
  - If these raw outputs match our code outputs, but not the exhibits in the paper → version control error

- Check your report **matches your own raw outputs** before submitting for verification

Version control: Solution in Flagships

In flagships, **many authors, chapters, and softwares** can lead to version mismatches.
A **well-structured workflow** helps avoid these issues.

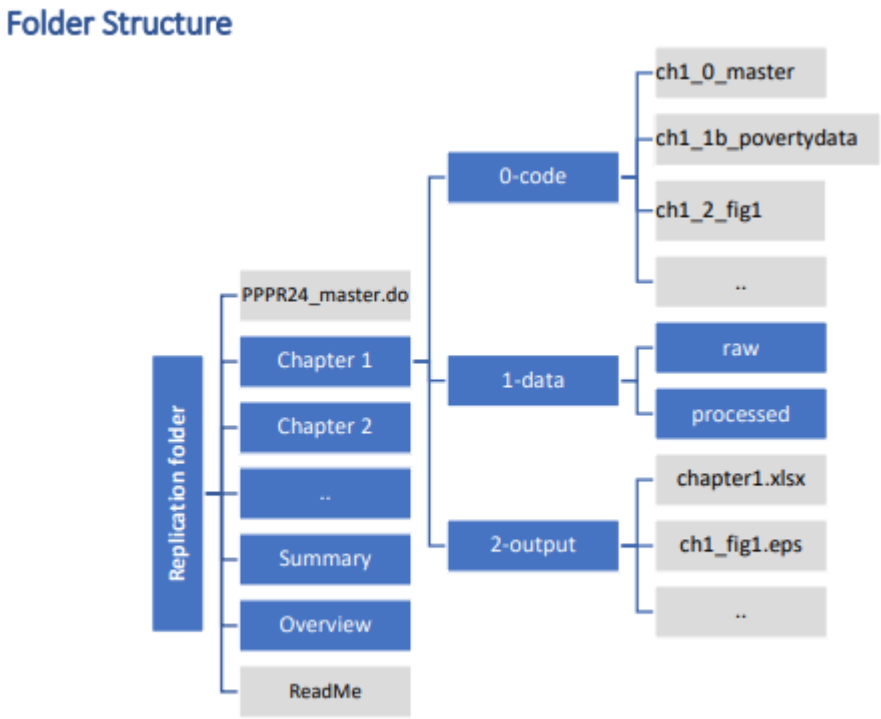
Best practices:

- 🗂️ **Organize code by chapter** with clear naming and one final master script

- 📃 **Keep a folder for final outputs** that are used in the report

- 🖥️ **Do a final run of the code** before submission
→ Make sure exhibits in the report match the latest code outputs

- 🔄 **Gold standard: Include GitHub in your workflow** to track changes and ensure alignment
→ Commit raw outputs and versioned figures/tables



https://reproducibility.worldbank.org/index.php/catalog/250

Version Control: Outputs

Commit raw outputs to GitHub to enable change tracking.

- **For figures:** use formats such as .png, .jpg, or .svg.
- **For tables:** use formats such as .tex, .txt, or .csv.

Reproducibility Failure: Manual steps not documented

- **Problem:** Code does not generate outputs that appear in the paper
  - Ex 1: Results displayed in console and manually copied into a table
  - Ex 2: Data is exported to Excel, then a chart is made in Excel

- **Minimum solution:** Document every step of the workflow
  - 1. What lines of code produce results used in each cell of the table?
  - 2. Was data exported in code or copy-pasted? Were any calculations made in Excel? Which columns and rows were used for the chart?

- **Best solution:** Automate the full workflow, export all outputs using statistical software [we can help!]
  - Reproducible visualizations libraries: Stata, R

Manual steps not documented: Solution in Flagships

🔍 **The Reality:**
Most flagships involve Excel-based workflows.
Ideally, everything would be coded—but we know that's not always possible.

Minimum solution

Setting up Excel sheet tabs:

To reproduce the graphs, follow these instructions:

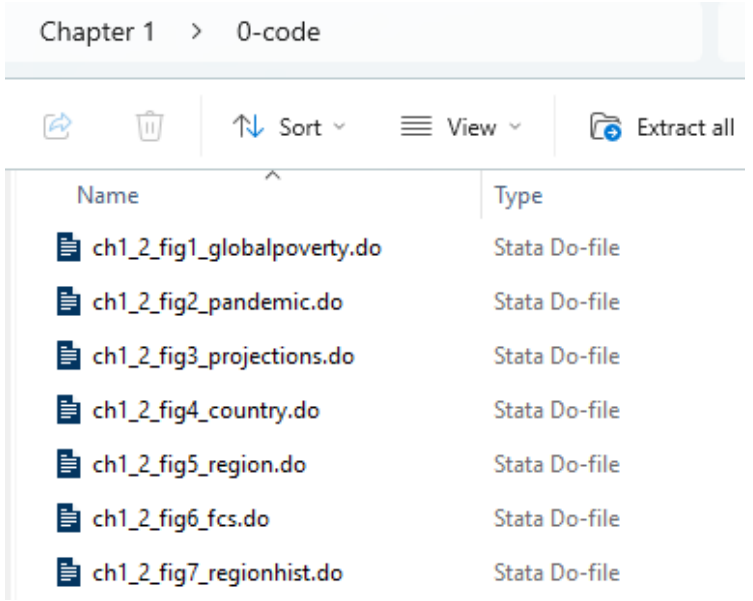
Sheet name	Details on setting up
CHL grown and inflation	Copy and paste the data "Quarterly growth data from Central Bank of Chile" in Columns E-H and rows 14-36 following the instructions for "Chile growth and inflation data" above.

raw G1.1 Pivot G1.1 G1.1	Follow instructions for "Global food trade data" source above. Generate pivot table over all data contained in 'raw G1.1.'. Select 'ProductCode' for rows, ReporterName and PartnerName for columns, and the sum of TradeValue as values. Copy ProductCodes into column A of 'G1.1'. and 'TradeValue' into column C. Copy world imports from all countries into column B, and world imports from Ukraine into column C. Copy Chile's imports from all countries into column F, and Colombia's imports from all countries into column H. Sort descending by column E (Ukraine's share in world trade).
"raw" tabs: G21., G3.1, G4.2, G 4.3, G5, G6, overview	Copy and paste these from the exported code output to the excel file "All result CHL COL note original.xlsx"

Figures can be found in the Excel **FINAL figures For MEX PEA.xlsx**

- Figure 1.
  - Figure 1 (left panel), Figure 1 (right panel). Data from sheet *Figure 1 data*. This data is generated by running the do file "1.0 Poverty Mexico - PovTrends.do" using data processed by SEDLAC.
- Figure 2.
  - Data from sheet *Figure 2 data*, and *Figure 2 data (2)*. This data is generated by running the do file "2 Poverty Mexico - Vulnerability and MC PovTrends.do" using data processed by SEDLAC. The poverty rate data comes from sheet *Figure 1 data*.

Best solution



https://reproducibility.worldbank.org/index.php/catalog/250

Reproducibility Failure: Package does not start from raw data or lacks data documentation

- Reproducibility packages should document all steps taken to get from the original data sources to the final exhibits in the paper
- Starting point for the reproducibility package **should** be a documented source dataset
- Starting from ‘analysis.dta’ without documentation of data cleaning and/or processing **does not** meet reproducibility standard

Solution in Flagships: Document & Archive Data From Day 1

📄 **In practice:**

Most flagships require **a lot of back and forth** on data access and documentation.
Avoid delays—**start documenting and organizing data from the beginning**.

Best solution

This folder contains files with data either downloaded from public datasets such as Databank, CEPALSTAT, etc., or prepared as part of a special request. Below is a detailed description of the files:

i. Publicly available files

- "P_Data_Extract_From_Population_estimates_and_projections": Total population data from the World Bank Databank.
  - Variable name: Population, total
  - Availability: Publicly available
  - Download link: https://databank.worldbank.org/source/population-estimates-and-projections#
  - Downloaded on: May 2024
- "Desigualdad_Datos": Data on inequality in Latin America and the Caribbean (LAC), measured by the Gini coefficient, sourced from the LAC Equity Lab.
  - Availability: Publicly available
  - Download link: https://www.bancomundial.org/es/topic/poverty/lac-equity-lab1/income-inequality/inequality-trends
  - Downloaded on: May 2024

https://reproducibility.worldbank.org/index.php/catalog/241

Reproducibility Failure: Coding mistakes or instability

- Verifier runs each package twice to ensure results do not change
- **Problem:**
  - If results do change → package **unstable** (usually due to uncontrolled random processes)
  - If results are consistent but differ from the paper (and it's not a simple version control problem) → **coding mistake**
- **Solution:**
  - Reproducibility team troubleshoots and work together with authors to resolve issues

Summary: Adapting Research Workflows for Reproducibility

Key Practice: Document & Define Your Data Clearly

- **Clearly define all source datasets.**
  - Document access details (date accessed, subset of data, etc).
  - Archive the version to be used for the analysis and ensure everyone on the team starts from that version
  - Keep track of figures from other papers and document sources

- **Document all changes made to the source datasets.**
  - Ensure clear documentation of how data is cleaned, transformed, and analyzed.
  - Modularize work: separate data cleaning scripts from analysis scripts.

Key Practice: Structure Code Files and Automate Workflows

- **Set up main scripts to manage analysis.**
  - Use a primary script to call sub-scripts (e.g., main-report → main-chapter1, main-chapter2).
- **Mixed (code and Excel) or Excel packages.**
  - Clearly state in the README if figures are generated by code or built manually in Excel.
  - Provide instructions on how Excel figures were created from raw or processed data.
- **Automate workflows as much as possible. Avoid copy-pasting!**
  - Contact reproducibility@worldbank.org for automation support.
  - Confirm that exhibits in final report match raw outputs in package.

Key Practice: Set Expectations Early

- **Assign a reproducibility lead for each chapter from day 1.**
  - Ensure all contributors know they need to document all data work starting from the raw input data.
  - Agree on a code and folder structure.

- **Require code files and list of data sources to be submitted along with the chapter draft**
  - Align with all contributors on data sources early.
  - If multiple chapters use the same dataset, ensure same version
  - Best practice: Include data citations for each chapter (with source, version, access date, etc.).

Important links

Consult our resources

SCAN ME

https://bit.ly/wb-rrr-resources

Verification Request

SCAN ME

https://bit.ly/wb-rrr-request

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

WORLD BANK GROUP
Academy • Development Data

Reproducible Research Repository

Thank you!

Questions?
reproducibility@worldbank.org

https://bit.ly/wb-rrr-resources