Data Analysis README

Contents

1. Project Overview

Purpose:
This project analyzes the main drivers of public sector workforce performance and explores how these factors influence countries' ability to achieve development objectives. The chapter is a collaborative effort, drawing on multiple contributors, diverse data sources, and a variety of analytical tools.

How to Use This Document:

- Start here to understand the overall structure and workflow.
- Follow the guidance in each section to ensure reproducibility and compliance.
- Refer to the folder-specific instructions for data access, processing, and output generation.

2. Folder Structure & Contents

The project is organized into several main directories. Each folder serves a specific purpose in the data analysis workflow.

A. Reproducibility Folder

Purpose:
Contains all materials required to recreate every figure and table in thereport. This ensures transparency and enables other researchers to replicate results.

Contents:

- Raw and processed datasets
- Scripts and code files (Stata, R, Excel)
- Analytical outputs (figures, tables, intermediate data)
- Documentation for datasets and workflows
- **This folder contains the "Synthesis Report_Reproducibility_tracker" document, which serves as a repository for all figures included in the report.** Certain figures—such as 3.1A—are created by manually combining data coming from different sources and analytical approaches. For the finalized figures, please consult this document.

Prerequisites

- **Software:**
  - https://www.stata.com/ (for survey data analysis)
  - https://www.r-project.org/ (for WWBI and GSPS analysis)

- https://www.microsoft.com/en-us/microsoft-365/excel (for data review and figure creation)

- **R Packages:**
  - All required R packages are listed in the project's renv environment.

Subfolders:

/01_Data:

/01_raw: Original, unmodified datasets collected by the team. These files are raw inputs and must not be edited or overwritten. All raw datasets are listed below under Table 1. While the survey data is stored under 01_data, the code that processes and analyzes these datasets lives in a separate folder- 02_code/stata/

/02_surveys_report_published: This folder stores published documents used to extract certain figures. These are not raw datasets but reference publications.

- Romania Report, "Selecting the right staff and keeping them motivated for a high-performing public administration in Romania" (link):
  - Figure 2.10 – This figure is using the data extracted from the Report (Figure 4, page.23)
  - Figure 3.1b – This figure is using the data extracted from the Report (para. 73, p.47)
- Brazil Report, "Ethics and Corruption in the Federal Public Service: Civil Servants' Perspectives" (link):
  - Figure 3_4 – This figure is using the data extracted from the Report (page 16)
- GovTech Maturity Index: "EQUITABLE GROWTH, FINANCE & INSTITUTIONS INSIGHT" (Link): Figure 4.1 uses the figure published as Figure 3.13 (page 47): Diffusion of selected government institutions and systems, 2022 (a) Diffusion of Government Systems (1984–2022).

/03_external_survey_datasets: Published external datasets for selected visualizations (e.g., GSPS, GEPD).

Dataset	Figures	Source	Notes
Global Survey of Public Servants (GSPS)	Figures 2_9, 2.13, 3.1a (data taken from the website is combined with	GSPS Dashboard	Data comes from a snapshot directly from **the online dashboard.**

Dataset	Figures	Source	Notes
	the survey data coming from Stata manually.)		https://www.globalsurveyofpublicservants.org/data-dashboard.
Global Education Policy Dashboard (GEPD)	2.7	Shared internally by the GEPD team	Dataset is shared by the GEPD team internally within the Bank. https://www.educationpolicydashboard.org/.

/04_non_survey_data: This folder contains datasets that are not survey-based but are used to make figures for the report (e.g., WWBI, IMF WEO, ISORA, Lightcast). All these sources are online and published.

- WB WWBI:
  - Figure 2.11 and ES3 are the same. They use the Public sector wage premium (compared to all private employees) indicator combined with GDP per capita 2015. . The folder includes the direct download file from this link (WB WWBI.xlsx) and the smaller excel sheet that just includes Public sector wage premium (compared to all private employees) indicator (WB WWBI- Wage premium.xlsx) used for the construction of the Figure.
  - Figure 2.16 uses “"Wage bill as a percentage of GDP” (BI.WAG.TOTL.GD.ZS) and “Wage bill as a percentage of Public Expenditure” (BI.WAG.TOTL.PB.ZS) data from the World Bank’s Worldwide Bureaucracy Indicators database as of Feb 2024. The folder includes the direct download file from this link (WB WWBI.xlsx) and the smaller excel sheet that just includes the two indicators (WB WWBI - Wage Bill as a share of GDP and Government Expenditures.xlsx) used for the construction of the Figure.
- IMF WEO:

- Figure 2.17 uses “general government net lending/borrowing” (GGXCNL_NGDP) indicator from the IMF’s World Economic Outlook Indicators database as of April 2025.The folder includes the direct download file from this link (IMF WEO.xlsx) and a smaller excel sheet that just includes the one indicator (IMF WEO - General government net lending-borrowing.xlsx) used for the construction of the Figure.

- ISORA:
  - Figure 2.18A uses data on “Salary cost as percent of operating expenditure” from the ISORA dataset. The folder includes the direct download file from this Link (IMF ISORA.xlsx) and a smaller excel sheet that just includes the one indicator (ISORA - Salary cost as percent of operating expenditure.xlsx) used for the construction of the Figure.
  - Figure 2.18B uses data from six separate variables “Population per FTE”, “Total tax administration FTEs – Derived”, “Number of taxpayers by tax type-Employers that withhold tax from employees-Registered employers”, “Number of taxpayers by tax type-Corporate income tax-Registered taxpayers”, “Number of taxpayers by tax type-Personal income tax-Registered taxpayers”, “Number of taxpayers by tax type-Value added tax-Registered taxpayers”, and “Number of taxpayers by tax type-Excise (domestic manufacturers / producers)-Registered taxpayers” from the ISORA dataset. The folder includes the direct download file from this Link (IMF ISORA.xlsx) and a smaller excel sheet that just includes the six indicator (ISORA - Registered Tax Payer and Population per FTE.xlsx) used for the construction of the Figure.
- Lightcast: Online job posting data. Figure 3.6A, 3.6B, and 3.7 use the above data. Data is scraped from this link. To access the data, users will need a paid subscription. These data capture the demand for climate-related or “green” skills in the public administration compared to the broader labor market.

Although some of these may later be used in R or Stata scripts, the data itself is stored here for a consistent structure.

/02_Code

The 02_code directory contains scripts used to clean, transform, analyze, and visualize the raw survey datasets included in the report and a part of the data analysis using Global Survey of Public Servants (GSPS) and Worldwide Bureaucracy Indicators (WWBI) data. The code is organized by software environment to ensure transparency, modularity, and ease of replication. Each subfolder includes self-contained scripts.

/Stata Scripts:

The master do-file (master.do) controls the complete data-cleaning and figure-generation pipeline for the survey figures. It sets directory paths, runs all country-level cleaning scripts, constructs the pooled dataset, and produces the listed figures and supplementary survey outputs:

Below code generates the figures: ES 4, 2.8, 2.9, 2.10, 2.11, 2.12,2.13, 3.1a (partly), 3.1b, 3.2, 3.3, 3_4, 3_5, 3_8, 4_2, 4.3, 4.4

The script executes the following steps in sequence:

1. Sets all project filepaths (main folder, input data, scripts, and output directories).: new user must update the folder path in line 12.
2. Runs country-specific data cleaning scripts for each survey:
    - Mali
    - Madagascar (has two different surveys: health and education sector)
    - Ethiopia
    - Croatia
    - Armenia
    - Uruguay
    - Philippines
    - Bosnia
    - Kosovo
    - Argentina
    - Estonia
    - Pakistan
    - Malaysia

Note: Some figures are made in their specific country do file. These are mentioned below under the Data List.

3. Creates a pooled cross-country dataset using append.do.
4. Generates the figures through graphs.do.
5. Processes additional external survey sources, including:
    - COVID-era surveys
    - World Values Survey

Each script relies on the folder structure defined at the top of the master do-file.

Table 1: Data list

Country / Source	Data Type	Notes
Mali	ADMIN_LEVEL_SURVEY, FACILITY_LEVEL_SURVEY	*Graphs.do script produces all figures*

Country / Source	Data Type	Notes
	Raw survey data	*We do not have permission to share the micro-data publicly from the government.*
Madagascar (Health)	ADMIN_LEVEL_SURVEY, FACILITY_LEVEL_SURVEY Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Madagascar (Education)	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Croatia (Health)	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Ethiopia	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Armenia	Raw survey data	*Figure 3_5 code in Armenia do-file* *We do not have permission to share the micro-data publicly from the government.*
Uruguay	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Philippines	Raw survey data	*Graphs.do script produces all figures*

Country / Source	Data Type	Notes
		We do not have permission to share the micro-data publicly from the government.
Bosnia	Raw survey data	*Figure 2.12 code in Bosnia do-file* *We do not have permission to share the micro-data publicly from the government.*
Kosovo	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Argentina	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Estonia	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Pakistan	Raw survey data	*Graphs.do script produces all figures* *We do not have permission to share the micro-data publicly from the government.*
Malaysia	Raw survey data	*Figure 4.4 code in Malaysia do-file* *We do not have permission to share the micro-data publicly from the government.*
COVID-19 Remote Work	Raw survey data	*Figure 3.2 code in COVID19 do-file*

World Values Survey	External dataset	*Figure3_3 code in the worldvaluessurvey.do file* Source: www.worldvaluessurvey.org

/R scripts and data
R folders for Figures produced in R

/WWBI
Contains the input, output and r scripts to produce some of the figures using the WWBI dataset.

- Data: Contains input and output data to create graphs. It includes WWBI data along with GDP per capita 2015 (link) data derived from World Bank databank.
- Code: Contains 3 r scripts:
    - 0-main: sets the working environment --> new users need to restore the environment and change the data_path (line 32)
    - 01_processing-data-graphs: loads the data, prepares the data for producing the graphs and produces figures.
    - 02-figues: exports all figures in png format into the output folder.
- Outputs: Contains figures 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.12, 2.14 and 2.15 in PNG format.

Figure 2.1: uses Public sector employment per 1000 people indicator

Figure 2.2: uses the following indicators:

- Core Public Administration as a share of public paid employees
- Public safety as a share of public paid employees
- Education as a share of public paid employees
- Health as a share of public paid employees
- Social security as a share of public paid employees

Figure 2.3: uses the public sector employment, as a share of a) paid employment and b) formal employment indicators

Figure 2.4: uses the Females, as a share of public paid employees and Females, as a share of private paid employees' indicators

Figure 2.5: uses the below indicators and GDP per capita 2015.

- Gender wage premium in the public sector, by industry: Education (compared to male paid employees)

- Gender wage premium in the public sector, by industry: Health (compared to male paid employees)
- Gender wage premium in the public sector, by industry: Public Administration (compared to male paid employees)

Figure 2.6: uses the following indicators:

- Public sector employment, as share of total employment
- Public sector employment, as share of paid employment
- Public sector employment, as share of formal employment

And classifies countries into FCV and non-FCV based on the following classification: https://www.worldbank.org/en/topic/fragilityconflictviolence/brief/classification-of-fragile-and-conflict-affected-situations

Figure 2.14: uses the Public sector wage premium, by education level: Tertiary Education (compared to formal wage employees) and public sector wage premium, by education level: Secondary Education (compared to formal wage employees) indicators.

Figure 2.15: uses the Pay compression ratio in public sector (ratio of 90th/10th percentile earners) and Pay compression ratio in private sector (ratio of 90th/10th percentile earners) indicators.

/GSPS

R folders for Figures produced in R

- Contains the input, output and r scripts to produce the ES2 figure coming from the GSPS data: https://www.globalsurveyofpublicservants.org/data-dashboard
- Data:
    - Global Survey of Public Servants.
    - Filename: GSPS_Indicators_Dataset_11_10_23.xlsx.
    - Source: Bureaucracy Lab, World Bank.
    - URL: https://www.globalsurveyofpublicservants.org/sites/default/files/2023-11/GSPS_Indicators_Dataset_11_10_23.xlsx
    - Access year: 2025.
- Code: contains 1 R script:
    - 0-analysis.R: cleans and analyzes the data to produce the graph.
- Outputs: Contains figure ES2 in PNG format.

Note: Please open the Rproject file and first run `renv::restore()`. This will install all the package dependencies required to reproduce the analysis.

/03_Outputs

The 03_outputs directory contains all analytical outputs generated from the Stata, R, and Excel scripts. These outputs include figures used in the chapter.

/Stata output

These outputs are created automatically when running the master Stata script (master.do) located in 02_code/stata/.

/figures/img: This folder stores all figures created using Stata: ***ES 4, 2.8, 2.9, 2.10, 11, 3.1a (partly), 3.1b, 3.2, 3.3, 3_4, 3_5, 3_8, 4_2, 4.3, 4.4***

/R output
This folder stores all figures created using R scripts located under 02_code/r scripts and data/. The R output folder includes two shortcut subdirectories:

- /GSPS

A shortcut linking to the GSPS R output subfolder under 02_code, allowing users to quickly access GSPS-generated figures without navigating through code directories.

Contains figure: **ES 2**

- /WWBI

A shortcut linking to the WWBI R output folder under 02_code.

These shortcuts ensure that users reviewing outputs can access all figures from a single location without duplicating files across directories.

Contains figures **2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.11, 2.12, 2.14 and 2.15** in PNG format.

/Excel

- GSPS_ Reproducability: Data comes from a snapshot directly from the online dashboard. These figures (**2_9, 2.13, 3.1a (partly)**) use the GSPS data. It is published: https://www.globalsurveyofpublicservants.org/data-dashboard.
- GEPD_Indicators_Reproducability: Dataset is shared by the GEPD team. It is published: https://www.educationpolicydashboard.org/. Figure **2.7** uses this data.
- WWBI_Reproducability: Figure **2.16** and Figure **ES 3/ Figure 2.11** use data from the World Bank's Worldwide Bureaucracy Indicators database as of Feb 2024. Raw data is in the WB WWBI.xlsx file under the 01/data/04_non_survey_data folder.

- WWBI-IMF WEO_Reproducability: Figure **2.17** uses “general government net lending/borrowing” (GGXCNL_NGDP) indicator from the IMF’s World Economic Outlook Indicators database as of April 2025. Raw data is in the IMF WEO.xlsx file under the 01/data/04_non_survey_data folder.
- ISORA_Reproducability:
    1. Figure **2.18A** uses data on “Salary cost as percent of operating expenditure” from the ISORA dataset. Raw data is in the IMF ISORA.xlsx file under the 01/data/04_non_survey_data folder.
    2. Figure **2.18B** uses data from six separate from the ISORA dataset. Raw data is in the IMF ISORA.xlsx file under the 01/data/04_non_survey_data folder.
- Lightcast: Online job posting data. Contains Figures **3.6A, 3.6B, and 3.7**.

/PNG Charts

This folder contains all the figures in the publishable format.

3. Data Access and Use

All datasets used in this study were accessed and analyzed in accordance with applicable data-sharing agreements, institutional policies, and legal requirements. The authors confirm that they had legitimate rights to access, use, and analyze all data included in this study for research purposes.

4. Synthesis Report — Reproducibility Tracker (Excel File)

This Excel file is used to track all figures included in the Synthesis Report. It serves as a consolidated reference that compiles every figure across data sources, helping ensure complete reproducibility and consistent documentation.