

README – Replication Files for Gender Gaps in Pay and Advancement Within Firms: Evidence from Ethiopia’s Financial Sector

Authors: Ketki Sheth, Toni Weis & Saba Yifredew

This version: June 2026

Data Availability Statement

The analysis uses multiple data sources from the Commercial Bank of Ethiopia (CBE). The core dataset consists of administrative human resources records covering the universe of the bank’s employees over the period 2019–2024. These records are complemented by two primary survey datasets: (i) a detailed, in-person employee and supervisor survey conducted among a sample of Addis Ababa–based employees, and (ii) a shorter online survey administered to employees across all regions. All datasets contain employee and employer identifiers. Access is restricted, and the data and surveys were obtained under an agreement between the CBE and the World Bank. Redistribution of the data is not permitted.

1. HR records of employees for 2019-2024. This is snapshot data of employees shared in various rounds in various years

Source: Commercial Bank of Ethiopia (CBE) (September 2022, July 2023, February 2024)

File location: "raw\emp_19_22.xlsx"

"raw\emp_2023.xlsx"

"raw\emp_2024.xlsx"

"raw\pay_2023.xlsx"

"raw\pay_2024.xlsx"

The HR records of employees cover the full universe of the bank's employees from 2019 to 2024. These records were acquired sequentially in multiple rounds and include demographics, HR records, and pay.

2. HR records of employees’ appraisal data. This is internal performance evaluation data of employees for 2020,2021 and 2022

Source: Commercial Bank of Ethiopia (CBE) (February 2023)

File location: "raw\Appraisal.xlsx"

The dataset contains employee and employer identifiers. Access is restricted and was obtained through an agreement between the CBE and the World Bank; redistribution is not permitted.

3. Separate HR record data on gender: This was separately shared during the cleaning process to rectify identified gender inconsistencies in the raw administrative files.
Source: Commercial Bank of Ethiopia (CBE) (July 2023)
File location: "raw\Gender.dta"
The dataset was obtained in excel but saved in Stata to merge with the full HR records dataset.
4. Surveys conducted by Gender Innovation Lab (GIL) World Bank
 - i. In-Person employees survey: This data set consists of primary data collected by GIL World Bank through two distinct survey instruments administered to a subset of the workforce in Addis Ababa branches. These data were received in May 2024.
File location: "raw\Allemp_final.dta"
"raw\perception_raw.dta"
 - ii. On-line employee survey: online.xlsx: A short online survey administered to employees across all regions of the bank through the bank's internal portal. This was shared in May 2024.
File location: " raw\online.xlsx"

All datasets contain employee and employer identifiers. Access is restricted and surveys were commissioned via an agreement between the CBE and the World Bank; redistribution is not permitted.

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.

Directory Structure

- The original, raw data are stored in the folder: raw
- All code is provided in the folder: do
- The intermediate and final datasets are in the data folder: data
- All tables are outputted into the folder: output
- Software Requirements: Stata version 15 or higher
- System specifications: the code was last run on an AMD Ryzen 7 8840HS w/ Radeon 780M Graphics 3.30 GHz, 16.0 GB RAM, 64-bit operating system on Windows 11. The entire sequence from data cleaning to final output generation took approximately **12 minutes** to complete.

Required user-written commands: To ensure the replication environment is fully functional, several user-written Stata packages are required. The master do-file is programmed to automatically check for these commands and install them from the

Statistical Software Components (SSC) archive if they are not already present on your system: APPENDFILE, ESTOUT, IETOLKIT, WINSOR2 and MISSINGS

Instructions for Replicators

All tables are created using Stata, and the replicator can run all cleaning, and analysis scripts sequentially from start to end without interruption using **0_master.do**. The reported results are generated using the two analysis do-files. Tables were generated from Stata as csv files, which are included in the replication package. For manuscript preparation, a second set of Stata do-files reads these .csv files and writes the corresponding LaTeX table code directly via file write commands.

Before running this script, you must make one edit to line 18 of **0_master.do**.

In Line 18: Define a global macro, **root**, that points to the directory containing this README file: for example, line 18 of the **0_master.do** should read something like the following:

global root "C:/Users/sabay/OneDrive/Desktop/Replication_gender_gap_in_pay"

If necessary, the replicator can install the user-written packages used in the process by running lines 32-64 in the **0_master.do**. Should one wish to run do-files individually, it is still necessary to always run lines 12 to 27 in **0_master.do** at the beginning of any session, in order to set up all paths globals. It is also important to run the files in the order of presentation in the master do file.

Do-File Structures

The **Master Do-file (0_master.do)** serves as the automated backbone of the replication package, ensuring all scripts are executed in the correct logical order. It begins by initializing the Stata environment and defining global macros for the project's directory structure. The script handles the installation of necessary user-defined commands. The master script calls all of the do-files in the directory in sequence to construct outcomes and produce all tables in the paper. The execution flow is divided into two main workstreams: the first manages the **Administrative Data** (cleaning annual records from 2019–2024 and appending them into a unified panel), while the second processes the **Survey Data** (covering in-person, perception and online data). The intermediate and final datasets are in the data folder. The four final datasets used in the analysis are:

- i. emp19_24.dta: Administrative data in panel structure
- ii. admin and inperson.dta: In_person survey data
- iii. perception_weighted.dta: In-person survey data on supervisor perceptions
- iv. online.dta: Online survey data used in the appendix.

The following describe the various do files in the do folder

- **1_employees_2019.do through 6_employees_2024.do:** These scripts process the raw annual Excel files, harmonizing variable names and formats across years.
- **7_panel.do:** Appends the cleaned annual records into a single longitudinal panel dataset and performs final data cleaning for the full universe of employees.
- **1s_cleaning_survey.do:** Processes the raw in-person survey data.
- **2s_online_cleaning.do:** Cleans the multi-language online survey data and merges it with relevant administrative records.
- **3s_perception_cleaning.do:** Cleans the specific survey instrument administered to managers.
- **4s_weights_merging.do:** Computes sampling weights for and merges the survey data and readies for analysis.
- **analysis_admin.do:** Produces the panel results, specifically Tables II and IV (the first two rows).
- **analysis_survey.do:** This script generates Table I, as well as III, IV/V, VI and Appendix tables.
- **results_export_tex.do:** This reads the csv output and exports as tex for the final tables.

List of Tables, Programs and Outputs

The provided code reproduces all tables and figures, both in the main body of the paper and in the appendix.

Exhibits	Do file/Program	Line	Output File
Table I	analysis_survey.do	32-162	table_I_unweighted_results.csv, table_I_weighted_results.csv
Table II, rows 1, 3, 5, and 7	analysis_admin.do	10-94	table_II_admin_results.csv
Table II, rows 2, 4, 6, and 8	analysis_survey.do	174-240	table_II_survey_results.csv
Table III	analysis_survey.do	251-295	table_III_results.csv
Table IV: Rows 1 and 2	analysis_admin.do	109-182	table_IV_panelA_admin_results.csv,
Table IV/V	analysis_survey.do	306-464	table_IV_all_panels_results.csv
Table VI	analysis_survey.do	529-625	table_VI_results.csv
Appendix A1	analysis_survey.do	636-702	table_A1_results.csv
Descriptive stats in the paper	analysis_survey.do	11-15, 473-519, 711	

Contact Information

For questions and/or comments, please contact the corresponding author for the paper Ketki Sheth (ksheth2@utk.edu) and/or the creator of this data replication package, Saba Yifredew (sgetnet@worldbank.org).