Replication kit for the paper: "Tax Administration vs. Tax Rates: Evidence from Corporate Taxation in Indonesia"

M. Chatib Basri, University of Indonesia
Mayara Felix, MIT
Rema Hanna, Harvard University
Benjamin A. Olken, MIT

This is the documentation for the replication kit for the paper "Tax Administration vs. Tax Rates: Evidence from Corporate Taxation in Indonesia." It contains a data availability statement, detailing the source and accessibility of the data used in the paper, as well as replication instructions.

1. Data Availability and Provenance Statement

Statement about Rights
The authors of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Summary of Availability
Administrative data, including accompanying auxiliary files, **cannot be made** publicly available. See DGT (2016), DGT (2017a), DGT (2017b), DGT (2018a), DGT (2018b), DGT (2018c), and DGT (2018d) for details.

Data on the Indonesian GDP deflator are provided in this repository. These data are publicly available, and can also be downloaded at: https://fred.stlouisfed.org/series/IDNGDPDEFAISMEI. See Organization for Economic Co-operation and Development (2018) for details.

Details on each Data Source

The project combines six main administrative data sources, outlined below. All administrative data were provided by Indonesia's Directorate General of Taxes (DGT), and contain either anonymized (but consistent) taxpayer identifiers or anonymized (but consistent) staff identifiers such that datasets can be linked.

1. Anonymized administrative data on tax office staffing – SIKKA database (DGT, 2016)

   These data are stored in raw files `LK01_MIT.dta`, `LK02_MIT.dta`, `LK03_MIT.dta`, `LK03A_MIT.dta`, `LK03C_MIT.dta`, `LK03D_MIT.dta`, `LK04_MIT.dta`, `LK16_MIT.dta`, and `NONLK_KeaktifanPegawai_MIT.dta`. See `0_1_Compiled_tables_and_materials_list.xlsx` for further details on the contents of each dataset, and the paper's Data Appendix for general information.

2. Anonymized administrative data on corporate income tax filings – Form SPT 1771 (DGT, 2017a)

   These data are stored in the raw file `SPT_1771.dta`. See the paper's Data Appendix for a detailed description.

3. Anonymized administrative data on tax payments – MPN database (DGT, 2017b, 2018a)

   These data are stored in raw files `MPN.dta` (received in 2017) and `MPN_detailed.dta` (received in 2018). The former dataset is aggregated by tax year for which the payment was made, whereas the latter includes both the tax year and the payment date. See the paper's Data Appendix for a detailed description.

4. Anonymized administrative data on employee income tax withholding – Form SPT 1721 (DGT, 2018b)

   These data are stored in the raw files `SPT_1721_02to08.dta` (2018b) and `SPT_1721.dta` (2018c). The former covers tax years 2002 through 2008, and the latter tax years 2009 through 2013. See the paper's Data Appendix for a detailed description.

5. Anonymized administrative data on tax audits (DGT, 2018c)

   These data are stored in raw file `audits.dta`. See the paper's Data Appendix for a detailed description.

6. Anonymized administrative data on tax assessments and disputes (DGT, 2018d)

   These data are stored in raw files `stpskp_ppn_skp_only.dta`, `stpskp_ppn.dta, and stp_skp_code.xlsx`. The first two datasets contain information on VAT tax assessments and disputes, whereas the third contains assessment code descriptions. See the paper's Data Appendix for a detailed description.

In addition, this paper uses auxiliary files with information from DGT documents, which also **cannot be made public**. See `0_1_Compiled_tables_and_materials_list.xlsx` for a description of each dataset.

Researchers interested in obtaining these data should contact DGT's Subdirectorate of Public Relations:

Subdirectorate of Public Relations
Email: humas@pajak.go.id
Phone Number (DGT Head Office): +6221 525 0208; +6221 525 1609; +6221 526 2880 (can be found on this page https://www.pajak.go.id/hubungi-kami)

Interested researchers will also need to contact the research division for a research permission approval, following the information at https://eriset.pajak.go.id/, but should do so after coordination with the public relations division.

Finally, this paper uses data on the Indonesian GDP deflator retrieved from the Federal Reserve Bank of St. Louis' website. See Organization for Economic Co-operation and Development (2018). We provide this dataset in this replication kit as file `raw/auxiliary/indonesia_gdp_deflator_FRED.csv`.

Dataset List

See `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Raw Datasets".

2. Replication Instructions

Software Requirements
All Stata .do files were run on batch mode in Stata 14.2. The following .ado packages were installed, and are provided in the subdirectory `code/ado/` for ease:

Package	Version number	Version date
binscatter	7.02	24-Nov-13
ebalance	1.5.4	29-Jan-15
gtools	1.0.1	23-Jul-18
ivreg2	4.1.10	9-Feb-16
ivreghdfe	1.0.0	7-Jul-18
reghdfe	5.7.3	13-Nov-19
renvars	2.4.0	22-Aug-05
rforest	1.7.2	Mar-20
unique	1.2.2	10-Nov-17

The R script `code/analysis/Plot_ETI_variation.R` was run in RStudio Version 1.2.1335 running R version 3.6.0 (2019-04-26). The following libraries were installed:

Library	Version
Plotly	4.9.4.1
Hmisc	4.2.0

Controlled Randomness

Some Stata .do files use commands that call on random numbers for sorting data. To keep the .do files reproduceable while also being able to run each code modularly, the seed has been set to 98765 at the top of each .do file. The file `0_master.do` calls each .do in the proper order. To make sure results replicate exactly, we advise running all code in the order they are called in `0_master.do`.

Memory and Runtime Requirements

Approximate time needed to reproduce the analyses on a standard desktop machine: between 8 to 12 hours depending on RAM. Approximately 65GB of storage is required.

Details

The `0_master.do` Stata file was last run on a PowerEdge R930 physical machine running RedHat 7.9 (Linux 3.10.0). The processor is Intel® Xeon® CPU E7-8880 v3 2.30GHz (x144). Memory: 1.48 TiB, Swap: 7.81 GiB. Total run time was 8 hours and 27 minutes.

The R script was last run on MacOS Catalina 10.15.7. This script takes roughly 5 minutes to run. Connection to plotly servers is not required. The graphs should be saved manually using the Viewer.

Replication Folder Structure

The replication folder is organized in the following subdirectories:

- The subdirectory `code/` includes all code necessary to replicate the analysis. Cleaning and preparation codes are saved in `code/prep/`, whereas code that conduct analyses are saved in `code/analysis/`. See Section "Instruction to Replicators" below for instructions on how to run these files.

- The subdirectory `raw/` is where some of the raw administrative datasets should be saved. See file `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Raw Datasets" for which datasets should be saved in this folder and in its subfolder `raw/auxiliary/`. With the exception of file "`raw/auxiliary/indonesia_gdp_deflator_FRED.csv`", which we provide to the public in this replication kit, **these files cannot be made available to the public.**

- The subdirectory `output/dta_raw/` is where some raw datasets should be saved. See file `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Raw Datasets", for which raw datasets should be saved in this subdirectory. **These files cannot be provided to be public.**

- The subdirectory `output/temp_data/` is an empty folder where temporary data produced by the code are saved. **These files cannot be provided to the public.**

- The subdirectory `output/analysis/` is where all intermediary and final analyses files produced by the code are saved. **These files cannot be provided to the public.**

- The subdirectory `output/analysis/regcoeffs/` is an empty folder where some intermediary data produced by the code are saved. **These files cannot be provided to the public.**

- The subdirectory `output/analysis/graphs/` is where all figures produced by the code are saved. See file `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Figures" for the correspondence between the figure names as saved by the code and the figure numbers in the paper. **We provide these files to the public.**

- The subdirectory `output/analysis/csv/` is where all .csv produced by the code are saved. These files should be pasted into their corresponding sheet names in file `0_1_Compiled_tables_and_materials_list.xlsx` to generate the formatted tables in the paper. See file `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of tables" for the correspondence between the .csv files, the excel sheet where they were pasted, and the paper tables they feed. **We provide these files to the public.**

Description of Code

All code necessary to replicate the analysis is provided in the subdirectory `code/`. See `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Scripts" for a description of each file.

Instructions to Replicators

The code provided constructs all the cleaned analysis datasets from raw datasets. Due to data restrictions, we cannot provide the raw datasets used, nor the derivative cleaned analysis datasets.

Steps for replication:

- Obtain the raw datasets from DGT and place them in their corresponding subdirectories, listed in `0_1_Compiled_tables_and_materials_list.xlsx`, sheet "List of Raw Datasets".
- Edit `code/0_master.do` to adjust the paths. Global ${output} should be the path leading to the `output/` subdirectory in this replication kit. Global ${raw} should be the path leading to the `raw/` subdirectory in this replication kit.
- Run the do file `0_master.do`. See above for run times.
- Once `0_master.do` has finished running:
  - Copy the .csv files saved to `output/analysis/csv` into their corresponding excel sheets in the file `0_1_Compiled_tables_and_materials_list.xlsx`. See sheet "List of Tables" for the sheet names where each .csv file should be pasted. Once pasted, these files feed the formatted tables used in the paper, named and listed accordingly in sheet "List of Tables".
  - Edit script `code/analysis/Plot_ETI_variation.R` to adjust the paths. The local inpath should be the path leading to the replication subdirectory `output/analysis/regcoeffs`, and the local outpath should be the path leading to the replication subdirectory `output/analysis/graphs`.
  - Run the R script `code/analysis/Plot_ETI_variation.R`. See above for run times.

List of Scripts, Figures, and Tables

The provided code reproduces all statistics, figures, and tables in the paper (upon access to the restricted data). See `0_1_Compiled_tables_and_materials_list.xlsx`, sheets "List of Tables", "List of Scripts", and "List of Figures" for the comprehensive list of these materials, along with descriptions.

3. References

DGT, 2016. "Anonymized SIKKA datasets". Received October 2016.

DGT, 2017a. "Anonymized form SPT 1771 dataset". Received April 2017.

DGT, 2017b. "Anonymized aggregated MPN dataset". Received May 2017.

DGT , 2018a. "Anonymized detailed MPN dataset". Received March 2018.

DGT, 2018b. "Anonymized form SPT 1721 datasets". Received March-May 2018.

DGT, 2018c. "Anonymized tax audit dataset". Received June 2018.

DGT , 2018d. "Anonymized tax assessments datasets". Received November 2018.

Organization for Economic Co-operation and Development, 2018. GDP Implicit Price Deflator in Indonesia [IDNGDPDEFAISMEI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/IDNGDPDEFAISMEI, November 27, 2018.