Replication Manual: "Identity verification standards in welfare programs: experimental evidence from India"

August 14, 2020

1 Overview

1.1 Description of content

The replication package contains a directory with four sub-directories and two master analysis files (one for code files written in R, one for in Stata). This manual describes in detail how one can replicate all analysis-based exhibits in the published version of "Identity verification standards in welfare programs: experimental evidence from India." "Analysis-based" exhibits include all tables and figures except for Figure 2 and Figure A.1, which are illustrations instead of results based on underlying data.

Section 4 describes the content in each sub directory in detail. The brief descriptions are as follows:

- *AnalysisCode* contains one Stata do-file or R code file for each exhibit. For example, file "TableA_3" produces Table A.3 in the paper. Note that the individual code files will not run without input from the master files (see below).
- *Compiled_exhibits* includes all the analysis-exhibits in the paper and that one should produce after running the two master files.
- *Output* is initially empty and stores the newly-generated exhibits after one runs the master files. If one re-runs the code, the new output file overwrites the existing file.
- *Data* contains all the data used in the analysis. See details in Section 2.

In addition, the base directory contains the master analysis files that create the computing environment in Stata or R to execute the analysis, a tex file that one can use to compile the newly produced exhibits in a pdf for easy of comparison, and a do file that shows analysis for in-text statistics not calculated from exhibits. .

- *master_analysis_stata.do* is the key file one should open to run all the do files. The user has to set a global macro pointing to this replication directory. After setting additional paths to data and output directories and loading utility functions, one can run each do file separately.
- *master_analysis_R.r* is the key file should open to run the R files. Similarly, the user has to set the root path to this replication directory. After setting paths to data and output folders and installing and loading necessary packages, one can run each R file separately.

1.2 Example

This section explains an example that replicates *Table A.8: Effects on food security*, generated in Stata, and *Table A.1: Representativeness within Jharkhand*, generated in R. Note that replication requires Stata version 13 and above and R version 3.5.3 and above.

1. Download the zip-archive containing replication package (thus all the content described above) and unpack it to, say, "C:/Users/testUser/ABBAReplication/."

2. Open the *master_analysis_Stata.do* file in Stata and enter the base directory to where the extracted contents of the zip-file are (such as "C:/Users/testUser/ABBAReplication"in this case [1]) in line 31 after "glo root". This path becomes the "root" path, and then set the paths for data, output, and code and load the programs by running lines 36 through 50.

3. Similary, open the *master_analysis_R.r* file in R and enter the base directory to where the extracted contents of the zip-file are (such as "C:/Users/testUser/ABBAReplication/"in this case) in line 13 after "root". This path becomes the "root" path, and then set the paths for data, output, and code and load the packages by running lines 15 through 64 (commands to install packages are original commented out to avoid issues with restarting R if one already has installed these packages).

4. To produce all the exhibits, run all the code files by executing the two master files.

5. To produce Table A.8, open TableA_8.do in Stata and execute the code. Alternatively, run the respective line of code in the master do file that executes TableA_8.do.

6. To produce Table A.1, open TableA_1.R in R and execute the code. Alternatively, run the respective line of code in the master do file that executes TableA_1.R.

7. Check the output in the *Output* folder.

    - One can also run the *To_compile.tex* file to compile all the exhibits from the *Output* folder into a pdf and then compare with the exhibits in draft (except for Figures 1 and A.2, which are illustrations)

2 Data

2.1 Description

The folder *Data* contains two subfolders: *Admin*, which stores cleaned data we obtained from the government's administrative records or publicly available data such as Census data, and *Survey*, which includes cleaned data from our household and dealer surveys. The data set *adherence_offtake_compare.dta* uses both administrative records and survey data and thus is a stand-alone file in the *Data* folder. The details of each data set are in Section 4.

2.2 Identifiers

We have removed all the original identifiers from the data that could identify individuals or small groups of individuals. We replaced these identifiers with randomly-generated ID - variables that replicate the original structure.

There are some key ID-variables which are used in almost all of the analysis:

- "rationcar_code" is the unique ration card id for the administrative beneficiary lists that contain the whole universe of PDS beneficiaries in the 10 study districts.
- "uid" is the unique ration card id for the 3960 sampled households. This id variable is present in all the household survey data sets.
- "fps_uid" is the unique id for the sampled fair price shops and is present in the dealer survey data sets.
- "admin_dealer_id" identifies the individual fair price shops from the administrative data obtained from the National Informatics Commission.

[1] Note the lack of trailing slash "/"

- "block_code" is the block-level unique id for the 132 study blocks and used in clustering the standard errors in most specifications.
- "district_code" is the district-level unique id for the 10 study districts.
- "Strata" is the Strata id and used to implement randomization fixed effects in most specifications.

3 Code

3.1 Structure

The folder *AnalysisCode* contains an individually code file for each exhibit, except for Figure 3 and Figure A.4, each of which has two code files with one for each panel. The name of the file corresponds to a table or figure in the paper. All files are written in a way that once the Stata or R environment is set - as described in section 1.2 - one can work with them independently. Each code file saves a table or figure in the *Output* folder. The *Compiled_exhibits* folder contains the exhibits used in paper and should match with those from the the *Output* folder. The *in_text_numbers.do* file executes the analysis for in-text numbers not available in exhibits.

The *AnalysisCode* folder also includes functions written by research assistants who worked on this project in *Lib* folder with two subffolders:

- *Ado*: ado files that create the multi-panel tables such as Table 2.
- *Functions*: user-written Stata programs that automatize parts of the analysis or provide better visual representation.

Overall, the *master_analysis_Stata.do* will take around 30-35 minutes to run, and the *master_analysis_R.r* will take around 5 minutes to run - the prediction based on an i7-processor with 16 GB RAM.

3.2 External packages

Throughout the analysis, we used a set of external packages which can be downloaded from SSC in Stata and CRAN/github in R. [2] We give a brief overview over the packages we use here:

Stata:

- *estout*: we use the "esttab" function to output most of our tables.
- *texsave*: To generate Table 12, we use this command that outputs the Stata data set itself as a tex file with a large number of formatting and customization options
- *dsconcat*: To calculate false discovery rate-adjusted q- values reported in brackets in many tables, we use this function to create a new data set in memory that concatenates the regression results and prepare for multiple hypothesis testing.
- *qqvalue*: We use function "qqvalue" from this package to calculate the q- values from the p-values stored in dataset prepared using the "dsconcat" function
- *dm79*: We use function "svmat2" to assist the formatting for Table 9 by storing the columns of a matrix as new variables

R:

- *ggplot2*: To plot most of visuals such as time-series graphs

[2] Packages in Stata can be downloaded using the following command: ssc install packagename. All packages but "starpolishr", "rgeos", and "rgdal" in R are downloaded by calling (incl. the quotation marks): install.packages("packagename"). "Starpolishr"can be downloaded by calling: devtools::install_github("ChandlerLutz/starpolishr"), and rgeos and rgdal can be downloaded by calling: install.packages("rgeos", type="source") and install.packages("rgdal", type="source").

- *stargazer*:To organize output regression results to tex files
- *starpolishr*: To enchance aesthestics of output from *stargazer*
- *RColorBrewer*, *colorspace*, and *crayon*: To provide various color palettes for plots
- *scales*: To create custom scales in figures
- *ggpubr*: To combine plots produced from *ggplot2* in customized positions and dimensions
- *foreign*: To load data files of different formats
- *readstata13*: To read and write Stata format files
- *xlsx*: To read and write xlsx format files
- *dplyr* and *plyr*: To manipulate data easily such as selecting subsets
- *maptools*, *rgdal*, *gpclib* and *rgeos*: To manipulate geographic data and produce Figure 1
- *plm*: To describe a panel data with an index attribute that describes its individual and time dimensions withh "pdata.frame" function to produce Figure C.1
- *reshape2*: To reconstruct data using "melt" function

3.3 User-written functions and programs

In order to automatize parts of the analysis or provide better visual representation, we defined some Stata programs and ado files which are repeatedly called. These are included in separate do and ado-files in the *Lib* directory under *AnalysisCode*.

Functions:

- *AttritionTable_program.do*: Given a list of variables of interest, this program creates a table with four columns (treatment control mean, regression-adjusted difference p-value) and is used for testing balance or attrition by treatment status. Additional arguments to the program are the arguments for strata fixed effect and clustered standard errors.
- *AttritionTable_program_svy.do*: This is the same as *AttritionTable_program.do* but takes into account the inverse sampling probability (sampling weights).
- *Balance_program.do*: Similar to *AttritionTable_program_svy.do* but one can specify the balance variable. This is used to compare sampled vs. new dealers in Table A.1 only.
- *StackingRegressionModels.do*: Given a set of regressions with a common dependent variable, this program stacks the regressions output in order to display a particular coefficient from each regression in a single cell.
- *Preferences_Perceptions_program.do*: This program formats Table 9 for the paper.

Ado:

- *MultiPartTabStart.ado* and *MultiPartTabEnd.ado* provide the header and ending line for most multi-panel tables
- *MultiPartTabPanelStart.ado* and *MultiPartTabPanelEnd.ado* provide formatting for most panels
- *MultiPartTabPanelStartEntitle.ado* and *MultiPartTabPanelEndEntitle.ado* provide formatting for the mean entitlement row such as that in Table 3
- *MultiPartTabPanel*Plain.ado* and *MultiPartTabPanel*noobs.ado* provide formatting for panels without the additional summary statistics and are only used for Table 2.

4 Detailed list of content

We describe the contents of every folder contained in the replication zip-folder:

- *AnalysisCode*: Includes all individual code-files that produce all exhibits in the paper except Figure 2 and Figure A.1, which are illustrations and not based on analysis data. See Section 3.3 for more details on the *Lib* subfolder.
- *Exhibits_compiled*: Includes all exhibits except Figure 2 and Figure A.1 in the paper. These should match with the newly-produced exhibits in the *Output* folder.
- *Output*: Initially empty, this folder will be populated as one runs each individual code file or the master files.
- *To_compile.tex*: A tex file with the preamble prepared to compile the exhibits from the *Output* folder into a pdf for easier comparison with exhibits in paper.
- *master_analysis_Stata.do*: A master do file that sets the environment for and can execute all individual do files in *AnalysisCode*.
- *master_analysis_R.r*: A master R file that sets the environment for and can execute all individual R code files in *AnalysisCode*.
- *Data*: Contains all data files used in the analysis.

  - adherence_offtake_compare.dta: Ration offtake quantity as reported by sampled households and recorded in ePOS machines; used to produce Figure C.2 Panel B

  Admin:

  - The *Map* subfolder: Includes geospatial files used in creating Figure 1 - the randomization map
  - adherence_disbursementstock_compare.csv: Cleaned Fair price shop-level disbursement and physical stock data for creating Figure C.1 and Figure C.2 Panel B
  - adherence_targetdisbursement.dta: Fair price shop-level comparison of actual disbursement and counterfactual disbursement calculated by assuming strict adherence to the reconciliation policy; used to plot Figure C.3
  - allocation_*level_*: Fair price shop or block-level disbursement data provided by the National Informatics Commission
  - BL_EL1_beneficiary_turnover_RC_BL_EL1: Rationcard - level beneficiary turnover data from May 2016 to October 2016
  - BLdisbursementdata.dta: Fair price shop-level disbursement data at baseline; used in creating baseline variables in most disbursement analysis
  - block_randomization_charas.dta: Block-level characteristics used in checking balance between treatment and control blocks in Table 2
  - blockLevel_strata.dta: Block-level information on strata; used in merging with main analysis files for conducting strata fixed effects
  - deletions_admin.dta: Rationcard-level status of card deletions in the 10 study districts across multiple periods

- dis_wheat_notzero_ind_fps.dta: Fair price shop-level indicators for whether the shop had been receiving wheat disbursement starting in January 2017; used to impute missing wheat stock balance values in Table A.10
- District_comparison.csv: District-level characteristics obtained from 2011 and 2001 Censuses, the PDS beneficiary list prior to baseline, and the 68^{th} Round of the National Sample Survey (NSS 68) to check for balance between study vs. non study districts in Table 1
- fps_allocation_2016_backups.csv: Fair price shop-level allocation data in 2016
- fps_allocations_upto21July.csv: Fair price shop-level allocation data up to July 2017
- FPS_sampling_prob.dta: Fair price shop-level sampling probability for each sampled shop
- JH_ePOS_HH_AadhaarStatus.dta: Ration card-level Aadhaar seeding status for the 10 study districts
- NICstock_cleaned_fpslevel.dta: Cleaned fair price shop-level stock data
- Statutory_Kerosene_Prices.dta: Monthly statutory kerosene prices from August 2016 to Novembet 2017
- total*_value_meandisbursement_reconciliation_time_series*.dta: Monthly average value of reconciled or unreconciled commodities disbursed to shops as predicted by the reconciliation specification (Equation 2 in paper); "treatment1" or "treatment0" suffix means the fitted values for treatment or control, respectively
- transaction_jun17: Fair price shop-level machine mode most frequently used for transactions in June 2017; only has records of dealers that conducted at least one transaction in that month

Survey

- deletions_svy: Rationcard - level status of card deletions for sampled households
- JH_ePOS_HH_DataforAnalysis: Cleaned household survey data at endline one with baseline variables needed for analysis
- JH_ePOS_HH_reconciliation_DataforAnalysis: Cleaned household survey data for endlines one, two, and three excluding endline one variables not used in reconciliation analysis
- JHePOS_BL_Dealer_data: Cleaned dealer survey data at baseline
- JHePOS_EL1_Dealer_data: Cleaned dealer survey data at endline one
- Single_BL_HH_Dataset_For_Validation: Cleaned household survey data at baseline with all survey responses
- Single_EL_HH_Dataset_For_Validation: Cleaned household survey data at baseline with all survey responses
- Single_FPS_Dataset_For_Validation: Cleaned dealer survey data at baseline with all survey responses
- Single_FPS_Dealer_Dataset_For_Validation: Cleaned dealer survey data at endline one with all survey responses
- total*_value_meanreceipt_reconciliation_time_series*.dta: Monthly average value of reconciled or unreconciled commodities received by sampled households as predicted by the reconciliation specification (Equation 2 in paper); "treatment1" or "treatment0" suffix means the fitted values for treatment or control, respectively