Replication manual: “Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India”

1. Overview

This package contains replication data for: "Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India" It contains analysis data from an RCT evaluating the impacts of a police reform, Women’s Help Desks (WHDs), in Madhya Pradesh, India. Also included in the package are the survey instruments used in the police, user, and citizen surveys, as well as the Stata code necessary to replicate the tables and figures from the paper and its supplement. For further details on the data or how to run the code, please see the readme file.

The abstract of the associated paper is as follows:
Gender-targeted police reforms are frequently proposed to tackle the global problem of rising yet under-reported gender-based violence (GBV) – but with mixed and often disappointing results. We explore this issue in India, a country with alarming rates of GBV and limited police capacity, by studying the impact of Women’s Help Desks (WHDs): dedicated spaces for women in local police stations, staffed by trained officers. Drawing on the largest randomized controlled trial of a police reform to date (180 police stations serving 23.4 million people), we find that officers in stations with WHDs are more likely to register cases of GBV, particularly where female officers run the desks. This suggests that, even in resource constrained and patriarchal environments, police responsiveness can be improved by focusing and mainstreaming attention to women’s cases and by greater gender representation within the police.

1.1. Folder structure:

- Data
    - Admin
    - CCTV
    - Citizen survey
    - Police Survey
    - Randomization_heterogeneity.dta
    - User survey
- Master.do
- Output:
    - Main_paper
        - Figures
        - Tables
    - Supplement
        - Tables
- Readme.pdf
- Scripts
    - Heterogeneity interactions.do
    - Main_paper
    - Supplement
- Survey questionnaires
    - Baseline

- Endline

1.2. Description of content

The replication archive contains a directory with 4 main sub-directories, one each for the data, output, scripts, and survey instruments used in the study. The “Data” is broken down into 5 sub-folders described in more detail below. The “Output” sub-directory stores figures and tables similar to the ones in the published version of “Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India.” Each script in the “Scripts” folder replicates exactly one table panel, table, or figure from the paper, and the scripts are organized by paper section (main vs. supplement) and whether they produce a table or figure. Finally, the “Survey questionnaires” folder contains the surveys used to produce the citizen, police, and user survey data.

1.3. Replication instructions

1. Open master.do and follow the instructions below (also in master.do)

2. Set the path to the top folder in the lines marked "SET DIRECTORY HERE" (Line 19)

3 Set the supplement global on line 22 to either 1, if you would like to replicate the supplemental tables, or 0, if not.

4. Run master.do

Note: The *heterogeneity interactions* do-file contains a part of code that is used to interact the treatment indicator with heterogeneity indices, including the police training score, community outreach strength, and a score for implementation quality. This do-file is often incorporated into other analysis do-files before regressions are run.

2. Code

2.1. Description

The do-files are written such that once the environment is set – as described in section 1.2 – one can work with them independently. All tables and figures are exported to the “Output” directory.

2.2. External package used

Throughout the analysis, the only external package we used is *esttab*, which provides commands such as *eststo*- and *esttab*- to create publication-style tables. The *esttab* package can be downloaded from SSC in Stata.

3. Data

The package contains data from five sources collected between November 2018 and December 2020: i) administrative data on crimes registered by the police; ii) CCTV data from the video feeds of cameras, present in all police stations, focused on the station entrance; iii) A user survey of members of the public who had visited study police stations, asking about their satisfaction with their visit; iv) A police survey, carried out at baseline and endline, of personnel in different roles and ranks in study stations; and v) A survey of citizens, carried out at baseline and endline on perceptions of safety, opinions of and contact with the police, and experiences of crimes. Additionally, we provide a dataset with police stations' fixed characteristics that are repeatedly included in the analysis, such as station code, district-sector code, the classification of treatment and control groups, and the first principal component score.

3.1. Description

With respect to the police survey and CCTV data, there are two sets of data provided: one contains all observations from baseline and the other contains observations with a corresponding entry at endline. The dataset with only baseline outcomes is named "[data type]_baseline data.dta". The dataset combining both baseline and endline outcomes is named "[data type]_full data.dta".

Particularly for the police survey, there is an additional set of data that contains information about the numbers of officers and all staff members at each station.

For administrative data, the analysis makes use of two data sets: one in the long format with data from May 2018 to May 2020; the other in the wide format with data from May 2018 to March 2019 (baseline), then from May 2019 to March 2020 (endline).

3.2. Identifiers

The data sets used for the paper have information such as administrative details of police stations and demographic characteristics of individuals that could potentially be used to identify a particular subject. Therefore, we have removed all the original identifiers from the data.

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

- ps_code is police station ID
- dist_id is district ID used as fixed effects in regressions with sector heterogeneity
- dist_urban is district-sector ID used to implement strata fixed effects in most specifications

There are also variables indicating the kind of treatment implemented at a police station, including:

- treatment: dummy variable (treatment group = 1, control group = 0)
- group: string variable showing the classification of treatment and control groups
- regular_whd: dummy variable (treatment arm 1, or regular WHD = 1, others = 0)
- women_whd dummy variable (treatment arm 2, or women-run WHD = 1, others = 0)

3.3. Detailed content of each data set

3.3.1. Randomization and heterogeneity.dta

The data set is at the station level. In addition to the identifiers mentioned in the previous section, this data set contains the following variables.

- total_assigned_officers: number of police officers (note: this variable is not the same as the staff data provided in the "Police survey" directory)
- total_fir_2017: number of FIR cases filed in 2017
- population: population served by the police station
- strat_pca: first principal component score
- urban: dummy variable (urban area = 1, rural area = 0)
- implement_quality: implementation quality index
- training_score: police training score
- comm_outreach_strength: community outreach strength index

3.3.2. Administrative data

Admin_long data.dta: data is at the station-month-year level. In addition to the identifiers, randomization and heterogeneity indicators, this dataset contains the following variables.

- fir_overall_count: number First Information Reports filed in all types of crime
- fir_caw_count: number of First Information Reports filed in CAW cases
- fir_bywomen_count: number of First Information Reports filed by women
- arrest_count: number of arrests made
- dir_count: number of Domestic Incidence Reports filed
- ncr_count: number of Non-cognizable Offence Reports filed
- dial100_count: number of Dial-100 calls
- month: month-year indicator

Admin_wide data.dta: data is at the station-month level with similar variables used in the long format of the data. The variables with the prefix "b_" are from baseline and "e_" from

endline. For example, b_dir_count refers to the number of Domestic Incidence Reports filed at baseline, and e_dir_count refers to the number of Domestic Incidence Reports filed at endline.

3.3.3. Police survey

Police_baseline data.dta: data is at the individual level. In addition to the identifiers, randomization and heterogeneity indicators, this dataset contains the following variables.

- b_uid: respondent ID at baseline
- uid: corresponding ID at endline
- gender: officer's gender
- b_wcase: baseline index for police's attention to women's cases (e_wcase is the corresponding variable at endline)
- b_effective: baseline index for perception of police's effectiveness in dealing with cases related to women (e_effectiveness is the corresponding variable at endline)
- b_pol_impt: baseline index for perceptions of tasks most important to the officer
- b_thana_impt: baseline index for perceptions of tasks most important to the station
- b_helpful: baseline index for perceptions of police's helpfulness
- b_sensitivity: baseline index for perceptions of police's sensitivity in dealing with cases related to women
- b_add_officer: baseline index corresponding to the survey question "Will hiring more officers make police more effective in women's cases?"
- b_add_female: baseline index corresponding to the survey question "Will hiring more female officers make police more effective in women's cases"
- b_female_better: baseline index corresponding to the survey question "Who is more effective in handling cases related to women?"

Police_full data.dta: data is at the individual level. The list of variables includes the identifiers, randomization and heterogeneity indicators similar to other data sets. Some variables with the prefix "e_" such as e_wcase, e_effective, etc. are the endline counterparts of the variables previously defined in the baseline version of the data. There are also variables reported uniquely at endline, which are named below.

- e_taken_seriously: index for the perceptions of whether female officers' opinions are taken seriously in the workplace
- e_prof_dev: index for the perceptions of workplace support in terms of professional development
- e_work_help: index for the perceptions of workplace support in terms of fulfilling duties at home

Police_personnel data.dta: data is at the station level. In addition to the identifiers, randomization and heterogeneity indicators, the data set contains the following variables.

- b_total_staff: count of all staff members in a station at baseline (e_total_staff is the corresponding variable at endline)
- b_female_staff: count of all female staff members in a station at baseline (e_female_staff is the corresponding variable at endline)
- b_male_staff: count of all male staff members in a station at endline(e_male_staff is the corresponding variable at endline)
- b_total_officers: count of all officers with rank ASI (Assistant sub-inspector) or higher (e_total_officers is the corresponding variable at endline)
- b_female_officers: count of all female officers with rank ASI (Assistant sub-inspector) or higher (e_female_officers is the corresponding variable at endline)
- b_male_officers: count of all male officers with rank ASI (Assistant sub-inspector) or higher (e_male_officers is the corresponding variable at endline)
- b_total_sampled: count of all staff members that were sampled (e_total_sampled_ is the corresponding variable at endline)
- b_female_sampled: count of female staff members that were sampled (e_female_sampled is the corresponding variable at endline)
- b_male_sampled: count of male staff members that were sampled (e_male_sampled is the corresponding variable at endline)
- b_total_surveyed: count of all staff members that were surveyed (e_total_surveyed is the corresponding variable at endline)
- b_female_surveyed: count of female staff members that were surveyed (e_female_surveyed is the corresponding variable at endline)
- b_male_surveyed: count of male staff members that were surveyed (e_male_surveyed is the corresponding variable at endline)
- b_male_weight: sampling weight for male officers surveyed (e_male_weight is the corresponding variable at endline)
- b_female_weight: sampling weight for female officers surveyed (e_female_weight is the corresponding variable at endline)

3.3.4. User satisfaction survey

User survey_endline data.dta: data is at the individual level. In addition to the identifiers, randomization and heterogeneity indicators, the data set contains the following variables.

- uid: respondent ID
- consented: consent to be surveyed (yes = 1, no = 0)
- gender: respondent's gender
- visitsats: respondent's level of satisfaction with their visit

- comfort: index for whether the respondent felt comfortable discussing their issues with the officer
- respect: index for respondent' perception of the level of respect they received
- resolution: index for respondent's expectation of a resolution
- fclitysats: index for respondent's level of satisfaction with the facilities available at the station

3.3.5. Citizen survey

Citizen_caw rates.dta: data is at the station level. In addition to the identifiers, randomization and heterogeneity indicators, the data set contains the following variables.

- b_caw_rate: CAW rate corresponding to a police station at baseline
- e_caw_rate: CAW rate corresponding to a police station at endline

Citizen_full data.dta: data is at the household level. Each household has one respondent at baseline and one at endline. In a few observations, the survey respondent at baseline was replaced with someone else in the same household at endline, leading to differences in the gender input. In addition to the identifiers, randomization and heterogeneity indicators, the data set contains the following variables.

- hh_id: household ID
- member_gender: gender of the respondent at baseline
- e_gender: gender of the respondent at endline
- b_visit: whether a person visited a station between January 2018 to before the baseline survey
- e_visit: whether a person visited a station between the baseline survey and endline survey
- b_pol_handling: index for perceptions of police's effectiveness in dealing with CAW (e_pol_handling is the corresponding variable at endline)
- b_safety: index for the level of safety a person felt at home, in their neighborhood and in the crowd (e_safety is the corresponding variable at endline)
- e_urja_knowledge
- spw: sampling weight
- attrited: dummy variable that turns one either when a household has no survey respondent at endline or when the baseline person was replaced at endline by someone from the opposite gender.

3.3.6. CCTV data

Cctv_baseline data.dta: data is at the station-day-shift level. In addition to the identifiers, randomization and heterogeneity indicators, the data set contains the following variables.

- dayofweek: day of week indictor (Monday to Sunday)
- timeofday: shift indicator (Morning, afternoon, and evening)
- bsum_all_duration: total footage length for all observations with valid visitor counts (after excluding extreme values)
- bavg_all: average number of visitors in a day-shift
- bavg_all_pixels: average pixel value from all footage with valid visitor counts (after excluding extreme values)
- bavg_all_fps: average frame per second value from all footage with valid visitor counts (after excluding extreme values)
- bsum_women_duration: total footage length for all observations with valid female visitor counts (after excluding extreme values)
- bavg_women: average number of female visitors in a day-shift
- bavg_women_pixels: average pixel value from all footage with valid female visitor counts (after excluding extreme values)
- bavg_women_fps: frame per second value from all footage with valid female visitor counts (after excluding extreme values)
- bsum_wprop_duration: total footage length for all observations with valid values of female visitor proportion (after excluding extreme values)
- bavg_wprop: average proportaion of female visitors in a day-shift
- bavg_wprop_pixels: average pixel value from all footage with valid female visitor proportion (after excluding extreme values)
- bavg_wprop_fps: average frame per second value from all footage with valid female visitor proportion (after excluding extreme values)

Cctv_full data.dta: data is at the station-day-shift level. In addition to the identifiers, randomization and heterogeneity indicators, the dataset contains variables with the prefix "e_" which are the endline counterparts of the variables previously defined in the baseline version of the data.

Please cite this data using the citation provided on the dataverse page (https://doi.org/10.7910/DVN/R75XVZ) if you reuse it.