Contents

[Logic for recoding variables for generating indices 1](#_Toc172896774)

[Police Station Facilities Survey 1](#_Toc172896775)

[Baseline Survey 1](#_Toc172896776)

[Randomisation 2](#_Toc172896777)

[Reflection Survey 2](#_Toc172896778)

[Endline Survey 2](#_Toc172896779)

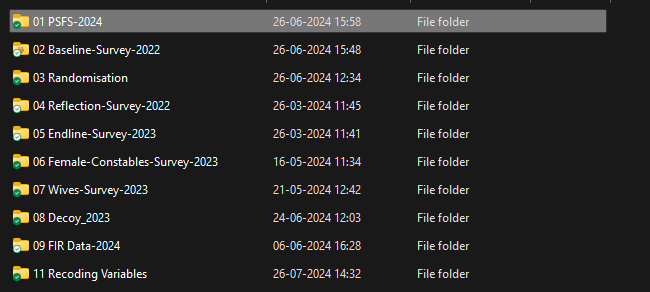
[Female Constables Survey 3](#_Toc172896780)

[Wives Survey 3](#_Toc172896781)

[Decoy Survey 3](#_Toc172896782)

[FIR Data 3](#_Toc172896783)

**Data analysis folder for project:** <https://www.dropbox.com/scl/fo/oe1fne6ybpz26paoz1d1m/AH93ftrnHtDFAyaqGJvwJZ8?rlkey=6h42yrzhjm5wy3zjqcu4uit81&dl=0>



# Logic for recoding variables for generating indices

For generating the indices for the datasets, we have to aggregate multiple variables into a single outcome. We have used the approach of [Jayachandran](../Dropbox/Debiasing%20Police%20in%20India/005-Data-and-analysis-2022/11%20Recoding%20Variables/Jayachandran%20-%20Haryana.pdf) (2018). The logic for creating dummy variables and reversal of some variables is given below:

1. **Primary and secondary outcomes -** <https://www.dropbox.com/scl/fi/jwss2pzrc6s9dg5m4bmtu/01-Officers_Survey_Outcomes_Mechanisms.xlsx?rlkey=ejjnxifyi7nh4v5rv78ud55ov&dl=0>
2. **Outcomes from female constables survey -** <https://www.dropbox.com/scl/fi/1ljr5hf69v99x9qbeyqk2/02-Female-Constables_Outcomes_Mechanisms.xlsx?rlkey=flcyft3y42nrsd01dt92tky28&dl=0>
3. **Outcomes from wives survey -** <https://www.dropbox.com/scl/fi/8v4b2jlgf60mm1oa81dg1/03-Wives_Outcomes_Mechanisms.xlsx?rlkey=cgcmim5pnl3o0wwuzw9j1t2j0&dl=0>
4. **Outcomes from decoy survey -** <https://www.dropbox.com/scl/fi/d2h1wd2q8uq4s3pwigb92/04-Decoy_Survey_Outcomes_Mechanisms.xlsx?rlkey=0ymc5xotljnywhzntesqjbhbt&dl=0>
5. **Indices for PSFS -** <https://www.dropbox.com/scl/fi/c90ylgk3eivd2kdab9mb3/05-PSFS_Survey_Indices.xlsx?rlkey=s5c3h5al52d2ytgyxwyy8tik4&dl=0>

# Police Station Facilities Survey

We start by importing and cleaning the PSFS data.

This has variables on officer strength, infrastructure, number of female officers, etc.

**Data:** <https://www.dropbox.com/scl/fi/ztp1hhct9kaaax61cc90j/03.PSFS_clean_deidentified_new.dta?rlkey=7y5jgbv2w35g8s57aahx7sn4j&dl=0>

**Codebook:** <https://www.dropbox.com/scl/fi/5rbt4auvu9jgf9c9d1ux7/Codebook.pdf?rlkey=l7no869c6fv8unfgn1jyl3kds&dl=0>

# Baseline Survey

We clean the baseline data.

We also make a preliminary attempt at creating the baseline indices using the Anderson method.

**Data:** <https://www.dropbox.com/scl/fi/kuihc1if4npu4v1o4bgof/03.officersurvey_clean_deidentified_indices.dta?rlkey=9e667a5h65s03h7bu0c67txjm&dl=0>

# Randomisation

Using the baseline and the PSFS data, we randomise the police stations into treatment and control group. We do this by each district, and then merge them to create a pooled randomisation .DTA file.

**Pooled .dta:** <https://www.dropbox.com/scl/fi/tgdam85zcu4yoy41vag7u/pooled_randomisation.dta?rlkey=3d8r00cqcza739voij6n736de&dl=0>

We also use QGIS to use raster data of population density to plot the average population density in a 10km radius around each PS (we have lat-long coordinates of each PS from the PSFS survey). We classify a police station as ‘Urban’ if the average population density is greater than 2000 in that 10km radius. This is merged with the randomisation .DTA file.

# Reflection Survey

We clean the reflection data for both treatment arms, and append the data to create a combined dataset.

**.DTA file:** <https://www.dropbox.com/scl/fi/m3czqi0sgokymdgljnmom/reflection_clean.dta?rlkey=l19qvigxs7nijbuphes68rhdq&dl=0>

# Endline Survey

* We clean the endline survey following the process used to clean the baseline survey as well.
* We then merge the baseline and endline survey in the following steps:

1. We merge the baseline data with the pooled randomisation .DTA file such that the baseline data now has variables of treatment, strata, population density (PS-level) merged into it. We also merge the PSFS data.
2. We do the same for the endline data.
3. We now merge the baseline and endline data. We do not drop non-merged observations. Now we have a combined dataset (n=4,166) that comprises officers who completed both surveys, officers who completed only baseline, and officers who completed only endline. This dataset contains PS-level variables such as treatment, strata, population density, and PSFS variables.
4. **Do-file:** <https://www.dropbox.com/scl/fi/g92p55kafo5p05efqrrgh/04.merging_endlinesurvey.do?rlkey=cw6v63r2y9f1x86l12ij31l2u&dl=0>

* We now create dummies that we will use for our analysis:

1. Transfers – We create a dummy for transfers, that is equal to 1 if the officer has been transferred anytime between the baseline and endline surveys.
2. Decompose (i) – We decompose the dummy in (i) into four parts: a) transfer from control to control b) transfer from treatment to treatment c) transfer from control to treatment and d) transfer from treatment to control
3. Out of sample – We create dummies for a) transfers from out of sample to control b) transfers from out of sample to treatment c) transfers from control to out of sample and d) transfers from treatment to out of sample

* We now drop all officers who have not completed both the baseline and endline surveys. This leaves us with the ‘clean’ sample of officers who have completed both the surveys (n=2571). We generate indices for the PSFS data, but only for baseline (coding is given [here](../Dropbox/Debiasing%20Police%20in%20India/005-Data-and-analysis-2022/11%20Recoding%20Variables/05%20PSFS_Survey_Indices.xlsx)).
* **Do-file:** <https://www.dropbox.com/scl/fi/rp41avgck3h3ohy8b2y2w/06.data-prep.do?rlkey=tind6xtsv4onssmcmpsdrtlks&dl=0>
* Since dropping the observations changes the means of the indices, we now redo the indices (for both baseline and endline) and normalise for the control group (assigned to the police station of the officer as per the baseline/endline survey).
* **Do-file:** <https://www.dropbox.com/scl/fi/yychwrqemy2tt7egaj0fx/07.final_index_prep.do?rlkey=no5zr27gdzahp2tj7wa8ohrvr&dl=0>
* **Dataset with n=4166 (all officers):** <https://www.dropbox.com/scl/fi/pc6r3t7mrj9951urxf7d1/endline_secondaryoutcomes.dta?rlkey=8iyx1bbgf6rb26as1wu7q99rg&dl=0>
* **Dataset with n=2571 (clean sample):** <https://www.dropbox.com/scl/fi/ndu2tih7a1eixcfxwis0k/combined_FINAL_indices.dta?rlkey=izl31r9dfrzm0q5zhohqsc0ea&dl=0>

# Female Constables Survey

We clean the female constables survey data.

**.DTA file:** <https://www.dropbox.com/scl/fi/h4h1lpcxg5vtcjiuvi23g/femaleconstables_indices.dta?rlkey=hya59shivub18xxqww2z0dtns&dl=0>

We create the indices as per the coding given [here](../Dropbox/Debiasing%20Police%20in%20India/005-Data-and-analysis-2022/11%20Recoding%20Variables/02%20Female%20Constables_Outcomes_Mechanisms.xlsx).

# Wives Survey

We clean the wives’ survey data.

**.DTA file:** <https://www.dropbox.com/scl/fi/umy98cqs9be8e4g1c8e53/wivessurvey_clean.dta?rlkey=wh3o2tyhk0waf77hlmxqxmre8&dl=0>

We create the indices as per the coding given [here](../Dropbox/Debiasing%20Police%20in%20India/005-Data-and-analysis-2022/11%20Recoding%20Variables/03%20Wives_Outcomes_Mechanisms.xlsx).

# Decoy Survey

* We clean the decoy survey data. We also reshape the data to have the data at a PS-level, with data for different visits being suffixed with \_visit1, \_visit2, \_visit3.
* **.DTA file:** <https://www.dropbox.com/scl/fi/19r9sqfuwl5dx1rewi8p5/decoy_indices.dta?rlkey=kw7blvx1x8gmcy6lmpputuy2u&dl=0>
* We create the indices as per the coding given [here](../Dropbox/Debiasing%20Police%20in%20India/005-Data-and-analysis-2022/11%20Recoding%20Variables/04%20Decoy_Survey_Outcomes_Mechanisms.xlsx).
* Since there were three visits to each police station, we create two types of indices:

1. **Visit-wise index** – For example, Empathy\_visit1, Empathy\_visit2, Empathy\_visit3
2. **Combined index** – Combining the variables across all three visits, Empathy\_

# FIR Data

* We scrape FIR data for the police stations in our sample from the SCRB [portal](https://scrb.bihar.gov.in/View_FIR.aspx).
* We run a code using Python to ascertain the gender of the complainant.
* **Do-files:** <https://www.dropbox.com/scl/fo/sm7b7ul87h5sl47jvbcuj/AA3X-SxTM6ugI0VQfIxwdHY?rlkey=yzv4r1qb6bs12beq3jxvj5eku&dl=0>
* There is a variable containing the Sections/Acts under which the FIR was filed. We exploit that to create a dummy variable for GBV crimes.
* We then use the variable containing the time and date of the FIR to collapse the dataset to a panel data form at the station and year-month level.
* **.DTA file:** <https://www.dropbox.com/scl/fi/pquhawwttkcyhxyi71i8e/ps_fir_stationmonth.dta?rlkey=jwjlaq20708x1m8ze6omksj38&dl=0>
* We further collapse the data using the station variable to get another panel data at simply the year-month level.
* **.DTA file:** <https://www.dropbox.com/scl/fi/9f638836opxszsiqw28am/ps_fir_collapsed.dta?rlkey=n9llwqanlhx5gmxzc5p7etp4g&dl=0>
* **Do-file:** <https://www.dropbox.com/scl/fi/lu54b0fszf48iuuwveiv1/02.fir_analysis.do?rlkey=lg4itmaqgvzdeplsawik4l3vm&dl=0>
* **Preliminary graphs:** <https://www.dropbox.com/scl/fo/hqcuqwfcgib86nfkd0nvc/ACsjeC51XR8rQ-u9tsoVzqk?rlkey=s3od6pvu8lgedxyt573pmh32d&dl=0>