**Making Moves Matter**

**Data and code guide**

**August 2018**

Instructions

* If you want to replicate the AER tables, run *2 Master Analysis*.do. The specific code for tables and figures is in the second half of this do file.
* If you also want to recode data from scratch (i.e. from the raw data), run *1 Master Compile* first before running *2 Master analysis*.
* Note you will need a Latex installation (e.g. MikTex) with the standalone.cls installed to make the tables. If the Latex output gives you an error message, just hit enter to continue.
* For a summary of the main analysis and datasets, please refer to the guide below.

Index

1. Sampling
2. Analysis overview
3. Cleaning
4. Main inputs used for analysis
5. Main analysis
6. Simulations
7. Base growth predictions
8. Marginal returns to effort
9. Other relevant folders: ranks and transfers data
10. Randomization inference
11. Notes about some specific datasets
12. **Sampling**

All randomization do files are located in Data\Do Files\Randomization.

Refer to the pdf called 150603TransfersPostingsBallotInfo for a complete description of the sampling procedure followed in Years 1 and 2.

Below is a summary:

1. There are two stages in the randomization: first we allocate circles into groups of 9-11 and then we randomly assign treatment and control status at the group level. Our stratifying variable is recovery at baseline.
2. Our sample consists of 525 circles located in Lahore and Multan. Initially we select 2 groups as pilot circles (~20 circles). For year 1, 115 circles were excluded from the ballot (including pilot circles). These were incorporated later for the ballot in year 2.



1. Treatment groups were randomly allocated to the “demand” or “revenue” scheme.
2. **Overview**

Ma

The main information we need for analysis is:

* Circle-level data on tax collection
* Inspectors’ preferences for circles
* Marginal returns to effort
* Transfers of inspectors from one circle to another after year 1.
1. **Cleaning**

The main cleaning do files can be found in Data/Do Files\Data prep/.

The most important ones are:

* 150525 Analysis Dataset Cleaning - Year 1 FULL SAMPLE & 150610 Analysis Dataset Cleaning Year 2

Merges collection data with treatment assignment for years 1 and 2, respectively.

* Baseline Preferences data prep

Prepares a circle-level dataset with average ranking and proportion of people that ranked each circle as first.

1. **Main inputs used for analysis**

The datasets are located in TP Analysis\Coded Data

* *150525 TP Analysis Data* & *TP Y2 Analysis Data w Y1 Treatment*: circle-level data with gross demand and recovery variables for years 1 and 2. Also includes inspector’s ID and treatment assignment.
* Baseline preferences: ranking of inspectors for each circle within their group (collected at baseline for both treatment and control groups).
* Y1 preferences: ranking of inspectors for each circle within their group (collected after Y1 for treatment group).
* You can find in *Matlab simulations* the simulation files used for analysis of heterogeneity of treatment effect by marginal returns to effort and average equilibrium effort.
1. **Main analysis**
* Basic treatment effect
* Regress collection variables (such as recovery) on its baseline value and treatment assignment.
* We do this for current, arrears and total collection.
* Sample: Y1 (all year 1 circles), Y2 (circles in year 2 that were part of the control group in year 1), Pooled (Y1+Y2).
* Standard errors always clustered by circle.

- Two variants that go in the Appendix:

1. Add more observations for year 2: include the 115 circles that were excluded in year 1 but that were randomized in year 2.
2. Add variables from the balance table: include the four variables as controls.
* Heterogeneity
* Regress the collection variables on treatment assigned interacted with marginal returns to effort / average equilibrium effort.
* Robustness:
1. Consider the equilibrium efforts under alternative cost and utility functions.
2. Consider the equilibrium efforts under alpha +/- 2 standard deviations and check that results are almost identical.
* Allocation
* For this analysis we use property level data, which gives us information about the characteristics of the circles: proportion of properties with residential/commercial use, average expenditure, average amount of bribes, etc.
* We have the ranking that inspectors gave to each circle within their group, so for **analysis 1** we will identify the characteristics that make a circle popular for all treated circles and for those from top inspectors only (columns 1 and 2). For **analysis 2**, we look at allocation instead of preferences.
* We do this for top 1 and 3 inspectors.
* Balance table
* We look for balance in four variables for years 1, 2 and pooled data.
* To get joint significance RI p-values, we follow Young’s version of multiple-equation RI testing, using equation (8) in this paper: <http://personal.lse.ac.uk/YoungA/ChannellingFisher.pdf>
1. **Simulations**
2. Base growth predictions

One important input for the simulations is the prediction of growth in recovery in year 1 (y0). We calculated y0 following a “leave-one-out” approach:

* For each control group *i*, regress performance in year 1 on past performance using as sample every control group different than *i*. Store the betas from all those regressions.
* In that way, each circle in the control group will use for prediction the betas obtained from regressions that excluded them.
* We will randomly assign a control group to each treatment group and we will use as prediction the betas coming from the regression that excludes that particular control group.
1. Average equilibrium effort

Please refer to the file “readme\_simulation\_eqeffort” located at “TP Analysis\simulation\_results” for a complete guide to the procedure followed.

Summary of steps:

* For a given 2\*alpha, calculate the absolute difference between 0.049 (empirical treatment effect) and the average equilibrium effort.
* Find the alpha that minimizes this moment using lsqnonlin function.
* For the moment minimizer alpha found above, we run the equilibrium effort-finder hundreds of times.
* Get the average equilibrium effort.
* Check for uniqueness of alpha and of the equilibrium effort.
1. **Other important folders: ranks and HR data**
* Ranks

The do files used to calculate ranks can be found in Data\Do Files\Data prep.

A brief summary of the steps is below:

* Get the list of circles where each inspector was during Year 1.
* Calculate the time spent by each inspector in every circle.
* Compute the growth (in recovery or demand, depending on the ranking method) of all circles.
* Weight the growth of the circle by the time spent by the inspector in that specific circle (fraction of time).
* Totalize the score corresponding to all the circles where the inspector worked in the year.
* Rank the inspector among those in his group.
* Transfers table

You can find data on where each inspector ended up after Year 1 in Data\Raw Data\FY 14 Transfers. These tables have information on the circles where each inspector was: (i) balloted (ii) working as of the end of year 1 (iii) transferred at the end of year 1 because of the scheme.

1. **Randomization inference (RI)**

You can find in TP Analysis\Coded Data\RI/rand\_code\_iter the input for RI. There are 2,000 “fake” year 1 and year 2 treatment assignments for each circle of our sample. I computed them using the same ballot code. You can find the latest version of the do file in TP Data\Do Files\RI/Ballot.

RI is part of the code that computes, among others, the basic treatment effect and the balance tables.

1. **Notes about some specific datasets**

There are some old coded files that we are using for which there is no corresponding code. The main ones are:

1. Cleaned baseline preferences for simulation OLD

Inspectors that did not rank circles (~6% of our sample) were randomly assigned ranks. This dataset was used as an input for the matlab simulations. Given that there is a “sortseed” missing in the code, we are using the old dataset when running the code that prepared the input simulations.

1. Baseline preferences files: Treatment Baseline Preferences for Analysis complete OLD
2. TP Year 1 Randomization
3. 150622 Complete Treatment List Year 2

To replicate the randomization, we use the original input files. These are located in Data\Randomization\Data for Ballot.