STARS Project Replication Files

This file summarizes the scripts used to produce the results in the paper, "Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools", by Clare Leaver, Owen Ozier, Pieter Serneels, and Andrew Zeitlin.

Setup

All data required to replicate the paper's results are available in the folder `1_data`. Figure and table outputs are written to the folders `2_tex/figures` or `2_tex/tables` respectively; where additional statistics are cited in the text these are noted in the summary table and outputs of the corresponding functions below. To conduct the replication, replicators should create a local copy of the file hierarchy contained in this repository (in particular, this requires the contents of top-level folders `0_scripts`, `1_data`, and `2_tex`, and all subfolders they contain).

Stata scripts in the top-level script folder (`0_scripts`) all define a machine-specific root folder, which should be the parent folder for the repository. Replicators should modify these paths to reflect their own machines prior to replication.

Matlab scripts assume the current working directory is the `0_scripts` folder, as will be the case without user modification if the project build scripts are run from that location on a Linux system.

Software requirements

Project code is written in Stata and in Matlab. Matlab scripts were run using Matlab version R2019b. Stata scripts were run on Stata 64-bit IC/16.1.

Randomization inference code takes substantial time to run, especially in the case of files 2.3 and 2.4 (calculation of RI-based confidence intervals, and to a lesser extent, p-values, is computationally intensive), and so this was run on a Linux cluster (Google Cloud Platform), with the SLURM workload manager used to manage parallel computing resources. This is strongly recommended for full replication.

Instructions

Paper results were produced on the Linux setup described above. On a Linux setup with the SLURM workload manager (our conditions), we recommend running the script `run_all_slurm.sh`, which will reproduce all figures, tables, and numbers quoted in the text. For replicators without this setup, we provide `run_all_alt_linux.sh` for Linux and `run_all_alt_windows.do` for Windows. These alternatives are capable of producing all paper results, but take substantially greater amounts of time to run the paper scripts sequentially.

It is also possible to replicate study results by individually executing the Stata and Matlab scripts described below, following instructions as noted in their individual descriptions.

Because of their greater execution time, two of the analytical scripts below (scripts 2.3 and 2.4, which create learning outcomes estimates and teacher performance metric estimates, Tables 3, 4, and A4) allow parameters to be passed to these by the calling script that tell them which parameters to estimate. Replicators wishing to replicate point estimates but *not* the computationally more costly p-values or (especially) confidence intervals can set the relevant lists of parameters to just `beta` (commenting out `p` and `ci`) as indicated in `run_all_slurm.sh`. On Windows machines, replicators can replicate a subset of results by setting the relevant parameters in the individual analytical files to execute only those parameters that they wish to estimate. Further instructions are provided in the discussion of those scripts that follows.

The replication repository also includes a LaTeX file, `2_tex/Exhibits.tex`, which can be compiled to display all tables and figures produced by the replication code for ease of comparison to the manuscript and appendices.

Datasets

- `Applicants.dta` : An applicant-level dataset, including both qualified and unqualified applicants (the latter have `qualification` of 'np'), and indicators for the districts in which they applied.
- `Applications.dta` (and `.csv`): An application-level dataset, including both qualified and unqualified applicants. Applicants are linked across applications by the variable `appplicant_id` .
- `Assessments.dta` (and `.csv`): Pupil assessment results at the student (`rosterid`) by round (`round`) by subject (`subject`) level. Pupils can be linked to teachers at the round-school-grade-stream-subject level.
- `GradeSubjectStreams.dta` (and `.csv`): Key file for linking teachers to pupils. Unit of observation is the round-schoolid-grade-subject-stream.
- `MarketAdjacencyMatrix.csv` : Symmetric adjacency matrix showing markets (at the district-by-subject-of-qualification level) that are adjacent to one another. Markets are defined as adjacenet if they share the same subject-of-qualification and they are in geographically contiguous districts.
- `tA.dta` (and `.csv`): Market-level realized assignment of Advertised P4P treatment.
- `tA0.dta` (and `.csv`): Set of *feasible* randomizations of Advertised P4P treatment, used in randomization inference.
- `tE.dta` (and `.csv`): School-level realized assignment of Experienced P4P treatment.
- `tE0.dta` (and `.csv`): Set of *feasible* randomizations of Experienced P4P treatment, used in randomization inference.
- `TeacherInputs.csv` : Results of surprise visits to observe teacher inputs into the classroom (4P metric). Unit of observation is the teacher-round-visit (since rounds correspond to study-years 1 and 2 for consistency with definition elsewhere, and there are two visits in round 1).
- `TeacherPanel.dta` (and `.csv`): Panel of incumbent and new-applicant teachers in study schools. Unit of observation is the `teacherid` by `round` .
- `TeacherValueAdded.dta` (and `.csv`): Derived teacher-level dataset containing estimates of value added. Produced by script `STARS_3_1_0_Estimate_TVA.do` .

Paper results and source files

Figure/table number in paper	Underlying output file	Corresponding script
Fig. 1	Manuscript .tex file	No data used
Fig. 2	ttcscore_outcomes.pdf	STARS_2_1_Impacts_Applications.m
Fig. 3a	PlacedTeacherAbility_outcomes.pdf	STARS_2_2_Impacts_Teacher_Types.m
Fig. 3b	PlacedDGsend_outcomes.pdf	STARS_2_2_Impacts_Teacher_Types.m
Fig. 4a	Recruits_TVA_byTx_yr1.pdf	STARS_3_1_2_TVA_by_P4P.m
Fig. 4b	Recruits_TVA_byTx_yr2.pdf	STARS_3_1_2_TVA_by_P4P.m
Table 1	Applications.tex	STARS_1_1_Descriptives_Applications.do
Table 2	Balance.tex	STARS_1_2_Balance.m
Table 3	LearningImpacts.tex	STARS_2_3_Impacts_Learning.m.m
Table 4	4Ps.tex	STARS_2_4_Impacts_4P_Metric.m
Table 5	Retention.tex	STARS_3_3_Teacher_Attrition.do
Table 6	Round2chars.tex	STARS_3_4_Teacher_Post_Chars.do
Fig. A1	Consort.tex	STARS_3_7_Consort_Diagram.do
Fig. A2	HiringRule_TTCscore.pdf	STARS_1_1_Descriptives_Applications.do

Figure/table number in paper	Underlying output file	Corresponding script
Table A2	TeacherInputs_P4P.tex	STARS_3_0_Descriptives_Inputs.do
Table A3	PlacedTeacherChars.tex	STARS_2_2_Impacts_Teacher_Types.m
Table A4	Learning_LM.tex	STARS_2_3_Impacts_Learning.m.m
Table A5	TeacherSatisfaction.tex	STARS_3_5_TeacherSatisfaction.m
Table A6	P4P_Opinion.tex	STARS_3_6_P4P_Opinion.do
Table C1	AppImpacts.tex	STARS_2_1_Impacts_Applications.m
Table C2	ApplicationSpillovers.tex	STARS_3_2_2_Application_Spillovers.m
Table D1	TVA_PWcorr.tex	STARS_3_1_1_Descriptives_TVA.do
Manuscript text: Rwanda labor market descriptives	-	STARS_1_3_Descriptives_NISR_LFS.do
Manuscript text: Application patterns across space descriptives	-	STARS_3_2_1_Application_Descriptives.do

Below, we outline each of the analytical scripts used in the project. For each script, we list its inputs (dependencies), which may take the form of datasets, output table templates, or 'subfunctions' (functional programs written for use across several of the analytical scripts below). Outputs corresponding to each script may either be a table/figure noted above, or a number provided in the text of the paper that does not appear in a figure or table.

STARS_1_1_Descriptives_Applications.do

Do file providing descriptive statistics on applications received and figure of empirical hiring rule.

Inputs:

Subfunctions

- `SetFilePaths.do` (sets relative file paths for the user)
- `ImportTxA.ado` (imports and cleans Advertised treatment assignment)
- `mattex.ado` (writes matrix to tex file)
- `graphexportpdf.ado` (by Gabriel Rossman, made available in the replication package. Writes graph to pdf on Linux systems.)

Data

- `tA.csv`
- `applications.dta`

Outputs:

- Table 1 (Tables/applications.tex)
- Text in Section II.A describing numbers of applications (with and without TTC qualifications) and unique applicants (with and without TTC qualifications).
- Fig. A2 (Figures/HiringRule_TTCScore.pdf)

STARS_1_2_Balance.m

Balance and descriptive stats for second-stage randomization.

As with other Matlab scripts that use our RI function, `ri_ci.m`, this allows users some options. The `RunParallel` option, if set to `true`, uses Matlab's parallel pool when computing RI p-values. This is the current setting. The `Noisily` option, if set to `true`, displays a progress bar and other diagnostic outputs during the RI calculation. This is set to `true` for Windows but `false` for Unix in the scripts as provided. The parameter `P` controls the number of permutations of the randomization to consider in RI; it is set to 2000 here, but could be set to a smaller number in order to save computational time if replicating point estimates. These values do not need to be changed by replicators.

Inputs

Subfunctions

- `ri_ci.m`. Wrapper function to conduct randomization inference.
- `SetFilePaths.m`
- `ImportTxA.m`. Function to import realized randomization of Advertised treatment status.
- `ImportTXA0.m`. Function to import feasible randomizations of Advertised treatment status.
- `ImportTxE.m`. Function to import realized randomization of Experienced treatment status.
- `PrepTeachers.m`
- `PrepTests.m`
- `LeftJoin.m`. Function to perform a left join while retaining sort order from left table.

Data

- `tA.csv`.
- `tA0.csv`.
- `tE.csv`.
- `tE0.csv`.

Outputs

- Tables/Balance.tex

STARS_1_3_Descriptives_NISR_LFS.do

Inputs

Data

- Government of Rwanda National Labor Force Survey (2017) data, freely downloaded (with registration required) from https://microdata.statistics.gov.rw/index.php/catalog/81. Reference ID RWA-NISR-RLFS-2017-v0.1.

Outputs

Employment rates and wages for those with TTC qualifications. Numbers cited in FN 1 and in final paragraph of Section I.A:

- percent of TTC graduates in teaching and non-teaching jobs
- average monthly salary of TTC graduates in teaching and non-teaching jobs.

The numbers quoted can be seen in the log file generated for this script generated by the parent/calling script for the paper.

STARS_2_1_Impacts_Applications.m

Script analyzing impacts of advertised treatment on applications. Convenience switches `switch_TestTx` and `switch_DescriptiveGraphs`, when set to value `true`, will produce, respectively, (a) a table of application impacts, and (b) an illustrative figure of TTC score distributions. Both are set to `true` in the versions provided and need not be changed by replicators.

Inputs

Subfunctions:

- `ri_ci.m`
- `RItable.m` to write tables with RI outputs.
- `SetFilePaths.m`
- `ImportTxA.m`
- `ImportTxA0.m`

Data

- `applications.csv`
- `tA.csv`
- `tA0.csv`

Outputs

- Table C1 (`Tables/AppImpacts.tex`)
- Figure 2 (`Figures/ttcscore_outcomes.pdf`)

STARS_2_2_Impacts_Teacher_Types.m

Matlab script to analyze impacts of Advertised P4P on attributes of recruited teachers at baseline. Switches `switch_PlotTeacherSkill`, `switch_PlotTeacherMotivation`, and `switch_TestAllChars`, all set to `true` in the script as provided, are convenience switches that allow replicators to choose which of the three outputs are produced. Point estimates and p-values referred to in Figure notes for Figures 3a and 3b refer to those produced in Table A3.

Inputs:

Subfunctions:

- `SetFilePaths.m`
- `ri_ci.m`
- `RItable.m`
- `leftjoin.m`
- `ImportTxA.m`
- `ImportTxA0.m`

Data:

- `TeacherPanel.csv`
- `tA.csv`
- `tA0.csv`

Outputs:

- Table A3 (`Tables/PlacedTeacherChars.tex`)
- Figure 3a (`Figures/PlacedTeacherAbility_outcomes.pdf`)

- Figure 3b (`Figures/PlacedDGsend_outcomes.pdf`)

STARS_2_3_Impacts_Learning.m

Matlab script to estimate learning-outcome effects of treatments. Because this file is computationally expensive to run (when applied to the linear mixed effects model in particular), we run it in parallel. Individual parameters are estimated in separate calls to this script, and randomization inference is conducted in parallel within each call to this script. The preferred method of doing so is within the SLURM script for the project as a whole; no modifications to parameters are required to produce the tables here.

On Linux, this script can be run as a standalone by commenting out other sections of the calling files, `run_all_slurm.sh` or `run_all_alt_linux.sh`. On Windows, this script can be run as a standalone either from the calling file `run_all_alt_windows.do`, or by running the Matlab script directly. Note that, in this last method (only), replicators must manually toggle the definition of the variable `estimator` between values `lme` and `lm` (line 35) to produce either the Linear Mixed Effects model of Table 3 or the Linear Model (OLS) of Table A4. No modification of this variable is required if using the calling scripts.

Parallelization is achieved by first copying the table template to the `2_tex/tables/` directory under the appropriate target table name, and then by overwriting individual cells with empirical results as they are produced. All of the project-build scripts provided undertake this copying as needed without further intervention required by the replicator. A replicator wishing to produce the results of this script by running the .m file interactively should first ensure that the table template has been copied to the relevant target files as described below.

Inputs

Table template

- `0_scripts/m/LearningImpacts_Template.tex`, which is copied to `2_tex/tables/LearningEffects_LME.tex` for LME model or to `2_tex/tables/LearningEffects_LM.tex`, for OLS estimates.

Subfunctions

- `ri_ci`
- `ri_estimate`
- `RItable.m`
- `leftjoin`
- `SetFilePaths.m`
- `ImportTxA.m`
- `ImportTxA0.m`
- `ImportTxE0.m`
- `PrepTests`

Data

- `tA.csv`
- `tE.csv`
- `tA0.csv`
- `tE0.csv`
- `TeacherPanel.csv`
- `Assessments.csv`

Outputs

- Table 3 (Tables/LearningImpacts_LME.tex)
- Table A4 (Tables/LearningImpacts_LM.tex)

STARS_2_4_Impacts_4P_Metric.m

Impacts on the teacher performance metric and its components. Use of this script requires the replicator to first copy a template table, `4Ps_template.tex` to the location of the target output file, as described below. This is done without additional user intervention required in the project calling scripts provided, but should be done manually by any replicator wishing to replicate the results of this script by running the .m file from within Matlab interactively.

To take advantage of parallelization with the SLURM workload manager, the preamble to this script defines four variables (`models` , `outcomes` , `rhsvars` , and `stats`) which determine the specific elements of the output table to be estimated. In Linux, these are passed to Matlab as environmental variables; a subset of these outputs to be estimated may be chosen in the calling scripts. On Windows, they may be toggled directly from within the .m file, lines 27--30. No modification is required of replicators using any of the project calling scripts provided.

Inputs

Table template

- `0_scripts/m/4Ps_template.tex` , which is copied to `2_tex/tables/4Ps.tex` .

Subroutines

- `PrepTeachers.m`
- `PrepTests.m`
- `BarlevyNeal.m`
- `ri_ci.m`
- `ri_estimates.m`
- `RItable.m`
- `leftjoin.m`
- `SetFilePaths.m`
- `ImportTxA.m`
- `ImportTxE.m`

Data

- `tA.csv`
- `tE.csv`
- `TeacherInputs.csv`
- `Assessments.csv`

Outputs

- Tables/4Ps.tex

STARS_3_0_Descriptives_Inputs.do

Descriptive statistics on teacher inputs into the classroom.

Inputs:

Subfunctions:

- `mattex.ado`

Data:

- `TeacherPanel.dta`

Outputs

- Tables/TeacherInputs_P4P.tex

STARS_3_1_0_Estimate_TVA.do

Estimate teacher value added, producing TVA estimates that are used as inputs in 3.1.1 and 3.1.2 below.

Inputs:

Subfunctions:

- `SetFilePaths.do`

Data

- `TeacherPanel.dta`
- `GradeSubjectStreams.dta`
- `Assessments.dta`

Outputs:

- `TeacherValueAdded.dta` (and `.csv`).

STARS_3_1_1_Descriptives_TVA.do

Descriptive statistics on teacher value added.

Inputs:

Subfunctions:

- `SetFilePaths.do`

Data

- `TeacherPanel.dta`
- `TeacherValueAdded.dta`

Outputs:

- Table D1 (Tables/tva_pwcorr.tex)

STARS_3_1_2_TVA_by_P4P.m

Plots of TVA by treatment arm in years 1 and 2; tests for differences across arm.

Inputs

Subroutines

- `ri_ci.m`
- `RItable.m`
- `SetFilePaths.m`

Data

- `TeacherPanel.csv`
- `TeacherValueAdded.csv`
- `tA.csv`
- `tA0.csv`

Outputs

- Figure 4a, 4b (Recruits_TVA_byTx_yr1.pdf, Recruits_TVA_byTx_yr2.pdf)
- Randomization inference test statistics using F test as reported in Figure 4 notes, reported in Matlab output/log file.
- Statement of percentile increase in TVA based on applying primary estimates of learning outcomes to TVA distribution among FW recruits.

STARS_3_2_1_Application_Descriptives.do

Descriptive statistics on application frequencies and geographic attributes.

Inputs

Subroutines

- `SetFilePaths.do`

Data

- `MarketAdjacencyMatrix.csv`
- `Applicants.dta`
- `Applications.dta`
- `tA.dta`

Outputs

Numbers cited in final paragraph of Section II.A.:

- Numbers of applications per applicant
- Estimated share of applicants applying to contiguous pairs of districts.

See Stata output/log file created by project calling scripts.

STARS_3_2_2_Application_Spillovers.m

Test for spillover effects in application decisions.

Inputs

Subroutines

- `ri_ci.m`

- `RItable.m`
- `SetFilePaths.m`

Data

- `applications.dta`
- `tA.csv`
- `tA0.csv`

Outputs

- ApplicationSpillovers.tex

STARS_3_3_Teacher_Attrition.do

Teacher attrition rates.

Inputs

Subroutines

- `ri_estimates.ado`

Data

- `tA.dta`
- `tA0.dta`
- `tE.dta`
- `tE0.dta`
- `TeacherPanel.dta`

Outputs

- Tables/Retention.tex
- Summary statistic in text: fraction of teachers retained, reported in Table 5 notes and Section III.C, Subsection "Retention Effects", Paragraph 4.

STARS_3_4_Teacher_Post_Chars.do

Attributes of retained teachers.

Inputs

Subroutines

- `ri_estimates.ado`

Data

- `tE.dta`
- `tE0.dta`
- `TeacherPanel.dta`

Outputs

- Round2Chars.tex

STARS_3_5_TeacherSatisfaction.m

Teacher job satisfaction by arm.

Inputs

Subroutines

- `ri_ci.m`
- `RItable.m`
- `SetFilePaths.m`

Data

- `TeacherPanel.csv`
- `tA.csv`
- `tA0.csv`
- `tE.csv`
- `tE0.csv`

Outputs

- TeacherSatisfaction.tex

STARS_3_6_P4P_Opinion.do

Teachers' opinions of performance pay.

Inputs

Subroutines

- SetFilePaths.do

Data

- `tA.dta`
- `tA0.dta`
- `tE.dta`
- `tE0.dta`

Outputs

- Tables/P4P_Opinion.tex

STARS_3_7_Consort_Diagram.do

Stata script to create CONSORT-style diagram of design and sample.

Inputs

Diagram template

- `0_scripts/do/Consort_template.tex`, which is copied to `2_tex/figures/Consort.tex` at the outset of this script. This contains the tikz code used to create the diagram, with placeholders that get updated to reflect descriptive statistics as appropriate.

Data

- `Applications.dta`
- `TeacherPanel.dta`
- `Assessments.dta`
- `tE.dta`

Outputs

- `Consort.tex`