Readme for the replication package of

Teacher Content Knowledge in Indian Secondary Schools and Its Relationship with Student Learning

Sabrin Beg, Anne Fitzpatrick, Jason T. Kerwin,
Adrienne Lucas, and Khandker Wahedur Rahman

February 19, 2026

Overview

This package contains the replication package for "Teacher Content Knowledge in Indian Secondary Schools and Its Relationship with Student Learning" by Sabrin Beg, Anne Fitzpatrick, Jason T. Kerwin, Adrienne Lucas, and Khandker Wahedur Rahman, published in *AEA Papers and Proceedings*, May 2026. It contains all data and code necessary for replicating the one figure and two tables in the paper. The data files are in Stata (.dta) format, and the replication code was written in Stata. Replication of the tables and figure will take less than 5 minutes on a standard desktop.

Data Availability and Provenance Statements

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

☑ All data **are** publicly available.
☐ Some data **cannot be made** publicly available.
☐ **No data can be made** publicly available.

Details on each Data Source

The data for this project were collected by the authors and their data collection partners during a randomized evaluation of Utkarsh, a secondary-school remedial-education program in Odisha, India. The program was implemented in Class 9 (ninth grade) in 300 schools during the 2019–2020 school year. Data collection was conducted in partnership with Transform Schools, People for Action, and the School and Mass Education Department of Odisha. The study is registered in the AEA RCT Registry, with identifier AEARCTR-0004138.

The two datasets in `01_data/` are from surveys and assessments administered as part of this evaluation. They are distributed as part of Beg et al. (2026) via the J-PAL Dataverse.

Data Name	Data Files	Location
"Teacher competency and demographics data"	`Teacher_Competency.dta`	`01_data/`
"Teacher-student linked data"	`Teacher_Competency_and_Individual_Student_Score.dta`	`01_data/`

All datasets were collected and compiled by the study team and are provided with the replication package. Each dataset is described in detail below.

Teacher competency and demographics data

`Teacher_Competency.dta` contains merged baseline and endline teacher-level data for the 899 teachers who were present during the second wave of the endline survey. The unit of observation is the teacher. It contains (among other variables) both the teacher content knowledge scores (English and math content knowledge scores from a grading-based assessment (following Bold et al., 2017)) and the teacher demographic variables (including Gender, age, highest education level, years of teaching experience, and years at the current school).

Teacher-student linked data

`Teacher_Competency_and_Individual_Student_Score.dta` contains data at the student-teacher-subject level, linking individual students to the content knowledge score of the teacher who taught them that subject. It is used to produce Table 2. It contains information on endline and baseline test scores, the content knowledge score of the matched teacher, and school treatment status.

Dataset list

Data file	Source	Notes	Provided
`01_data/Teacher_Competency.dta`	Beg et al. (2026)	Teacher-level data; primary input for Table 1 and Figure 1	Yes
`01_data/Teacher_Competency_and_Individual_Student_Score.dta`	Beg et al. (2026)	Teacher-student linked data; input for Table 2	Yes

Computational requirements

Software Requirements

The replication code was written in Stata, and last run in Stata 16. The following user-written packages are required:

- Stata
  - `reghdfe` — used for high-dimensional fixed effects regressions in Table 2
  - `ftools` — required dependency for `reghdfe`
  - `require` — required dependency for `reghdfe`
  - `estout` — used for regression table output (`estout` and `esttab` commands) in Tables 1 and 2
  - `swindex` — used to construct the teacher gender-equality index
- Mac M1 with MacOS 15.7.3

Memory and Runtime Requirements

Summary Approximate time needed to reproduce the analyses on a standard (2026) desktop machine:

☑ <10 minutes
☐ 10-60 minutes
☐ 1-2 hours
☐ 2-8 hours
☐ 8-24 hours
☐ 1-3 days
☐ 3-14 days
☐ > 14 days
☐ Not feasible to run on a desktop machine, as described below.

Details The code was last run on a Mac M1 with MacOS 15.7.3 and took less than 5 minutes to finish.

Description of programs/code

All do-files are located in the `02_analysis/` folder.

- `0_run_analysis.do`: The master do-file. Sets global file paths (keyed by Stata username), opens the main data file (`Teacher_Competency.dta`), opens a log file, and sequentially calls the three analysis scripts below. In order to run a full replication, this is the only script that needs to be run directly. See Instructions to Replicators for how to set the file path.
- `1_ssrp_p_and_p_data_cleaning.do`: Data preparation and variable creation.
- `2_ssrp_tables_p_and_p.do`: Produces all tables in the paper.
- `3_ssrp_figures_p_and_p.do`: Produces all figures in the paper.

A Stata log from the most recent run of the analysis is provided at `03_logs/p_and_p_log.smcl`.

Instructions to Replicators

- Edit line 6 of `02_analysis/0_run_analysis.do` to include the path of the replication package in your local environment.
- Run `02_analysis/0_run_analysis.do`.

All output tables will be saved to `04_output/tables/` and figures to `04_output/figures/`. These folders must exist before running the code; they are included in the package as provided.

List of tables and programs

The provided code reproduces:

☐ All numbers provided in text in the paper
☑ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below.

Note: Table 2 in the paper is produced as three separate output files (`table2_a.tex`, `table2_b.tex`, `table2_c.tex`) corresponding to Panels A, B, and C, respectively, which are intended to be assembled into a single table in the paper's LaTeX source.

Figure/Table	Program	Line	Output file
Table 1	`2_ssrp_tables_p_and_p.do`	36	`table1_regs_Eg_Ma.tex`
Table 2 (Panel A)	`2_ssrp_tables_p_and_p.do`	68	`tables/table2_a.tex`
Table 2 (Panel B)	`2_ssrp_tables_p_and_p.do`	92	`tables/table2_b.tex`
Table 2 (Panel C)	`2_ssrp_tables_p_and_p.do`	115	`tables/table2_c.tex`
Figure 1	`3_ssrp_figures_p_and_p.do`	36	`figures/figure1_kdensity_math_eng_uncondit`

Table 3: Programs and output files

References

Beg, Sabrin, Anne Fitzpatrick, Jason Kerwin, Adrienne Lucas, and Khandker Wahedur Rahman. 2026. "Data for: 'Teacher Content Knowledge in Indian Secondary Schools and Its Relationship with Student Learning."' Distributed by the J-PAL Dataverse. [DOI to come]

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives*, 31(4): 185–204.