

README

2026-07-06

Replication Package

Overview

Title: How to Build a Reader: Evidence from a Scalable Literacy Intervention in Ghana

Authors: Erik Andersen, Simon Graffy, Jason Kerwin, Monica Lambon-Quayefio

This replication package contains all code and data needed to reproduce the tables, figures, and in-text statistics reported in the paper. The analysis evaluates a scalable literacy intervention in Ghanaian primary schools using an Early Grade Reading Assessment (EGRA).

Data Availability and Provenance Statements

Rights

The authors have legitimate access to and permission to use all data included in this package.

Summary of Availability

Some data cannot be made publicly available due to personally identifiable information (PII). Raw data containing PII is not included. This study is currently ongoing, so the authors have embargoed the data until such time as the project is complete.

Ghana EGRA Data

Student-level data were collected via custom survey instruments administered to students, teachers, and school leaders during the evaluation. Data cleaning code is not included; a summary of the cleaning process is available from the authors upon request.

Uganda Comparison Data

Observational EGRA data from Uganda, de-identified.

Pilot Data

Data from a small scale pilot run prior to the main experiment.

Dataset List

Provided Data (data/raw/)

Data file	Source	Notes
data/raw/ghana_full.dta	Ghana EGRA evaluation	Student-level baseline+endline data; master file for analysis
data/raw/pilot.dta	Pilot cohort	Pilot treatment effects by duration (Figure 2)
data/raw/uganda.dta	Uganda EGRA survey	Comparison sample for Table A.25

Derived Data (data/derived/, generated by 01_prep_data.do)

Data file	Source	Notes
data/derived/ghana_endline.dta	Generated from ghana_full.dta	Endline-only sample (present == 1); main analysis dataset
data/derived/ghana_endline_school_level.dta	Generated from above	School-level aggregates for maps and quality regressions

Computational Requirements

Software Requirements

- **Stata** ≥ 18.5 (SE, MP, or BE editions)
 - 02_analyze_egra.do requires **Stata 18.5+** for multi-variable `vce(cluster ...)` on `areg` (used in the enumerator-heterogeneity tables; this feature was added in 18.5)
 - 04_random_forest.do requires **Stata 19** (uses the `cate` command introduced in Stata 19)
 - **No SSC packages need to be installed.** The required user-contributed Stata packages — `qreg2`, `reghdfe` (+ `ftools`), `coefplot`, `ritest`, `ivreg2` — are bundled in `code/ado/` and loaded automatically via `adopath ++ "$root/code/ado"` at the top of each `.do` file. This guarantees every replicator uses the *exact same package versions* the authors used, regardless of what is or isn't installed on their machine. The bundled versions take precedence over any local installation.
- **R** $\geq 4.x$
 - R packages are managed via `renv`. Run `renv::restore()` to install exact package versions from `renv.lock`.

Controlled Randomness

All random-number-generating programs use set seeds. Seeds are documented in the code. Randomization inference (via `ritest`) uses explicit `seed()` options.

Random forest reproducibility (Stata/MP only): Stata 19's `cate` command parallelizes random-forest construction across processors. When the processor count differs between machines, tree-build order changes and IATE predictions drift slightly — enough to reassign a few students across the `gate_group` quartiles used in Tables 4, 5, and 6, producing small differences in group-composition statistics. To prevent this, `04_random_forest.do` forces single-threaded execution via `capture set processors 1` at the top of the file. Keep this setting to exactly reproduce Tables 4–6. The trade-off is runtime: on Stata/MP this is slower than full parallelization, but the output is guaranteed bit-identical across machines. Stata/SE users can ignore this — SE always runs single-threaded.

Memory and Runtime Requirements

- **Hardware used by authors:** macOS, Apple M3, 16 GB RAM
- **Recommended:** 8+ CPU cores, 16+ GB RAM
- **Storage:** ~100 MB for data and outputs

Wall-Clock Runtime by Script

Script	Approximate runtime
<code>01_prep_data.do</code>	< 1 minute
<code>02_analyze_egra.do</code>	~80 minutes
<code>03_uganda.do</code>	~5 minutes
<code>04_random_forest.do</code>	~8 hours
<code>05_create_graphs.R</code>	< 1 minute
<code>06_endogenous_stratification.R</code>	~10 minutes
Total	~9.5 hours

Description of Programs/Code

Directory Structure

```
replication_package/
├── README.md
├── LICENSE.md
├── run_all.R                # Master script
├── .here                   # R here package sentinel
├── .Rprofile               # renv bootstrap
├── renv.lock               # R package versions
├── renv/                   # renv library
├── manuscript.pdf         # Final manuscript
└── data/
```

```

├── raw/                                # Provided source data
│   ├── ghana_full.dta
│   ├── pilot.dta
│   └── uganda.dta
├── derived/                            # Generated by 01_prep_data.do
│   ├── ghana_endline.dta
│   └── ghana_endline_school_level.dta
├── code/
│   ├── 01_prep_data.do
│   ├── 02_analyze_egra.do
│   ├── 03_uganda.do
│   ├── 04_random_forest.do
│   ├── 05_create_graphs.R
│   ├── 06_endogenous_stratification.R
│   └── ado/                            # Bundled SSC packages (loaded via
adopath)
│   ├── c/coefplot.{ado,sthlp}
│   ├── f/ftools.{ado,mata,...}
│   ├── i/ivreg2*.{ado,sthlp}
│   ├── l/{livreg2,lmoremata*}.mlib
│   ├── q/qreg2.{ado,sthlp}
│   └── r/{reghdfe*,ritest}.{ado,mata,sthlp}
├── results/                            # Created by master script
│   ├── tables/
│   └── figures/

```

Program Overview

Program	Description	Output
01_prep_data.do	Derives endline-only and school-level datasets from baseline+endline master file	data/derived/ghana_endline.dta, data/derived/ghana_endline_school_level.dta
02_analyze_egra.do	Main EGRA analysis: balance, treatment effects, quantile and distribution regressions, survey outcomes, A/B tests	Tables 1-3, 7-8, A.1-A.24; Figures 3-5 (distribution regression and QTE panels)
03_uganda.do	Uganda comparison analysis	Table A.25
04_random_forest.do	Causal forest CATE estimation and GATE tables	Tables 4-6; Figures 7-8
05_create_graphs.R	QTE-vs-distribution-regression CDF, pilot growth, treatment effect bar charts, lesson quality, teacher/student quality	Figures 2, 9-13

Program	Description	Output
06_endogenous_stratification.R	Endogenous stratification with local polynomial smoothing	Figures 6, A.1

Instructions to Replicators

1. Install R ($\geq 4.x$) and Stata (≥ 18). **Stata 19** is required for 04_random_forest.do.
2. **No Stata package installation is required.** The needed user-contributed packages (qreg2, reghdfe + ftools, coefplot, ritest, ivreg2) are bundled in code/ado/ and loaded automatically. If you have older versions of these packages already installed on your machine, the bundled versions in this archive will be used in preference — no action needed on your part.
3. Open run_all.R and set root to the full path of this replication package directory.
4. Set stata_executable to match your Stata installation (e.g., "stata-se").
5. Set stata19_executable to your Stata 19 binary. If your default Stata is already version 19, set it to the same value as stata_executable.

6. In R, run:

```
source("run_all.R")
```

7. The master script will:
 - Restore R packages via `renv::restore()`
 - Create output directories (results/tables/, results/figures/, data/derived/)
 - Run 01_prep_data.do to derive analysis datasets from source data
 - Run all Stata and R analysis scripts in sequence
 - Log wall-clock runtime for each script
8. All outputs will be in results/tables/ and results/figures/.

List of Tables and Programs

Main Tables

Table	Program	Output file
Table 1: Balance Table	02_analyze_egra.do	table_1.tex
Table 2: EGRA Treatment Effects	02_analyze_egra.do	table_2.tex

Table	Program	Output file
Table 3: Gender Heterogeneity	02_analyze_egra.do	table_3.tex
Table 4: GATE (Student)	04_random_forest.do	table_4.tex
Table 5: GATE (Teacher)	04_random_forest.do	table_5.tex
Table 6: GATE (Leader)	04_random_forest.do	table_6.tex
Table 7: Quality and Compliance	02_analyze_egra.do	table_7.tex
Table 8: AB Quality and Compliance	02_analyze_egra.do	table_8.tex

Appendix Tables

Table	Program	Output file
Table A.1: Attrition	02_analyze_egra.do	table_A1.tex
Table A.2: Post-Attrition Balance (Students)	02_analyze_egra.do	table_A2.tex
Table A.3: Post-Attrition Balance (Teachers)	02_analyze_egra.do	table_A3.tex
Table A.4: Post-Attrition Balance (School Leaders)	02_analyze_egra.do	table_A4.tex
Table A.5: Post-Attrition Balance (Schools)	02_analyze_egra.do	table_A5.tex
Table A.6: Robustness	02_analyze_egra.do	table_A6.tex
Table A.7: Lee Bounds	02_analyze_egra.do	table_A7.tex
Table A.8: PCA Weights	02_analyze_egra.do	table_A8.tex
Table A.9: Quality and Compliance (Mechanical Zero)	02_analyze_egra.do	table_A9.tex
Table A.10: Quality Multi 1	02_analyze_egra.do	table_A10.tex
Table A.11: Quality Multi 2	02_analyze_egra.do	table_A11.tex
Table A.12: Compliance Multi	02_analyze_egra.do	table_A12.tex
Table A.13: Compliance (Mechanical Zero)	02_analyze_egra.do	table_A13.tex
Table A.14: AB Test	02_analyze_egra.do	table_A14.tex
Table A.15: AB Test Interaction	02_analyze_egra.do	table_A15.tex
Table A.16: AB Quality Multi	02_analyze_egra.do	table_A16.tex

Table	Program	Output file
1		
Table A.17: AB Quality Multi	02_analyze_egra.do	table_A17.tex
2		
Table A.18: AB Compliance Multi	02_analyze_egra.do	table_A18.tex
Table A.19: SISO Enumerator Heterogeneity	02_analyze_egra.do	table_A19.tex
Table A.20: Teacher Enumerator Heterogeneity	02_analyze_egra.do	table_A20.tex
Table A.21: Pupil Aspirations	02_analyze_egra.do	table_A21.tex
Table A.22: Teacher Survey	02_analyze_egra.do	table_A22.tex
Table A.23: Pupil Home Survey	02_analyze_egra.do	table_A23.tex
Table A.24: Pupil Quality Belief	02_analyze_egra.do	table_A24.tex
Table A.25: Uganda EGRA	03_uganda.do	table_A25.tex

Main Figures

Each figure is saved in **both PNG and PDF formats**. The filenames below list the .png versions; a .pdf with the same stem is produced alongside each one.

Figure	Program	Output file(s)
Figure 1: Study Sample Schools	<i>Not replicated (requires school GPS coordinates, which are PII)</i>	—
Figure 2: Quantile vs Distribution Regression for Letter Sound Knowledge	05_create_graphs.R	figure_2.png
Figure 3: Distribution Regression Estimates for Basic Skills	02_analyze_egra.do	figure_3_clspm_dist_reg.png, figure_3_initial_sounds_dist_reg.png, figure_3_clspm_qte.png, figure_3_initial_sounds_qte.png
Figure 4: Distribution Regression Estimates for Word Reading	02_analyze_egra.do	figure_4_cfwpm_dist_reg.png, figure_4_cnwpm_dist_reg.png, figure_4_cfwpm_qte.png,

Figure	Program	Output file(s)
Figure 5: Distribution Regression Estimates for Advanced Skills	02_analyze_egra.do	figure_4_cnwpm_qte.png figure_5_cwpm_dist_reg.png, figure_5_reading_comp_dist_reg.png, figure_5_cwpm_qte.png
Figure 6: Treatment Effect Heterogeneity by Counterfactual Untreated Basic Skills	06_endogenous_stratification.R	figure_6_letter_sounds.png, figure_6_initial_sounds.png, figure_6_familiar_words.png, figure_6_non_words.png, figure_6_orf.png, figure_6_reading_comp.png
Figure 7: Predicted Individual Average Treatment Effects	04_random_forest.do	figure_7.png
Figure 8: Predicted Group Average Treatment Effects	04_random_forest.do	figure_8.png
Figure 9: Lesson Quality Elements	05_create_graphs.R	figure_9_oral_language_{1-3}.png, figure_9_phonics_{1-3}.png, figure_9_reading_{1-3}.png, figure_9_writing_{1-3}.png
Figure 10: Teacher and Student Quality Elements	05_create_graphs.R	figure_10_teacher_{1-5}.png, figure_10_student_{1-5}.png
Figure 11: Learning Gains vs Months of Treatment	05_create_graphs.R	figure_11.png
Figure 12: Basic Skills Treatment Effects	05_create_graphs.R	figure_12_letter_sounds.png, figure_12_initial_sounds.png
Figure 13: Advanced Skills Treatment Effects	05_create_graphs.R	figure_13_orf.png, figure_13_reading_comp.png

Appendix Figures

Figure	Program	Output file(s)
Figure A.1: Endogenous Stratification (Additional)	06_endogenous_stratification.R	figure_A1_letter_names.png, figure_A1_listening_comp.png, figure_A1_egra_index.png

License

See LICENSE.md for details. Code is licensed under CC-BY 4.0. Data is provided for replication purposes only.

Acknowledgements

Simon Graffy is the co-founder & CEO of Inspiring Teachers. Andersen: Department of Economics, University of Washington; Graffy: Inspiring Teachers; Kerwin (corresponding author): Department of Economics, University of Washington (jkerwin@uw.edu), IZA, and J-PAL; Lambon-Quayefio: University of Ghana and J-PAL. We thank Noam Angrist, Peter Bergman, Alex Eble, Owen Ozier, Rebecca Thornton, Terrence Dai, Matthew Grimm, Alfredo Effendy, Eric Wang, and seminar audiences from GSF, the Gates Foundation Global Education Team and the Prevail Fund for helpful comments and suggestions. Refine.ink provided comments, error-checking, and copy-editing on the paper. The authors used ChatGPT 5.2 Thinking to accelerate the drafting and revision of the paper. This project would not have been possible without the efforts of Dr. Emmanuel K. Larbi Mantey, Regina Mensah, Mark Valentine Aikins, Philip Semaha, Isaac Gaisey, Victor Ayanful, Tom Lewis, Lukasz Labedzki, and the entire Inspiring Teachers staff. Charles Oloa provided excellent data collection project management, enumerator training, and fieldwork supervision, with support from Emmanuel Bentsi Parker Jr. and Chloe Lewis; we are also thankful for the hard work of our data collection team. We are indebted to the teachers and students who generously gave their time to take part in the study. Inspiring Teachers received funding from the IDP Foundation and the Global Schools Forum to run the intervention, and broader support from the Peter Cundill Foundation, Windfall Trust, Mulago Foundation, Pace Able, the Bisconti Family Foundation, Prevail Fund. Research activities were supported by the Bill and Melinda Gates Foundation through the Abdul Latif Jameel Poverty Action Lab (J-PAL)'s Learning for All Initiative and by the Jacobs Foundation through Innovations for Poverty Action's Partnership for Technology Program. This study was reviewed and approved by the University of Ghana Ethics Committee for the Humanities (protocol # ECH 278/ 23-24), the University of Washington Human Subjects Division (protocol # STUDY00021036), and Heartland IRB (protocol # 062424-573); it is registered with the AEA RCT Registry under registration number [AEARCTR-0014627](#).