

REPRODUCIBILITY PACKAGE

The Impact of Procurement Training for Women Entrepreneurs - An Impact evaluation of Sourcing2Equal (S2E) Colombia

Purpose:

The full reproducibility package takes clean data from three rounds of survey - baseline, midline, and endline - construct variables merges datasets, and reproduced 21 tables that appear in the paper.

Impact Evaluation Team:

- Rachel Cassidy
- David McKenzie
- Kalyan Neelamraju
- S. Raaga Akkineni

Last updated:** July 16 2026

Stack: Stata 18 (settings harmonized via `ieboilstart`), WBG DECDI `iefolder` folder conventions

1. Project overview

Randomized evaluation of a capacity-building program, Sourcing2Equal, on the procurement readiness, knowledge, and attitudes** of women-led/owned micro, small, and medium enterprises (MSMEs) in Colombia, and on their engagement in corporate procurement.

- Design: RCT rolled out in **waves** (Wave 1, 2, 3). Randomization was stratified by `City`, `SME`, `Some procurement knowledge`, and `Procured a corporate client`.
- Program activities (treatment): Masterclass, Bootcamp, Mentorship.
- Survey rounds: Baseline → Midline (~6 months) → Endline (~12 months).

- Primary outcomes: Knowledge, Attitude, and Readiness indices + Procurement Activities (Registering to portals, Submitting corporate bids, Winning corporate bids)
- Secondary outcomes: Number of corporate bids submitted, number of corporate bids won, revenue from corporate bids won, number of employees hired, submitted bids to other MSMEs, won bids from other MSMEs, Submitted bids to Public sector, Won bids from Public sector.
- Analytical estimators used in the paper: ANCOVA (ITT), LATE, and CWTE; additionally conducted 6-vs-12-month, midline-vs-endline, pooled-sample comparisons. Also includes heterogeneity checks and Lee-bounds robustness.

> Scope note: Only the ****endline analysis**** produces the tables

> in the final paper. Baseline and midline are run solely to construct indicators and build the merged baseline–midline–endline panel used by the endline analysis. No baseline/midline tables appear in the paper.

2. Repository structure

Two top-level branches under `DataWork/` :

- EncryptedData/ — PII / identifiable data. Raw survey exports, randomization inputs, and the PII-cleaning do-files that de-identify the data. Access-restricted.

- MasterData/ — De-identified (`NO\_PII`) working data, construction, merging, and analysis. Everything downstream of de-identification lives here

DataWork/

```

└─ Project_MasterDofile.do    # top-level orchestrator (PART 0–3)

└─ global_setup.do           # constant globals / control varlists (called by master)
|
└─ EncryptedData/            # PII — restricted
|
|   └─ 01_baseline/ (00_Randomization, 01_datasets, 02_dofiles, 03_output)
|   └─ 02_midline/ (01_datasets, 02_dofiles, 03_output)
|   └─ 03_endline/ (01_datasets, 02_dofiles, 03_output)
|

```

- └─ MasterData/ # de-identified working data + analysis
 - └─ 01_baseline/
 - | └─ masterdata_baseline_MasterDofile.do
 - | └─ 01_datasets/ # *_data_clean.dta, *_data_indicators.dta, randomized waves
 - | └─ 02_dofiles/ # S2EColombia_masterdata_baseline_construct.do
 - | └─ 03_output/
 - └─ 02_midline/
 - | └─ masterdata_midline_MasterDofile.do
 - | └─ 01_datasets/ (01_raw, 02_inter, 03_final)
 - | └─ 02_dofiles/
 - | | └─ 01_cleaning/ S2EColombia_masterdata_midline_cleaning.do
 - | | └─ 02_construct/ ..._midline_construct.do, ..._baseline_midline_merge.do, ...
 - | | └─ 03_analysis/ (midline analysis — not used in paper)
 - | └─ 03_output/ (01_figures, 02_tables)
 - | └─ 04_reports/
 - └─ 03_endline/
 - └─ masterdata_endline_MasterDofile.do
 - └─ S2E_Colombia_endline_README.txt
 - └─ 01_datasets/ (01_raw, 02_inter, 03_final, ops team request)
 - └─ 02_dofiles/
 - | └─ 01_cleaning/ S2EColombia_masterdata_endline_cleaning.do
 - | └─ 02_construct/ ..._endline_construct.do,
 - ..._baseline_midline_endline_merge.do,
 - | | ..._6vs12mnth_dataupdate_1.do, ..._quantitative_data.do
 - | └─ 03_analysis/ Analysis/ Figures/ (produces the paper tables & figures)

```

└─ 03_output/ (01_figures, 02_tables/tex files) # .tex inputs consumed by the paper
└─ 04_reports/
    └─ exhibit_map.csv

```

Each round follows the WBG DECDI `iefolder` dataset convention: `01\_raw` → `02\_inter` → `03\_final`.

> Structural asymmetry to be aware of: midline and endline have the full

> `01\_cleaning` / `02\_construct` / `03\_analysis` do-file split. Baseline does not — its MasterData

> side has only a *construct* do-file, because baseline cleaning is a PII step under

> `EncryptedData/01\_baseline`.

3. How to run

1. Open `DataWork/Project\_MasterDofile.do`.

2. ****Set your user number**** in PART 1:

```
```stata
```

```
global user 2 // 2 = [Your name]
```

Your `projectfolder` path is selected from the `if \$user == ...` block. Add/verify your own absolute path there before running.

3. Run the file. It executes in four parts:

- **\*\*PART 0\*\*** — installs required packages and harmonizes settings (`ieboilstart`, Stata 18).

- **\*\*PART 1\*\*** — folder-path globals (`\$dataWorkFolder`, `\$mastData`, `\$encryptFolder`).

- **\*\*PART 2\*\*** — `global\_setup.do` (constant globals, control varlists).

- **\*\*PART 3\*\*** — calls the round master do-files via toggles:

```
```stata  
  
local importDo_baseline 1  
  
local importDo_midline 1  
  
local importDo_endline 1  
  
```
```

Set a toggle to `0` to skip that round. Rounds are normally run individually; run all three together only for a full replication.

## Pipeline

baseline - construct

↳ s2e\_colombia\_baseline\_data\_indicators.dta

midline - clean → midline construct → baseline–midline merge

↳ s2e\_colombia\_midline\_data\_indicators.dta

↳ s2e\_colombia\_panel\_baseline\_midline\_merge.dta

endline - clean → endline construct → baseline–midline–endline merge → 6v12 update

↳ s2e\_colombia\_endline\_data\_indicators.dta

↳ s2e\_colombia\_panel\_baseline\_midline\_endline\_merge.dta

↳ s2e\_colombia\_panel\_baseline\_midline\_endline\_6vs12mnth\_update.dta

↳ .tex tables in 03\_output/02\_tables/tex files → tables included paper

## 4. Software & package requirements

- **\*\*Stata 18\*\*** (version pinned by `ieboilstart, version(18.0)`).

- User-written commands (auto-installed by PART 0 if missing):

```
`randtreat`, `ietoolkit`, `iefieldkit`, `winsor`, `pdslasso`, `ivreg2`, `ranktest`,
`leebounds`, `asdoc`, `reghdfe`, `ftools`, `require`.
```

- To update outdated versions, use `adoupdate` (the installer does **\*\*not\*\*** upgrade already-installed commands)

## 5. Key datasets

| Stage                  | File                                                               | Location |
|------------------------|--------------------------------------------------------------------|----------|
|                        |                                                                    |          |
| Baseline (clean)       | `s2e_colombia_baseline_data_clean.dta`                             |          |
|                        | `MasterData/01_baseline/01_datasets/`                              |          |
| Baseline (indicators)  | `s2e_colombia_baseline_data_indicators.dta`                        |          |
|                        | `MasterData/01_baseline/01_datasets/`                              |          |
| Midline (indicators)   | `s2e_colombia_midline_data_indicators.dta`                         |          |
|                        | `MasterData/02_midline/01_datasets/03_final/`                      |          |
| Baseline–Midline panel | `s2e_colombia_panel_baseline_midline_merge.dta`                    |          |
|                        | `MasterData/02_midline/01_datasets/03_final/`                      |          |
| Endline (indicators)   | `s2e_colombia_endline_data_indicators.dta`                         |          |
|                        | `MasterData/03_endline/01_datasets/03_final/`                      |          |
| B–M–E panel            | `s2e_colombia_panel_baseline_midline_endline_merge.dta`            |          |
|                        | `MasterData/03_endline/01_datasets/03_final/`                      |          |
| 6-vs-12 panel          |                                                                    |          |
|                        | `s2e_colombia_panel_baseline_midline_endline_6vs12mnth_update.dta` |          |
|                        | `MasterData/03_endline/01_datasets/03_final/`                      |          |

Raw survey data and randomization inputs live under the respective  
`EncryptedData/`

folders and are de-identified into `NO\_PII` datasets before entering MasterData.

## 6. Outputs → paper

Endline analysis writes LaTeX fragments to

`MasterData/03\_endline/03\_output/02\_tables/tex files/`.

Representative mappings:

Figures are written to `.../03\_output/01\_figures/`. No Figures are included in the paper.

## 7. Data handling & conventions

- **\*\*PII discipline:\*\*** identifiable data stays in `EncryptedData/`. Only `NO\_PII` datasets are

promoted to `MasterData/`. Access to EncryptedData is restricted.

- **\*\*File naming:\*\*** `s2e\_colombia\_<round>\_<content>` for datasets/tables;

`S2EColombia\_masterdata\_<round>\_<step>` for do-files.

## 8. Notes: AI assisted and Manual data categorization

- Updating WSME Sector

-----

- WSMES were asked to described their main good and services but most self selected "Others".

To ensure robustness of the regression analysis, the IE team them individually went through their answers to divide them into the following sector:

1) Manufacturing

2) Agriculture

3) Services

## **Changes since Treatment**

Treated participants recorded the changes they experienced since participating in treatment.

Verbatim answers were put through excel copilot using the following prompt:

"In the "Project Data" worksheet, please create columns for each of the categories listed in column C in the "Inc Consolidated Kywd Themes" as "change\_"categoryname"" and mark it as 1 if any of the corresponding keywords in column D of "Inc Consolidated Kywd Themes" are mentioned in column K in "Project Data". Be inclusive of accent marks and mindful of other typos. Observations can be marked for multiple categories as long as they apply. If an observation does not match any of the categories, look for the closest match based on the meaning of the observation. Highlight any cells where manual review is needed. Only run this command where Participant in treatment = 1. Do not review other observations. If any new keywords are identified, add the new keywords to Column D in Inc "Consolidated Kywd Themes" in font color orange."

Observations which were tagged for manual review (as per prompt) were then manually selected to key themes identified.

## **Challenges since Treatment**

Treated participants recorded the challenges they experienced since participating in treatment.

Verbatim answers were put through excel copilot using the following prompt"

"In the "Project Data" worksheet, please create columns for each of the categories listed in column C in the "Challenge Keyword Themes" as "challenge\_"categoryname"" and mark it as 1 if any of the corresponding keywords in column D of "Challenge Keyword Themes" are mentioned in column L in "Project Data". Be inclusive of accent marks and mindful of other typos. Observations can be marked for multiple categories as long as they apply. If an observation does not match any of the categories, look for the closest match based on the meaning of the observation. Highlight any cells where manual review is needed. Only run this command where Participant in treatment = 1. Do not review other observations. If any new keywords are identified, add the new keywords to Column D in Inc "Challenge Keyword Themes" in font color orange."

Observations which were tagged for manual review (as per prompt) were then manually selected to key themes identified.

**END OF README**