

README: Replication for “Testing the promise of digital scaling: in-person vs. app-based training for women entrepreneurs” (Cassidy et al 2026)

Overview:

The code in this replication package constructs all tables and figures from the paper using Stata and R. The package consists of four Stata do-files and 2 R scripts. All figures and tables but three can be constructed using the code from the primary Stata do-file. Two raw datasets are included in the package, one for the main panel, and one for the screening applicants. The “clean” do file will generate variables needed for analysis, and two new datasets with these variables. The replicator should expect the code to run for 2 hours or less.

Data and Code Availability Statement:

This paper uses data from the baseline and endline survey of an impact evaluation of an app-based v/s in-person business training for women entrepreneurs. These can be shared publicly in this reproducibility package. These are also available in Microdata library at:
<https://microdata.worldbank.org/catalog/6422/data-dictionary>

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.

Summary of availability:

- The data used for the reproducibility package is publicly available on the Microdata Library at the link provided above.

Data Sources:

Two raw datasets are included: the main panel survey including a baseline and endline phone survey, carried out with women entrepreneurs, and a screening survey, carried out with women entrepreneurs who were contacted to determine eligibility into the study. Additional information on each dataset can be found in the table below:

File name:	Location:	Description:
application_data_anon.dta	`user'/Data	N=6036 women entrepreneurs. Phone survey consisting of some characteristics and other variables needed to determine eligibility for the app-based training
elearning_bl_fu_anon_clean.dta	`user'/Data	N=2000 women entrepreneurs

		Raw merged dataset that includes in person baseline survey data, administrative data from the trainings, and the endline phone survey data.
--	--	---

Software Requirements:

This replication package requires Stata (code was last run with version 14), and R (code was run with R 4.4.1) The Stata do-file will require the following packages to be installed (which the master do file should take care of):

Winsor; mat2txt; estout; ietoolkit; kmatch; ritest; moremata; ivreg2; ranktest; pdslasso; sg97_5; st0085_2; lassopack

Description of code:

This package consists of four Stata do-files. All figures and data in the paper can be generated using only the *master_replication.do* file. The files and their purposes are as follows:

- Code in *master_replication.do*, specified directories, installs packages, and generates all figures and tables in the paper. It references *clean.do*, *analysis.do*, and *appendix.do* and is the only Stata file that needs to run for the replication.
- Code in *clean.do* generates the variables and datasets used for analysis from the two raw data files described above. It is called in *master_replication.do* and does not need to be run separately.
- Code in *analysis.do* generates one figure and four tables in the paper. It is called in *master_replication.do* and does not need to be run separately.
- Code in *appendix.do* generates figures and tables in the online supplementary appendix of the paper. It is called in *master_replication.do* and does not need to be run separately.
- Code in *generic_inperson.R* generates Table S1.18.
- Code in *generic_elearning.R* generates Table S1.19.

Instructions for replicators:

- Download the two datasets in the replication package and place in the 'Data folder'

STATA:

- Edit the file path in line 33 in the *master_replication.do* file to reflect the folder that will be used for replication.
- Install required packages as necessary.
- Run only *master_replication.do* to generate all tables and results.
- NOTE: One table and two Figures in the online appendix are constructed by hand, and are therefore not included in this package. See additional details in the below table:

R:

- ONLY run the R scripts once clean.do has run which will generate the elearning_bl_fu_indicators.dta file that the R scripts use
- To produce Table S1.18 and Table S1.19, run generic_inperson.R and generic_elearning.R respectively. Edit the file path in line 1 in both scripts to reflect the folder that will be used for replication
- We record all the dependencies in R using the R package renv. Please download renv using install.packages("renv") and then run renv::activate() and renv::restore(). R should now download all the necessary dependencies.
- Run generic_inperson and generic_elearning independently for each table. This will take a long time to run, as the GenericML machine learning process is computationally heavy.

Figure/Table #	Program	Line Number	Output file	Notes
Figure 1	analysis.do	149	Fig_1_Course_hist.png	
Table 1	analysis.do	194	Table_1_Impacts_takeup.tex	Table is manually formatted for optimal size
Table 2	analysis.do	291	Table_2_Impacts_all.tex	Table is manually formatted for optimal size
Supplemental Appendix				
Figure S1	N/A	N/A	N/A	Manually generated in excel
Figure S2	N/A	N/A	N/A	Manually generated in excel
Figure S3	appendix.do	303	Fig_S3_quantile_reg_app.png	
Figure S4	appendix.do	308	Fig_S4_quantile_reg_person.png	
Table S1.1	N/A	N/A	N/A	Manually generated in Word
Table S1.2	appendix.do	330	Table_S2_application_ttests.tex	Table is manually formatted for

				optimal size
Table S1.3	appendix.do	429	Table_S3_balance.mat.tex	Table is manually formatted for optimal size
Table S1.4	appendix.do	438	Table_S4_balance_all4.tex	Table is manually formatted for optimal size
Table S1.5	appendix.do	497	Table_S5_attrition.tex	Table is manually formatted for optimal size
Table S1.6	appendix.do	511	Table_S6_balance_end.tex	Table is manually formatted for optimal size
Table S1.7	appendix.do	550	Table_S7_Impacts_takeup_4g.tex	Table is manually formatted for optimal size
Table S1.8	appendix.do	601	Table_S8_Impacts_primary_4g.tex	Table is manually formatted for optimal size
Table S1.9	appendix.do	673	Table_S9_Impacts_secondary_4g.tex	Table is manually formatted for optimal size
Table S1.10	appendix.do	718	Table_S10_Impacts_primary_nocontrol.tex	Table is manually formatted

				for optimal size
Table S1.11	appendix.do	770	Table_S11_Impacts_secondary_nocontrol.tex	Table is manually formatted for optimal size
Table S1.12	appendix.do	815	Table_S12_Impacts_primary_log.tex	Table is manually formatted for optimal size
Table S1.13	appendix.do	842	Table_S13_Impactsmatching.tex	Table is manually formatted for optimal size
Table S1.14	appendix.do	950	Table_S14_Impactshetero_services.tex	Table is manually formatted for optimal size
Table S1.15	appendix.do	950	Table_S15_Impactshetero_profitsmed.tex	Table is manually formatted for optimal size
Table S1.16	appendix.do	950	Table_S16_Impactshetero_digskillsmed.tex	Table is manually formatted for optimal size
Table S1.17	appendix.do	1011	Table_S17_Impactshetero_post_secondary.tex	Table is manually formatted for optimal size
Table S1.18	generic_inperson.R		Table_S18_ml_inperson	Table is manually

				formatted for optimal size
Table S1.19	generic_elearning.R		Table_S19_ml_elearning	Table is manually formatted for optimal size
Table S1.20	appendix.do	1058	Table_S20_Takeup_predict.tex	Table is manually formatted for optimal size