Read Me
Data and Code for: "Skill Versus Voice in Local Development"
Casey, Glennerster, Miguel, and Voors

22 June 2021

Before running the replication:

- Download and unzip the "SvV_replication" file folder
- Note that this replication file was created using Stata version 17

What to expect:

- The "SvV_replication" folder contains the de-identified datasets. These data files are in Stata (.dta) format. It also contains a do-file called "SvV_replication.do" which uses these datasets, performs all the analysis, and outputs the tables and figures from the manuscript and appendix. This do-file also reproduces estimates stated in the main text of the article.
- To run the replication, first set the directory at the top of "SvV_replication.do" to match the location where you have saved the "SvV_replication" (unzipped) folder on your device.
- The first part of the do-file formats and compiles the main variables used in the analysis. It generates the main data analysis files, entitled "analysis.dta" and "mc_cap.dta", which will appear in the replication folder after the program runs. It also generates several temporary intermediate datasets that are deleted at the end of the section.
- The second part of the do-file runs the analyses and outputs tables and figures into the replication folder.
- Note that the replication folder already contains the outputted tables and figures, which will be overwritten as you run the .do file.

Other things to note:

- Figures 4 and Appendix Figure A1 present the results of merging proposal submission survey answers to the managerial capital testing and community nomination data directly on nominees' names. Since this would be identifying, the replication file builds the figures from a separate file containing the de-identified results of the merge, "public_submission_matched.dta".
    This data was originally matched via a fuzzy string merge using the "matchit" user-written command, with entries considered matched if name strings for given nominees could be identified across datasets with greater than 80% unweighted bigram Jaccard-distance similarity.
- Respondent managerial capital test scores are potentially sensitive, so we have masked their identifiers. The dataset "public_man_capital.dta" has the test scores redacted, and the dataset "public_mancap_results.dta" contains the actual test scores with a different village-level identifier for use in analysis. All relevant tables and estimates are preserved.
    - To avoid the original identifiers being inferred, we have also included in "public_mancap_results.dta" a masked version of the variable "zscore_HII_sq" (a standardised index of proposal scores). Some details about this:
    - The masking procedure is called Enhanced Generalised Additive Data Perturbation (Muralidhar & Sarathy, 2005). The paper describing this process can be found at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5414.2005.00082.x
    - The procedure adds random noise to the relevant variables while preserving its covariance structure with a given set of other data, allowing for the recovery of point estimates that would have been found without the masking.

- Included in the replication directory is the main analysis data file from the paper "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan" by Casey, Glennerster, and Miguel (https://doi.org/10.1093/qje/qje027). This file is titled "gbf_analysis.dta".
- Some of the included data files correspond directly to one of the survey instruments used in data collection. These survey instruments can be found in the "Instruments" subfolder and are .pdf files with the questions and answer prompts given to respondents. Next to each question is a variable name corresponding to the variable name in the relevant data file. The data files correspond to survey instruments as below:

Instrument	Corresponding Data
Expert Scoring.pdf	public_expscore.dta
Manager selection – tally sheet enumerator A.pdf	public_manager_selection.dta
Manager selection – tally sheet enumerator B.pdf	
Managerial Capital Test.pdf	public_man_capital.dta
Submission Survey.pdf	public_submission_survey.dta
Technical Scoring.pdf	public_techscore.dta

Detailed Inventory:

"svv-replication.do" uses the following de-identified datasets:

- public_vill_overview.dta
- public_techscore.dta
- public_sullay_data.dta
- public_submission_survey.dta
- public_manager_selection.dta
- public_man_capital.dta
- public_expscore.dta
- public_contentproposals.dta

And the following de-identified intermediate datasets:

- public_submission_matched.dta
- public_mancap_results.dta

And the following public replication dataset:

- gbf_analysis.dta

And generates the following outputs relating to tables and figures in the main text and appendix:

- Table 1: table1.rtf
- Table 2: table2A.xlsx, table2B.xlsx
- Table 3: table3.rtf
- Figure 2: fig2.png
- Figure 3: fig3.png
- Figure 4: fig4A.png, fig4B.png
- Appendix Table A1: tableA1.rtf
- Appendix Table A2: tableA2.rtf
- Appendix Table A3: tableA3.rtf
- Appendix Table A4: tableA4.rtf
- Appendix Table A5: tableA6.rtf
- Appendix Table A6: tableA6.rtf
- Appendix Table A7: tableA7.rtf

- Appendix Table A8: tableA8.rtf
- Appendix Table A9: tableA9.rtf
- Appendix Table A10: tableA10.rtf
- Appendix Table A11: tableA11.csv
- Appendix Table A12: tableA12.rtf
- Appendix Table A13: tableA13.rtf
- Appendix Figure A1: figA1A.png, figA1B.png
- Appendix Figure A2: figA2.png

The citation for this data is: [INSERT]