README

Manuscript title: "Barriers to entry: Decomposing the gender gap in job search in urban Pakistan"
Authors: Elisabetta Gentile, Nikita Kohli, Nivedhitha Subramanian, Zunia Tirmazee, Kate Vyborny
Date: 10 July 2025

Overview:

This replication package contains the code and data used to generate 52 tables, 4 stats files and 37 figures in the paper, including those in the online appendix. The code, written in Stata, constructs the analysis data files by integrating multiple data sources. These sources include platform data, and survey data collected from households, jobseekers, and firms.

The file ***master.do*** provides the structure for organizing the replication process. The code is divided into three main subsidiary do-files which perform the following tasks:

1. Setup
2. Code to generate datasets
3. Code to replicate all tables and figures

Code files are organized into four sub folders: ***dataprep, oneoff stats***, ***figures*** and ***tables***. The ***dataprep*** folder contains 11 do-files to prepare datasets. The ***oneoff stats*** folder contains 4 do-files to generate the stats used in the body of the paper. The ***figures*** and ***tables*** folders contain 19 and 32 do-files respectively to generate figures and tables in the paper, including those in the online appendix.

The replicator should expect the code to run for approximately 20-25 hours.

Data availability and provenance statements:

The data used in this analysis originates from a combination of administrative sources and surveys collected directly from households, firms, and jobseekers connected to a job search platform based in Lahore, Pakistan. All collected data has been de-identified to ensure confidentiality. This research received ethics approval from Duke University (#2019-0067).

Additionally, external data from Pakistan's Labor Force Survey (LFS), which is publicly available and maintained by government statistical authorities, was incorporated into the analysis.

The details of the data sources used as part of analysis are mentioned below:

- Jobseeker data sources:
    - Household data:
        - Household data comes from the initial household sign-up, which includes a basic roster, elicitation of interest in the Job Talash platform, and basic neighborhood characteristics. Only those who expressed interest in the platform in that survey were approached for enrollment in the study.
            - Both semi-cleaned and cleaned data have been provided.
    - Jobseeker baseline data:
        - Baseline data comes from the signup instrument targeting jobseekers who expressed an interest in signing up with the platform as part of the household survey.
            - Both semi-cleaned and cleaned data have been provided.
    - Jobseeker application data:
        - The application data come from the screener survey, where job seekers indicate which jobs they want to apply for after hearing details about the list of jobs they were matched with in a given round.
            - Both semi-cleaned and cleaned data have been provided.
- Firm data:
    - Firm baseline data & detailed data on the vacancies listed by firms on the platform:
        - Baseline firm data come from the signup survey, where firms were visited in person, and their decision to sign up was recorded. The survey also collected detailed information on each firm's hiring process. Additionally, information on the vacancies firms wanted to list on the platform was recorded either at the time of signup or afterward through the same survey.
            - Both semi-cleaned and cleaned data have been provided.
    - Incentivized Binary Choice Experiment
        - Conducted with firms in the Job Talash sample. We show employers on the platform a series of pairs of CVs and in each pair ask the respondent to select the one that they would be most likely to hire,

with the incentive that this could help inform the applicant pool sent to them through the Job Talash platform.
        - Both semi-cleaned and cleaned data have been provided.
- Administrative data:
    - Jobseeker-vacancy pair (or "match") data:
        - After signing up, jobseekers are matched to each listed vacancy using a simple algorithm: the jobseeker must have at least the required years of education and experience, match any gender requirement, and have indicated interest in this occupation. We refer to each jobseeker-vacancy pair, for which the respondent qualifies and has indicated interest in the occupation, as a match.
            - Both semi-cleaned and cleaned data have been provided.
    - Hypothetical jobseeker-vacancy pair (or "match") data:
        - We also construct matches that the jobseeker would have received if they had retained their original job preferences, education and years of experience, rather than updating this information at some time after their initial registration.
            - Cleaned data has been provided.
- Analysis datasets:
    - We combine data from multiple sources, as mentioned above, and generate a dataset at the match level which is mostly used as part of analysis. We further collapse this dataset at alternative units of observation, such as the job level, jobseeker level, and jobseeker*round level.
        - All datasets have been provided.
- Miscellaneous datasets:
    - These are one-off datasets that are either used in analysis do-files or as part of the cleaning do-files to obtain specific variables.
        - All cleaned datasets have been provided.

Details on each data source

All relevant datasets used in the analysis presented in this paper are available in the directory: *BTE_Replication_Package\intermediate data* and *BTE_Replication_Package\analysis data*.

Public use data sourced

Data on the Pakistan Labor Force Survey 2018-2019 were downloaded from the Pakistan Bureau of Statistics (2018). Data can be downloaded from https://www.pbs.gov.pk/content/lfs-2018-2019-microdata. The data are in the public domain.

Software requirements

- The replication package contains one or more programs to install all dependencies and set up the necessary directory structure.
- The code is based on Stata (version 15 and above). All necessary packages can be installed from the *master.do* file.
- The output either consists of tables and figures. All tables are saved as .tex files, therefore a program that can read .tex files is required.
- The code uses backslashes in file paths, which will run on Windows machines but not on Linux or maxOS machines.

Memory, runtime, storage requirements

- Approximate time needed to reproduce the analysis on a standard desktop machine: 20-25 hours.
- Approximate storage space needed: 12 GB

Description of programs/code

Program in *master.do* is designed to reproduce all tables and graphs included in the paper following the given sequence:

1. Set globals and install packages
2. Running do-files that prepare intermediate datasets to generate final analysis datasets
3. Runs individual table and figure files that create the paper exhibits and one-off stats

List of tables and programs

The provided code reproduces:

1. All numbers provided in text in the paper
2. All tables and figures in the paper
3. Selected tables and figures in the paper, as explained and justified below

Exhibits in the main paper and appendix		
Figure/ Table #	**Program**	**Output file**
Table 1	pipeline2.do	pipeline2.tex
Table 2	gender_selection2_jul11.do	gender_selection2_jul11.tex
Table 3	price_cv_sample2.do	price_cv_sample2.tex
Table 4	pipeline_ed2.do	pipeline_ed2.tex
Table 5	t2andt3_a.do, t2andt3_a_educ.do	t2andt3_a.tex, t2andt3_a_educ.tex
Table 6	fulldyad_hypo.do	fulldyad_hypo.tex
Figure 1	vacancy_firm_gender_ed.do	jobchars_grapha_o.eps, jobchars_graphb_o.eps, firmchars_grapha_o.eps, firmchars_graphb_o.eps
Figure 2	qual_matched_salary_ed.do	qualify_gender_salary_edu_low.eps, qualify_gender_salary_edu_med.eps, qualify_gender_salary_edu_high.eps
One-off stats referred to in the body of the text		
	oneoff_stats_dyad	ljs_stats_oneoff_dyad.tex
	oneoff_stats_irr	ljs_stats_oneoff_irr.tex
	oneoff_stats_firm_vacancy	ljs_stats_oneoff_firm_vacancy.tex
	oneoff_stats_indiv	ljs_stats_oneoff_indiv
Appendix A		
Table A.1	lfs_table.do	summstats_compare.tex summstats_compare_female.tex summstats_compare_male.tex
Table A.2	firmselection_table.do	firm_balance_panel1_o, firm_balance_panel2_o, firm_balance_panel3_o
Table A.3	Built directly into the paper	
Table A.4	pipeline2_2mo.do, pipeline2_ed2_2mo.do	pipeline2_2mo.tex, pipeline2_ed2_2mo.tex

Table A.5	educ_selection_v3.do	educ_selection_v3.tex
Table A.6	gender_selection_firmcomp2.do	gender_selection_firmcomp2.tex
Table A.7	interview_allFE_may15.do	interview_allFE_may15.tex
Table A.8	gender_selection_ed2_jul11.do	gender_selection_ed2_jul11.tex
Table A.9	jobseeker_selection_clean.do	jobseeker_selection_clean.tex
Table A.10	gender_selection2_jul11_entry.do	gender_selection2_jul11_entry.tex
Table A.11	gender_selection_ed2_jul11_entry.do	gender_selection_ed2_jul11_entry.tex
Table A.12	jobseeker_selection_clean_entry_js.do	jobseeker_selection_clean_entry_js.tex
Table A.13	apply_allFE.do, apply_allFE_ed.do	apply_allFE.tex, apply_allFE_ed.tex
Table A.14	mech_integrationcosts2_fes2.do	mech_integrationcosts2_fes2.tex
Table A.15	gender_ads_may9.do, gender_5ads_may9.do	gender_ads_may9.tex, gender_5ads_may9.tex
Figure A.1		Sms_screenshot.jpg
Figure A.2	gender_ads_feb18_ind_o.do	gender_ads_feb18_ind_o.eps
Figure A.3	gender_ads_feb18_occp_o.do	gender_ads_feb18_occp_o.eps
Figure A.4	edu_firms_irr_fullsample.do, firmcomp_irr_fullsample.do, edu_firmspanelb.do	edu_firms_irr_fullsample.eps, firmcomp_irr_fullsample.eps, edu_firmspanelb.eps
Figure A.5	firm_comp_o.do	firm_comp_o.eps

Figure A.6	reasons_womennowork_o.do	reasons_womennowork_mcq_bar_o.eps, reasons_womennowork_r_combined.eps
Figure A.7	ad_open_ed.do	ad_open_ed.eps
Figure A.8	vacancy_firm_gender_ed.do	jobchars_grapha_FE_o.eps
Figure A.9	vacancy_firm_gender_ed.do	firmchars_grapha_FE_o.eps
Figure A.10	indiv_bl_search.do	indiv_bl_search.eps
Figure A.11	indiv_currentwork_hh.do	indiv_currentwork_hh.eps
Figure A.12	selection_apply_both.do	selection_apply_both.eps
Figure A.13	qual_matched_salary_ed.do	qualify_educ_exp_salary_edu_low.eps, qualify_educ_exp_salary_edu_med.eps, qualify_educ_exp_salary_edu_high.eps
Appendix B		
Figure B.1		Cv_1.png
Table B.1	Built directly into the paper	
Table B.2	balanceTable.do	balanceTable.tex
Table B.3	balanceTable_b4.do	firm_balance_panel1cv.tex, firm_balance_panel2cv.tex, firm_balance_panel3cv.tex
Table B.4	price_cv_sample2_nofe.do	price_cv_sample2_nofe.tex
Table B.5	cv_chosen_cv_noweights_ads_nofe.do	cv_chosen_cv_noweights_ads_nofe.tex

Figure B.2	edu_firms_irr_fullsample_nofe.do, firmcomp_irr_fullsample_nofe.do, edu_firmspanelb_nofe.do	edu_firms_irr_fullsample_nofe.eps, firmcomp_irr_fullsample_nofe.eps, edu_firmspanelb_nofe.eps
Table B.6	price_cv_sample1.do	price_cv_sample1.tex
Table B.7	cv_chosen_cv_noweights_ads.do	cv_chosen_cv_noweights_ads.tex
Table B.8	balanceTable.do	balanceTable_repad.tex
Table B.9	balanceTable_b4.do	firm_balance_panel1cv_oth.tex, firm_balance_panel2cv_oth.tex, firm_balance_panel3cv_oth.tex
Table B.10	price_cv_sample2_other.do	price_cv_sample2_other.tex
Figure B.3	edu_firms_irr_fullsample_other.do, firmcomp_irr_fullsample_other.do, edu_firmspanelb_other.do	edu_firms_irr_fullsample_other.eps, firmcomp_irr_fullsample_other.eps, edu_firmspanelb_other.eps
Appendix C		
Table C.1	oaxaca.do	oaxaca_matched_bootstrap_esttab.tex
	oaxaca.do	oaxaca_matched_noqg_bootstrap_esttab.tex
Table C.3	oaxaca.do	oaxaca_planinterview_bootstrap_esttab
Table C.4	oaxaca.do	oaxaca_planinterview_noqg_bootstrap_esttab
Table C.5	oaxaca.do	oaxaca_inviteinterview_bootstrap_esttab
Table C.6	oaxaca.do	oaxaca_inviteinterview_noqg_bootstrap_esttab

Instructions to replicators

- Download and unzip the files to one directory

- Run master.do to run the full sequence of data preparation and analysis, which will generate all tables and figures listed above