Read Me File for: Labor Market Returns to Personality: A Job Search Approach to Understanding Gender Gaps

Christopher Flinn, Petra Todd and Weilong Zhang

Data Availability Statement

The main sources of confidential data used in the paper are:

- (Main dataset) **SOEP-Core V33.1:** The data are provided by the German Institute for Economic Research (DIW Berlin). Because the SOEP data are subject to data protection legislation, researchers have to sign a contract with the SOEP providers first (free of charge). Data access is usually established within three to five days.
  - You can find more information about the SOEP and data access here: https://www.diw.de/en/diw_01.c.601584.en/data_access.html
  - If you have questions, please contact the SOEPhotline (email: soepmail@diw.de , phone: +49 30 89789-292).

Other datasets (only for figure 2):

- **IZA Evaluation Dataset Survey**: The data are provided by IZA Institute of Labor Economics. The IZA-ED data are also subject to data protection legislation. Researchers have to sign a contract with the data providers first (free of charge). Data access is usually established within three to five days.
  - You can find more information about the IZA-ED access here: https://dataverse.iza.org/dataset.xhtml?persistentId=doi:10.15185/izadp.7971.1
  - For more information see https://ed.iza.org or contact idsc@iza.org
- **British Household Panel Survey Waves 1-18:** These data can be obtained from UK Data Archive (UK Data Archive Study Number 5151). To access the data, research need to create an account at UK Data Service first.
  - You can find more information about the BHPS and data access here: https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200005

Description of the replication code

Part 1: description of the Stata Codes (do files)

Because of the confidentiality issues discussed above, this replication package does not contain the original data from GSOEP, IZA-ED or BHPS. However, the following is a summary of the data structure needed to run our replication code (after obtaining access to the data):

All analyses were implemented in Stata 17.

Pre-requisite packages:

```
ssc install asdoc, replace

ssc install Oaxaca, replace
```

1. Step 1: Step 1: Transfer the necessary files (artkalen.dta, cognit.dta, pequiv.dta, pgen.dta, pl.dta) from SOEP-Core V33.1 to the input_data directory.
2. Step 2: Configure the file path in "*Run data_construction.do*." For our example, it was set to "**D:\Dropbox\Personality and Gender Wage Gap**\replication file\data preparation". Update the bold part with the location of your replication file
3. Step 3: Execute "data_construction.do." This Stata script generates summary statistics and prepares the sample for estimation in Part II.

Input data files (from SOEP-Core V33.1)

- *artkalen.dta*: the spell file ARTKALEN is collected from Individual Questionnaires as a calendar-matrix of months of the previous year and respective statuses of 15 categories (for example Question 118 in 2015). The information from each annual Individual Questionnaire is attached to the information of previous surveys in a way that same statuses in consecutive months were treated as one continuous spell. The generated spell file starts with the year before the entrance into the sample and ends with a respondent's last observation.
- *cognit.dta*: the file provided variables related cognitive achievement potentials.
- *pequiv.dta*: this file contains individual-level and household-level data to ensure a harmonized cross-country comparison. Each observation is a person ID (PID) and survey year (SYEAR) combination.
- *pgen.dta*: this file contains individual-level generated data to ensure complete and consistent information over time. Each observation is a person ID (PID) and survey year (SYEAR) combination.

- *pl.dta*: this file contains individual-level data for respondents by answering the annual individual questionnaire. It is keyed on PID (Person ID) and SYEAR (Survey Year)

Output files:

(a) Results for summary statistic tables:

- Table_1_summary_statistics: Provides essential information for Table 1 in the paper.
- results_log: Contains a log from running "data_construction.do." You will find results for Table 2, Table 3 and Table 4 here. You will also find data for calculating columns (2) and (3) in Table 10. To obtain the reported figures in Table 10, these need to be divided by the gender wage gap (0.216).
- Employment_duration.eps and Unemployment_duration.eps: These are figures used to create Figure 3 in the paper.

(b) Data files used to estimate the structural model:

- employment_history.xls: Job history for each individual.
- characteristics.xls: Individual characteristics (education, cognition, personality, cohort, prior employment, and unemployment experiences).
- characteristics_non_duplication: This dataset includes the same sample but restricts it so that each worker has only one observation.
- characteristics_female.xls: Individual characteristics for female workers only.
- characteristics_male.xls: Individual characteristics for male workers only.

Part 2: description of the matlab Codes (m files)

The code requires Matlab R2020b or a later version, and the Optimization Toolbox must be installed to utilize the "particleswarm" global optimizer.

Note that the structural model estimation should only be performed after running the Stata code described above and generating the data files (employment_history.csv, characteristics.csv). However, it is possible to run the Counterfactual_experiment.m code directly to generate the simulation results presented in Tables 7, 8, 9, and 10 (column 1) of the paper.

(a) Structural estimation:

- **Main_server.m**: This is the primary file for running the structural estimation.

- Change the savepath to the current location of the replication file
- **Flags 1, 2, 4** correspond to three different specifications: (1) homogeneous, (2) without personality, and (3) fully heterogeneous, as detailed in Table 5.
- Initial parameter values are read in from `p_boundary1.csv`, `p_boundary2.csv`, and `p_boundary4.csv`.
- There is a call to the `MLE_estimation.m` function to perform maximum likelihood estimation.
- This program performs structural estimation and stores the estimates in the following files: (1) for the homogeneous case: result1.mat, result1_std.mat, result1_simu.mat; (2) for the case without personality: result2.mat, result2_std.mat, result2_simu.mat; (3) for the fully heterogeneous case: result4.mat, result4_std.mat, result4_simu.mat;
- The program outputs parameter estimates (as reported in Table 5), the distribution of search parameters (Table 5), and model goodness of fit (presented in Figures 6 and 7).

- **MLE_estimation.m**: Aggregates individual likelihood values.
- **individual_estimator.m**: Calculates the likelihood value for each individual using the datasets prepared in Part I (`employment_history.csv`, `characteristics.csv`).
- **Likelihood_Type_K_alternative.m**: Calculates the likelihood value for each individual based on a specified set of parameters.

(b) Counterfactual experiments:

- **Counterfactual_experiment.m**: Generates model simulation results using estimated parameters from the fully heterogeneous model (stored in `result4.mat`) and individual characteristics files (`characteristics_non_duplication.csv`, `characteristics_female.csv`, `characteristics_male.csv`). Produces results presented in Tables 7, 8, 9, and 10 (column 1).
- **individual_simulation.m**: Simulates labor market outcomes in the steady state for all individuals.
- **Simulation_Type_K2.m**: Simulates an individual's job history based on a given set of parameters and initial conditions (initial employment status, individual heterogeneity).

(c) Other functions and files:

- **reservation_ojs.m**: Solves for the reservation wage.
- **randpdf.m**: Generates random numbers from a user-specified distribution.
- **hatchfill2.m**: A graph function that draws bars and fitted curves.

Part 3: Code for Figures 1 and 2

Before executing the scripts, configure the file path in the .do files. For example, in our project, it was set to "**D:\Dropbox\Personality and Gender Wage Gap**replication file\data preparation". Please update the highlighted part with the location of your own replication file.

gender_difference_GSOEP.do: Generates Figures 1 and 2 (GSOEP segment).

- Place pl.dta (from GSOEP) in the input_data folder.
- Outputs gender_difference_GSOEP.doc, the dataset for Figures 1 and 2 (GSOEP segment).
- Includes a log file figure12_GSOEP.smcl from executing the script.

gender_difference_BHPS.do: Generates Figure 2 (BHPS segment).

- Place bo_indresp.dta (from BHPS) in the input_data folder.
- Outputs gender_difference_BHPS.doc, the dataset for Figure 2 (BHPS segment).
- Includes a log file figure1_BHPS.smcl from executing the script.

gender_difference_IZA.do: Generates Figure 2 (IZA-ED segment).

- Place the following files (From IZA-ED) in the input_data folder:
  IZA_ED_Survey_SUF_iw_download_08_qs,
  IZA_ED_Survey_SUF_w1_download_04_qs,
  IZA_ED_Survey_SUF_w1_download_08_qs,
  IZA_ED_Survey_SUF_w2_download_08_qs, and
  IZA_ED_Survey_SUF_w3_download_08_qs.
- Outputs gender_difference_IZA.doc, the dataset for Figure 2 (IZA segment).
- Includes a log file figure1_IZA.smcl from executing the script.