Code and Data Documentation: Worker Beliefs About Outside Options

Simon Jäger (MIT)
Christopher Roth (U Cologne)
Nina Roussille (MIT)
Benjamin Schoefer (UC Berkeley)

Our analysis uses three main data sets:

1. German administrative employer-employee data from the Institute for Employment Research (IAB) (Integrated Employment Biographies - IEB 1975-2019) merged to the German Socio-Economic Panel (SOEP version 2020).

2. An information experiment, which we ran in summer 2022.

3. A robustness survey, which we ran in summer 2021.

The matched administrative-survey data underlying our analysis can be accessed for replication purposes through the IAB. Access is regulated by §75 of the German Social Code, Book X. Charges may apply. For a description of how to obtain linked IAB-SOEP data, and a more extensive explanation of the data, please visit:
https://fdz.iab.de/en/pd_hd/soep-cmi-adiab-7520-v1-en/.

This linked data description refers to the "7520" version of the data, which was not available to us at the time of the empirical analysis and covers 1975-2020. Our analysis instead uses a customized "7519" version of the data, which ends in 2019. Both data sets are constructed analogously. However, an important difference is that the "7519" version we use additionally includes panel data on *coworkers* of SOEP respondents (while the "7520" version does not). Coworker information is necessary to construct the outside option proxies.[1]

[1]The additional coworker information in our dataset consists of all workers in an establishment who overlap with a SOEP respondent who consented to linkage. For these coworkers of SOEP respondents, the data includes the set of spells at the establishment at which the overlap occurred as well as data (up to) 5 years before joining and after leaving that establishment. The coworker data is necessary for the construction of wage changes for coworker movers. The coworker data is stored separately from the SOEP data with only the constructed outside option measures being merged with the survey data. Access to the "7519" version of the data for replication purposes is regulated by IAB.

The SOEP data are available from the DIW Berlin after signing a data distribution contract (https://www.diw.de/en/diw_02.c.222829.en/access.html). As the SOEP data are personal data, a data distribution contract is mandatory to access the data. After signing the contract, the data can be accessed by other researchers in the same manner by which the authors obtained them. Further information on the SOEP, including the sample structure, subsamples, and panel attrition can be found online at https://paneldata.org/soep-core. Further information on the SOEP Innovation Sample (SOEP-IS), which is part of the SOEP and included the survey we fielded in 2019 and 2020, can be found online at https://paneldata.org/soep-is/. All procedures, measures, variables, and their coding can be found at https://www.diw.de/sixcms/detail.php?id=diw_02.c.298578.en. All of our estimates are based on the SOEP 2020.

We supply the data from the information experiment and the robustness survey in the replication package.

In addition to these three main data sets, we also use information on the median wage in 2018 per detailed 5-digit occupation categories (*Klassifikation der Berufe 2010*). The data was obtained from the Federal Employment Agency's Statistics Group on the universe of full-time employed subject to social security, can be used without signing a data distribution contract, and is hence provided in the replication package.

Note that we do not provide documentation on purely conceptual exhibits that are not based on any data (Figure I, Figure VIII, Figure X, Table A.1, Figure B.8).

Software requirements: Stata 15 (researcher's local device), Stata 14 (IAB server).

Estimated run time of the local code (tested on a Personal Computer with Microsoft Windows 10 and 16 GB RAM): less than 2 minutes.

Estimated run time of the IAB code on server (tested in November 2023): below 3 days.

General Instructions to Run the Code:

1. The do files in the `iab_code/` folder prepare and analyse the data on the IAB server, which have to be provided to the IAB in order to replicate the results. The subfolder `iab_code/ado/` contains user-written ado files which have to be submitted with the analysis. The median wage file in `occupation_wage/wage_information_kldb_2010` also has to be provided for the analysis. Files from the SSC are automatically installed during the code run from a local repository that IAB updates about twice a year from the SSC. For replication purposes, note that we also supply the SSC ado files as of 2023-12-11 in the subfolder `iab_code/20231212_ssc_ado/` in case ado files change in the future.

2. The do files in the `local_code/` folder prepare and analyze the provided information experiment and robustness survey data, as well as analyses which only require the SOEP itself (not linked to the IAB data). Files from the SSC are automatically installed during the code run. For replication purposes, note that we also supply the SSC ado files as of 2023-12-11 in the subfolder `local_code/20231212_ssc_ado/`.

3. Before running the `local_code/0_master.do` file which runs the entire local code, note that you have to provide a folder path to the replication package in the do file.

4. The replication package contains the empty folders `tables/`, `figures/`, and `temp/`, which are populated by the exhibits and temporary files by the local code run.

IAB Data Preparation and Analysis

The code in `iab_code/` must be run on the IAB server using the master do file `iab_code/0_master.do`, which ensures that the script runs on the folder structure of the data center. The do files create all output which is based solely on the integrated employment biographies (IEB) of the IAB or the IAB-SOEP linkage. In addition to do files which create the exhibits of the paper, we also provide the do file with which we calculated the outside options for our information provision experiment `iab_code/cell_calculations_exp.do`.

The researcher must provide the IAB with the following:

1. All do files in `iab_code/`.

2. All user-written ado files in `iab_code/ado/`.

3. The occupation medium wage data in `occupation_wage/wage_information_kldb_2010`.

4. The ado files downloaded from the SSC in December 2023 in `iab_code/20231212_ssc_ado/` are optional, as the master do file automatically installs these commands. However, we provide them to guarantee reproducible estimates as their version in the SSC may change.

The do files use the following data sets which are provided by the IAB and are placed in the `$orig/` folder (the exact path is set in the master do file) and do not need to be provided by the user:

- `$orig/SOEP-ADIAB_7519_v1.dta` (the linked SOEP-IAB data, containing the identifiers for the SOEP and the IAB, 1975-2019)
- `$orig/coworker_fullhist_IEB.dta` (the complete employment biographies for matched SOEP-IAB observations and their coworkers, 1975-2019)
- `$orig/Inno19F_IVER.dta` (SOEP-IS data, 2019)
- `$orig/Inno20F_IVER.dta` (SOEP-IS data, 2020)
- `$orig/pgen.dta` (SOEP education data for heterogeneity analysis, 1984-2020)

Data Preparation:

- `iab_code/profile.do` sets globals on the IAB server side.

- `iab_code/1_data_prep.do` performs some basic data preparation on the IAB data.
- `iab_code/2_coworker_wagechanges.do` creates the different versions of the coworker wage change variable.
- `iab_code/3_firm_turnover.do` creates the firm-year level turnover rates used in Figure A02.
- `iab_code/4_occupationrank.do` creates the objective occupation rank of workers to contrast the subjective rank elicited in the SOEP-IS.
- `iab_code/5_machinelearning.do` creates different versions of the machine learning benchmark and validates it with a split-sample procedure.
- `iab_code/6_soep_clean.do` cleans the SOEP-IS data. It mirrors the SOEP data clean in the local folder `local_code/6_soep_clean.do`.
- `iab_code/7_finalize.do` merges the files above to create data sets used for the analysis detailed below.

Analysis: Each do file corresponds to an exhibit in the paper.

- `iab_code/FigureII.do`
- `iab_code/FigureIIIab.do`
- `iab_code/FigureIV.do`
- `iab_code/FigureVab.do`
- `iab_code/FigureVIabcd.do`
- `iab_code/FigureVIIabcdef.do`
- `iab_code/FigureB01.do`
- `iab_code/FigureB02.do`
- `iab_code/FigureB04ab.do`
- `iab_code/FigureB05ab.do`
- `iab_code/FigureB07abcd.do`
- `iab_code/FigureE01.do`
- `iab_code/TableI.do`

- iab_code/TableA02.do
- iab_code/TableD01.do

Local (Non-IAB) Data Preparation and Analysis: Information Experiment, Robustness Survey, and Stand-Alone SOEP

The code which can be run locally and without access to the IAB data is located in local_code/. Run local_code/0_master.do to recreate the exhibits which are not required to be linked to the IAB data (i.e., information experiment, robustness survey, and stand-alone SOEP).

Before running the code, you must place the following raw data files from the SOEP (which are not included, see above) into the local_code/soep_data folder:

- Inno19F_IVER.dta, Inno19F_PE.dta (SOEP-IS 2019, the latter data set contains wage information – the wage information is only used for local_code/TableF03.do)
- Inno20F_IVER.dta, Inno20F_PB.dta (SOEP-IS 2020)

We provide the following datasets for the information experiment and the robustness survey:

- local_code/exp_data/raw_exp.dta (anonymized main raw dataset for experiment)
- local_code/exp_data/3_digit_occupation.xlsx, local_code/exp_data/kreis18.xlsx (crosswalks needed for data construction)
- local_code/exp_data/hypothesis_guess_1.xlsx (hypothesis guessing data)
- local_code/exp_data/hypothesis_guess_2.xlsx (hypothesis guessing data)
- local_code/exp_data/hypothesis_guess_3.xlsx (hypothesis guessing data)
- local_code/exp_data/hypothesis_guess_4.xlsx (hypothesis guessing data)
- local_code/exp_data/raw_robust.dta (anonymized raw dataset from robustness survey)
- occupation_wage/wage_information_kldb_2010 (median wages on occupation level)

Data Preparation:

- local_code/1_occupation_crosswalk.do creates a crosswalk between the occupation codes used in the information experiment and the occupation classifications used in the SOEP and IAB data.

- `local_code/2_region_crosswalk.do` creates a crosswalk between the region codes used in the information experiment and the region classifications used in the SOEP and IAB data.
- `local_code/3_exp_clean.do` cleans the data of the information experiment.
- `local_code/4_robust_clean.do` cleans the data of the robustness survey.
- `local_code/5_soep_clean.do` cleans the SOEP-IS data. Note that these data are not the linked data used on the IAB server, but the stand-alone SOEP-IS data. Also note that this do file performs more extensive data cleaning than strictly necessary for the few exhibits which rely on the not matched SOEP data. It mirrors `iab_code/6_soep_clean.do`.

Analysis:

- `local_code/FigureIXab.do` (information experiment)
- `local_code/FigureB03.do` (information experiment, SOEP)
- `local_code/FigureB06.do` (SOEP)
- `local_code/FigureE02abcdef.do` (robustness survey)
- `local_code/TableII.do` (information experiment)
- `local_code/TableA03.do` (information experiment)
- `local_code/TableA04.do` (information experiment)
- `local_code/TableE01.do` (robustness survey)
- `local_code/TableF01.do` (information experiment)
- `local_code/TableF02.do` (information experiment)
- `local_code/TableF03.do` (SOEP)

Additional Programs Underlying Information Treatment:

The individual-specific information treatment is included in the data files above. In addition, as supplementary material, we provide the programs (Qualtrics survey files or *.qsf) needed to implement the experiment survey and the robustness survey. These experiments were run using Qualtrics and can be found at the following paths:

- `survey_programs/Experiment.qsf` (information experiment)

- survey_programs/Robustness.qsf (robustness survey)

Experiment.qsf features the prediction for wage changes based on coefficients from a regression run on IAB data, as documented in iab_code/cell_calculations_exp.do. As described in the paper, the do file implements a regression of log daily wages on a female dummy, 5-digit occupation dummies, labor market region dummies, education dummies, an age cubic, and education interacted with the age cubic. The portion of the full log file that corresponds to the regression that produced the coefficients for prediction can be found in local_code/exp_data/cell_calculations_exp_regression_result.log (run on March 21st, 2022). The java code for these predictions is placed in the question confidence_prior in the qsf file.