

REPLICATION PACKAGE FOR:

College, cognitive ability, and socioeconomic disadvantage: policy lessons from the UK in 1960-2004

by Andrea Ichino, Aldo Rustichini, and Giulio Zanella

The Review of Economic Studies

October 17, 2025

1. Data Availability Statement

The data used in the article are publicly available and freely accessible to any researcher:

- The 1970 British Cohort Study (BCS70) is a public data set, and the raw data files can be freely downloaded from the [1970 British Cohort Study section at UK Data Service](#) after registering and agreeing with their terms and conditions.
- Understanding Society (USoc) is a public data set, and the raw data files can be freely downloaded from the [Understanding Society section at UK Data Service](#) after registering and agreeing with their terms and conditions.
- The University Statistical Record (USR), Undergraduate Records, is a public data set, and the raw data files can be freely downloaded from the [USR section at UK Data Service](#) after registering and agreeing with their terms and conditions.
- The list of all Royal Charters granted in the UK since the 13th century is publicly available at the [Privy Council website](#).
- The Consumer Price Index (all items) annual series released by the UK Office for National Statistics (ONS) that we use to compute real earnings are publicly available, and can be downloaded at the ["Inflation and price indices" section of the ONS website](#).

To reproduce our results, the raw data files downloaded from these sources should be placed in the appropriate folders as described below.

2. Computational Requirements

For the empirical analyses, simulations, and computations reported in the article, we used Stata[®] (StataNow 18.5 MP-Parallel Edition) and Matlab[®] (R2020b-R2024b). The Stata[®] do-files run in few minutes on ordinary computers, including laptops. The following Stata[®] additional packages must be installed (by typing `ssc install [package]`): `dm79.pkg`, `svmat2`, `grc1leg2`, `coefplot`. The Matlab[®] m-files take longer, ranging from a few minutes for basic plots to up to 35 hours to compute the bootstrap standard errors for a single college cohort. These more intensive computations were performed on a Scientific Linux 7.9 HPC cluster consisting of seven compute nodes: four HPE BL460 Gen9 servers (each with 2×Intel Xeon E5-2697 v4 CPUs at 2.30 GHz, 18 cores per CPU, and 128 GB RAM) and three HPE DL380 Gen10 servers (each with 2×Intel Xeon Gold 6240R CPUs at 2.40 GHz, 24 cores per CPU, and 768 GB RAM), all accessing a shared 3.5 TB NFS disk.

3. Overview of replication files

The data files, Stata® do-files (.do), Matlab® m-files (.m), and ancillary files needed to replicate the results in the article and its Online Appendix are in folder /IRZ_replication_package, which also contains the present README.pdf file. This folder is organized into seven subfolders whose content is summarized and described in detail below:

- /data_bcs70 contains the .do file that creates the final .dta file for our analysis of data from the 1970 British Cohort Study (BCS70).
- /data_usoc contains the .do files that create .dta and .csv files for our analysis of data from Understanding Society (USoc).
- /data_usr contains the .do files that create .dta files for our analysis of data from the University Statistical Record (USR).
- /section_2 contains the .m files that produce the figures in Section 2 of the article.
- /section_3 contains the .do files that produce the tables and figures in Section 3 of the article and in the Online Appendix to Section 3.
- /section_4 contains the .do files and .R files that produce the tables and figures in Section 4 of the article. Some .R files in this folder also produce ancillary data files.
- /section_5 contains the .do files and .m files that produce the tables and figures in Section 5 of the article and in the Online Appendix to Section 5.

4. Order of execution

The .do files in subfolders /data_bcs70, /data_usoc, and /data_usr, as well as the .R files in subfolder /section should be executed first (in any order), to create the analysis data files. The remaining files can be executed in any order to reproduce a particular table or figure.

5. Detailed description of replication files

Content of /data_bcs70

- setup_bcs70_data.do – This .do file uses the BCS70 raw data files and produces clean BCS70 data files for the analyses in Section 3.3 of the article.
- setup_bcs70_data.log – The log file from our execution of setup_bcs70_data.do
- /bcs70_raw – This is an empty folder where the raw BCS70 files (to be obtained as indicated in Section 1 of this document) should be placed to reproduce our final BCS70 dataset.
- /bcs70_edited – This is the destination folder of the final output file produced by setup_bcs70_data.do. The output file in this folder is: BCS70_cleaned.dta, which is the input file for our analysis of BCS70 data in Section 3.3 of the article and its Online Appendix.

Content of /data_usoc

- setup_usoc_data.do – This .do file uses the USoc raw data files (to be obtained as indicated in Section 1 of this document) and produces clean Usoc data files for the analysis

of USoc data in the article and in the Online Appendix. This .do file also performs the PCA whose output is reported in Tables A-1 and A-6 in the Online Appendix.

- `setup_usoc_data.log` – The log file from our execution of `setup_usoc_data.do`
- `/usoc_raw` – This folder contains `cpi_ons_gov_uk_series-280122.xls`, which is the Consumer Price Index (all items) annual series released by the UK Office for National Statistics (ONS). Make sure that this folder also contains the other raw data files (to be obtained as indicated in Section 1 of this document), which should be placed into:
 - subfolders named `ukhls_w[wave]`, which will contain the USoc raw data files from the UK HLS (Household Longitudinal Study) waves 1 to 11;
 - subfolders `bhps_w[wave]`, which will contain the BHPS (British Household Panel Survey) waves 1 to 18.
- `/usoc_edited` – This is the destination folder of the final output files produced by `setup_usoc_data.do`. These final output files are:
 - `usoc_w3_final.dta`: the input file for our descriptive analyses based on USoc
 - `usoc_w3_structural.dta`: the input file for our structural analyses in Stata®
 - `usoc_w3_structural.csv`: the input file for our structural analyses in Matlab®

Content of `/data_usr`

- `setup_usr_data.do` – This .do file uses the USR raw data files and produces clean USR data files for the analysis of USR data in the Online Appendix.
- `setup_usr_data.log` – The log file from our execution of `setup_usr_data.do`
- `/usr_raw` – This is an empty folder where the raw USR files (to be obtained as indicated in Section 1 of this document) should be placed to reproduce our final USR dataset. The USR raw data files that we used consist of 220 files (one for each of the 22 years between 1972 and 1993, for 10 different types of data files) whose names can be inferred from `setup_usr_data.do`
- `/usr_edited` – This is the destination folder of the final output file produced by `setup_usr_data.do`. The final output file contained in this folder is:
 - `usr_stock_start.dta`, which is the input file for our descriptive analyses based on USR data in Section 4 of the article.

Content of `/section_2`

- `figure_1.m` – This .m file simulates population data and uses these simulated data to produce Figure 1: Status quo in Society 1 or Society 2, effects of education-biased technological change (EBTC), and effects of three expansion policies. Users must activate specific lines to produce the different panels of the figures. Instructions are provided inside the .m file.
- `ExcessDemand_Simulation_manygroups.m` – This .m file is used by `figure_1.m` to solve numerically for the equilibrium in the simulation.

Content of `/section_3`

- `figure_2_table_A3.do` – This .do file uses BCS70 clean data file and produces:
 - Figure 2: The effect of higher education on cognitive ability.
 - Table A-3: Cognitive ability of college and high school graduates at different ages

- `figures_A1_A4.do` – This `.do` file uses the final USoc data file and produces:
 - Figure A-1: Distribution of cognitive ability in the USoc sample.
 - Figure A-4: Distribution of socioeconomic disadvantage in the USoc sample
- `figure_A2.do` – This `.do` file uses the final USoc data file and produces Figure A-2: Evolution of the cognitive ability score with different standardizations.
- `figures_A3_A5.do` – This `.do` file uses the final USoc data file and produces:
 - Figure A-3: PCA vs factor analysis measures of cognitive ability
 - Figure A-5: PCA vs factor analysis measures of socioeconomic disadvantage
- `table_1.do` – This `.do` file uses the final USoc data file and produces the summary statistics in Table 1: The UK Understanding Society sample.
- The `.log` files from our execution of each of the `.do` files in this folder.
- `tables_A1_A2_A3_A4_A6.txt` – This `.txt` file explains which log files contain the tabulations reported in Online Appendix Tables A-1, A-2, A-3, A-4 and A-6.

Content of /section_4

- `pratt_poly_uni.xlsx` – This `.xlsx` file contains the university and polytechnics enrollment data collected by Pratt (1997), “The Polytechnic Experiment: 1965-1992,” Bristol: Society for Research into Higher Education, Open University Press.
- `royal_charter_list.xlsx` – This `.xlsx` file contains the list of all Royal Charters granted in the UK since the 13th century.
- `prepare_uni_dist.R` – This `.R` file reads the Royal Charters `.xlsx` data file and produces a `.dta` file that contains county distances from colleges.
- `uk_population_deaths_age_1961_2017.xlsx` – This `.xlsx` file contains UK population data from the UK ONS.
- `prog_read.R` – This `.R` file reads the UK population `.xlsx` data file and produces a `.dta` file `uk_pop_age.dta`
- `usoc_w3_structural.csv` – this `.csv` file contains the final Usoc data for our structural analyses in Matlab®
- `figure_3.m` – This `.m` file uses the final USoc data file and produces Figure 3: Empirical joint distribution of ability and disadvantage by education
- `figure_4.do` – This `.do` file uses the final USoc data file and produces Figure 4: Evolution of ability and disadvantage by education, and group shares
- `figure_5.do` – This `.do` file uses the final USR data file and produces Figure 5: Higher education enrollment and distance to the closest college
- `figure_6.do` – This `.do` file uses the final USR data file and produces Figure 6: Criteria for admission to a university
- `figure_7_8.do` – This `.do` file uses the final USoc data file and produces
 - Figure 7: Pop. shares of HS and college graduates by ability and associated odds
 - Figure 8: DPV of lifetime earnings by ability and associated college premium
- `table_2.do` – This `.do` file uses the final USoc data file and produces Table 2: Joint distribution of ability and disadvantage groups in the college population
- `table_3.do` – This `.do` file uses the final USoc data file and produces Table 3: Probability of college graduation as a linear function of ability and disadvantage
- The `.log` files from our execution of each of the `.do` files in this folder

Content of /section_5

- /tables_4_A8_A9_A10 – This subfolder contains:
 - three .m files named md_oneshot_terciles_[cohort].m (one for each college cohort) that perform minimum distance estimation of the structural parameters once, in the actual USoc sample.
 - three .m files named md_bootstrap_terciles_[cohort].m (one for each college cohort) that perform minimum distance estimation of the structural parameters 1,000 times, in USoc bootstrap samples.
 - six log files (Matlab® diary files) from out execution of these .m files, named diary_md_[estimation]_terciles_[cohort].txt, for each college cohort and type of estimation (once or 1,000 bootstrap replications). These log files contain the point estimates reported in Table 4 and Online Appendix Tables A8 and A9
 - six Matlab® .out files from out execution of the .m files in this folder, named md_[estimation]_terciles_[cohort].txt, for each college cohort and type of estimation (once or 1,000 bootstrap replications) that contain the log from Matlab's Command Window during our execution.
 - three .xlsx files named est_terciles_[cohort].xlsx (one for each college cohort) containing the 1,000 bootstrap estimates
 - three .m files named md_oneshot_2b2_[cohort].m (one for each college cohort) that perform minimum distance estimation of the structural parameters from the extended “2-by-2” model (described in the Online Appendix to Section 2) once, in the actual USoc sample.
 - three .m files named md_bootstrap_2b2_[cohort].m (one for each college cohort) that perform minimum distance estimation of the structural parameters from the extended “2-by-2” model (described in the Online Appendix to Section 2) 1,000 times, in USoc bootstrap samples.
 - six log files (Matlab® diary files) from out execution of these “2-by-2” .m files, named diary_md_[estimation]_2b2_[cohort].txt, for each college cohort and type of estimation (once or 1,000 bootstrap replications). These log files contain the point estimates reported in Online Appendix Table A-10.
 - six Matlab® .out files from out execution of these “2-by-2” .m files in this folder, named md_[estimation]_2b2_[cohort].txt, for each college cohort and type of estimation (once or 1,000 bootstrap replications) that contain the log from Matlab's Command Window during our execution.
 - ExcessDemand_terciles.m – the .m file that computes numerically the model's labor market equilibrium at each point of the parameters grid by the .m files in this subfolder
 - ExcessDemand_2b2.m – the .m file that computes numerically the model's labor market equilibrium in the extended “2-by-2” model (described in the Online Appendix to Section 2) at each point of the parameters grid by the .m files in this subfolder
 - usoc_w3_structural.csv – the Usoc data input file for the .m files in this subfolder
- estimates_grid.dta – This .dta file stores structural estimates and a Theta-Lambda grid for the plot in Figure 10: Estimated study effort cost shifts

- `estimates_grid_2b2.dta` – This `.dta` file stores structural estimates and a Theta-Lambda grid for the plot in Figure A-11: Estimated study effort cost shifts, “2-by-2” model
- `table_5_figure_9.do` – This `.do` file uses the final USoc data file and produces
 - Table 5: Implied $a(k,j)$ parameters of the production function
 - Figure 9: Education and cognitive ability biases of technological change
- `figure_10.do` – This `.do` file uses the structural estimates stored in `estimates_grid.dta` and produces Figure 10: Estimated study effort cost shifts
- `figure_11.m` – This `.m` file uses the final USoc data file and produces Figure 11: Effects of actual and counterfactual expansion policies on the 1990-2004 cohort
- `figure_A6.do` – This `.do` file uses the bootstrap estimates and produces Figure A-6: Distribution of MD estimates across 1,000 bootstrap samples
- `figure_A7.m` – This `.m` file uses the final USoc data file and the structural estimates and produces Figure A-7: 2D sections of criterion function, log scale
- `figure_A8_A9.m` – This `.m` file uses the final USoc data file and the structural estimates and produces
 - Figure A-8: 3D sections of criterion function for cohort 1960-1974, log scale
 - Figure A-9: 3D sections of criterion function for cohort 1990-2004, log scale
- `figure_A10.do` – This `.do` file uses the final USoc data file and produces Figure A-10: DPV by ability and disadvantage, and college premium, “2-by-2” model
- `figure_A11.do` – This `.do` file uses the structural estimates stored in `estimates_grid_2b2.dta` and produces Figure A-11: Estimated study effort cost shifts, “2-by-2” model
- `figure_A12.m` – This `.m` file uses the final USoc data file and produces Figure A-12: Effects of actual and counterfactual expansion policies, “2-by-2” model
- `figure_A13.do` – This `.do` file uses the final USoc data file and produces Figure A-13: Estimated study effort cost shifts under alternative definitions of ability and disadvantage that include the “Big Five” (B5) personality traits
- `table_A7.do` – This `.do` file uses the final USoc data file and produces Table A-7: Initial estimates of policy parameters
- `table_A10.do` – This `.do` file uses the final USoc data file and produces Table A-10: Min-distance estimates of technology and policy parameters, “2-by-2” model
- `table_A11.do` – This `.do` file uses the final USoc data file and produces Table A-11: Correlation η between alternative measures of cognitive ability Theta and socioeconomic disadvantage Lambda in USoc
- `ExcessDemand_terciles.m` – the `.m` file that computes numerically the model’s labor market equilibrium at each point of the parameters grid by the `.m` files in this folder
- `ExcessDemand_2b2.m` – the `.m` file that computes numerically the model’s labor market equilibrium in the extended “2-by-2” model (described in the Online Appendix to Section 2) at each point of the parameters grid by the `.m` files in this folder
- `usoc_w3_structural.csv` – the USoc data input file for the `.m` files in this folder
- The `.log` files from our execution of each of the `.do` files in this folder