This document describes all the programs and data sources that one would need to replicate the empirical analysis in Blundell, Green and Jin (2021).

Most of the raw data can be downloaded from the internet, sometimes after online registration. Section 1 describes all the data sources. All the analysis was done in STATA. All the do-files are provided. Section 2 lists all the outputs (tables, figures, numbers in text) and where each was produced. Section 3 describes the empirical work step by step.

1 Data Availability Statements

1. UK Labour Force Survey (End-User-License)

Most analysis of UK data in this paper is based on the End-User-License version of the LFS 1993Q1-2016QQ4. "End-User-License" means they can be freely downloaded from UK Data Service for any academic user who register and agree to their conditions. You can find them all at this link `https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000026#!/access-data`, from 1975 to 2020 in fact. At that link, you can download the datasets quarter by quarter within "GN 33246" (from 1992Q2) and annually within "GN 33132" (before 1992). Because there are more than 100 individual datasets within these two groups, I will not cite them individually here. As an example, the citiation for 2014Q2 LFS is listed in the references Northern Ireland Statistics and Research Agency, Central Survey Unit,Office for National Statistics, Social Survey Division (2019).

2. UK General Household Survey (GHS) 1972-1991 The GHS is also downloadable from the UK data service (End-User-License). You can find all the datasets (year by year) within "GN 33090" at this link: `https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200019#!/access-data`. Each year of GHS has a study number. For example, for 1972., the citation is: Simister, J. (2000).

3. US Current Population Survey Outgoing Rotation Group samples

We have downloaded the MORG annual data from `https://data.nber.org/morg/annual/`, from 1979 to 2017.

4. Workplace Employment Relations Survey (WERS): 1998-2011: Secure Access

We access this data through the UK Data Service's Secure Lab. One has to submit an application describing the research project, how the secure data will be used, and why the EUL version of the data isn't sufficient. When using this data, any output or file can only be released after they are checked by the staff. The citation for this study is: Advisory, Conciliation and Arbitration Service, National Institute of Economic and Social Research, Policy Studies Institute, Department for Business, Innovation and Skills. (2018).

5. Quarterly Labour Force Survey, 1992-2018: Secure Access

We access this data through the UK Data Service's Secure Lab. The application process and restrictions are the same as for the WERS data above. The citation for this study is: Office for National Statistics, Social Survey Division, Northern Ireland Statistics and Research Agency, Central Survey Unit. (2021).

6. Cross-country statistics on graduate proportions and wage premiums, from Table A1.4 and Table A8.2a in Education at a Glance: OECD Indicators, 2012. OECD (2012a), OECD (2012b). The tables can be downloaded from https://www.oecd-ilibrary.org/education/education-at-a-glance-2012/trends-in-educational-attainment-25-64-yea eag-2012-table9-en and https://www.oecd-ilibrary.org/education/education-at-a-glance-2012/trends-in-relative-earnings-total-populati eag-2012-table75-en. We save the relevant statistics in ednedif.dta, which is provided in the replication package.

7. Lee-Lee Data Set on Long-term Educational Attainment by Country, Lee and Lee (2016). We download the dataset covering 1870-2010 acorss 89 countries, from https://barrolee.github.io/BarroLeeDataSet/OUPdownload.html

8. HESA In section B.4 in the paper, we compare the BA proportions measured from the LFS with that from the Higher Education Student Statistics (HESA). The latter stats can be downloaded from this link: https://www.hesa.ac.uk/data-and-analysis/publications. There are dozens of tables for each academic year, and excel table 'HESA stats

summary' puts together all the relevant numbers.

9. deflator from OECD statistics The excel file "GDP, nominal, real and deflator, UK and US, 2018July edition" contain the time series of real and nominal GDP downloaded from OECDstat, at this link `https://stats.oecd.org/Index.aspx?DatasetCode=SNA_TABLE1`

2 Outputs

Table 1 lists all the figures, which do-file created them, and their file name. Table 2 does the same for all tables in the paper. Table 3 does the same for all other numbers quoted in text.

3 Instruction for replication, step by step

All analysis is conducted in STATA. I have used STATA packages listtex, outreg2 and estimates. They are installed from the net. Fior example, just type 'help listtex' in the STATA command window, if you haven't installed it, the help window will show a link which allows you to install it. This section describes all the do-files necessary to reproduce all the outputs in the paper. There are 4 steps to be carried out in order. Within each step, there are multiple do-files.

3.1 Step 1, setting up the data and summary statistics

- 'set up LFS rawdata.do' extracts key variables from the quartely LFS. You only need to change the path `$RawData` to where you download the EUL versions of LFS data near the start of the do-file.
- 'set up GHS.do' extracts key variables from the GHS year by year and save the microdata as 'GHS7291.dta'. You should change the path `$GHS` to where you download the EUL versions of LFS data. The path `$BGJ` is the project folder, containing several subfolders

where intermediate datasets and outputs are saved. Other do-files will refer to folders relative to `$BGJ` .

- 'deflate data.do' imports GDP deflators from the OECDstats, and save deflated microdata as `'GDPdeflated_all.dta'`, one for the US and one for the UK.
- 'UK time sreies from GHS and LFS' summarizes the LFS and GHS to create a small STATA file `'propBA_year_GHS_LFS.dta'`, which will be used later to produce Figure 1.

3.2 Step 2, descriptive analysis

There are a few do-files, each implements some descriptive analysis and produces some figure or tables. Here I describe the content do-file by do-file, and within each do-file I'll describe what it does in the order that appears in the do-file.

- 'over year figures.do' first defines a couple of programs to aggregate data and to estimate year effects in the aggregated data. It summarizes UK LFS to create a small STATA file `UK_by_age_EDU_year16.dta` and summarizes US microdata to create a small STATA file `US_by_age_EDU_year.dta` Using `UK_by_age_EDU_year16.dta`, this do-file makes figure 2 (`lnmedratio_year_UK.png`. This do-file also merges `'propBA_year_GHS_LFS.dta'` and `US_by_age_EDU_year.dta` in order to plot all the UK and US time series in figure 1 (`propBA_year_GHS_LFS_US.png`).
- to demonstrate the robustness of the stylized facts for the UK, we have tried various subsamples and alternative definitions of variables to produce the trends akin to figures 1-2. This is all done in 'over yer figures.do'.
  - we plot the trends separately by gender in `BAprop_year_by_sex.png` and `lnmedratio_year_by_sex.png` (Figure 3);
  - we exclude the public sector to produce `lnmedratio_year_private.png` (Figure 8 in online appendix)
  - we exclude postgrades, to produce `PGprop_year.png` (Figure 6 in online appendix)

  - We define the High-School group as A-Levels+ instead of GCSE+, to produce `wratio_year_HSdefine.png` (left graph in Figure 2 in online appendix)
  - We classify education by age left full-time education rather than the qualifications obtained, to produce `lnmedratio_year_UK_1718.png` (right graph in Figure 2 in online appendix).
  - we exclude immigrants to produce `BA_UKnational.png` and `lnmedratio_year_UKnational.png` (Figure 7 in online appendix)

- finally, 'over year figures.do' computes year effects in employment rate from `UK_by_age_EDU_year16.dta` and exports the graph as `yeareffect_emprate.png` (Figure 9 in online appendix)

- 'UK occupations.do' adjusts LFS data so that the occupation classification is consistent over time, then computes the occupational statistics in `OCC00_30_19932016.tex` (table 2 in the paper). 'UK occupations.do' also computes time series of the share of managers in graduates and the share of graduate in managers, which are `BA_year_managers_30_34` and `manager_year_BA_30_34.png` (Figure 5).

- 'compare HESA and LFS.do' computes times series from editted LFS microdata[1], merges in the HESA stats from the excel file, and produces `HESA_LFS_compare.png` (Figure 3 in online appendix)

- 'crosscountry stats.do' examines the time series on graduate proportion and wage premium across some OECD countries. It runs simple regressions to get the time trends for each country, and the results are reported in table 2 in the online appendix section10. Some estimates are described in text in section 10 in the online appendix as well.

- 'cohort figures.do' summarizes UK LFS by birth cohort and age, computes the cohort effects and saves the data points in a small

[1] the LFS data was editted in Step 1 by 'set up LFS rawdata.do'

STATA file UK_by_cohort_plot. Then the do-file produces 'mediangapbycohort.png' and cohorteffect_prop.png (Figure 4 in online appendix), and cohorteffect_percentiles.png and cohorteffect_lower (Figure 5 in online appendix). 'cohort figures.do' also conducts a bounding exercise to adjust the wage distribution using the 1965 cohort as the reference point. The adjusted statistics are saved as adjmed_cohort1965.dta. The program then produces graphs adjratio_cohort1965.png and adjratio__coef.png (Figure 10 in online appendix). Finally, it computes the BA proportions for the 1965 and 1975 cohorts, for both the UK and the US, which are discussed in text on page 28, around footnote 19.

3.3 Step 3, checking the Skill-Biased Technical Change hypothesis

In 'SBTC analysis.do' has 6 sections, clearly labeled in the do-file.

1. We summarize UK LFS to the level of year and 5-year age band, compute the relevant variables such as log skill ratios and log wage ratios, and merge in macroeconoimc variables such as TFP, and save the resulting data as toregress_TFP_age.dta

2. We do the same at the level of region, 5-year-age-band and 3-year-period, and save the stats as toregress_region.dta

3. Using toregress_region.dta , we compute long difference in wages and skill ratios and plot them in regionalvariation_Dwage_DlnSUgjt_30.pn (Figure 4)

4. We construct instrumental variables for $\ln S_{gt}/U_{gt}$ and $\ln S_{gjt}/\ln U_{gjt}$.

5. We run all the regressions reported in tables 1,7 in the paper and table 1 in online appendix.

6. We test some inequality restrictions predicted by the framework. The results are saved in 'Inequality constraints.xls' and discussed in text in section 2 in online appendix. Some of the estimates for the determinant of the Hessian of the production function are

also mentioned in text on page 37. All the numbers that are mentioned in the paper have been highlighted yellow in 'Inequality constraints.xls'.

7. We do a calibration exercise at the year-ageband level. Starting with `toregress_TFP_age.dta`, we assume some parameters to calculate the implied thetas, and plots the ratios and one time series as `lnthetaratio.png` and `lntheta_ut.png` (Figure 1 in online appendix)

3.4 Step 4, analysis of workers' autonomy

The following do-files need to be imported into the Secure Lab in order to run on Secure Access datasets. To comply with data security, the do-files do not contain folder paths within the secure lab. But the do-files are self-explanatory enough that the reader should know which datasets are used.

- 'set up WERS data.do' selects and edits relevant variables from the raw WERS data.
- 'compute influence index.do' uses a range of influence variables to establish one index of worker's influence.
- 'set up LFS data.do' extracts relevant variables from LFS, aggregate it to TTWA level and construct instrumental variables from 1992-93 data.
- 'WERS-LFS analysis at workplace level.do' merges LFS area information into workplace-level WERS data. It summarizes the data to produce statistics in table 3, and runs regressions that are reported in tables 4 and 5.
- 'WERS-LFS analysis at TTWA level.do' merges LFS and WERS at the TTWA level, and runs regressions that are reported in table 6.

References

Advisory, Conciliation and Arbitration Service, National Institute of Economic and Social Research, Policy Studies Institute, Department for Business, Innovation and Skills. (2018). Workplace Employment Relations Survey: 1998-2011: Secure Access. *UK Data Service. SN: 6712*. http://doi.org/10.5255/UKDA-SN-6712-5.

Lee, J.-W. and Lee, H. (2016). Human capital in the long run. *Journal of Development Economics*, 122:147–169.

Northern Ireland Statistics and Research Agency, Central Survey Unit,Office for National Statistics, Social Survey Division (2019). Quarterly Labour Force Survey, April - June, 2014. *UK Data Service. SN: 7557*. http://doi.org/10.5255/UKDA-SN-7557-6.

OECD (2012a). Trends in educational attainment: 25-64 year-olds (1997-2010). Technical report.

OECD (2012b). Trends in relative earnings: Total population (2000-10).

Office for National Statistics, Social Survey Division, Northern Ireland Statistics and Research Agency, Central Survey Unit. (2021). Quarterly Labour Force Survey, 1992-2020: Secure Access. [data collection]. 22nd Edition. *UK Data Service. SN: 6727*. http://doi.org/10.5255/UKDA-SN-6727-23.

Simister, J. (2000). General Household Survey, 1972. *UK Data Service. SN: 1406*. http://doi.org/10.5255/UKDA-SN-1406-1.

Table 1: List of all figures and programs that generated each

Output in paper	Programme generating it	Filename or number in text
Figure 1	'over year figures.do'	propBA_year_GHS_LFS_US.png
Figure 2	'over year figures.do'	lnmedratio_year_UK.png
Figure 3	'over year figures.do'	BAprop_year_by_sex.png and lnmedratio_year_by_sex.png
Figure 4	'SBTC analysis.do'	regionalvariation_Dwage_DlnSUgjt
Figure 5	'UK occupations.do'	BA_year_managers_30_34 and manager_year_BA_30_34.png
	Output in the online appendix	
Figure 1	'SBTC analysis.do'	lnthetaratio.png and lntheta_ut.png
Figure 2	'over year figures.do'	wratio_year_HSdefine.png and lnmedratio_year_UK_1718.png
Figure 3	'compare HESA and LFS.do'	HESA_LFS_compare.png
Figure 4	'cohort figures.do'	mediangapbycohort.png cohorteffect_prop.png
Figure 5	'cohort figures.do'	cohorteffect_percentiles.png and cohorteffect_lowerend.png
Figure 6	'over year figures.do'	PGprop_year.png
Figure 7	'over year figures.do'	BA_UKnational.png and lnmedratio_year_UKnational.png
Figure 8	'over year figures.do'	lnmedratio_year_private.png
Figure 9	'over year figures.do'	yeareffect_emprate.png
Figure 10	'cohort figures.do'	adjratio_cohort1965.png and adjratio__coef.png

Table 2: List of all tables and programs that generated each

number in paper	Programme generating it	Filename
Table 1	'SBTC analysis.do'	OLSreg_regior
Table 2	'UK occupations.do'	OCC00_30_1993
Table 3,4,5	'WERS-LFS analysis at workplace level.do'	no tex or Excel f
Table 6	'WERS-LFS analysis at TTWA level.do'	no tex or Excel f
Table 7	'SBTC analysis.do'	IVreg_b45_1st
Table 1 in online appendix	'SBTC analysis.do'	IVreg_cons_re
Table 2 in online appendix	'SBTC analysis.do'	IVreg_cons_re

*Note: results in tables 3-6 were generated in the Secure Lab. We are not allowed to take excel files out, only results in a written-up format.

Table 3: List of all numbers in text and programs that generated each

number in text	position in paper	Programme generating it
11%	end of page 1	'deflate data.do' line 50
0.13	page 8	'over year figures.do' line 188
25%	page 12	'over year figures.do' line 250
0.16,0.34	page 14	'cohort figures.do' line 382
0.05, -0.15	page 14	'cohort figures.do' line 364
1.1, 1.5, 10%,0.15	page21	'SBTC analysis.do' line 307
22%,18%,15%,	page 28 and footnote17	'crosscountry stats.do' line 137
12%,24%,16%,27%,34%,32%	page 28	'cohort figures.do' line 380-410
.007, -.033,-.16,-.12	page 37	'SBTC analysis.do' line 867
25%,50%,23%,19%	page 38, simply describing Figure 5	'UK occupations.do' line 282-314
1.8	page 47	'crosscountry stats.do' line 108
1.5	page 48	'crosscountry stats.do' line 28
	Output in the online appendix	
-.16,0.10	page 4	'SBTC analysis.do' line 867
2 log points, 5 log points	page 6	'SBTC analysis.do' line902,927
0.07	page 12	'cohort figures.do' line 97
under 5%, above 10%	page14	'over year figures.do' line427
0.03	page 17	'over year figures.do' line259
4%,15%	page 24, oline appendix .	'cohort figures.do' line 364