Skill-Biased Structural Change - Replication Package

Francisco Buera, Joseph Kaboski, Richard Rogerson, and Juan Ignacio Vizcaino

This readme file describes the structure of the directories of the replication package containing the data and code needed to replicate the results in "Skill-Biased Structural Change". When necessary, readme files within the subdirectories provide additional information about the individual files and procedures used.

Folders in the Replication Package

1. Bridging_WorldKLEMS

This folder contains the files used to *bridge* or produce consistent time series for labor compensation, hours worked, and employment in the World KLEMS data. The correction is needed as a consequence of the changes in the measurement of educational attainment and in industry classification applied by the CPS in 1992 and 2002, respectively. The CPS is the main data source of the World KLEMS.

- **bridging_worldKLEMS_employment.do:** This .do file produces the *bridged* series for total employment.
    - **Inputs:**
        - *excel_files/usa_wk_apr_2013_labour.xlsx* -> the 2013 version of the World KLEMS' labor input file .
    - **Output:**
        - *dta_files/emp_bridged.dta* -> a .dta file containing the bridged time series for employment.
        - *Figures/Emp_bridged.pdf* and *Figures/Emp_raw.pdf* -> two figures containing the bridged and raw time series for employment aggregated at the educational attainment level.
- **bridging_worldKLEMS_hours.do:** This .do file produces the *bridged* series for average hours worked per person.
    - **Inputs:**
        - *excel_files/usa_wk_apr_2013_labour.xlsx* -> the 2013 version of the World KLEMS' labor input file .
    - **Output:**
        - *dta_files/h_emp_bridged.dta* -> a .dta file containing the bridged time series for hours.
        - *Figures/H_Emp_bridged.pdf* and *Figures/H_Emp_raw.pdf* -> two figures containing the bridged and raw time series for hours worked aggregated at the educational attainment level.
- **bridging_worldKLEMS_compensation.do:** This .do file produces the *bridged* series for average compensation per hour worked.
    - **Inputs:**

        - *excel_files/usa_wk_apr_2013_labour.xlsx* -> the 2013 version of the World KLEMS' labor input file .
    - **Output:**
        - *dta_files/comp_bridged.dta* -> a .dta file containing the bridged time series for compensation.
        - *Figures/Comp_bridged.pdf* and *Figures/Comp_raw.pdf* -> two figures containing the bridged and raw time series for compensation aggregated at the educational attainment level.

The *bridged* files serve as intermediate inputs for the exercises carried out in the following folders **Data for Calibration Shift_Share_KM_Methodology - World KLEMS**.

---

2. Table1

The exercises performed through the files in this folder constitute the analysis described in Table 1 of Section 2.4.

A. Files

- **MakeTable1.do:** runs the regressions needed to produce the results in Table 1 and creates Table 1. It also produces the results using monthly data and the results without controls. Additional files that are created are intermediates data files, *CEXdemographicdata.dta* and *CombinedVADataforRegressions.dta*, and the log file *monthlyresults.log*, which contains the monthly results.
    - **Inputs:**
        - *dta_files/CEXmicrodata/intrvw12/
            - ./fmli131.dta
            - ./fmli121x.dta
            - ./fmli122.dta
            - ./fmli123.dta
            - ./fmli124.dta

          Contain the demographic and monthly expenditure data, respectively, for the CEX 2012 interview sample, downloaded August 2013.
        - *VAdataforCEXmerge.dta:* Contains the final mapping of high-skill value-added per dollar of expenditures for disaggregated CEX expenditure categories as described below in B.
    - **Output:**
        - *Table1Results.log:* an ASCII log file containing the regression results in Table 1. * *Table1.csv* and *Table1.tex*: Table 1 in the paper in *.csv* and *tex* format.
- **MakeTable1Analog_DiaryData.do:** runs the regressions needed to produce the analog of the results in Table 1 for the CEX diary sample, a smaller sample with a survey that focuses on higher frequency expenditures. Additional files that are created are intermediates data files, *dta_files/CEXdemographicdataDiary.dta* and *dta_files/CombinedVADataforRegressions_Diary.dta*, and the log file *Table1resultsDiaryAnalog_Monthly.log*, which contains the monthly results.

- **Inputs:**
    - *dta_files/CEXmicrodata/diary12/
        - ./fmld121.dta
        - ./fmld122.dta
        - ./fmli122.dta
        - ./fmld124.dta

      Contain the demographic and monthly expenditure data, respectively, for the CEX 2012 diary sample, downloaded January 2016.

    - *VAdataforCEXmerge.dta:* Contains the final mapping of high-skill value-added per dollar of expenditures for disaggregated CEX expenditure categories as described below in B.
- **Output:**
    - *Table1resultsDiaryAnalog_Quarterly.log:* an ASCII log file containing the regression results for the diary sample referenced in Section 2.4.

- **CEXtoVAdocumentation.xlsx:** Contains all of the necessary matrices and cross-walks to reproduce the analysis in Section 2.2, including Table 1, as explained below in B. The explanation of each worksheet is given in the first sheet entitled *ReadMe*.
- **MatlabManipulations.m:** Matlab matrix manipulation codes for constructing the matrix *dataforCEXmerge* in *CEXtoVAdcoumentation.xlsx* as explained below in B.

B. Instructions for Mapping U.S. Industrial Skill Intensity Data to U.S. CEX Consumption Data through the U.S. Input-Output Structure

There are several important steps to mapping the data together.

We start by defining a binary variable for high-skill intensive industries as well as a continuous measure of the fraction of labor that is high-skill intensive. The definitions for high-skill are done at a level that allows us to match up with the industry classifications in the EU KLEMS "Basic Tables" data. We define *"Financial Intermediation"*, *"Business Services*", *"Education"*, and *"Health and Social Work"*. We map the disaggregate components of these SIC industries to the 69 industries defined by the U.S. input-output classification to create a dummy variable for which of the 69 NIPA industries are high-skill intensive. These final decisions are all given in the file in the worksheet *'HS_Cutoff for CEX'* in *"CEXtoVAdocumentation.xlsx"*. (Red indicates low-skill intensive industries, while green indicates high-skill intensive industries.)

The second thing that needs to be done is to translate 76 personal consumption expenditure categories (NIPA lines) into the 71 Input-Output commodity categories for the IO system (69 industries plus "used" and "other"). This involves two steps (1) attributing the expenditure to a final goods/services producing industry, and (2) attributing the appropriate transportation and distribution margins for the final goods/services into the appropriate industries. We start this by using the data in worksheet *'PCEBridge 2012'* in *"CEXtoVAdocumentation.xlsx"*. It maps NIPA lines to commodity codes and it also gives the totals of producer value (produced in the appropriate commodity code), as well as transportation, retail, and wholesale margins for the year 2012. Using these data, we can construct the fraction of a dollar in final expenditures (purchaser value) that goes to each industry. However, the table does not tell the breakdown of transportation costs

across various transportation industries, nor does it appropriate retail and wholesale costs. We impute transportation costs at a finer classification by assuming the same split we observe across five distribution industries in the 2002 benchmark data, which are in worksheet *'2002 PCE_Bridge_Summary'* in *"CEXtoVAdocumentation.xlsx"*. We also choose a mapping for retail costs into subindustries of retail (automotive, food and drink, general merchandise and other retail). (1)

We thus form a matrix, in which each column represents a NIPA line from the personal consumption expenditures and each row is an industry. Element *ij* represents the total (current value) final output demanded for an industry (or commodity) *i* from expenditures in NIPA line *j*. After normalizing columns to some to one, we get the value of demand for final output for each industry (or commodity) that stems from one dollar of final expenditure in a given PCE category.

The next thing to do is translate these demands for final output into demands for total output by industry using the total requirements table for 2012, which is the worksheet *'IxC_TR_2012'*, which is the '2012' sheet in the summary BEA table *"IxC_TR_1997-2012_Summary.xlsx"*. Each element of this table shows the total amount of industry *i* output required in the production of one dollar of industry *j* output. This industry-by-commodity table translates final output of commodity *i* into total output required from each industry. Finally, we get the value-added demand from output in that industry using the direct requirements table, the worksheet *'CxI_DR_2012'*, which is *'2012'* sheet in the summary BEA table *"CxI_DR_1997-2012_Summary.xlsx"*. Rows correspond to factors and intermediate inputs, so that element *ij* is the (current value) value of commodity/factor *i* used to produce one dollar of output industry *j* output. We the sum of V001 (compensation by employees), V002 (taxes on production and imports, less subsidies), and V003 (gross operating surplus) to get value added as a fraction of total output in each industry, which we then use to get the value added from each industry in a dollar of each PCE NIPA line. This is a 69x76 (industry x PCEline) matrix saved in the worksheet *'VA Bundle Matrix'* in the file *"CEXtoVAdocumentation.xlsx"*.

Lastly, we multiply this matrix of value added in each industry per dollar of PCE commodity by the vector of high skill-intensive dummies for each industry (i.e., "hsect" in *'HS Cutoff for CEX'*) to get value-added in high-skill intensive sector. This results in a vector that tells us the value added in the high skill-intensive sector that is embodied in one dollar of final output for each of NIPA PCE lines. This resulting vector is saved in the sheet 'Data for CEX Merge' in the file *"CEXtoVAdocumentation.xlsx"*. This is the output of the Matlab code *"MatrixManipulations.m"*.

The final step is to create a correspondence to allow us to merge this vector with the CEX data. We start with the household level Consumer Expenditure Survey (CEX) data for the United States from 2012. We restrict ourselves to months in the primary interview sample (spread out over contained in the files *mtbi121x.dta*, *mtbi122.dta*, *mtbi123.dta*, *mtbi124.dta*, and *mtbi131.dta*) and each observation is a household-month observation. We combine these expenditure data with demographic data from the 'fmli' files (*fmli121x.dta*, *fmli122.dta*, *fmli123.dta*, *fmli124.dta*, and *fmli124.dta*) to get household composition, education levels, and after tax income.

Merging the data requires mapping the "UCC" codes on expenditures in the CEX to the *"PCE codes"*. We do this using the BLS-provided Concordance Mapping of CE UCC codes to PCE Series from Table 2.4.5U, which is worksheet titled *'UCC-PCE Item Mapping'* in *"CEXtoVAdocumentation.xlsx"*.

---

(1) NIPA lines 5-7 (new autos, used autos, and automotive parts) are mapped to auto retailers. NIPA lines 27-28 (food and non-alcoholic drink for off premises consumption and alcoholic drinks for off premise consumption) are mapped into food and drink retailers. Everything else is mapped into miscellaneous retailers, with nothing mapped into general merchandise stores.

There are three caveats, however.

- First, the BLS Concordance mapping attributes different weights to CEX expenditures in some cases. These are recorded as a variable “weight”, which is used in aggregating expenditures.
- Second, several UCC categories are not mapped because they cannot be fully mapped into disaggregate PCE series. For many of these, the expenditures would all fall under the same category at the level of aggregation (76 PCE lines) we use in our input-output analysis, and so we assign these by hand. Our original hand mapping was less complete than the one later constructed by Comin, Lashkari, and Mestieri (2019), which included our mappings plus additional hand mappings as well as those deemed as “expert judgement” from a BEA’s updated correspondence. We have therefore adopted their mapping.
- Third, this table only maps UCC codes into PCE Series names and using designations in Table 2.4.5U. The sheets ‘Table 2.4.5U’ is needed to map it into PCE lines, and the ‘Table 2.4.5U to Table 2.4.5 Map’ (also BLS-provided) is needed to map these into the final NIPA lines of BEA NIPA Table 2.4.5. The final mapping, including all stages, is available in the worksheet *‘VAdataforCEXmerge’* in the file in the *“CEXtoVAdocumentation.xlsx”* and the Stata file *‘VAdataforCEXmerge.dta’* . The column “source” indicates the source of the mapping.

---

3. HS_Sector_Definition

This folder contains the files needed to compute the shares of high-skill labor by industry under different employment measures.

- **HS_Shares_by_Industry.do:** This .do file computes the share of High-Skill Labor's Compensation, Employment, and Hours Worked by Industry, for the US, between 1977 and 2005. These shares are used to define the High-Skill Intensive Sector in Section 2.2 of the paper. It is also used to compute the supporting numbers provided in Section 2 of the Online Appendix.
    - **Inputs:** *excel_files/usa_wk_apr_2013_labour.xlsx* -> the 2013 version of the World KLEMS' labor input file .
    - **Output:** *dta_files/HS_labor_shares_1977_2005.dta* -> a .dta file containing the time series and the average for the period of the high-skill labor shares for the 31 Industries in the World KLEMS database.

---

4. Figures1&2

This folder contains the files needed to produce Figure 1 in section 2.3 and Figure 2 in section 2.4 of the paper. Each subfolder contains the corresponding *.do* file needed to produce the corresponding figure. They take as inputs files contained in the folders called *excel_files* and/or *dta_files*.

Subfolder: Figure 1

- **Figure1.do:** produces Figure 1 in Section 2.3. It also produces the two alternative versions of Figure 1 in the Online Appendix under two broader definitions of the high-skill sector. Additionally, this routine

computes the regressions in the figure and their corresponding R^{2}s with and without country-specific fixed effects.

- **Inputs:**
  - *excel excel_files/basic_files_09/'i'_output_09I.xls* -> one excel spredsheet per country '*i*' containing the basic files in the EUKlems database, November 2009 release, revised in June 2010.
  - *dta_files/pwt90.dta* -> the Penn World Table, version 9.0.
- **Output:**
  - *Figure1/figure1.dta:* a file containing the data in Figure 1.
  - *Figure1/Figures/figure1_benchmark.pdf:* Figure 1 in the paper.

Subfolder: Figure 2

- **chain_price_indice.do:** produces the Chain Price Indices for the High- and Low-Skill Sector and a Relative Price Index for the 15 countries in Figure 2.
  - **Inputs:**
    - *excel excel_files/basic_files_09/'i'_output_09I.xls* one excel spredsheet per country '*i*' containing the basic files in the EUKlems database, November 2009 release, revised in June 2010.
  - **Output:**
    - *dta_files/Prices/HS_Sector_Benchmark/P_indices_combined.dta:* a *.dta* file containing time series of the chain-weighted price indices for the countries in Figure 2.
- **Figure2.do:** produces Figure 2 in Section 2.4. It also produces the two alternative versions of Figure 2 in the Online Appendix under two broader definitions of the high-skill sector. Additionally, this routine computes the regressions in the figure and their corresponding R^{2}s with and without country-specific fixed effects.
  - **Inputs:**
    - *dta_files/Prices/HS_Sector_Benchmark/P_indices_combined.dta:* a *.dta* file containing time series of the chain-weighted price indices for the countries in Figure 2.
    - *excel_files/basic_files_09/'i'_output_09I.xls* -> one excel spredsheet per country '*i*' containing the basic files in the EUKlems database, November 2009 release, revised in June 2010.
    - *dta_files/pwt90.dta* -> the Penn World Table, version 9.0.
  - **Output:**
    - *dta_files/figure2.dta:* a file containing the data in Figure 2.
    - *Figure2/Figures/figure2_benchmark.pdf:* Figure 2 in the paper.

---

5. Data_for_Calibration

This folder contains the files needed to compute the time series used to calibrate the model, as described in Section 4 of the paper.

- **data_for_calibration.do:** This .do file produces the time series for the variables used to calibrate the model. The time series computed, all for the US, are the following:

    1. The share of the HS Sector in Total VA
    2. The Share of HS Labor in the Labor Compensation of the HS Sector
    3. The Share of HS Labor in the Labor Compensation of the LS Sector
    4. The chained relative price index of the HS Sector relative to the LS Sector
    5. GDP per-capita
    6. The Skill Premium (w_HS/w_LS)
    7. The share of Efficiency Units that Correspond to HS Labor
    8. The Share of HS Sector's Labor Compensation in Total Labor Compensation

    - **Inputs:**

        - *excel_files/usa_wk_apr_2013.xlsx* -> the World KLEMS database for the US, 2013, basic file: used for VA shares by sector, the chained relative price index, and the labor productivity indices.

        - *excel_files/usa_wk_apr_2013_labour.xlsx* -> used for VA shares by sector, the chained relative price index, and the labor productivity indices.

        - *dta_files/emp_bridged.dta* & *dta_files/h_emp_bridged.dta* & *dta_files/comp_bridged.dta*-> these files correspond to the *bridged* version of the World KLEMS labor input file (see section Bridging World KLEMS). Used to produce compensation share of the high- and the low-skill intensive sectors, the skill intensity by sector, and the skill-premium.

        - *dta_files/pwt90.dta* -> the Penn World Table, version 9.0. Used to compute GDP growth for the US.

    - **Outputs:**

        - *Data_for_Calibration.dta:* a *.dta* containing the eight series used in the calibration.
        - *data_2020_05_15_1950_2010.csv:* a *.csv* containing the eight series used in the calibration.

---

6. Cross-Country

This folder contains the files needed to produce the results in Section 8 of the paper and Section 5 of the Online Appendix.

Subfolder: Cross_Country_Calibration_Targets

- *CC_cali_targets.do:* This .do file produces the times series used to perform the calibration of the technology parameters for the ten OECD countries in the Cross-Country Analysis presented in Section 8 of the paper and Section 5 of the Online Appendix. The series computed are the same as those in section **Data for Calibration**. The countries used in Figure 1/Figure 2 that have labor input files in the EUKlems database (Austria, Australia, Belgium, Denmark, Spain, Germany, Italy, Japan, The Netherlands, United Kingdom, and the United States).

    - **Inputs:**

        - *excel_files/LI_files/'i'_labour_input_08I.xls:* a series of Excel spreadsheets containing the the EUKLEMS labor input file for each country *i* in the analysis.
        - *excel_files/basic_files_09/'i'_output_09I.xls:* a series of Excel spreadsheets containing the the EUKLEMS basic files for each country *i* in the analysis.
        - *excel_files/benchmark_1997.xls:* an excel spreadsheet containing Internationally comparable PPP for value-added (double deflated) for each country for 1997.
    - **Outputs:**
        - *dta_files:* '*i*'*.dta* a series of *.dta* files containing the time series above mentioned for each country '*i*' in the analysis.
        - *csv_files: data_*'*i*'*.csv* a series of *.csv* files containing the time series above mentioned for each country '*i*' in the analysis.

Subfolder: Cross-Country_Model_Fit

- *cross_country_figures.do:* Produces Figures 7 and 8 in Section 5 the Online Appendix. These figures measure the ability of the model to fit the actual series for the skill premium and the share of the high-skill intensive sector in the ten countries mentioned above.
    - **Inputs:**
        - *excel_files/model_fit_crosscountry.xlsx:* and excel spreadsheet containing series of model-simulated and actual data for the countries in the cross-country analysis.
    - **Outputs:**
        - *dta_files/cross_country_model_fit.dta:* a *.dta* file containing the data in Figures 7 and 8 in Section 5 of the Online Appendix.
        - *Figures/figure7.pdf:* & *Figures/figure8.pdf:* Figures 7 and 8 in Section 5 of the Online Appendix.

---

7. Sensitivity

This folder contains the files for the sensitivity exercises carried out in Section 7 of the paper. In particular, we here provide the workfiles for the exercises carried out in Subsections 7.1 and 7.2.

Subfolder: Consumption_Investment_VA

This folder contains the files used to carry out the sensitivity analysis in Section 7.1 of the paper, called "Consumption Value Added vs Investment Value Added".

In this exercise, we use input-output data for the US to compute the time series of the High- and Low-Skill Sectors' shares of Value Added in producer's prices that are generated by final expenditure on consumption and investment (in producer's prices). Data are available every five years between 1977 and 2002 and annually between 1997 and 2005. For simplicity, and to be consistent with the period used in our benchmark results, we use data for 2005, 2002, 1997, 1992, 1987, 1982, 1977.

1977

IOprocessing1977.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *2digit_tables1977/IOuse.xlsx:* the industry by commodity use table, which is built by unstacking columns A, B, and C from the file */2digit_tables1977/77IO85-levelexcel/1977 Transactions 85-level Data.xls* together with the industry labels available in the file */2digit_tables1977/77IO85-levelexcel/1977 85-level Industry Code descriptions.rtf* that we copy and paste into the excel spreadsheet */2digit_tables1977/77IO85-levelexcel/1977_85_level_Industry_Code_Descriptions.xlsx*.
    - *2digit_tables1977/IxCTR.xlsx:* the industry by commodity total requirement table, which is built by unstacking columns A, B, and G from the file */2digit_tables1977/77IO85-levelexcel/1977 Transactions 85-level Data.xls* together with the industry labels available in the file */2digit_tables1977/77IO85-levelexcel/1977 85-level Industry Code descriptions.rtf* that we copy and paste into the excel spreadsheet */2digit_tables1977/77IO85-levelexcel/1977_85_level_Industry_Code_Descriptions.xlsx*.
- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *1977* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

1982

IOprocessing1982.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *2digit_tables1982/ndn0125/82-2dall.txt*: a file containing all the transactions to build the Industry by Commodity Total Requirements and Industry by Commodity Use Tables.
- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *1982* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

1987

IOprocessing1987.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *2digit_tables1987/ndn0019/IOUSE.xlsx:* the industry by commodity use table, which is built by combining the files *ndn0019/TBL1-87.DTA* to *ndn0019/TBL8-87.DTA*.
    - *2digit_tables1987/ndn0019/IXCTR.xlsx:* the industry by commodity total requirements table, which is built by combining the files *ndn0019/TBL1-87.DTA* to *ndn0019/TBL8-87.DTA*.

- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *1987* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

1992

IOprocessing1992.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *2digit_tables1992/ndn0180/IXCTR.txt:* a file containing the data for the total requirements table.
    - *2digit_tables1992/ndn0180/IOUSE.txt:* a file containing the data for the industry by commodity use table.
    - *2digit_tables1992/ndn0180/io-code.txt:* a file containing the industry codes to build the tables.
- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *1992* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

1997

IOprocessing1997.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *summarytables1997/IOUseSummary.xlsx:* a file containing the data for the industry by commodity use table.
    - *summarytables1997/IndByComTRSum.xlsx:* a file containing the data for the industry by commodity total requirements table.
- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *1997* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

2002

IOprocessing2002.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *summarytables2002/2002_Requirements_summary.xlsx:* a file containing the data for the industry by commodity total requirements table.
    - *summarytables1997/2002_IOMakeUse_summary.xlsx:* a file containing the data for the industry by commodity make and use tables.

- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *2002* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

2005

IOprocessing2005.m: a Matlab code that takes data from the input-output table and produces data on value added by industry generated by final expenditures in consumption and investment.

- **Inputs:**
    - *summarytables2005/IxC_TR_2005_AR_PROD_SUM.xls:* a file containing the data for the industry by commodity total requirements table.
    - *summarytables2005/IO_Use_2005_AR_PROD_SUM.xls:* a file containing the data for the industry by commodity use tables.
    - *summarytables2005/IO_Make_2005_AR_PROD_SUM.xls:* a file containing the data for the industry by commodity make tables.
- **Output:** A table with the data on the value added generated by consumption and investment expenditures by industry. This table is exported to the excel spreadsheet *Historical_IO_Results_Summary.xlsx* sheet *2005* where industries are classified into High- and Low-Skill and the consumption and investment value added shares are computed.

Historical_IO_Results_Summary.xlsx: this spreadsheet takes the data on value added generated by final expenditures on consumption and investment, assigns industries to the High- and Low-Skill sector and computes the share of value added generated by consumption and investment expenditures that corresponds to the High-Skill sector. Time series with the results are presented on the worksheet *Summary and Figures*.

Subfolder: Trade

- *US Trade Data.xlsx:* this Excel file contains the calculations performed to adjust the shares of value added in the high- and low-skill sector for sectoral net trade flows, as described in Section 7.2 of the paper.
    - **Inputs:** the exercise uses annual data on trade of goods and services for the U.S. (Balance of Payment Basis) from the U.S. Census Bureau, which is stored in the worksheet *Trade*. It also requires a time series on the U.S. GDP, which we obtain from the Bureau of Economic Analysis and store in the sheet U.S. GDP.
    - **Output:** two time series of the net trade flows for the high- and low-skill sector. Due to data limitations, to compute this series we assume that the net trade in services corresponds to the net trade flow in the high-skill sector, while the net trade in goods represents the trade flow in the low-skill sector.

      The Bureau of Economic Analysis provides data for the U.S. trade in services by type of service since 1999. We use these data to validate our assumption that the net trade in services is close to the net trade flow for the high-skill sector. Columns I to L in the Sheet Calculated Series confirm that that is indeed the case.

8. Shift_Share_KM_Methodology

In this folder, we perform a series of shift-share analyses following the methodology proposed in Katz and Murphy 1992. The results are presented in Section 6.1 of the paper, and more extensively in Section 1.4 of the Online Appendix.

Subfolder: World KLEMS

In this subfolder, we perform a series of shift-share analyses using bridged World KLEMS data for the US. The exercises performed in the *.do* files here differ in either the way human capital is measured (BKRV or KM efficiency units), the way sectors are measured (value added or employment), and/or the level of sectoral aggregation (2 or the 31 sectors). These different features are reflected in the name of the *.do* files. Additionally, the period of study and the set of industries in the high-skill sector can be easily changed within each file. All the *.do* files in the folder use the following files as inputs:

Inputs: the bridged files for the US's World KLEMS 2013 labor input files (see folder Bridging_WorldKLEMS), contained in the following files: * *dta_files/comp_bridged.dta* * *dta_files/h_emp_bridged.dta* * *dta_files/emp_bridged.dta*

- *Shift_Share_KM_WK2013_Eff_KM_2sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in KM efficiency units, aggregating industries at the two-sector level, and using employment measured in KM efficiency units as a measure of sector size. The result of this exercise is presented in Row (i) of Table 7 of the paper and in Column H of Table 2 in the Online Appendix.
    - **Output:**
        - *Table7_row_i.smcl:* a *log* file that containes the output of the code.
- *Shift_Share_KM_WK2013_VA_KM_2sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in KM efficiency units, aggregating industries at the two-sector level, and using value added as a measure of sector size. The result of this exercise is presented in Row (ii) of Table 7 of the paper and in Column N of Table 3 in the Online Appendix.
    - **Input:**
        - *dta_files/VA_shares.dta:* a file containing time series with the share of value added for the high- and low-skill sectors.
    - **Output:**
        - *Table7_row_ii.smcl:* a *log* file that containes the output of the code.
- *Shift_Share_KM_WK2013_VA_BKRV_2sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in BKRV efficiency units, aggregating industries at the two-sector level, and using value added as a measure of sector size. The result of this exercise is presented in Row (iii) of Table 7 of the paper and in Column M of Table 3 in the Online Appendix.
    - **Input:**

        - *dta_files/VA_shares.dta:* a file containing time series with the share of value added for the high- and low-skill sectors.

    - **Output:**

        - *Table7_row_iii.smcl:* a *log* file that containes the output of the code.

- *Shift_Share_KM_WK2013_Eff_BKRV_2Sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in BKRV efficiency units, aggregating industries at the two-sector level, and using employment measured in BKRV efficiency units as a measure of sector size. The results of this exercise are presented in Column G (1977-2005) and Column F (1980-1990) of Table 3 in the Online Appendix.

- *Shift_Share_KM_WK2013_Eff_BKRV_31Sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in BKRV efficiency units, aggregating industries at the 31 sector level, and using employment measured in BKRV efficiency units as a measure of sector size. The result of this exercise is presented in Column E (1980-1990) of Table 3 in the Online Appendix.

- *Shift_Share_KM_WK2013_Eff_KM_31Sectors.do:* computes the between-industry relative demand shift in favor of high-skill labor using World KLEMS data for the US, measuring human capital in KM efficiency units, aggregating industries at the 31 sector level, and using employment measured in KM efficiency units as a measure of sector size. The result of this exercise is presented in Column D (1979-1989) of Table 3 in the Online Appendix.

Subfolder: IPUMS

In this subfolder we perform a series of shift-share analyses using IPUMS US data. We do this exercise under two different grouping criteria for workers: experience (as in KM) and age (as in the World KLEMS database). The files for each of these exercises are presented in the folders Experience and Age, respectively. The results of these exercises are presented in Columns B and C of the Online Appendix.

Note: the results reported in the Online Appendix require a dataset that is too heavy (around 1.3 GB). We include here a five percent sample of the data instead (using Stata's sample 5 command). The *.do* files *Age/Demand_Shift_KM_IPUMS_E_KM_Age.do* and *Experience/Demand_Shift_KM_IPUMS_E_KM_Experience.do* provide a description on how to download the data from the IPUMS website. The dataset is also available upon request.

Age

- *Age/KM_count_data.do:* this file creates the count sample in KM 1992 (file *KM_count_data.dta*), following the description in section II of their paper. It uses as input the file *usa_80_90_5pct_sample.dta* (see Note above).
- *Age/KM_wage_data.do:* this file creates the wage sample in KM 1992 (file *KM_wage_data.dta*), following the description in section II of their paper. It uses as input the file *usa_80_90_5pct_sample.dta* (see Note above).
- *Age/KM_skill_prem_age_ipums.do:* this file computes the skill premium under KM's methodology (see footnote 20 in KM 1992), using as inputs the count and the wage samples created by the files above (*KM_count_data.dta*,*KM_wage_data.dta*).

- *Age/Demand_Shift_KM_IPUMS_E_KM_Age.do:* this *.do* file computes the between industry demand shift in favor of high-skill labor following the methodology in KM 1992. It uses as inputs the count and wage samples created above (*KM_count_data.dta*,*KM_wage_data.dta*), and the *.do* file *bridge_ind1990_eu31.do*, which bridges the NAICS industries in IPUMS into the 31 industries in the World KLEMS data. The results obtained are presented in Column C of Table 1 in the Online Appendix.

Experience

- *Experience/KM_count_data.do:* this file creates the count sample in KM 1992 (file *KM_count_data.dta*), following the description in section II of their paper. It uses as input the file *usa_80_90_5pct_sample.dta* (see Note above).
- *Experience/KM_wage_data.do:* this file creates the wage sample in KM 1992 (file *KM_wage_data.dta*), following the description in section II of their paper. It uses as input the file *usa_80_90_5pct_sample.dta* (see Note above).
- *Experience/KM_skill_prem_age_ipums.do:* this file computes the skill premium under KM's methodology (see footnote 20 in KM 1992), using as inputs the count and the wage samples created by the files above (*KM_count_data.dta*,*KM_wage_data.dta*).
- *Experience/Demand_Shift_KM_IPUMS_E_KM_Experience.do:* this *.do* file computes the between industry demand shift in favor of high-skill labor following the methodology in KM 1992. It uses as inputs the count and wage samples created above (*KM_count_data.dta*,*KM_wage_data.dta*), and the *.do* file *bridge_ind1990_eu31.do*, which bridges the NAICS industries in IPUMS into the 31 industries in the World KLEMS data. The results obtained are presented in Column B of Table 1 in the Online Appendix.

Subfolder: EUKLEMS

In this subfolder, we perform a series of shift-share analyses using EUKLEMS data for the US. We here produce four different between-industry demand shifts and demand-shift contributions. They vary in the way employment is measured, which could be either in BKRV or KM efficiency units and on if we measure the size of sectors using the corresponding employment measure or value added. The results are presented in Columns I and J of Table 2 and in Columns O and P of Table 3 in the Online Appendix.

- *Shift_Share_KM_EUKLEMS_2Sectors.do:* this *.do* file computes the between-industry demand shift and its contribution to the total demand shift in favor of high-skill labor using EUKLEMS data for the US. Industries are aggregated at the two-sector level, at the cell level workers are grouped according to age, and the analysis is performed for our benchmark period, 1977-2005.

- *Inputs:*

    - *excel_files/usa-naics_labour_input_08I.xls:* the EU KLEMS labor input file for the US, release March 2009.
    - *excel_files/usa-naics_output_09I.xls:* the EU KLEMS basic file for the US, released in November 2009 and revised in June 2010.

- *Output:* The file produces four different between-industry demand shifts:

    - *log_Delta_Xd_BKRV:* measuring employment in BKRV efficiency units and using employment in BKRV efficiency units as a measure of sector size. This result is presented in row (i) of Column I in Table 2 of the Online Appendix.

- *log_Delta_Xd_KM:* measuring employment in KM efficiency units and using employment in KM efficiency units as a measure of sector size (BKRV Value Added). This result is presented in row (i) of Column J in Table 2 of the Online Appendix.
- *log_Delta_Xd_VA_BKRV:* measuring employment in BKRV efficiency units and using value added as a measure of sector size (BKRV Value Added). This result is presented in row (i) of Column O in Table 3 of the Online Appendix.
- *log_Delta_Xd_VA_KM:* measuring employment in BKRV efficiency units and using value added at fixed prices as a measure of sector size (KM Value Added). This result is presented in row (i) of Column P in Table 3 of the Online Appendix.

This *.do* file also produces two auxiliary *.dta* files that are required to compute the between-industry demand shift contribution to the total change in demand of high-skill workers: *eff_KM_BKRV_1977_2005.dta* and *wage_premiums.dta*.

These files have the data to compute the log change in the relative supply of skills under the KM and the BKRV measure and the log change in the skill-premium, which are presented in rows (ii) and (iii), respectively, of columns I, J,O, and P.

---

9. mfiles

This subdirectory contains the codes used to run the simulations reported in the paper and in the Online Appendix.

There are 15 files in the main directory, including six (6) Matlab functions that are called by the programs running the simulations. In Addition, there are four subdirectories with additional files which are described below.

Files in the main directory:

1. **"main_cal_sim.m"**: Calculate the model and perform the simulations used to produce the main tables in the paper (tables 2, 3, 4, 5, 6 in the main text, tables A1 and A2 in the appendix, and Tables 5 and 6 in the Online Appendix). This program uses the following Matlab functions that are provided in the same directory: *eq_chbar_t.m*, *eq_w0.m*, *eq_wt_cal.m*, *eq_Ah.m*, *eq_wt.m*, *eq_Ah_frp.m*. Depending on the sample used, the program requires one of three data files that are described next.

2. **"data_2020-05-15_1950_2010"**: This is the data file with the moments used to calibrate the model in the main results.

3. **"data_2020-05-15_1950_2010_+RE&Chem"**: Alternative data used to produce results in Table 5 in the Online Appendix. It also uses WORLDKLEMS data, but with real estate & chemical in the high skill sector.

4. **"data_2020-06-08_1970_2010_+RE&Chem&Elec&PubAdm"**: Alternative data used to produce results in Table 5 in the Online Appendix. It also uses WORLDKLEMS data, but with real estate, chemical, electrical, and public administration in the high skill sector.

5. **"data_USA"**: Alternative data used to produce results in Table 5 in the Online Appendix. It uses EUKLEMS, NAICS data.

6. **“homothetic_model.m”**: This file runs the simulation for the homothetic version of the model. These simulations are inputs to do the calculation in the excel file *‘Comparison_w_Leonardi.xlsx’* which is described next.

7. **“Comparison_w_Leonardi.xlsx”**: This excel file uses the simulation results from the homothetic version of the model that are calculated with the file *‘homothetic_model.m’*, together with the benchmark that are calculated with *‘main_cal_sim.m’*, to calculate the contribution of Leonardi's channel in our model, i.e., following his procedure for calculating the contribution of his channel in our model. These are the numbers are reported in Section 6.2.

8. **“sim_alt_rp.m”**: This program is used to run the simulations alluded to in Section 7.3., where we use a hypothetical, alternative series for the evolution of the price of services relative to goods.

9. **“explore_growth.m”**: This program is used to run the simulations reported in the last two columns of Table 6 in the Online Appendix.

Files in the subdirectory ‘Trade’:

1. **“sim_nx.m”**: This program is used to run the simulations reported in Section 7.2., where we allow for trade. This program uses different functions, with the same names as in the main directory (*eq_chbar_t.m*, *eq_w0.m*, *eq_wt_cal.m*, *eq_Ah.m*, *eq_wt.m*, *eq_Ah_frp.m*), which are provided in the subdirectory where this function is located, i.e., *'trade*'. In addition, it uses the data for the calibration of the benchmark model, *'data_2020-05-15_1950_2010*', and data on the net exports for "goods" and "services" in the file *'data_NX.txt*'.

2. **“data_NX.txt”**: This file contains data on the net exports for "goods" and "services".

Files in the subdirectory ‘endogenous_skills’:

1. **“sim_endo_fh.m”**: This program is used to run the simulations reported in Section 7.4., where we endogenize the supply of skills. It uses the functions and a data file in the main directory, except for the functions *'eq_wt.m*' and *'eq_Ah_frp.m*', which are called from the same directory. In addition, it uses the data for the calibration of the benchmark model, *'data_2020-05-15_1950_2010’*.

Files in the subdirectory ‘world’:

1. **“sim_cross_country.m”**: This program is used to run the simulation for the cross-country analysis reported in Section 8. This program uses different functions, with the same names as in the main directory (*eq_Ah.m*, *eq_chbar_t.m*, *eq_w0.m*, *eq_wt_cal.m*, *eq_wt.m*), which are provided in the subdirectory where this function is located, i.e., 'world'. The reason is that for some countries the consumption of market services might be zero. In addition, it uses data files with the moments needed to calibrate individual countries, ‘data_i’, where i=AUS, AUT, BEL, DNK, ESP, GER, ITA, JPN, NLD, UK, USA.

2. **“sim_cross_country_fixed_ah.m”**: This program is used to produce the model simulations reported in figures 7 and 8 of the Online Appendix. This program uses the following data files: data_USA, data_AUS, data_DNK, data_ESP, data_GER, data_ITA, data_JPN, data_NLD, data_UK, data_BEL. The program also uses the following Matlab functions: *eq_chbar_t.m*, *eq_w0.m*, *eq_wt_cal.m*, *eq_Ah.m*, *eq_Ah_ppp.m*, *eq_w00.m*. The main output is the file *output_fixed_ah.txt*, which is use to populate columns C-I of the excel file *model_fit_crosscountry.xlsx*.

Files in the subdirectory 'heterogeneous_rh':

1. **"sim_hetero_rh.m"**: This program is used to run the simulations reported in Table 7 of the Online Appendix. This program uses different functions, with the same names as in the main directory (*eq_Ah.m*, *eq_chbar_t.m*, *eq_w0.m*, *eq_wt_cal.m*, *eq_wt.m*), which are provided in the subdirectory where this function is located, i.e., *'heterogeneous_rh'*. In addition, it uses the data for the calibration of the benchmark model, *'data_2020-05-15_1950_2010'*.

Software Requirements

- **Stata**: code was ruen with Stata/SE 16.1 for Windows (64-bit x86-64).
    - `estout` : to install the package run:

```
ssc install estout, replace
```

- **Matlab**: code was run using MATLAB R2019b Update 4, 64-bit (maci64).
- **Excel**: we used Microsoft Excel for Microsoft 365 MSO.

Data Availability Statement

1. World KLEMS, United States, Labor File, April 2013 Release

These data are publicly available at http://www.worldklems.net/data/input/usa_wk_apr_2013_labour.xlsx . We use them in section 2.2 to compute the share of high-skill labor in total compensation by industry and year, the average share of high skill labor compensation in total compensation for the period of 1977 to 2005, and the average industry rank in terms of skill intensity for the period of 1977 to 2005. The average shares of high-skill labor in total compensation and the corresponding average ranks are used to determine which industries belong to the high-skill sector in the papers. The highest shares and ranking are discussed in Section 2.2 of the paper. A detailed Table with average high-skill labor shares and ranks under different employment measures can be found in Section 2 of the Online Appendix.

2. World KLEMS, United States, Labor File, April 2013 Release

These data are publicly available at http://www.worldklems.net/data/input/usa_wk_apr_2013_basic.xlsx .

3. EUKLEMS database, Basic File, November 2009 release.

These data are publicly available at http://www.euklems.net/euk09I.shtml . Fourteen (14) Countries Used: Australia (AUS), Austria (AUT), Belgium (BEL), Denmark (DNK), Spain (ESP), France (FRA), United Kingdom (UK), Germany (GER), Greece (GRC), Ireland (IRL), Italy(ITA), Japan(JPN), Sweden (SWE), United States-NAICS(USA NAICS) These data are used to produce Figure 1 in Section 2.3 of the paper using the .do file Figure1.do and

the two alternative versions of Figure 1 using two broader definitions of the High-Skill sector presented in Section 2 of the Online Appendix.

4. EUKLEMS database, Labor Input Files, March 2008 release.

These data are publicly available at http://www.euklems.net/euk08I.shtml . These data are used to produce the data used in the Cross Country Analysis in Section 8 of the paper. In particular, labor input data is used to compute the share of compensation of high-skill labor in the high- and the low-skill sector and the skill-premium, which are used as inputs in the calibration of the technology parameters for eleven (11) countries in the Cross Country validation exercise performed in Section 8. The countries used in this exercise are Australia (AUS), Austria (AUT), Belgium (BEL), Denmark (DNK), Spain (ESP), United Kingdom (UK), Germany (GER), Italy(ITA), Japan(JPN), the Netherlands (NLD), and United States-NAICS(USA NAICS). Compared to the countries in figure1 and figure2, we lose France, Greece, Ireland, Luxembourg, and Sweden because they do not have a labor input file available.

5. Penn World Tables 9.0

The data are described in Feenstra et al. (2015) and are publicly available at https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt9.0 . Cross-country data from this dataset on real GDP (rgdpna) and population (pop) are used in the .do files Figure1.do and Figure2.do to produce Figures 1 and 2.

6. US Bureau of Economic Analysis, Table 1.1.5, Gross Domestic Product

The Excel spreadsheet containing the annual series for US GDP is in Table 1.1.5 and publicly available at https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=2# . We use this data to compute net goods and net service exports as a percentage of GDP to perform the sensitivity analysis in Section 7.2 of the paper called "Allowing for Trade".

7. US Census Data, U.S. Trade in Goods and Services, Balance of Payment Basis

Annual data on goods and services exports and imports and the corresponding trade balance is publicly available at the US Census' Website at the link https://www.census.gov/foreign-trade/statistics/historical/gands.pdf . We use this data to compute net good and net service exports as a percentage of GDP to perform the sensitivity analysis in Section 7.2 of the paper called "Allowing for Trade".

8. US Bureau of Economic Analysis, Table 2.1, U.S. Trade in Services, by Type of Services

Annual data on services exports and imports by type of services are publicly available at the BEA's website (https://apps.bea.gov/iTable/iTable.cfm?reqid=62&step=6&isuri=1&tablelist=245&product=4). This data is available since 1999. We use it to check if service exports are close to the actual exports of the high skill sector for the period where both series are available (1977-2005). It turns out that services net exports are very close to the net exports of the high-skill sector.

9. US Input-Output Data

The exercise requires data from two tables, the Industry-by-Commodity Total Requirements table, and the use Use of Commodities by Industries table. In both cases, we use data *After Redefinitions* and at *Producer's Prices*. The data used for each year together at publicly available at the links in the table below.

Year	Industry-by-Commodity Total Requirements	Use of Commodities by Industries
2005	IxC Summary Level Tables 1997-2005	CxI Use Summary Level Tables 1997-2005
2002	IxC Summary Level Tables 1997-2005	CxI Use Summary Level Tables 1997-2005
1997	IxC and CxI Use - Summary Level Tables	
1992	IxC and CxI Use - Summary Level Tables	
1987	IxC and CxI Use - Summary Level Tables	
1982	IxC and CxI Use - Summary Level Tables	
1977	IxC and CxI Use - Summary Level Tables	

10. US Consumer Expenditure Surveys

The consumer expenditure data used to produce the results in Table 1 is publicly available at the BLS website (https://www.bls.gov/cex/pumd_data.htm#stata).

References

Feenstra, R. C. and R. Inklaar(2016): "Penn World Table 9.0", *Groningen Growth and Development Centre*, (https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt9.0).

Herrendorf, B., R. Rogerson, and A. Valentinyi(2013): "Two Perspectiveson Preferences and Structural Transformation", *American Economic Review*, 103,2752–89.

Dale W. Jorgenson, Mun S. Ho, and Jon Samuels, "A Prototype Industry-Level Production Account for the United States, 1947-2010" *Second World KLEMS Conference, Harvard University*, August 9, 2012.

Ruggles S., Flood S., Foster S., Goeken R, Pacas J., Schouweiler M., and Sobe M. (2021). "IPUMS USA: Version 11.0" [dataset].Minneapolisv,MN:IPUMS,2021. (https://doi.org/10.18128/D010.V11.0).

O'Mahony, M. and M. P. Timmer(2009): "Output, Input and Productiv-ity Measures at the Industry Level: The EU KLEMS Database", *The Economic Journal, Oxford University Press (OUP)*, (https://doi.org/10.1111/j.1468-0297.2009.02280.x), 119, F374–F403.