Do Elite Universities Overpay Their Faculty?*

César Garro-Marín†, Shulamit Kahn‡, and Kevin Lang§

This version: January 29, 2026

Overview

The code in this replication package produces all the paper's tables and figures using confidential data from the Survey Doctorate Recipients (SDR). The package uses both Stata and R.

One main file transforms all the data and produces the 2 figures and 4 tables in the main body paper, as well as the 17 tables and 6 figures in the Online Appendix.[1] The replicator should expect the code to run for about 5 days.

We use restricted-use data from the US National Center for Science and Engineering Statistics (NCSES), which is housed on a remote server in the NORC Data Enclave. The publicly available version of this package includes most of the code needed to reproduce the analyses in the paper, except for a few files that we cannot publish under NCSES disclosure rules. We list all the undisclosed scripts in Table 4. A full version of the package can be provided to researchers with SDR licenses. See https://ncses.nsf.gov/licensing for details on the license application process.

Data Availability and Provenance Statements

Statement about Rights

- Shulamit Kahn and César Garro-Marín have legitimate access to and permission to use the data used in this manuscript.
- All authors of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Data Sources and Availability

This paper primarily utilizes restricted-access data from the US National Center for Science and Engineering Statistics (NCSES), which cannot be shared publicly. Therefore, we do not include this data in this package. Table 2 and the accompanying description below provide a summary of all the datasets used in the paper and indicate whether they are included in the package.

- **Survey of Doctorate Recipients (SDR):** data on faculty job history and characteristics comes from the restricted-use files of Survey of Doctorate Recipients (SDR) from the National Center for Science and Engineering Statistics (NCSES, 1993–2019). This is a representative panel that contains demographic, salary, and employment information for STEM PhD recipients from US universities.

*This readme file is based on the Social Science Data Editors template created by Vilhuber et al. (2022).

†University of Edinburgh, email: cgarrom@ed.ac.uk
‡Boston University, email: skahn@bu.edu
§Boston University, email: lang@bu.edu

[1]Table A1 in the Online Appendix does not use any data and was done manually.

We use restricted-use files that cannot be made publicly available, but they can be accessed via a license agreement with the NCSES. Interested researchers can apply for access following the procedure described on this website: https://ncses.nsf.gov/licensing.

We use the complete files of all SDR waves between 1993 and 2019 and place them in the folder `data/raw/`. Panel A in Table 2 lists all the SDR waves we use, along with the data file names assumed by our code.

- **Integrated Post-Secondary Education Data System (IPEDS):** data on university and college characteristics comes from the IPEDS surveys from the National Center of Education Statistics (NCES, 2001, 2005, 2012, 2017). We use the complete files for the years 2001, 2005, 2012, and 2017. These datasets are publicly available and are included in the replication package. Interested readers can also download the complete files using this link: https://nces.ed.gov/ipeds/use-the-data/download-access-database.

  We placed entire csv data files for each wave in `data/raw/` using the names listed in Panel B of Table 2.

- **University and College Rankings:** our main sources for the institution rankings are the Times Higher Education 2017 World University Rankings and the Wall Street Journal-Times Higher Education 2017 College rankings (Times Higher Education, 2017*a*,*b*). In the paper, we refer to these two institution rankings simply as *THE rankings*. They are publicly available at https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking and https://www.timeshighereducation.com/rankings/united-states/2017, respectively.

  We supplement the THE rankings with data from the following US News rankings (US News & World Report, 2021*a*,*d*,*c*,*b*; U.S. News & World Report, 2022, 2023):

  - 2021 Best National Universities (newest version available at https://www.usnews.com/best-colleges/rankings/national-universities).
  - 2021 Best Liberal Arts Colleges (newest version available at https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges).
  - 2021 Regional Universities: North, South, Midwest, and West (newest versions available at https://www.usnews.com/best-colleges/rankings/regional-universities).
  - 2021 Regional Colleges: North, South, Midwest, and West (newest versions available at https://www.usnews.com/best-colleges/rankings/regional-colleges).
  - 2022 Best Graduate Schools: Biological Sciences (newest version available at https://www.usnews.com/best-colleges/rankings/national-universities)
  - 2022 Best Graduate Schools: Engineering Schools (newest version available at https://www.usnews.com/best-colleges/rankings/national-universities)

  We provide these rankings in the replication package. The US News Website only provides the latest versions of each ranking. Researchers can download the versions we used using the Internet Archive (https://web.archive.org/).

- **Consumer Price Index:** We convert all nominal monetary values into real 2020 dollars using the CPI from the US Bureau of Labor Statistics (U.S Bureau of Labor Statistics, 2023). We use the yearly-average CPI for All Urban Consumers, All Items, US City Average. We provide this dataset in the replication package, but it can also be downloaded at https://fred.stlouisfed.org/series/CPIAUCSL.

Additional Datasets Required for the Replication

Table 2 contains a list of additional datasets required for the replication. These datasets are additional to those listed in Table 1. Most of these datasets are derived from restricted-use SDR and result from case-by-case data cleaning. Due to the NCSES disclosure rules, we cannot publish them in this replication package, but they can be provided to researchers with SDR license agreements in the restricted-use version of this package.

Computational requirements

Replication of the paper requires using Stata and R. Below is a list of all the required packages. These packages are automatically installed in the code files `code/stata_setup.do` and `code/install_packages.R`. The code was last run using Stata 16 and R version 4.3.2 in the NORC Data Enclave servers.

Software Requirements

- Stata (code was last run with version 16). The program "`code/stata_setup.R`" automatically installs the latest available versions of these packages available in the NORC Data Enclave Server. The code was last run using the package versions available in the Data Enclave server as of January 1, 2026.
  - `estout`
  - `labutil`
  - `colrspace`
  - `ftools`
  - `reghdfe`
  - `rscript`
  - `grstyle`
  - `regsave`
  - `coefplot`
  - `parmest`
  - `palettes`
  - `unique`
  - `gtools`
  - `erepost`
  - `binscatter`
- R (4.3.2): The file "`code/install_packages.R`" installs the latest available versions of these packages available in the NORC Data Enclave server.
  - `pacman` (0.5.1)
  - `stringdist` (0.9.12)
  - `tidyr` (1.3.1)
  - `stringr` (1.5.1)
  - `tidyverse` (2.0.0)
  - `openintro` (2.5.0)
  - `igraph` (2.0.3)
  - `lfe` (3.0.0)
  - `reshape` (0.8.9)
  - `lattice` (0.22.5)
  - `gridExtra` (2.3)
  - `ggplot2` (3.5.0)
  - `gmm` (1.8)
  - `dplyr` (1.1.4)
  - `testit` (0.13)
  - `Rfast` (2.1.0)
  - `ff` (4.0.12)
  - `plyr` (1.8.9)
  - `broom` (1.0.6)

Memory and Runtime Requirements

This project *can only be run* in the remote servers of the NORC Data Enclave. Data use license agreements allow remote access from personal computers located anywhere within the United States.

The whole code was last run on Janurary 15, 2026 on the NORC Data Enclave Server (NCSES Desktop) with Windows 10.0.17763.6054 with 32 GB of RAM.

The main portion of the data cleaning takes up approximately 3.2 hours(`code/build_database/master_build.do`). However, computing the variance correction of the institution effects (`code/build_database/correct_KSS.do`) can take up to 5 days. Creation of tables and figures takes approximately 10 minutes.

Description of programs/code

The do file `code/master_do_file.do` runs all the code necessary to clean and transform the data, and produce all tables and figures. This program executes the following files:

- `code/stata_setup.do` loads three programs for configuring Stata, which are executed immediately after:
  - `set_global_vars` defines global variables used throughout the program.
  - `install_stata_dep` installs all the stata packages.
  - `clean_folders` sets the folder structure assumed by the code. When the erase option is set to yes, it erases all temporary and output files. If you do not want to start from scratch, set the option erase to no (`clean_folders, erase(no)`)
- `code/install_R_packages.R` installs all the required R packages.
- `code/build_data/master_build.do` executes all the data cleaning, produces the databases used for the analysis, and saves some of the regression estimates. The do file calls subprograms in the `code/build_data/` directory, which sequentially execute the data-cleaning steps.
- `code/build_data/correct_KSS.do` performs the Kline, Saggio and Sølvsten (2020) variance correction of the institution fixed effects. This is the most time-intensive portion of the code; it can take up to 5 days.
- `code/data_analysis/master_tables_and_figures.do` produces all the tables and figures from the paper. This do file sequentially calls the sub-programs in the code/data_analysis/, which produce the tables and figures included in the paper.

Instructions to Replicators

1. Edit the working directory and R_library path at the top of `code/master_do_file.do`. The lines that must be edited are appropriately indicated at the top of the code.
2. Edit the working directory and default R package library path in `code/R_setup.R`. The lines that must be edited are appropriately indicated at the top of the code.
3. Run `code/test_R_config.do` to test that the R configuration in Stata works. If everything is okay, the program will display the message **"R was configured successfully"**.
   While we have made every effort to facilitate replication by running all routines in Stata, the link between Stata and R on the NORC server can break, particularly when NORC updates the R version on the server. If `code/test_R_config.do` fails, we have included a troubleshooting guide (`code/fixing_rscript_problems.txt`) on ways to fix the most likely problems (based on our experience).
4. Execute `code/master_do_file.do` to run all steps in sequence.

Execution Details

The file `code/master_do_file.do` executes all required replication programs. This file executes all the routines from both Stata and R. Each program called in this do file can be run separately, **as long as they are executed in the order they appear in the file**.

- `code/stata_setup.do` configures Stata and installs or updates the required packages.
- `code/install_R_packages.R` installs all the required R packages.
  - Execution of the program from Stata using `rscript` will open a DOS window. **Do not close this window**. Although it might at times look like the program is hung, it is still running.
- `code/build_data/master_build.do` cleans and transforms the data. It calls all the required data cleaning programs.
  - This do file also executes R code and opens a DOS window several times. **Do not close this window**. Although it might at times appear to be hung, the program is still running.

  - If you wish to, the programs called by this do file can be run separately. However, **they have to be run in order.**
- `code/build_data/correct_KSS.do`: executes the programs required for the correction of fixed effect variances. This program is slow and computationally intensive. It can take up to 5 days.
  - This do file also executes R code and opens a DOS window. **Do not close this window**. Although it might at times appear to be hung, the program is still running.
  - If you wish, you can also run this code directly in R.
  - The results from this correction are mentioned in the body of the paper, but they are not required for the tables and figures. The replicator can also comment out this line and create the tables and figures without a problem.
- `code/data_analysis/master_tables_and_and_figures.do`: calls all the programs needed to create tables and figures.
  - These programs can be run individually in any order.

List of tables and programs

The programs listed in Table 3 create all tables, figures, and numbers in the text of the paper, except for Table A1, which was written manually and does not involve any data processing.

Table 4 lists all the programs not contained in the replication package. The NCSES disclosure rules prevent us from releasing these programs to the public, but they can be provided to SDR data license holders.

Table 1: Data sources list

Source name	Data files	Location	Type	Provided	Citation
A. Survey of Earned Doctorates					
1993 wave	esdr93.dta	data/raw/	Restricted use	No	NCSES (1993)
1995 wave	esdr95.dta	data/raw/	Restricted use	No	NCSES (1995)
1997 wave	esdr97.dta	data/raw/	Restricted use	No	NCSES (1997)
1999 wave	esdr99.dta	data/raw/	Restricted use	No	NCSES (1999)
2001 wave	esdr01.dta	data/raw/	Restricted use	No	NCSES (2001)
2003 wave	esdr03.dta	data/raw/	Restricted use	No	NCSES (2003)
2006 wave	esdr06.dta	data/raw/	Restricted use	No	NCSES (2006)
2008 wave	esdr08.dta	data/raw/	Restricted use	No	NCSES (2008)
2010 wave	esdr10.dta	data/raw/	Restricted use	No	NCSES (2010)
2013 wave	esdr13.dta	data/raw/	Restricted use	No	NCSES (2013)
2015 wave	esdr15.dta	data/raw/	Restricted use	No	NCSES (2015)
2017 wave	esdr17.dta	data/raw/	Restricted use	No	NCSES (2017)
2019 wave	esdr19.dta	data/raw/	Restricted use	No	NCSES (2019)
B. Integrated Post-Secondary Education Data System					
2001 wave	IPEDS_2001.csv	data/raw/	Publicly available	Yes	IPEDS (2001)
2005 wave	IPEDS_2005.csv	data/raw/	Publicly available	Yes	IPEDS (2005)
2012 wave	IPEDS_2012.csv	data/raw/	Publicly available	Yes	IPEDS (2012)
2017 wave	IPEDS_2017.csv	data/raw/	Publicly available	Yes	IPEDS (2017)
C. University and College Rankings					
THE World University Rankings	THE_WUR.xlsx	data/raw/	Publicly Available	Yes	Times Higher Education (2017*a*)
WSJ-THE College Rankings	THE_college.xlsx	data/raw/	Publicly Available	Yes	Times Higher Education (2017*b*)
USNWR National University Rankings	USNWR_national_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*a*)
USNWR Liberal Arts College Rankings	USNWR_liberal_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*d*)
USNWR Regional University Rankings, North	USNWR_north_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*c*)
USNWR Regional University Rankings, South	USNWR_south_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*c*)
USNWR Regional University Rankings, Midwest	USNWR_midwest_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*c*)
USNWR Regional University Rankings, West	USNWR_west_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*c*)
USNWR Regional College Rankings, North	USNWR_north_college_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*b*)
USNWR Regional College Rankings, South	USNWR_south_college_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*b*)
USNWR Regional College Rankings, Midwest	USNWR_midwest_college_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*b*)
USNWR Regional College Rankings, West	USNWR_west_college_rank.csv	data/raw/	Publicly Available	Yes	US News & World Report (2021*b*)
USNWR Best Graduate Schools: Biological Sciences	USNWR_field_rankings_bio.xlsx	data/raw/	Publicly Available	Yes	U.S. News & World Report (2022)
USNWR Best Graduate Schools: Engineering Schools	USNWR_field_rankings_eng.xlsx	data/raw/	Publicly Available	Yes	U.S. News & World Report (2023)
D. Other Datasets					
Consumer Price Index	CPIAUCSL.xls	data/raw/	Publicly available	Yes	U.S Bureau of Labor Statistics (2023)

Notes: The table lists the source and location of all the datasets used in the paper. The Data files column lists the file names *assumed by our code* and not the original names of the data files.

Table 2: List of datasets needed for replication

Data file	Source	Notes	Confidential	Provided
A. Manual modifications files				
cuny_corrections.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
drop_schools_renamed.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
final_institution_list_medical.dta	Generated from SDR data	Contains manual corrections necessary for data cleaning. Flags list of some schools dropped from the analysis	Yes	No
inconsistent_movers_shu.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
leave_check_v4_renamed.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
manual_check_v1_processed.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
mn_corrections.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No

Continues on next page

Continued from previous page

Data file	Source	Notes	Confidential	Provided
modification_list_old_part_shu.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
old_panelid_key.dta	Generated from SDR data	Contains a crosswalk of person ids needed to ensure consistency with older version of the code	Yes	No
people_to_check_bug.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
people_to_check_medical.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
people_to_check_new.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
people_to_check_new_proc.dta	Generated from SDR data	Contains manual corrections to individual data necessary for data cleaning	Yes	No
reintroduced_schools.dta	Generated from SDR data	Contains manual corrections to some instution codes	Yes	No
switcher_file_coding_mistake.dta	Generated from SDR data	Contains manual corrections needed to ensure consistency with older version of the code	Yes	No
things to be changed.xlsx	Generated from SDR data		Yes	No

Continues on next page

Continued from previous page

Data file	Source	Notes	Confidential	Provided
B. Other datasets				
inst_labels.dta		Associates institution names to IPEDS codes	No	Yes
iped_dataset_v1.dta	IPEDS	Contains institution-level characteristics	No	Yes
osep_to_iped_rev2014.dta		Cross-walk between OSEP and IPEDS institution codes	No	No
unranked_list.xlsx	Created by authors	Manual corrections to some institution codes	No	Yes

Notes: This table lists data files required for the replication that are additional to those in Table 1. All these files must be placed in `data/raw`.

Table 3: List of Tables, Figures, and Programs

Exhibit	Panel (If Applicable)	Program	Output files
Main Figures			
Figure 1	(a)	create_figure_main_binscatter.do	figure_salary_ranking.png
Figure 1	(b)	create_figure_binscatter_additional.do	figure_binscatter_weighted_movers.png
Figure 2		create_figure_mobility_summary.do	figure_hist_l_change_ranking_within.png
Main Tables			
Table 1		create_table_summary_stats.do	table_summary_stats_clean.tex
Table 2		create_table_variance_decomp.do	table_variance_decomp_clean.tex
Table 3		create_table_premiums_rankings.do	table_premiums_rankings_clean_mixed.tex
Table 4		create_table_premiums_endowment.do	table_premiums_endowment_clean_mixed.tex
Appendix Figures			
Figure C1		create_figure_mobility_summary.do	figure_hist_l_change_ranking_within_tenured.png
Figure C2		create_figure_mobility_summary.do	figure_binscat_d_salary_vs_d_rankings.png
Figure C3	(a)	create_figure_event_studies.do	figure_event_prestige_panel_A_universities.pdf
Figure C3	(b)	create_figure_event_studies.do	figure_event_prestige_panel_B_colleges.pdf
Figure C4	(a)	create_figure_main_binscatter.do	figure_fe_ranking.png
Figure C4	(b)	create_figure_binscatter_additional.do	figure_binscatter_grouped.png
Figure C5	(a)	create_table_field_specific_results.do	binscatter_uw_field_all_bio.png
Figure C5	(b)	create_table_field_specific_results.do	binscatter_uw_field_bio.png
Figure C5	(c)	create_table_field_specific_results.do	binscatter_uw_field_all_eng.png
Figure C5	(d)	create_table_field_specific_results.do	binscatter_uw_field_eng.png
Figure C6		create_figure_event_studies.do	figure_event_coworker.pdf
Appendix Tables			
Table B1		create_table_summary_stats.do	table_summary_stats_raw.tex
Table B2		create_table_variance_decomp.do	table_variance_decomp_raw.tex
Table B3		create_table_premiums_rankings.do	table_premiums_rankings_raw_mixed.tex
Table B4		create_table_premiums_endowment.do	table_premiums_endowment_raw_mixed.tex
Table B5		create_table_transition.do	table_transition_probability.tex
Table B6		create_table_transition.do	table_transition_probability_tenured.tex
Table B7		create_table_transition.do	table_transition_salary.tex
Table B8		create_table_transition.do	table_transition_salary_tenured.tex
Table B9		create_table_AKM_first_stage.do	table_AKM_first_stage_nosen.tex
Table B10		create_table_tenured.do	table_tenured_clean_nosen.tex
Table B11		create_table_field_specific_results.do	table_field_specific_results_condensed_nosen.tex
Table B12		create_table_one_step_time_varying.do	table_one_step_time_varying.tex
Table B13		create_table_one_step_estimates_w_origin.do	table_pv-r.tex
Table B14		create_table_job_satisfaction.do	table_jobsat_rankings_sat_combined_mixed.tex
Table B15		create_table_transition_coworker.do	transitions_table_salary_coworker.tex
Table B16		create_table_transition_coworker.do	transitions_table_probability_coworker.tex
Table A2		create_table_ranking_imputation.do	table_ranking_imputation.tex

Notes: all code for producing tables and figures is located in the folder `code/data_analysis/`. Output files from all figures are located in the folder `results/figures/`, while tables are located in `results/tables/`.

Table 4: List of Undisclosed Programs

File Name	Location	Description

Notes: The table lists the programs we are unable to publish due to the NCSES disclosure rules. These programs are included in the restricted-use version of this package that can be provided to SDR license holders.

References

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. "Leave-Out Estimation of Variance Components." *Econometrica*, 88(5): 1859–1898.

National Center for Education Statistics (NCES). 2001, 2005, 2012, 2017. "Integrated Postsecondary Education Data System (IPEDS)." Retrieved from https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?year=-1&sid=4dda9a06-7adf-4ad1-9eaa-a07908708f04&rtid=7, April 24, 2023.

National Science Foundation, National Center for Science and Engineering Statistics (NCSES). 1993, 1995, 1997, 1999, 2001, 2003, 2006, 2008, 2010, 2013, 2015, 2017, 2019. "Survey of Doctorate Recipients."

Times Higher Education. 2017*a*. "World University Rankings 2017." Retrieved from https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking, July, 7, 2021.

Times Higher Education. 2017*b*. "WSJ/THE College Rankings 2017." Retrieved from https://www.timeshighereducation.com/rankings/united-states/2017, July 7, 2021.

U.S Bureau of Labor Statistics. 2023. "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL]. retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL, April 24, 2023."

US News & World Report. 2021*a*. "Best National University Rankings 2021." Retrieved from https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges, May 7, 2021.

US News & World Report. 2021*b*. "Best Regional Colleges Rankings 2021." Retrieved from https://www.usnews.com/best-colleges/rankings/regional-colleges, May 7, 2021.

US News & World Report. 2021*c*. "Best Regional Universities Rankings 2021." Retrieved from https://www.usnews.com/best-colleges/rankings/regional-universities, May 7, 2021.

US News & World Report. 2021*d*. "National Liberal Arts Colleges Rankings 2021." Retrieved from https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges, May 7, 2021.

U.S. News & World Report. 2022. "Best Graduate School: Biological Sciences Rankings." Retrieved from https://www.usnews.com/best-graduate-schools/top-science-schools/biological-sciences-rankings, January 30, 2025.

U.S. News & World Report. 2023. "Best Graduate Schools: Engineering Rankings." Retrieved from https://www.usnews.com/best-graduate-schools/top-engineering-schools/eng-rankings, January 30, 2025.

Vilhuber, Lars, Marie Connolly, Miklós Koren, Joan Llull, and Peter Morrow. 2022. "A Template README for Social Science Replication Packages." Available at: https://social-science-data-editors.github.io/template_README/.