README: Data and Code Why Does Education Increase Voting? Evidence from Boston's Charter Schools

Sarah Cohodes and James Feigenbaum

July 2025

Overview

The code in this replication package constructs the analysis file from charter school lottery files and Massachusetts Department of Elementary and Secondary Education student-level data using Stata. A master data file runs all of the relevant do files, which first do a basic clean of the relevant datasets (Section 1), then generate the analysis file for the project (Section 2), and finally analyze the data (Section 3).

Data Availability and Provenance Statements

The data for this project are confidential, but may be obtained with Data Use Agreements with the Massachusetts Department of Elementary and Secondary Education (DESE) and Massachusetts charter schools. Researchers interested in access to the data may contact research@doe.mass.edu and www.doe.mass.edu/research/contact.html, see also www.doe.mass.edu/research/agenda.html "Office of Planning and Research: Research Agenda") for access to the enrollment, student characteristic, test score, and teacher data. Access to charter school lottery records comes from collecting lottery records from individual charter schools over many years. See Table A1 for the list of participating charter schools. Contact information for individual charter schools is available through the Massachusetts DESE website (http://www.doe.mass.edu/charter/ – see the page "Charter School List"). It can take some months to negotiate data use agreements and gain access to the data.

Sarah Cohodes (scohodes@umich.edu) will maintain an archive of the data until 12 months after publication of the article at the Education Policy Initiative (EPI) at the University of Michigan Ford School of Public Policy (UMich), at which time we are required to destroy the Massachusetts DESE data per the UMich agreements with DESE. The charter lottery records, programs, and

documentation will be archived and stored indefinitely on the UMich secure server. She will provide reasonable assistance to requests for clarification and replication.

This document provides instructions on how that archive can be replicated if the data are obtained by another researcher or accessed via EPI's secure server. If working in EPI's secure server, the only manual setup is to create a folder with your username in the `charter_voting` directory and to update file paths in the first do file. You will need a version of Stata with capability to import SPSS datasets. Alternatively, Stat/Transfer can be used to convert SPSS files to Stata.

DESE provided data on student demographics and school enrollment (SIMS), test scores (MCAS), AP and SAT scores, college enrollment (NSC), and courses (SCS). See the text for decisions around sample inclusion and data coding.

The master do file `0_voter_master.do` constructs the relevant file paths. Updating the paths via local and global macros at the top of the master do file will point them to your directory.

Statement about Rights

☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

Summary of Availability

☐ All data **are** publicly available.
☐ Some data **cannot be made** publicly available.
☑ **No data can be made** publicly available.

☑ Confidential data used in this paper and not provided as part of the public replication package will be preserved for one year after publication, in accordance with the authors' DUA with the Massachusetts Department of Elementary and Secondary Education.

Details on each Data Source

- Student Information Management System (SIMS) (DESE, 2023c):
  - DESE provided each end of year ("eoy") data file in SPSS format as SIMSD**.sav where ** is for years 02-23. The program converts the .sav files to a Stata data file, saving the files as simseoy**.dta in the SIMS folder. The EOY data is the main data used for student information, but for high school graduation, we also pull from simsoct**.dta, which is the October version of the file.

  - The SIMS data handbook is available at: https://www.doe.mass.edu/infoservices/data/sims/sims-datahandbook.docx
  - The data should be saved in raw/SIMS.

- Massachusetts Comprehensive Assessment System (MCAS) (DESE, 2023a):
  - DESE provided MCAS files in SPSS format as MCAS **** megafile.sav where **** is for years 2002-2023 (some file names have additional words in them like "official"). The program converts the .sav files to a Stata data file, saving the files as mcas**.dta in the MCAS folder where ** is 02-23. Note that some years of MCAS data are omitted due to adoption of a new test (by different districts at different times) and the Covid-19 pandemic.
  - General information on the MCAS is available at: https://www.doe.mass.edu/mcas/results.html. DESE provides an Excel spreadsheet with a codebook along with the MCAS data.
  - The data should be saved in raw/MCAS.

- National Student Clearinghouse (NSC)(NSC, 2023):
  - DESE provided NSC data in various formats. We use all files ever provided by DESE, as NSC is continually updating its databases and a student previously submitted may match in a later year. This includes students who graduated high school in 2003-2023 (or should have, but left the data).
  - The two most common files cover several years of graduates and non-graduates each. Each of these files includes high school graduates in the relevant class years and in the case of nongrads, 9th grade students who do not graduate but would have in the relevant high school classes had they remained in Massachusetts public high schools. These students may have dropped out, moved out of state, or switched to a private school.
  - There are also several one-off NSC matches with unique file names, detailed in 1.4_nsc.do
  - Information on NSC data fields is available at: https://nscresearchcenter.org/workingwithourdata/.
  - The data should be saved in raw/NSC.

- Student Course System (SCS) (DESE, 2023b):
  - DESE provided each SCS file in SPSS format SCS**.sav where ** is for years 10-23. (Note that the SCS system began in 2010.) The program converts the .sav files to a Stata data file, saving the files as as scs**.dta in the SCS folder.
  - The SCS data handbook is available at: https://www.doe.mass.edu/infoservices/data/scs/scs-datahandbook.docx.
  - The data should be saved in raw/SCS.

- Advanced Placement (AP) (College Board, 2023a):

  - DESE provided each AP file in Excel spread format ap_20xx_student_level_data.xlsx where ** is for years 07-22. (2023 was provided with slightly different nomenclature.) The program inputs the Excel data and saves it in Stata format.
  - General information on the AP test program is available here: https://apstudents.collegeboard.org/courses. Each subject exam is scored from 1 to 5.
  - The data should be saved in raw/AP.

- SAT (College Board, 2023b):
  - DESE provided each SAT file in Excel spread format sat_20xx_student_level_data.xlsx where ** is for years 07-23. The program inputs the Excel data and saves it in Stata format.
  - General information on the SAT test program is available here: https://satsuite.collegeboard.org/sat. The SAT has changed the test and scoring over the years. For consistency over years, we only use the Math (800) and Verbal (800) SAT scores for a total SAT Reasoning score of (1600).
  - The data should be saved in raw/SAT.

- Voting Records:
  - States vary in the accessibility of voter files. Massachusetts maintains a centralized record, but access is limited; see https://www.mass.gov/info-details/mass-general-laws-c51-ss-47c. Commercial vendors collect this information from authorized parties, add their own analytics, and make these files available for purchase for political purposes.
  - Voting files came from various sources, including a 2018 file provided by DESE, a 2015 file gathered by the researchers from NationBuilder, and 2018 and 2020 files from L2. Parent information came from a 2009 voter file from Catalist. Contact individual vendors to negotiate access. Alternatively, many universities have access to these files via their libraries.
  - See the Online Appendix for details on the fuzzy matching process used to match student records to voter records.

- Additional Data Sources:
  - Data from Charter School Lottery Records: Each participating charter school provided yearly lottery data, typically in Excel, which we compiled and formatted into school-year datasets and matched to state student IDs (SASIDs) based on name, date of birth, and other characteristics.
  - Descriptive information on colleges comes from the US Department of Education's Integrated Postsecondary Education Data System (IPEDS), accessed via the Urban Institute's Data Explorer which automatically downloads the relevant files. We have provided this file

as part of the data repository in case the Data Explorer is no longer available in the future.

Computational requirements

A master data file `0_voting_master.do` runs all of the relevant do files, which first do a basic clean of the relevant datasets (Section 1), then generated the analysis file for the project (Section 2), and finally analyze the data (Section 3). The master do file sets up appropriate directories and installs the relevant data packages.

Software Requirements

☑ The replication package contains one or more programs to install all dependencies and set up the necessary directory structure.

- Stata (code was last run with version 18.5)
  - `ivreg2`
  - `ranktest`
  - `reclink`
  - `center`
  - `unique`
  - `jarowinkler`
  - `blindschemes`
  - `grstyle`
  - `libjson`
  - `educationdata`
  - `reghdfe`
  - `coefplot`
  - `ftools`
  - `regsave`
  - The program "`0_voting_master.do`" will install all dependencies locally. For reasons unknown to the authors, the user-written command `ivreg2` sometimes produces the error `struct ms_vcvorthog undefined`. If this happens, delete the file livreg2.mlib from the ado directory and reinstall `ivreg2`. A similar problem occurs when using the user-written command `educationdata`. If you run into an error using this package, delete all of the files from the ado directory associated with the `educationdata` and `labutil` packages and reinstall them.

Controlled Randomness

☑ Random seed is set at line 28 of program `0_voting_master.do`

☐ No Pseudo random generator is used in the analysis described here.

Memory, Runtime, Storage Requirements

Summary Approximate time needed to reproduce the analyses on a standard 2024 desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☐ 2-8 hours
- ☐ 8-24 hours
- ☑ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days

Approximate storage space needed:

- ☐ < 25 MBytes
- ☐ 25 MB - 250 MB
- ☐ 250 MB - 2 GB
- ☐ 2 GB - 25 GB
- ☑ 25 GB - 250 GB
- ☐ > 250 GB
- ☐ Not feasible to run on a desktop machine, as described below.

Details The code was last run on a virtual Windows desktop (64-bit x86-64) in a secure server environment with a AMD EPYC Processor (with IBPB) 2.90 GHz and 56 GB of RAM. The server users Windows Server 2022 Datacenter.

Description of programs/code

- Program `0_voting_master.do` sets up directory paths, installs necessary programs, and runs all of the programs to clean, merge, and analyze data for this project. Be sure to update the directory paths here to direct them to your replication directory.
- Programs in Section 1 with prefix `1.x_` clean and merge datafiles of the same type together and prepare them to be used in the project. For example, programs in this section clean and merge multiple years of SIMS data.
- Programs in Section 2 with prefix `2.x_` clean and merge datafiles specific to this project on the relevant sample of students. This includes constructing the analysis file.
- Programs in Section 3 with prefix `3.x_` conduct the analysis for the paper.

Instructions to Replicators

- Unzip the programs from the replication file and that should create the `do` folder with all the necessary Stata programs.

  - Create folders for `data_raw` and `data_lotteries` if importing raw data to a new replication environment. This is not necessary if replicating within the EPI secure server.
- Edit `0_voting_master.do` to adjust the default paths to the code and raw data. Set up folders to match those in the globals defined in the program.
- Run `0_voting_master.do` which will install software extensions and run the entire set of code to clean, merge, and analyze data. You may comment in and out the relevant programs to run them one-by-one but must run them in order.
- Make sure all the switches are turned on to "1" in each program after the master program.

Details

Configuration

- `do/0_voting_master.do`: will create all global paths, install needed ado packages, and run the entire set of code for the project.

Section 1: Data Setup

- `do/0_data_setup.do`: Wrapper file for all of the data set up code.
- `do/1.1_sims.do`: Cleans individual years of SIMS data and merges them together.
- `do/1.1.1_sims.do`: Cleans individual years of additional SIMS data and merges them together.
- `do/1.2_mcas.do`: Cleans individual years of MCAS data and merges them together.
- `do/1.3_nsc.do`: Cleans individual years of NSC data and merges them together. As this program involves many millions of records, it can take some time and memory to run. If running on a server environment, it is ideally run in the evening or weekend to avoid hogging memory for other users.
- `do/1.4_ap.do`: Cleans individual years of AP data and merges them together.
- `do/1.5_sat.do`: Cleans individual years of SAT data and merges them together.
- `do/1.6_lottery.do`: Cleans individual years and schools of lottery data and merges them together.
- `do/1.7_sims_susp.do`, `do/1.7_ssdr.do`, `do/1.7.1_combine_susp.do`: Cleans individual years of discipline data and merges them together. There are multiple sources for discipline data.
- `do/1.8_hsgrad.do`: Cleans individual years of SIMS data to generate high school graduation information and merges them together. `do/1.9_collegechars.do`: Cleans college characteristics data.
- `do/1.10_vocal.do`: Cleans school climate survey data.
- `do/1.11_clean_state.do`: Prepares data for full state used for descriptive

statistics.
- `do/1.12_ipeds.do`: Downloads and cleans IPEDS data.
- `do/1.13_scs.do`: Cleans individual years of SCS data and merges them together.

Section 2: Sample-Dependent Data Setup

- `do/2.1_voter.do`: Cleans voter file data.
- `do/2.2_voter_lottery.do`: Selects appropriate sample from lottery data.
- `do/2.3_baseline.do`: Uses combined SIMS data to identify baseline student characteristics.
- `do/2.4_hsgrad.do`: Uses combined hsgrad data created in `do/1.8_hsgrad.do` to identify high school graduation. This must be run after the identification of the analysis sample as it counts high school graduation within years of the lottery application.
- `do/2.5_nsc.do`: Uses combined NSC data created in `do/1.3_nsc.do` to generate college enrollment and college graduation indicators. This must be run after the identification of the analysis sample as it counts college within years of projected high school graduation year based on lottery application.
- `do/2.6_votermatch.do`: Matches voter data to student data.
- `do/2.7_voter_setup.do`: Combines all of the DESE data with the lottery data, creates necessary variables, and generates the main analysis file.
- `do/2.8_daveleip.do`: Cleans community voter participation information.
- `do/2.9_bps.do`: Combines all of the DESE data together, creates necessary variables, and generates the Boston descriptives file based on 9th grade year.

Section 3: Analysis

- `do/3.1_analysis_voter.do`: Conducts all the student-level data analysis to create tables and figures for the project, including appendix tables and figures.
- `do/3.2_analysis_parent.do`: Conduct parent analysis using a different set of lottery data.

List of tables and programs

The provided code reproduces:

☑ All numbers provided in text in the paper
☑ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below.

The table below lists the exhibits and location in the relevant programs. For concision, we use nicknames for the do files that create the exhibits: 3.1_analysis_voter.do is referred to in the table as 3.1 and 3.2_analysis_parent.do is referred to as 3.2.

Exhibit #	Program	Line Number	Output file
Table 1	3.1	496	tn_means.pdf
Figure 1	3.1	2,468	f_voting.pdf
Figure 2	3.1	3,594	f_sub_combined.pdf
Figure 3	3.1	3,328	f_civic.pdf
Figure 4	3.1	1,727	f_peers.pdf
Figure A1	3.1	2,635	f_close.pdf
Figure A2	3.1	3,705	f_cog_noncog.pdf
Tables A1-A5	3.1	1,223	tn_academic.tex, tn_hsprogress.tex, tn_attainment.tex, tn_attend.tex, tn_susp.tex
Table A6	3.1	1,643	tn_instate_outstate.tex
Table A7	3.1	2,170	tn_voter.tex
Table A8	3.1	3,529	tn_voter_sub.tex
Table A9	3.1	2,635	tn_voter_close.tex
Table A10	3.1	3,705	tn_index_sub.tex
Table A11	3.1	3,459	tn_mission.tex
Table A13	3.1	3,180	tn_community.tex
Table B1	3.1	4,435	tn_votertab.tex
Table C1-C2	3.1	722	tn_lottorecords1.tex, tn_lottorecords2.tex
Table 1	3.1	722	tn_lottorecords1.tex
Table C3	3.1	427	tn_covbal.tex
Table C4	3.1	323	tn_voter_pred.tex
Table C5	3.1	913	tn_simsmatch.tex
Table C6	3.1	995	tn_attrition.tex
Table C7	3.1	1,064	tn_firstage.tex
Figure E1	3.1	2,253	f_overtime.pdf
Figure E2	3.1	4,198	f_omit.tex
Table E1	3.1	4,285	tn_robust.tex
Table E2	3.1	2,870	tn_voter_detailed.tex
Table E3	3.1	3,032	tn_voter_imputed.tex
Table F1	3.2	783	t_parent_vote.tex
Table F2-F3	3.2	630	t_covbal_parent1.tex, t_covbal_parent2.tex

Please note that some tables include some rounding to improve the formatting of the tables. The Stata output will include more significant digits. Table A12 does not have data analysis and is thus not reported in this table.

References

Massachusetts Charter School Lottery Records. 2002–2014. Collected by the authors and their research collaborators from individual schools. Multiple

electronic files.

Massachusetts Department of Elementary and Secondary Education. 2001-2023a. Massachusetts Comprehensive Assessment System (MCAS). Commonwealth of Massachusetts. Multiple electronic files.

Massachusetts Department of Elementary and Secondary Education. 2011-2023b. Student Course System (SCS). Commonwealth of Massachusetts. Multiple electronic files.

Massachusetts Department of Elementary and Secondary Education. 2001-2023c. Student Information Management System (SIMS). Commonwealth of Massachusetts. Multiple electronic files.

National School Clearinghouse. 2003-2023. College Enrollment Records for Massachusetts Students. Provided by the Massachusetts Department of Elementary and Secondary Education, Commonwealth of Massachusetts. Multiple electronic files.

The College Board. 2007-2023a. Advanced Placement (AP) Scores for Massachusetts Students. Provided by the Massachusetts Department of Elementary and Secondary Education, Commonwealth of Massachusetts. Multiple electronic files.

The College Board. 2007-2023b. SAT Scores for Massachusetts Students. Provided by the Massachusetts Department of Elementary and Secondary Education, Commonwealth of Massachusetts. Multiple electronic files.

US Department of Education Integrated Postsecondary Education Data System (IPEDS), Education Data Portal (Version 0.22.0), Urban Institute, accessed January 2025, https://educationdata.urban.org/documentation/, made available under the ODC Attribution License.

Acknowledgements

This README file was based on the template README file from the Social Science Data Editors available at: https://social-science-data-editors.github.io/template_README/template-README.html.