This README file describes the data and analysis conducted for:

"Price Regulation, Price Discrimination, and Equality of Opportunity in Higher Education: Evidence from Texas"

Published in *American Economic Journal: Economic Policy*

by

Rodney J. Andrews (The University of Texas at Dallas , rodney.j.andrews@utdallas.edu) and

Kevin Stange (The University of Michigan, kstange@umich.edu)

April 2019

Data access

This project uses administrative student and workforce micro data housed at the Education Research Center at the University of Texas - Dallas.   This is non-public data that we are not permitted to share. Any research must be approved by an oversight board, done on-site in Dallas, with output cleared through a formal review process. The data and procedure to access it is described here: https://www.utdallas.edu/research/tsp-erc/access.html.

Dataset construction

To construct our student level dataset (*student_data.dta*), we began by collecting all graduates from public high schools in Texas between 1996 and 2011, from the PEIMS (Public Education Information Management System) graduation report, dropping any students with invalid PEIMS identification codes, SSN identification codes, with multiple PEIMS identification codes associated with the same SSN identifier, with multiple SSN identification codes associated with the same PEIMS identifier, or those who showed up as graduating multiple times. To this base sample, we merged student demographic information using the PEIMS demograd file to obtain student gender and ethnicity, and using the PEIMS enrollment files for the two school years prior to graduation, in order to obtain student's free or reduced price lunch eligibility, limited English proficiency status, and special education status. Next, we merged high school exit exam scores in mathematics and reading from the TAAS exit exam files for the 1996 to 2003 cohorts, or from the TAKS exit exam files, for the 2004 through 2011 cohorts. For each student, we extracted the first valid test score, either from the initial administration, or from the subsequent re-test administrations if the initial administration was missed; these exam scores were then normalized by exam type and year of administration. We then followed this sample's entry in to the post-secondary educational system, tracking students' application behavior and acceptance decisions using the THECB application dataset. From here, we tracked students' enrollment into junior and senior colleges for all students between 1996 and 2016 using the THECB junior and senior enrollment files. This data was used to construct indicators for whether or not a student had enrolled in a junior college within 2-years of high school graduation, whether or not a student had enrolled in a senior college within 2-years of high school graduation, as well as to track students' school of initial enrollment and their initially declared major. For students enrolled in senior colleges, we went on to track their continued enrollment, as well as which school they were enrolled in and what their declared major was.

Using the THECB Financial Aid Data (FAD), we were able to merge on financial aid information for those students enrolled in senior colleges, keeping track of grant and scholarship awards by school and year. Subsequently, we merged on graduation data from the THECB graduation reports, keeping track of each degree earned, the institution where it was earned and the subject it was earned in. Finally, for the full set of high school graduates, we compiled quarterly earnings data. We went on to use a transformed version (taken from the residuals of a regression of the logged earnings on year and quarter fixed effects) of the earnings at 10 years after high school graduation for those students who graduated between 2000 and 2002 to create a measure of predicted earnings for each school and major combination in Texas, using as a counterfactual those high school graduates who had not enrolled in a post-secondary education in Texas. Additionally, this data was merged with pricing data collected outside of the ERC. The final dataset is a student level data set which tracks student's educational outcomes and includes their demographic information as well as the predicted earnings associated with their first program (institution X major) enrolled.

To construct our program-level dataset (*program_level_data.dta*), we created a dataset which tracked senior college instructor characteristics between 2000 and 2016 from the THECB class and faculty reports, keeping information on instructor rank, salary, ethnicity, gender, age, number of unique courses and sections taught, class enrollment, and the department/school they were teaching for. Additionally, this data was merged with pricing data collected outside of the ERC. The final data set contains the characteristics of the courses, sections, instructors, and students in a given program over time.

Data analysis

The following Stata do files conduct our main analysis.

Andrews_Stange_Combined_Analysis_Part1.do uses *student_data.dta* to conduct a series of descriptive regressions of the form reg `outcome' poor post poor_post `controls' (possibly for some subsamples), saving the results in *descriptive_regression_results.dta*.

Andrews_Stange_Combined_Analysis_Part2.do uses *student_data.dta, program_level_data.dta*, and *descriptive_regression_results.dta* to produce all tables and figures contained in the paper and appendix materials.

Andrews_Stange_tuition_descriptives.do uses *tuition_CIP_matches.dta,* the hand-collected tuition pricing data to create time trends in the dispersion of program prices, shown in Figure1.

Andrews_Stange_control_state_analysis.do uses six different datasets (contained in this zip file) to generate the control state analysis contained in Appendix C.

Datasets included in this zip file

1. *tuition_CIP_matches.dta* - this dataset has an observation for each program (institution X major) in each year. It was hand-collected by Kevin Stange and Jeongeun Kim from course catalogs and archival sources and captured separately for each identifiable program (with a distinct tuition or fee), residency status, undergraduate level, academic year, entering cohort, and number of credit hours. Price data was then converted to the CIP2 level for merging onto the THECB data. Our tuition measure is "tuition14", which is the sticker price for in-state juniors taking 15 SCH in a given year

2. ipeds_fouryear.dta - four-year institutions from IPEDS

3. IPEDS_TX_2002 - list of Texas institutions in our analysis
4. ump_rate_0010.dta - unemployment rate by state by year
5. MERGED2011_12_PP.dta - data extracted from College Scorecard for 2011-12 (earnings by institution)
6. Barrons7208.dta - Barron's codes for institutions
7. pell_merged_1974_to_2012.dta - number of Pell recipients by institution from FSA