Data Appendix for 'Student Portfolios and the College Admissions Problem'

Hector Chade[*] Gregory Lewis[†] Lones Smith[‡]

November 5, 2013

Data Source: We use data from the Freshman Survey, run annually by UCLA's Higher Education Research Institute (HERI). The data is public use archive data, downloadable here:

https://heri.gseis.ucla.edu/data-archives

The survey is given to entering freshman by participating colleges, and the results are pooled and analyzed by HERI. There are over 700 participating two- and four-year colleges. Our data is for the subsample of four-year colleges only. As a result, it is representative only of those students who successfully matriculate (i.e. are admitted and attend) at a four-year college. The public use data available to us spans the period 1975–1999 (additional data is available for more recent years by application). However, a survey question asking about SAT scores was only added in the 1986 survey, so we end up using data from 1986-1999.

Data Cleaning and Analysis The variables we employ in the analysis are income, number of applications sent, SAT scores (both verbal and math) and ACT score. The number of applications is top-coded at 7, and we impute a mean number of

[*] Arizona State University, Department of Economics, Tempe, AZ 85287.
[†] Harvard University, Department of Economics, Cambridge, MA 02138.
[‡] University of Wisconsin, Department of Economics, Madison, WI 53706.

applications by assuming that the distribution is geometric with success probability $\frac{1}{3}$— an approximation that fits the observed data well, and is validated by additional data from 1997 onwards, when the survey instrument was changed to top-code only at 12.

To get a single measure of intellectual aptitude that is consistent across years, we do a few conversions. First, we convert ACT to SAT using the 2008 concordance table. Next, we converted all the SAT scores from before the 1995 re-centering to the modern equivalent, using the conversion tables published by the College Board. Finally, we took the sample of students who took both tests in a single year, and if our SAT score converted from the ACT appeared to be systematically different because of different standards in that year (i.e. the SAT mean was higher than the ACT mean that year, or vice versa), we moved all the scores up or down accordingly. Then we got a single SAT score by taking the average of the two scores for students who took both, and otherwise taking one or the other. As a sanity check, the mean SAT scores do not vary much across years after all these conversions.

Finally, income was only reported in categories, with categories that changed over time. To address this, we assumed that income was log normal with year specific means and common variance, and imputed incomes based on the censored data using an interval regression. Notice that the specifics of the imputation are not all that important here, since we only used quartiles of the income distribution in the analysis.

The two figures shown in the introduction are outcomes of local polynomial regressions, where the x-variable is SAT score, and the y-variable is the number of applications sent and matriculation outcome (binary) respectively. The figures were obtained from the "lpolyci" command in Stata, using a polynomial of degree 1, an Epanechnikov kernel and a bandwidth of 40.

Included files: The zip archive includes three files:

- "cleaning.do" is a Stata do file that takes three large raw data files (available on request) and processes them as described above to produce...
- "college_cleaned.dta", a Stata data file
- "graphs.do" is a Stata do file that produces the graphs used in the papers