Students' Heterogeneous Preferences and the Uneven Spatial Distribution of Colleges

Chao Fu, Junjie Guo, Adam J. Smith, and Alan Sorensen*

January 2021

Abstract

We estimate a model of high school students' college choices, allowing for rich heterogeneity in students' preferences for college attributes. We use data on students' enrollment decisions and application decisions—i.e., the sets of colleges to which they applied—to identify the distribution of students' preferences. We use our estimates to quantify differences in a student's expected value upon college application that result from the uneven spatial distribution of colleges. As with other aspects of economic opportunity, we find that place matters: students with otherwise identical characteristics can have very different expected values depending on where they live. The importance of location reflects differences across states as well as differences across counties within a state. For students with low parental incomes and low SAT scores, over 70% of the variation is within-state across counties, while for students with high parental incomes and high SAT scores, 66% of the variation is across states.

*Fu: University of Wisconsin & NBER, *cfu@ssc.wisc.edu*. Guo: University of Wisconsin, *jguo27@wisc.edu*. Smith: University of Wisconsin, *ajsmith26@wisc.edu*. Sorensen: University of Wisconsin & NBER, *sorensen@ssc.wisc.edu*

1 Introduction

Some of young adults' most consequential decisions are about whether and where to attend college. The weight of these choices reflects both the importance of college as an economic investment, since the choice of college can substantially influence both near-term costs and lifetime earnings,[1] and the fact that college is also an expensive consumption good, since a student's enjoyment of the multi-year college experience depends on the match of college attributes to her preferences. However, the uneven spatial distribution of colleges in the United States means that not all students are endowed with equal access. Given that students typically face much lower in-state tuition than out-of-state tuition, cross-state differences in the quality of public colleges directly translate to differences in students' ex ante expected net returns to a college education depending on which state they live in.[2] Moreover, to the extent that students prefer to attend college close to home, they face unequal access to colleges even within a state.

Table 1 shows the relevance of both types of spatial dispersion faced by college-bound students surveyed in the Educational Longitudinal Study 2002.[3] The first row shows the cross-student distribution of the quality of one's home state's flagship college, as proxied by the median SAT score of admitted freshmen.[4] At the lower end, 5% of the students are from states where the flagship colleges have a median SAT score of 1080 or lower; at the upper end, 5% are from states where the flagship colleges have a median SAT score of 1325 or more. The remaining rows summarize the cross-student distribution of the number of four-year colleges within a 250 kilometer radius of the student's home. Some students have over 200 colleges nearby, including over 50 high-SAT colleges; some have fewer than 10 colleges nearby, with none of them being high-SAT colleges.

However, whether the heterogeneity described in Table 1 should cause concerns about "education deserts" depends on students' preferences, as would policies aimed at addressing such concerns. For example, if students care little about distance, then it will not matter that students from Wyoming have to travel greater distances from home to attend college. Similarly, if students do not have strong attachment to their home states, then cross-state differences in the quality of public colleges could be mitigated via tuition subsidies that offset the out-of-state vs. in-state tuition difference for students who attend high-quality out-of-state public colleges.[5] However, if students do care about proximity, and to different degrees, then such subsidies will do little to level the playing

[1] See, for example, Brewer, Eide, and Ehrenberg (1999) and Black and Smith (2006).

[2] Another major source of inequality, one that has been the focus of a large literature, is credit constraints. See Monge-Naranjo and Lochner (2012) for a review.

[3] We classify as "college-bound" any student who applied to at least one four-year college.

[4] Our data report the 25^{th} and 75^{th} percentiles of SAT scores, but not the 50^{th}. We compute the average of the 25^{th} and 75^{th} percentiles and refer to it as the median for expositional simplicity.

[5] This statement is made in a partial equilibrium (individual optimality) sense. Large-scale policies such as cross-state reciprocal tuition agreements can serve a similar purpose, but likely stimulate general equilibrium responses.

Table 1: Heterogeneity in college access

	Percentiles				
	5	25	50	75	95
Median SAT of home state's flagship college[a]	1080	1160	1195	1260	1325
# colleges within 250km[b]	6	24	54	93	210
# private colleges within 250km	3	14	36	70	148
# public colleges within 250km	3	11	18	27	63
# top-quartile-SAT colleges within 250km	0	6	16	26	56

[a]The average of the 25th and 75th percentiles of SAT scores of the college reported in IPEDS.

[b]250km radius around the centroid of one's home zip code.

SOURCES: U.S. Department of Education, National Center for Education Statistics, Educational Longitudinal Study (ELS) 2002 and Integrated Postsecondary Education Data System (IPEDS) 2004.

field for students whose willingness to pay to stay in their home states exceeds the out-of-state vs. in-state tuition difference; instead, these subsidies will disproportionately benefit students who value college quality over proximity. Moreover, if the latter group tend to have more advantaged family backgrounds, these subsidies will be regressive in nature, leading to serious equity concerns. More generally, the efficiency and equity implications of education policies clearly depend on the *distribution* of student preferences for various college attributes.

This paper aims at recovering a richer characterization of students' preferences for college attributes by incorporating information about the sets of colleges to which they applied, which we will refer to as students' *application sets*.[6] The essential idea is that when we observe the set of colleges a student applied to, the strength of her preference for a given attribute is reflected in the similarity of that attribute across colleges in that set. For example, conditional on observables, a student who applies only to colleges near her home may have very different preferences than her counterpart who applies only to academically competitive colleges: the former appears to care mostly about geographic proximity while the latter mostly about academic quality. Intuitively, recovering the distribution of preferences is then based on observing the fractions of students who appear to care a lot about the given characteristic.

Our model of student preferences follows the approach that is common in Industrial Organization studies of differentiated product markets, casting student utility as a function of college characteristics. Heterogeneity in preferences is incorporated by allowing student-specific coefficients on those characteristics. Application sets are most informative about students' preferences—i.e., their vectors of coefficients for college characteristics—if we fully utilize comparisons of all colleges included *and* excluded from these sets. Given the large number of colleges to choose from (and hence combi-

[6]Some studies have attempted to quantify which factors are influential in students' college choice decisions (e.g., Manski and Wise 1983, Avery and Hoxby 2004, Long 2004, Dillon and Smith 2017); while another set of studies has focused specifically on the impact of tuition or financial aid on college choices (e.g., Curs and Singell 2000, Dynarski 2003, Avery and Hoxby 2004, Kane 2007 and Deming and Walters 2018).

natorially large number of possible application sets), empirically modeling the optimal application decision becomes a daunting task. However, there are useful properties that the optimal set must obey,[7] which we utilize in our empirical approach: we derive necessary conditions for optimality of students' observed application sets, and base our estimator on these conditions.

We estimate our model with data from the Educational Longitudinal Study (ELS) 2002, the National Postsecondary Student Aid Study (NPSAS), and the Integrated Postsecondary Education Data System (IPEDS). The ELS data provide information on application sets, admission and enrollment outcomes, and binary indicators of whether financial aid was received at each of the colleges to which a student was admitted. We supplement ELS with more detailed information from NPSAS about financial aid amounts. The IPEDS data provide information on college attributes.

We use our estimates of students' preferences to answer two questions about college choices. First, we quantify the implications of the uneven spatial distribution of colleges for student welfare. Following in the spirit of Chetty *et al* (2014) and several other recent papers that have emphasized the geography of opportunity,[8] we use our estimates to calculate ex ante welfare for the same student were she to live in different counties across the U.S. We find that geographic variation in student welfare is considerable; that the variation is more pronounced for high-SAT students; and that the geographic patterns are quite different for high-SAT students vs. low-SAT students. For example, we find that across U.S. counties the interquartile range of the ex ante expected utility for an average high-SAT, low-income college-bound student is over 2,500 tuition dollars, compared to about 1,700 tuition dollars for her low-SAT, low-income counterpart. There is important variation both across states and across counties within a state: for low-SAT low-income students, over 70% the variation is within-state across counties, while for high-SAT high-income students, 66% of the variation is across states. We discuss the broader implications of these findings in our concluding section.

Second, we predict the substitution patterns that would result if a student were to face out-of-state tuition rates in all states. Peltzman (1973) argues that subsidies in the form of lower tuition for in-state students can perversely lead to a reduction in education—the idea being that inexpensive public colleges may attract students who otherwise would have attended costlier but higher-quality colleges. As a preliminary, partial-equilibrium investigation of such a hypothesis, we use our estimated model to simulate the choices of students if they had to pay the out-of-state tuition at their home-state public colleges. We find that while high-income students with high SAT scores would enroll in colleges with higher SAT scores on average, across all students the average quality of the

[7] See Chade and Smith (2006) for a theoretical analysis.

[8] See, for example, Abbott and Gallipoli (2017); Corak (2019); Berger (2018); Berger and Engzell (2019); and the follow-up paper by Chetty and Hendren (2018).

chosen college goes down, largely because many students simply switch to lower quality in-state universities that charge lower out-of-state tuition than their higher-quality counterparts. Our findings suggest that based on substitution effects alone, increasing in-state tuition would have a very limited effect in pushing students toward higher quality institutions.

Our paper contributes to the broad literature on the economics of higher education, especially the branch that studies the college market through the lens of structural models. For instance, Arcidiacono (2005) and Howell (2010) estimate structural models of students' choices and use them to address questions about affirmative action policies. Epple, Romano and Sieg (2006), Fu (2014), Bodoh-Creed and Hickman (2018), Fillmore (2018), and Cook (2019) estimate equilibrium models of the college market in which both students and colleges make strategic decisions.

The geography of college opportunity has been analyzed in the sociology literature, where researchers such as Turley (2009) and Hillman (2016) have documented geographic disparities in college availability. These studies emphasize that most students choose colleges in close proximity to their homes, and the number of nearby colleges varies considerably depending on where a student lives. Moreover, this variation is correlated with race and socioeconomic status, with minorities and lower-income students having fewer nearby colleges on average. Hillman (2016) contemplates whether some locations should be described as education deserts. Our estimated model allows us to quantify such geographic disparities not just in terms of proximity but also incorporating other college characteristics that students value.

Our estimation method, which exploits necessary conditions for optimality of students' application sets, is similar to approaches other authors have used in the IO literature. For example, Ellickson, Houghton, and Timmins (2013) use profit inequality conditions to estimate the strength of network economies for retail chains like Walmart and Target. As in our application, it would be infeasible to characterize the exact optimal choice of where these chains should locate their stores; but estimation can be based on necessary conditions for the optimality of those choices. Our use of data on application sets is somewhat similar to the use of survey data by Avery, Glickman, Hoxby, and Metrick (2013) to construct a revealed preference ranking of U.S. universities. They surveyed high school seniors to determine the set of colleges to which each student was admitted, as well as the single college the student chose to enroll in. Knowing the admissions set enables them to characterize each student's chosen university as the winner of a small tournament, and their overall ranking of colleges is essentially an aggregation of the preference rankings implied by these tournaments.

2 Data

We analyze a sample of college applicants from the Educational Longitudinal Study (ELS) 2002 run by the National Center for Education Statistics (NCES). The ELS 2002 surveyed a nationally representative sample of students as 10th graders in 2002 and as 12th graders in 2004, and also conducted follow-up surveys of the same students in 2006 and 2012. For our purposes, the important survey questions are about the students' college application and enrollment decisions: for each student, we know which colleges they applied to, where they were admitted, whether they received financial aid at each of the colleges to which they were admitted, and where they chose to enroll. We limit our sample to the respondents who reported applying to college while still in high school, which yields a sample of 7,410 students,[9] whose characteristics are summarized in Table 2.

Table 2: Summary of student characteristics ($N = 7,410$)

			Percentiles		
	Mean	Std. Dev.	10	50	90
High school GPA	3.12	0.58	2.30	3.19	3.84
SAT score	1,037	201	780	1,030	1,300
Family income	79,650	60,090	22,500	62,500	150,000
Female	0.55				
Black	0.12				
Hispanic	0.09				
College-educated Parents	0.56				

SOURCE: U.S. Department of Education, National Center for Education Statistics, Educational Longitudinal Study (ELS) 2002.

Our data on college characteristics come from NCES's Integrated Postsecondary Education Data System (IPEDS) for the academic year 2004-2005, to match the year when the students in our sample would have been entering college. In estimating our college choice model, we include only colleges that offer four-year degrees, and we exclude the five U.S. service academies and colleges whose Carnegie classification is "Special Focus Institution".[10] The resulting sample includes 1,337 four-year colleges, whose characteristics are summarized in Table 3.

The cost of attending a college includes both tuition and fees.[11] For public colleges, the cost often depends on a student's state of residency due to differences between in-state and out-of-state tuition. Among the 492 public universities in the data, 479 charge higher tuition for out-of-state students than in-state students, with out-of-state tuition on average over $7,400 higher. At least 54 of these public colleges have reciprocity agreements that allow neighboring states' students to

[9] This sample size is rounded to the nearest 10, at the request of the NCES.

[10] These are mostly seminaries/theology schools, technical colleges, and specialized medical schools.

[11] The tuition numbers reported in Table 3 include fees, and throughout the paper when we say or report "tuition" we mean "tuition plus fees."

pay discounted tuition. However, many of the most prestigious flagship universities opt out of their states' reciprocity agreements. For example, UC Berkeley and the University of Michigan do not offer in-state tuition to students from neighboring states even though other colleges in California and Michigan do.[12]

Table 3: Summary of college characteristics ($N = 1,337$)

			Percentiles		
	Mean	Std. Dev.	10	50	90
Tuition: Public In State	5,088	2,023	2,955	4,658	7,891
Tuition: Public Out of State	12,504	3,779	8,354	12,384	17,097
Tuition: Private	18,830	5,977	11,610	18,230	27,703
SAT of admitted students	1,065	124	930	1,050	1,225
# of freshmen	938	1,126	157	504	2,270
# of full-time undergraduates	4,205	5,334	629	2,034	10,984
Fraction women	0.58	0.13	0.47	0.57	0.71
Fraction Black	0.12	0.19	0.01	0.06	0.24
Fraction Hispanic	0.06	0.09	0.01	0.03	0.13
NCAA Division 1 sports*	0.09	0.28	0.00	0.00	0.00

* This is an indicator equal to one if the college has an NCAA Division 1 football team.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS) 2004.

A final data source is the 2004 wave of the NCES National Postsecondary Student Aid Study (NPSAS), which we use to augment the information from the ELS about students' financial aid outcomes. While the ELS survey only indicates whether a student received any financial aid at each college to which she was admitted, the NPSAS data also include information on the amounts and sources of financial aid received. As we explain below, we use these data from NPSAS to estimate the distribution of aid amounts conditional on receiving aid.

Before outlining our model, we first describe several key facts and patterns in the data. Table 4 shows the distribution of application set sizes (i.e., how many colleges a student applies to). An important and perhaps surprising fact is that 30% of students apply to only one college. Applying to multiple colleges is more common for students who have higher family income and higher SAT scores.

Table 5 shows some examples of "overlaps"—namely, colleges that tend to appear together in a student's application set. In some cases the overlaps reflect similarity in quality—for example, students who applied to Harvard also tended to apply to Yale, Princeton, and UPenn. But more often the overlaps reflect geographic proximity. For example, students who applied to the University of Georgia also commonly applied to Georgia State, Auburn, and Georgia Tech. This suggests

[12]Our data on reciprocity agreements were obtained from a survey conducted in 2001 by the Cornell Higher Education Research Institute.

Table 4: Distribution of the number of college applications

	Number of colleges applied to				
	1	2	3	4	5+
All students	.30	.24	.17	.11	.18
Low income, low SAT	.39	.30	.16	.07	.08
Low income, middle SAT	.34	.26	.16	.10	.14
Low income, high SAT	.27	.15	.15	.14	.28
Middle income, low SAT	.41	.27	.18	.07	.07
Middle income, middle SAT	.32	.25	.19	.11	.14
Middle income, high SAT	.24	.22	.17	.13	.24
High income, low SAT	.27	.30	.19	.13	.10
High income, middle SAT	.24	.22	.15	.12	.27
High income, high SAT	.15	.15	.17	.14	.39

Cells indicate fractions. Low (high) income students are those whose parents' total family income is 35,000 or less (100,001 or more). Low (high) SAT students are those whose SAT score is 950 or less (1,130 or more).

SOURCE: U.S. Department of Education, National Center for Education Statistics, Educational Longitudinal Study (ELS) 2002.

most students prefer to attend colleges close to their homes, which means that differences in the availability of nearby colleges (as described above in Table 1) could translate into economically important differences in the ex ante value of students' college choice sets.

Table 5: Examples of application overlaps

College	Three most common overlaps
UC Berkeley	UCLA, UC San Diego, UC Davis
U Georgia	Georgia State, Auburn, Georgia Tech
UNC Chapel Hill	NC State-Raleigh, Duke, Elon U
U Wisconsin-Madison	U Minn.-Twin Cities, Marquette, U Wisconsin-Milwaukee
U Oregon	U Washington, Oregon State, UC Berkeley
New York U	Boston U, Columbia, Boston College
Harvard	Yale, Princeton, Penn
Stanford	UC Berkeley, UC San Diego, UCLA
Notre Dame	Miami U-Oxford, Marquette, Boston College

Overlaps are the additional colleges most commonly applied to by students who applied to the college listed in the left column. Overlaps are listed starting with the most common.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Educational Longitudinal Study (ELS) 2002.

3 Model

Our purpose is to estimate high school students' preferences for college attributes using a framework that leverages not only those students' enrollment decisions (which college they choose to attend),

but also their application decisions (which colleges they choose to apply to). As explained above, knowing the full set of colleges to which a student applied—what we call the student's *application set*—should improve estimation of preference heterogeneity, since similarities in the applied-to colleges reflect the strength of the student's preferences for certain characteristics.

In this section we outline the structure of our model, specifying the decisions students make and the uncertainties they face when making those decisions. Details of how various functions are parameterized for estimation are described in Section 4.

3.1 Primitives

There are J (four-year) colleges, each characterized by a vector W_j of attributes including location, academic quality, public/private dummy, and college athletics. Each student i is characterized by a vector of observable characteristics X_i (including location, demographics, family background, and test scores) and a vector of unobservable tastes (β_i) associated with the various college characteristics. Each student makes two decisions in our model: which colleges to apply to, and—conditional on the admissions and financial aid outcomes—which college to enroll in.

3.1.1 Admissions and Net Tuition

Students face uncertainty over the outcomes of admissions and financial aid. The probability that student i is admitted to college j is assumed to be a function of student and college observable characteristics, given by

$$p_{ij} = P\left(X_i, W_j\right). \tag{1}$$

A student may obtain financial aid to attend college, the amount of which is a stochastic function of student characteristics, college characteristics and gross tuition t_j. The net tuition t_{ij} for student i attending college j is given by

$$t_{ij} = f\left(X_i, W_j, t_j\right) + \eta_{ij}, \tag{2}$$

where η_{ij} is a random shock that is realized *after* the student makes her application decisions.[13] Students know the admissions probabilities and the distribution of financial aid amounts when they make their application decisions.

[13] Financial aid includes both government aid (e.g., Pell grant) and college-specific aid.

3.1.2 Student Preferences

Students care about the net tuition cost t_{ij} and college characteristics W_j, and both the sign and strength of student preferences for these characteristics may vary with their own characteristics X_i and taste vector β_i. Student i's utility from attending college j is given by

$$u_{ij} = U\left(X_i, W_j, t_{ij}; \beta_i\right)$$

There is an outside option available to all, the value of which is normalized to zero ex ante. After applications are submitted, the outside option is subject to a shock u_{i0} that captures unforeseen events that change the opportunity cost of attending college (e.g., getting a job offer).

3.2 Student Problem

Student i faces a two-stage decision problem. In the first stage she chooses a set of colleges to apply to, after which admissions, financial aid outcomes, and the shock to the outside option are realized. Then, in the second stage, she chooses to enroll in one of the colleges that admitted her, or the outside option. To characterize students' optimal choices, we begin with the second stage enrollment decision and work backward.

Given a set of admissions and financial aid outcomes, student i chooses her most preferred college within the set O_i of colleges that admit her, or the outside option, i.e.,

$$v\left(O_i, X_i, \beta_i, \eta_i, u_{i0}\right) \equiv \max\{\{u_{ij}\}_{j \in O_i}, u_{i0}\} \tag{3}$$

Denoting the ex-ante value of being admitted to O_i as $\overline{v}(O_i, X_i, \beta_i) \equiv E\left[v\left(O_i, X_i, \beta_i, \eta_i, u_{i0}\right)\right]$, we can write the value of an application portfolio $Y \subseteq \mathcal{J}$ for student i as

$$V(Y, X_i, \beta_i) \equiv \sum_{O \subseteq Y} \Pr(O|X_i)\overline{v}(O, X_i, \beta_i) - C(|Y|),$$

where $\Pr(O|X_i)$ is the probability that i is admitted to the set of colleges O. $|Y|$ is the number of colleges in Y and $C(|Y|)$ is the application cost. Denoting the set of J colleges as $\mathcal{J}$, the student's application problem is therefore

$$\max_{Y \subseteq \mathcal{J}} \{V(Y, X_i, \beta_i)\}. \tag{4}$$

3.2.1 Simplification

Note that uncertainty about admissions makes a student's application decision (4) a complicated portfolio problem rather than one of simply listing the colleges she most wishes to attend. For example, admissions uncertainty creates incentives for students to include "safety schools" in their application sets.[14] Moreover, the complexity of this portfolio problem increases combinatorially with the number of colleges (J). Other studies that examine students' college choices have typically restricted J to be a small number, either by allowing for only a small number of colleges in the choice set (e.g., Arcidiacono (2005) and Cook (2019)) or by grouping colleges into a small number of types (e.g., Epple, Romano and Sieg (2006) and Fu (2014)). Since the goal of this paper is to gain a more precise understanding of students' heterogeneous preferences over college attributes, we treat each college as a unit (instead of grouping them) and allow for a large number of colleges in the consideration set ($J = 80$ in our empirical application), which makes solving the full problem (4) a daunting task.

However, notice that (4) can be viewed as a two-layer problem, where a student chooses the best portfolio of a given size n in the inner layer and optimizes over n in the outer layer, i.e.,

$$\max_{n\in\{1,..J\}}\{\max_{Y\subseteq\mathcal{J} \text{ s.t. } |Y|=n}\{V(Y,X_i,\beta_i)\}\}. \tag{5}$$

To simplify our analysis we focus on the inner layer of (5) and solve a student's problem taking the observed application set size n as given. The cost of this simplification is that we cannot estimate the application cost function $C(|Y|)$. This also means that in the counterfactual simulations below we must hold each student's n fixed at the value we observe in the data.

Even taking n as given, with $J = 80$ (as in our empirical application) it is computationally infeasible for an estimator to find the exact optimal set of colleges to include in the application set. For example, if $n = 4$ there would be over 1.5 million possible sets to check. The following assumption greatly facilitates the search for a tractable estimator.

Assumption 1: Conditional on observables, student i's admissions outcomes are independent across colleges, i.e.,

$$\Pr(O|X_i) = \prod_{j\in O} p_{ij} \prod_{j'\in Y\setminus O} (1-p_{ij'}). \tag{6}$$

Assumption 1 is not entirely innocuous: it would be violated if multiple colleges receive similar information about student i beyond X_i and interpret it in similar ways. In order to make As-

[14] See Chade, Lewis, and Smith (2014) for discussion and analysis of the student's portfolio choice problem.

sumption 1 as realistic as possible, in our empirical analysis we include a rich set of observables in the admissions probability function $P(X_i, W_j)$, and we assume the independence of admissions outcomes conditional on those observables.

Under Assumption 1, we can form an estimator based on necessary conditions for optimality of the application set, as stated in the following Proposition.

Proposition 1. *Given Assumption 1, a necessary condition for the optimality of application set Y_i among sets of the same size is that for all $y^* \in Y_i$ and all $k \notin Y_i$,*

$$\begin{aligned} & p_{iy^*} \sum_{\{O_i'\} \subseteq Y_i \setminus y^*} \Pr\left(O_i'|X_i\right) \overline{v}\left(\left\{O_i', y^*\right\}, X_i, \beta_i\right) - p_{ik} \sum_{\{O_i'\} \subseteq Y_i \setminus y^*} \Pr\left(O_i'|X_i\right) \overline{v}\left(\left\{O_i', k\right\}, X_i, \beta_i\right) \\ \geq \; & (p_{iy^*} - p_{ik}) \sum_{\{O_i'\} \subseteq Y_i \setminus y^*} \Pr\left(O_i'|X_i\right) \overline{v}\left(O_i', X_i, \beta_i\right) \end{aligned}$$

The proof of this proposition is in Appendix A. In essence, the proposition says that for the observed application set to be optimal, it must be that all possible pairwise swaps—of one college outside the set for one of the colleges in the set—would weakly reduce the expected utility. Our estimator utilizes these necessary conditions for optimality and involves checking these pairwise swaps, which is tractable because for a student who applied to n colleges, we only need to check $n(J-n)$ conditions instead of comparing all $\binom{J}{n}$ possible application sets.

4 Estimation

Our primary objective is to structurally estimate the distribution of students' preferences for college characteristics, rather than colleges' preferences for students. As such, we estimate parameters governing admissions probabilities and financial aid distribution outside of the model. In this section, we will first briefly describe our estimation of these two components, and then we will describe our empirical specification for student preferences and how we estimate them within the model.

4.1 Admissions Probabilities and Financial Aid

Admissions Probabilities are estimated via probit regressions in which student i's probability of admission at college j is a function of the student's characteristics, the college's characteristics, and their interactions. In the interest of flexibility, we estimate the model separately for six categories

of colleges defined by (public vs. private) $\times$ (tercile of SAT_j^c), where SAT_j^c (the median SAT score of students in college j) is a proxy for college quality that we obtain from IPEDS.[15] In each case, the included covariates are student high school GPA; student SAT score; median SAT of the college SAT_j^c; an indicator for whether student i's SAT score is below the 25th percentile of SAT scores in college j; an indicator for whether college j is in the student's home state; an indicator for whether the student has taken any Advanced Placement course; indicators for female, black, and Hispanic; an indicator for whether the student is from a single-parent family; an indicator for whether at least one of the student's parents graduated from college; and indicators for 7 family income categories.

Importantly, the probit regressions deliver predicted admissions probabilities that exhibit reasonable patterns (e.g. they are increasing in student's GPAs and SAT scores) and cover a sensible range (e.g. low-SAT students' predicted probabilities of being admitted to Harvard are around 3 percent, and high-SAT students' predicted probabilities of being admitted to non-competitive public universities are above 90 percent). Additional details and fit statistics are available in an online appendix.

Financial Aid includes both government aid (the Pell grant) and college-specific aid. We compute the Pell grant following the government-specified formula, where the amount of grant depends mainly on one's expected family contribution (EFC) and the cost of attendance. For college-specific aid, we model the probability of receiving aid in a way that mirrors the admissions probabilities, with probit regressions run separately for the six different college types. In addition to the covariates listed above for the admissions model, we also allow the probability to depend on the college's tuition and the student's EFC. This yields a predicted probability that student i will receive aid at college j for any i-j pair.

To estimate the *amount* of college-specific aid received, conditional on receiving any, we use the NPSAS data (described in Section 2). We model the log of aid received as a truncated normal with the upper truncation point set at 1.2 times the maximum observed amount of aid,[16] and the mean being a linear function of covariates including the college's gross tuition, the student's EFC, sex and race dummies, student SAT score, college median SAT score, an indicator for whether the student is in the same state as the college, and a few interactions among these variables. Full details are in the online appendix referenced above. The NPSAS data introduce a possible selection bias because they only report aid amounts at students' chosen colleges—i.e., the colleges where they chose to enroll. If students tend to enroll in colleges that offer more aid, then the aid amounts of enrolled

[15] Each college in IPEDS reports the 25th and the 75th percentiles of SAT scores of its enrollees; we take the average of these two percentiles as SAT_j^c.

[16] We found that if we simply model aid amounts as being log-normally distributed without any upper bound, our estimator for student preferences would sometimes draw simulated aid amounts that were unrealistically high—i.e., out in the long tail of the log-normal distribution.

students will tend to be higher than the aid amounts offered to admitted students, so our model may slightly overpredict aid amounts.[17] Fortunately, selection is not a problem in our model at the aid vs. no-aid margin, since the ELS data report whether any aid was received at *all* colleges to which the student applied.

4.2 Student Preferences

Empirical Specification Student i's utility at college j is given by

$$\begin{aligned} u_{ij} = & -\left(\gamma_1 LowInc_i + MidInc_i + \gamma_2 HighInc_i\right) t_{ij} \\ & + \alpha_0 + \alpha_1 (SAT_i - SAT_j^c)_+^2 + \alpha_2 (SAT_i - SAT_j^c)_-^2 + \alpha_3 Black_i + \alpha_4 Hispanic_i \\ & + \exp\left(\beta_{1,i}\right)\left[SAT_j^c + \delta_1 (SAT_j^c - 1200)_+\right] + \beta_{2,i}\left[\ln\left(Dist_{ij}\right) + \delta_2 OutState_{ij}\right] \\ & + \beta_{3,i} Private_j + \beta_{4,i} NCAAI_j. \end{aligned} \tag{7}$$

The first component of this function reflects the student's sensitivity to net tuition (t_{ij}), which may differ across students from different family income groups. We categorize a student i's family income as low ($LowInc_i = 1$) if it is less than \$35,000, as high ($HighInc_i = 1$) if it is above \$100,000, and as middle ($MidInc_i = 1$) otherwise. The parameters γ_1 and γ_2 measure how the price sensitivity of low- and high-income students (respectively) differ from that of students in the middle income group. We normalize the tuition coefficient for middle-income students to 1, so student preferences for various college attributes are measured in tuition dollars.

The parameter α_0 represents the overall attractiveness of attending a 4-year college relative to the outside option for an average student; α_3 and α_4 are introduced to capture potential differences in preferences among black and Hispanic students. To allow for the possibility that a student may prefer colleges that closely match her own academic ability, we introduce parameters α_1 and α_2 to measure students' preference for the difference between her own SAT (SAT_i) and the median SAT score at the college (SAT_j^c), allowing for asymmetry in the preference for over-match vs. under-match.

For our purposes, the most important components of the utility function (7) are the college characteristics over which students have heterogeneous preferences, as reflected by the student-specific $\beta_{k,i}$ coefficients. First, students are allowed to have heterogeneous preferences for a college's academic quality, as measured by SAT_j^c. Since these quality differences are most likely to be meaningful for

[17] To check whether this selection effect is likely to be important, we examined data from the National Longitudinal Survey of Youth (NLSY 97), which reports aid even for unaccepted offers. We estimated models for aid amounts using both the full sample of all offers and the selected sample of accepted offers, and found that the latter predicted aid amounts only slightly higher than the former.

colleges toward the upper end of the distribution, we allow the slope to differ depending on whether SAT_j^c is above or below 1200. Student-specific preferences for proximity are represented by $\beta_{2,i}$, where we use a distance index that combines actual distance ($Dist_{ij}$) and an indicator for whether j is out of student i's home state. We measure $Dist_{ij}$ as the distance in kilometers between college j and the centroid of student i's home zip code. Finally, students have heterogeneous preferences over whether or not the college is private ($Private_j \in \{0,1\}$), and for whether or not the college has an NCAA Division I football team ($NCAAI_j \in \{0,1\}$). The latter serves as a proxy for whether major sporting events are an important aspect of college life in j.

Student-specific preference parameters $\beta_{k,i}$ are drawn from the following normal distribution,

$$\beta_{k,i} = \mu_k(X_i) + \epsilon_{k,i}\sigma_k, \text{ with } \epsilon_{k,i} \sim N(0,1).$$

The mean tastes for college SAT scores ($\beta_{1,i}$), distance ($\beta_{2,i}$), and private colleges ($\beta_{3,i}$) are allowed to vary with family income, while the mean tastes for Division I sports ($\beta_{4,i}$) are common across students, such that

$$\mu_k(X_i) = \begin{cases} \mu_{k,0} + \mu_{k,1} LowInc_i + \mu_{k,2} HighInc_i & \text{for } k = 1,2,3 \\ \mu_{k,0} & \text{for } k = 4 \end{cases}$$

Students are subject to post-application shocks to their outside option, drawn from a normal distribution:

$$u_{i0} = \epsilon_{0,i}\sigma_0(X_i), \text{ with } \epsilon_{0,i} \sim N(0,1).$$

The dispersion of shocks is allowed to be different for low-income and/or low-SAT students, such that

$$\sigma_0(X_i) = \exp\left[\lambda_0 + \lambda_1 LowInc_i + \lambda_2 I(SAT_i \leq 950)\right]. \tag{8}$$

We allow this layer of flexibility to better fit the data: conditional on admissions and financial aid outcomes, low-income and/or low-SAT students have a much lower enrollment rate than other students, which holds even if we compare students with similar application behaviors. Such patterns can arise, for example, if low-income households are subject to higher income volatility (unemployment), which would be captured by a larger dispersion of post-application shocks faced by these students.

Estimation Procedure At a high level, the goal of our estimation approach is to choose parameters that maximize the likelihood of students' observed application sets and enrollment decisions. Two complicating factors are that (1) our model does not admit a closed-form solution to the portfolio problem of choosing an application set, and (2) the number of colleges in the U.S. is

quite large. As explained above, our solution to the first problem is to base our estimator on the necessary conditions for optimality of the application set, as described in Proposition 1.

To address the second issue, instead of including the full set of J colleges in each student's choice set, we draw a subset $\mathcal{J}_i$ of 80 colleges for each student i: $\mathcal{J}_i$ always includes colleges in student i's observed application set Y_i, and the remaining colleges are drawn from $\mathcal{J} \backslash Y_i$ in a way that accounts for both variety and relevance in terms of geography, school type (public vs. private) and school quality. The sampling scheme, which we describe in more detail in Appendix B, draws colleges proportionally from bins defined by public vs. private ownership, in-state vs. out-of-state, and academic quality (as measured by SAT_j^c). The scheme guarantees inclusion of at least one academically competitive public university from the student's home state, since the flagship university of a student's home state is almost certainly in her consideration set. Importantly, the sampling rules are common across students and independent of Y_i.

Once we have constructed choice sets $\mathcal{J}_i$ for each student, we hold those sets fixed during the estimation. We construct the quasi-likelihood function using a simulation procedure that (1) simulates M copies of each student i, each with different preference "shocks" $\epsilon_{k,i}$ that lead to different preference coefficients $\beta_{k,i}$; (2) uses these simulated students to compute a kernel-smoothed probability that the chosen application set is better than all possible one-for-one swaps (Proposition 1); (3) computes a smoothed probability that the enrollment decision is optimal given the admissions and financial aid outcomes; and (4) combines the probabilities from (2) and (3) to construct the quasi-likelihood for student i's observed choices (application set and enrollment decision). The details of this procedure are explained in Appendix C.

5 Results

5.1 Parameter Estimates

Table 6 reports the parameter estimates and associated standard errors.[18] The estimated tuition sensitivity declines with family income, which is in line with findings from the previous literature. The coefficients on the distance between one's own and a college's quality as proxied by SAT_j^c, α_1 and α_2, are both negative—which is consistent with the idea that students prefer to fit in academically rather than overmatch or undermatch.

[18] We estimate the information matrix as the sum of the outer products of the scores: $\hat{I} = \sum_i g_i g_i'$, where g_i is the score function for student i. We estimate the Hessian matrix as $\hat{H} = \sum_i h_i$, with h_i being the Hessian for student i. The standard errors are then computed as the square roots of the diagonal elements of $\hat{H}^{-1}\hat{I}\hat{H}^{-1}$.

Table 6: Utility parameter estimates

Variable	Parameter	Estimate	Std. Error
Tuition × Low income	γ_1	1.364	0.106
Tuition × High income	γ_2	0.269	0.029
Constant	α_0	7.677	1.349
$(SAT_i - SAT_j^c)_+^2$	α_1	-0.574	0.053
$(SAT_i - SAT_j^c)_-^2$	α_2	-0.267	0.045
Black	α_3	0.493	0.919
Hispanic	α_4	-2.861	0.963
$(SAT_j^c - 1200)_+$	δ_1	2.692	0.250
SAT_j^c	$\mu_{1,0}$	0.991	0.054
SAT_j^c × Low income	$\mu_{1,1}$	0.155	0.052
SAT_j^c × High income	$\mu_{1,2}$	-0.348	0.048
Out of state	δ_2	1.461	0.151
Distance	$\mu_{2,0}$	-2.753	0.107
Distance × Low income	$\mu_{2,1}$	-1.270	0.222
Distance × High income	$\mu_{2,2}$	1.125	0.089
Private	$\mu_{3,0}$	-2.236	0.300
Private × Low income	$\mu_{3,1}$	-0.693	0.574
Private × High income	$\mu_{3,2}$	0.370	0.393
NCAA Division 1	$\mu_{4,0}$	0.766	0.207
Std dev. of SAT preferences	σ_1	0.278	0.039
Std dev. of distance preferences	σ_2	5.253	0.677
Std dev. of Private preferences	σ_3	1.177	0.099
Std dev. of NCAA Div. 1 preferences	σ_4	4.722	0.490
Std dev. of outside option	λ_0	2.628	0.042
Std dev. of outside option (low income)	λ_1	0.458	0.069
Std dev. of outside option (low SAT)	λ_2	0.284	0.041

The main parameters of interest are the ones related to heterogeneity in preferences—i.e., the distributions of student-specific coefficients for various college attributes (academic quality, distance, in-state vs. out-of-state, public vs. private and college athletics). To better understand these estimates, we report, in Table 7, the change in a student's utility, measured in thousands of tuition dollars, associated with a given change of an attribute. The middle column reports the impact for a student with the mean $\beta_{k,i}$. The first and third columns report the same effects for students with $\beta_{k,i}$'s one standard deviation below or above that mean, respectively. Since we estimate different tuition coefficients for different family income levels, we report the effects separately for each income group. Students from high-income households are estimated to have a lower coefficient on tuition ($\hat{\gamma}_2 = 0.269$), so the heterogeneity in their preferences for non-tuition college characteristics is amplified when expressed in terms of tuition dollars.

Two points stand out from the table. First, there is considerable heterogeneity in how much students value academic quality, as proxied by the college's median SAT score. A middle-income

Table 7: Preference heterogeneity

	Preference at:		
	$\hat{\mu} - \hat{\sigma}$	$\hat{\mu}$	$\hat{\mu} + \hat{\sigma}$
Increase SAT 1000 to 1100			
Low income	1.75	2.31	3.04
Middle income	2.04	2.69	3.56
High income	5.36	7.07	9.34
Increase SAT 1300 to 1400			
Low income	6.45	8.51	11.24
Middle income	7.53	9.95	13.13
High income	19.78	26.11	34.47
Increase distance from 10km to 500km			
Low income	-20.74	-11.54	-2.34
Middle income	-23.32	-10.77	1.78
High income	-70.34	-23.68	22.99
Out of state vs. in state			
Low income	-7.75	-4.31	-0.87
Middle income	-8.71	-4.02	0.67
High income	-26.27	-8.84	8.59
Private vs. public			
Low income	-3.01	-2.15	-1.28
Middle income	-3.41	-2.24	-1.06
High income	-11.31	-6.94	-2.56
NCAA Division I sports			
Low income	-2.90	0.56	4.02
Middle income	-3.95	0.77	5.49
High income	-14.70	2.85	20.40

student at the high end of the preference distribution (roughly the 85th percentile) would be willing to pay $13,130 more in tuition to attend a college with median SAT scores of 1400 vs. 1300 (rough examples would be UCLA vs. Loyola Marymount, or NYU vs. Rutgers), whereas a student at the low end of the distribution would be willing to pay only $7,530. Second, a majority of students have strong preferences for attending colleges close to home. For a middle-income student with average preferences—i.e., with the mean value of $\beta_{2,i}$—an increase in distance from 10 to 500 kilometers is equivalent to a nearly $11,000 increase in tuition. However, some students appear to prefer being further from home. Our estimates imply that 10 percent of low-income students and 31 percent of high-income students have positive coefficients on distance. Similarly, most students exhibit strong home-state biases (for reasons beyond tuition and distance), while a small fraction of students

prefer to go out of their home states.

5.2 Model Fit

As discussed in Section 3.2.1, all of our simulations take as given the observed number of colleges $|Y_i^o|$ a student applied to. To evaluate how well our model fits the data, we simulate each student i's optimal application set given size $|Y_i^o|$ by solving the inner layer of problem (5) and then deriving her optimal enrollment decision given the admissions and financial aid outcomes for the applied-to colleges. Panel A of Table 8 shows the average characteristics of the colleges students actually applied to and enrolled in as well as the average characteristics of the colleges our model predicts they would apply to and enroll in. Some of these college characteristics are common across students—such as college median SAT, private, and NCAA—while others are college-student specific, such as admissions and aid probabilities, aid amount, tuition (because tuition for public colleges depends on in-state status and reciprocity agreements with other states), differences between the student's SAT and the college's median SAT, whether the college is out of state, and home-college distance. For enrollment, each row is a simple average across college enrollees. For application, since some students applied to more than one college, we first take the average of (college-student-specific) characteristics across the colleges a student applied to, and then average across students.[19] Overall, the model fits the data well. However, it underpredicts the tuition and home-to-college distance for both applied colleges and enrolled colleges.

Panel B reports model fits for the fraction of students admitted to any college, and the fraction of college enrollees among those with at least one offer. Table 12 and Table 13 in the appendix show model fits by family income and by student SAT, respectively.

6 Counterfactual Simulations

Using our estimated model, we explore two questions about higher education. First, we examine the implications of the uneven spatial distribution of colleges in the U.S. for students' choices and welfare. Then, we examine the substitution patterns that would result if public universities' in-state subsidies were eliminated.

[19] For example, to obtain the entry in Row 1 of the Application column, we first calculate the college-student-specific admissions probability p_{ij} for student i at each of the colleges she applied to, and take the average across $j \in Y_i$, yielding an average $\overline{p}_i \equiv \frac{1}{|Y_i|}\sum_{j\in Y_i} p_{ij}$ for the student; then we take the average across students, i.e., $\frac{1}{I}\sum_{i=1}^{I}\overline{p}_i$.

Table 8: Model fit

Panel A: College characteristics

	Data		Model	
	Application	Enrollment	Application	Enrollment
Admission probability	0.74	0.80	0.76	0.77
Tuition ($1,000)	11.34	11.51	10.52	9.97
Aid probability	0.49	0.50	0.52	0.51
Aid amount ($1,000)	7.76	7.69	8.15	7.82
$(SAT_i - SAT_j)_+^2$	0.70	0.81	0.62	0.63
$(SAT_i - SAT_j)_-^2$	2.74	1.36	2.45	1.95
Median SAT (100)	11.03	11.08	11.04	11.14
Private	0.32	0.33	0.38	0.34
Distance (100 km)	3.47	3.38	2.74	2.49
Out of state	0.27	0.26	0.17	0.14
NCAA Division I sports	0.34	0.35	0.32	0.36

Panel B: Admission and enrollment rates

	Data	Model
Admission rate	0.90	0.92
Enrollment rate	0.84	0.88

The admission rate is the fraction of students who were admitted to at least one of the colleges they applied to, and the enrollment rate is conditional on being admitted to at least one college.

6.1 Geographic Differences in Student Welfare

Given our estimated student preferences, the uneven spatial distribution of colleges in the U.S. may lead to different outcomes and welfare levels for otherwise identical students depending on where they live. To quantify these differences, we use our estimates to simulate the outcome and welfare for the same student were she to live in different counties across the U.S. Since locations may matter more depending on students' backgrounds, we conduct the cross-county comparison separately for 9 hypothetical students, each representing a group defined by SAT (low, middle, high) and family income (low, middle, high). The representative student in each group is assigned the average characteristics of the students in that group.[20] For each of the 9 representative students, we place her into each U.S. county and simulate her application and enrollment outcomes in each county. We use the same draws of random preference coefficients and shocks (to financial aid and the outside option) in all counties, so that all differences across simulations for the same representative student are attributable to the county of residence.

Figure 1 summarizes the geographic variation in students' ex ante welfare upon college application with heat maps for each combination of family income (low, middle, high shown from the top

[20]To construct the group averages, we use means for continuous variables and medians for categorical variables (like family income).

to the bottom) and student SAT tercile (low, middle, high shown from the left to the right).[21] The differences shown in the figure reflect a variety of factors. Obviously the main driver is that most students have relatively strong preferences to attend a nearby college. This preference for proximity, combined with the substantial heterogeneity in students' preferences for other college characteristics, makes it quite valuable to live in a place with a wide variety of nearby colleges. This is especially true for high-SAT students, since their higher chances of admission mean that more of the nearby colleges will be realistic options for them. Indeed, the largest geographic differences shown in Figure 1 are for students with high SAT scores. The values of high-SAT students' choice sets differ sharply across regions, with higher values in the eastern half of the country. A high-performing student is meaningfully better off if she lives in Virginia instead of Nevada, for instance. By contrast, geographic heterogeneity for students with low SAT scores is far less pronounced. With the exception of some remote areas in states like Wyoming and Montana, low-SAT students' ex ante expected values are roughly the same regardless of where they live. This suggests that the supply of non-selective colleges in the U.S. has a spatial distribution that mostly matches demand.

Besides the welfare differences shown in Figure 1, the uneven spatial distribution of colleges can also lead to substantial differences in the same student's likelihood of enrollment, and the characteristics of the enrolled colleges, depending on her county of residence. The magnitudes of these differences are summarized in Table 9, which shows the interquartile range of welfare and interquartile ranges of four predicted outcomes across counties. For instance, the enrollment probability of a low-income student with a middle SAT score varies by 5.47 percentage points between the 25th and 75th percentile counties, and the median SAT score of the enrolled college varies by 53 points.

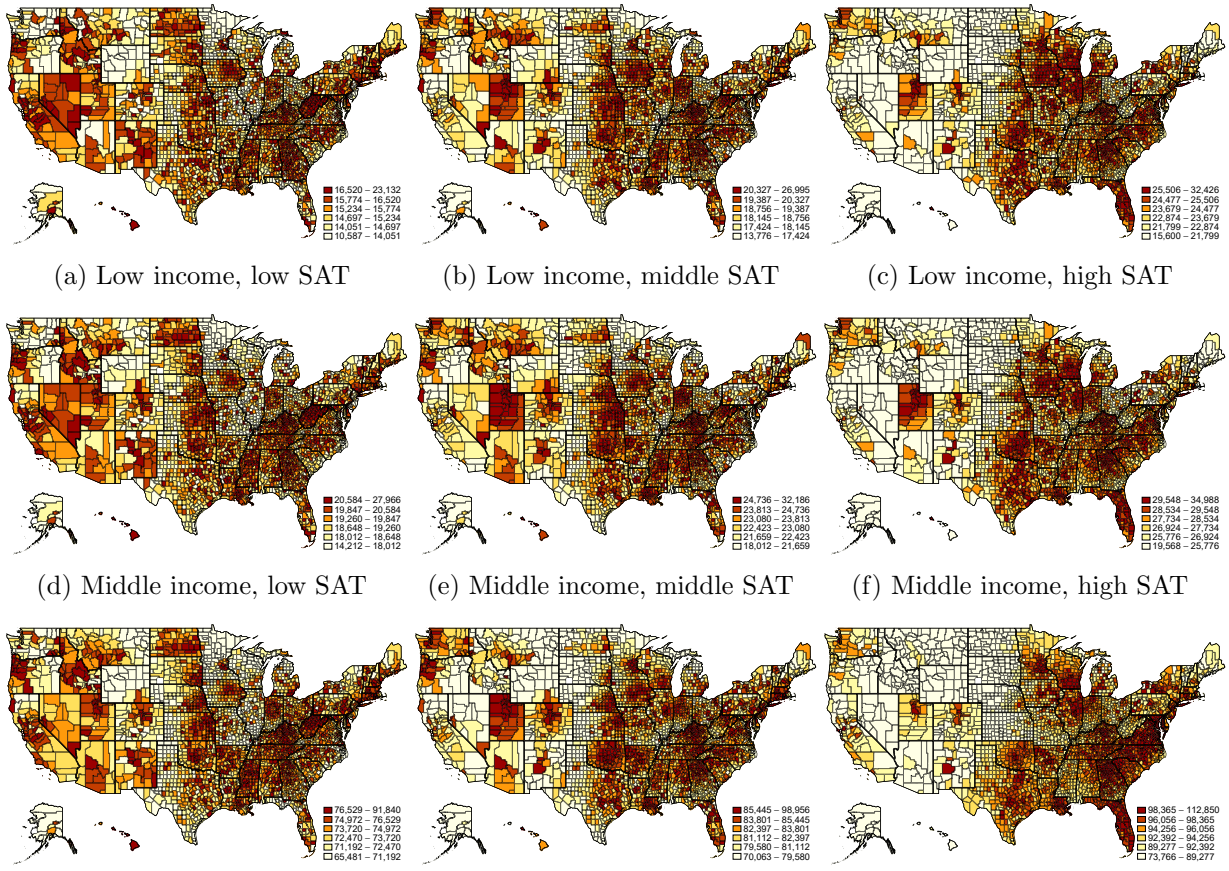
Table 9: Interquartile Ranges of Simulated Outcomes

	Expected	Enrollment	Characteristics of enrolled colleges		
Student group	utility ($)*	Prob (%)	SAT	Distance (km)	Net Tuition ($)
Low income, Low SAT	1,710	5.71	49	188	1,533
Low income, Middle SAT	2,027	5.47	53	145	1,514
Low income, High SAT	2,537	4.29	63	176	2,390
Middle income, Low SAT	1,855	4.90	41	205	1,356
Middle income, Middle SAT	2,161	4.15	45	169	1,450
Middle income, High SAT	2,577	3.23	53	205	2,366
High income, Low SAT	3,790	2.78	31	254	1,841
High income, Middle SAT	4,167	2.30	33	219	2,021
High income, High SAT	5,875	2.18	25	370	3,660

* Expected utility values are divided by the relevant tuition coefficient in order to express utility in terms of tuition dollars.

[21] Student welfare is measured by the ex ante value $\max_{Y \subseteq \mathcal{J} \text{ s.t. } |Y|=n} \{V(Y, X_i, \beta_i)\}$, where n is the number of colleges applied to by a representative student in each of the 9 groups. Values are divided by the tuition coefficient corresponding to the student's income group, so they are expressed in tuition dollars.

Figure 1: Expected utility of college education across U.S. counties



(a) Low income, low SAT (b) Low income, middle SAT (c) Low income, high SAT

(d) Middle income, low SAT (e) Middle income, middle SAT (f) Middle income, high SAT

(g) High income, low SAT (h) High income, middle SAT (i) High income, high SAT

Expected utility is measured in tuition dollars. From top to bottom, the family income for the three rows are $\leq 35,000$, $(35,000, 100,000]$, and $> 100,000$, respectively. From left to right, the student SAT for the three columns are ≤ 950, $(950, 1130)$, and $\geq 1,130$, respectively. Each map has 6 colors, each representing 1/6 of the counties.

There are also meaningful correlations between the same student's welfare and her enrollment outcome across counties, as shown in Table 10. Each row in the table refers to one of the 9 representative students; the four columns show how each representative student's expected utility is correlated with her college enrollment probability and the characteristics of enrolled college (quality, distance and net tuition) when she resides in different counties in the U.S. Not surprisingly, expected utility is highly correlated with enrollment probability. Among the three characteristics of the enrolled college, the expected utility is most strongly correlated with distance, but there is some heterogeneity across student groups. For example, regardless of income, a high-SAT student's expected utility is strongly correlated with the quality of her enrolled college, while this correlation is either negative or near zero for low- and middle-SAT students.

Table 10: Correlations with the expected utility

Student group	Enrollment Probability	Characteristics of enrolled college			R^2
		SAT	Distance	Net Tuition	
Low income, Low SAT	0.93	-0.23	-0.68	-0.38	0.27
Low income, Middle SAT	0.89	-0.00	-0.72	-0.15	0.19
Low income, High SAT	0.86	0.44	-0.68	0.08	0.41
Middle income, Low SAT	0.93	-0.15	-0.65	-0.36	0.29
Middle income, Middle SAT	0.87	0.04	-0.69	-0.24	0.25
Middle income, High SAT	0.83	0.48	-0.66	0.14	0.45
High income, Low SAT	0.95	-0.14	-0.52	-0.25	0.29
High income, Middle SAT	0.91	0.09	-0.65	-0.06	0.26
High income, High SAT	0.90	0.52	-0.72	-0.30	0.66

Cells in columns 2-5 report correlations of the indicated outcome with ex ante expected utility across counties. The last column reports the R^2 from a regression of county expected utilities on state fixed effects.

A natural question to ask is whether the geographic differences mostly reflect state-level variation, or whether variation across counties within a state is also important. To answer this question, we regress a student's county-specific welfare on state fixed effects; the R^2 from this regression is shown in the last column of Table 10. State fixed effects generally explain between one fifth to two thirds of the cross-county variation, implying that both between- and within-state variation in college access are important, but to different extents depending on family income and SAT. In particular, as a student's family income and SAT increase, especially the latter, the student's state of residence becomes more and more relevant for her utility. For example, for a student with high family income and high SAT, 66% of the geographic dispersion of expected utility reflects cross-state variation.

6.2 Student Responses to In-state Tuition Subsidies

Most public universities are heavily subsidized, charging much lower tuition for in-state residents. Peltzman (1973) argues that such subsidies might actually cause students to choose in-state universities instead of unsubsidized but higher quality institutions for which they would qualify, thus reducing their educational attainment. At a deeper level, this is an argument against the tuition subsidy policy that applies to all in-state students; and an evaluation of counterfactual policies would require an equilibrium model that takes into account the supply side responses. However, before conducting such a full-blown investigation, a pre-requisite is to understand how an individual student would respond.

To this end, for each student, we simulate her choice when facing the actual/baseline tuition schedules, and separately simulate her choice if she were to face counterfactual tuition schedules

in which she has to pay out-of-state tuition at her home-state institutions. To ensure that the comparison of these two simulations isolates the effect of tuition changes, for each student the two simulations use the same consideration set, the same draws of random preference coefficients, and the same draws of the random shocks to financial aid and the outside option.[22] Table 11 reports the differences in average outcomes between the two simulations (counterfactual minus baseline).

Table 11: Simulated changes when in-state subsidies are removed

	Characteristics of enrolled college			
Student group	SAT	% out of state	Distance (km)	% Private
Low income, Low SAT	-5.79	5.68	56.39	24.83
Low income, Middle SAT	-7.47	3.85	52.09	32.57
Low income, High SAT	-6.77	4.36	71.71	33.52
Middle income, Low SAT	-7.33	9.23	104.62	25.20
Middle income, Middle SAT	-8.85	7.73	94.23	29.99
Middle income, High SAT	-3.21	8.04	105.25	29.90
High income, Low SAT	-2.17	4.40	61.06	8.16
High income, Middle SAT	1.56	5.39	78.17	9.64
High income, High SAT	6.27	5.00	79.33	9.49
All students	-3.11	6.97	88.75	23.91

Naturally, we find that eliminating subsidies leads students to substitute away from their home states' universities. Overall, students are 6.97 percentage points more likely to attend a college outside of their home states, and the average distance to the enrolled college increases by 88.75 kilometers. Some of the substitution is consistent with Peltzman's hypothesis, as higher income students with high SAT scores on average enroll in colleges with higher SAT scores. Over all students, however, the average quality of the enrolled college goes down. Some students simply switch to lower quality in-state universities that charge lower out-of-state tuition than their higher-quality counterparts. Others switch to lower quality private colleges where they are more likely to get in and receive aid.[23,24]

These simulations are at best a crude evaluation of Peltzman's hypothesis, but they suggest that substitution effects resulting from the removal of in-state tuition subsidies would do little to push students toward higher quality institutions. Perhaps a stronger argument for increasing in-state

[22] Tuition is an input into our model of college-specific financial aid, therefore, financial aid amounts are adjusted accordingly.

[23] If students in the baseline are not allowed to re-optimize in response to the elimination of the in-state tuition subsidy, the average net tuition of the enrolled college across all students would increase from $2,105 to $7,413, compared to $5,191 in the counterfactual where re-optimization is allowed (not shown in the table). The smaller increase in the counterfactual reflects the switch to colleges with lower out-of-state tuition and more generous financial aid.

[24] The predicted enrollment rate (not shown in the table) also drops by 3 percentage points, but our simulation may underpredict the drop in enrollment because we do not model the extensive margin of college application: if application costs are high enough, an increase in tuition levels may discourage some students from applying to colleges at all.

tuition would be that the increased tuition revenue could be used to improve the quality of public universities.

7 Conclusion

A central purpose of this study was to develop and estimate a model that allows for rich heterogeneity in students' preferences for college characteristics. From a modeling standpoint, allowing for heterogeneity in preferences is nothing new: estimating choice models with random coefficients has long been a standard approach to estimating demand systems in product markets. From a data standpoint, our key innovation is to use data on students' application sets as a way of credibly identifying preference heterogeneity. The modeling challenge is to incorporate these data in estimation without having to fully solve the computationally intractable portfolio problem of students choosing which colleges to apply to. We achieve this by exploiting necessary conditions for optimality that respect the subtleties introduced by admissions uncertainties (e.g. the "safety schools" problem).

Our estimates confirm considerable heterogeneity in students' preferences for college attributes. Most students prefer to attend colleges close to home, and for many students this preference is quite strong. Preferences for other college characteristics are more variable: for instance, some students appear to care a lot about academic quality, others very little. Given the uneven spatial distribution of colleges in the United States, the combination of strong preferences for proximity and variable preferences for other characteristics implies substantial differences in the expected values of students' choice sets depending on where they live. These differences are especially large for high-performing students.

The fact that most students have strong preferences for proximity also means that even large changes in tuition may not meaningfully change their choices. Peltzman (1973) hypothesized that tuition subsidies for in-state students might inefficiently distort their choices away from higher-quality colleges outside their home states, but our simulations indicate that if students were forced to pay out-of-state tuition at their home state public colleges, most would simply switch to cheaper colleges that are still close to home. Only high-performing students with higher incomes appear to substitute toward higher-quality colleges that are further away.

Many policies and programs already aim to equalize opportunity in higher education, such as private scholarship funds and government financial aid programs that specifically help low-income students. Our results suggest these policies could also consider equalizing geographic differences in opportunity, for instance by subsidizing students in locations where colleges are sparse, or by making investments to raise the quality of academic institutions in targeted locations.

References

Abbott, Brant, and Giovanni Gallipoli. "Human capital spill-overs and the geography of intergenerational mobility." *Review of Economic Dynamics* 25 (2017): 208-233.

Arcidiacono, Peter. "Affirmative action in higher education: How do admission and financial aid rules affect future earnings?" *Econometrica* 73, no. 5 (2005): 1477-1524.

Avery, Christopher, and Caroline Minter Hoxby. "Do and should financial aid packages affect students' college choices?" In *College choices: The economics of where to go, when to go, and how to pay for it*, pp. 239-302. University of Chicago Press, 2004.

Avery, Christopher, Mark Glickman, Caroline Hoxby, and Andrew Metrick. "A revealed preference ranking of US colleges and universities." NBER Working Paper 10803, 2004.

Berger, Thor. "Places of persistence: Slavery and the geography of intergenerational mobility in the United States." *Demography* 55, no. 4 (2018): 1547-1565.

Berger, Thor, and Per Engzell. "American geography of opportunity reveals European origins." *Proceedings of the National Academy of Sciences* 116, no. 13 (2019): 6045-6050.

Brewer, Dominic J., Eric R. Eide, and Ronald G. Ehrenberg. "Does it Pay to Attend an Elite Private College?" *The Journal of Human Resources* 34, no. 1 (1999): 104-123.

Black, Dan A., and Jeffrey A. Smith. "Estimating the returns to college quality with multiple proxies for quality." *Journal of Labor Economics* 24, no. 3 (2006): 701-728.

Bodoh-Creed, Aaron L., and Brent R. Hickman. "College assignment as a large contest." *Journal of Economic Theory* 175 (2018): 88-126.

Chade, Hector, and Lones Smith. "Simultaneous search." *Econometrica* 74, no. 5 (2006): 1293-1307.

Chade, Hector, Gregory Lewis, and Lones Smith. "Student Portfolios and the College Admissions Problem." *Review of Economic Studies* 81, no. 3 (2014): 971-1002.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. "Where is the land of opportunity? The geography of intergenerational mobility in the United States." *The Quarterly Journal of Economics* 129, no. 4 (2014): 1553-1623.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. "Is the United States still a land of opportunity? Recent trends in intergenerational mobility." *American Economic Review* 104, no. 5 (2014): 141-47.

Cook, Emily. "Competing Campuses: An Equilibrium Model of the U.S. Higher Education Market." Working paper, University of Virginia (2020).

Corak, Miles. "The Canadian geography of intergenerational income mobility." *The Economic Journal* (2018).

Curs, Bradley, and Larry D. Singell Jr. "An analysis of the application and enrollment processes for in-state and out-of-state students at a large public university." *Economics of Education Review* 21, no. 2 (2002): 111-124.

Dale, Stacy Berg, and Alan B. Krueger. "Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables." *The Quarterly Journal of Economics* 117, no. 4 (2002): 1491-1527.

Deming, David J., and Christopher R. Walters. "The impact of price caps and spending cuts on US postsecondary attainment." NBER Working Paper 23736, 2017.

Dillon, Eleanor Wiske, and Jeffrey Andrew Smith. "Determinants of the match between student ability and college quality." *Journal of Labor Economics* 35, no. 1 (2017): 45-66.

Dynarski, Susan M. "Does aid matter? Measuring the effect of student aid on college attendance and completion." *American Economic Review* 93, no. 1 (2003): 279-288.

Ellickson, Paul B., Stephanie Houghton, and Christopher Timmins. "Estimating network economies in retail chains: a revealed preference approach." *The RAND Journal of Economics* 44, no. 2 (2013): 169-193.

Epple, Dennis, Richard Romano, and Holger Sieg. "Admission, tuition, and financial aid policies in the market for higher education." *Econometrica* 74, no. 4 (2006): 885-928.

Fillmore, Ian. "Price discrimination and public policy in the US college market." *Employment Research Newsletter* 23, no. 2 (2016): 2.

Fu, Chao. "Equilibrium tuition, applications, admissions, and enrollment in the college market." *Journal of Political Economy* 122, no. 2 (2014): 225-281.

Hillman, Nicholas W. "Geography of college opportunity: The case of education deserts." *American Educational Research Journal* 53, no. 4 (2016): 987-1021.

Howell, J. "Assessing the Impact of Eliminating Affirmative Action in Higher Education." *Journal of Labor Economics*, 28, no. 1 (2010): 113-166.

Long, Bridget Terry. "How have college decisions changed over time? An application of the conditional logistic choice model." *Journal of Econometrics* 121, no. 1-2 (2004): 271-296.

Manski, Charles F., and David A. Wise. *College choice in America*. Harvard University Press, 1983.

Monge-Naranjo, Alexander, and Lance Lochner. "Credit Constraints in Education," *Annual Review of Economics*, Vol. 4 (2012): 225-256.

Peltzman, Sam. "The effect of government subsidies-in-kind on private expenditures: The case of higher education." *Journal of Political Economy* 81, no. 1 (1973): 1-27.

Turley, R. N. L. (2009). "College Proximity: Mapping Access to Opportunity." *Sociology of Education*, 82(2), 126-146.

Appendices

A Proof of Proposition 1

For convenience, we will drop the X_i and β_i arguments from the $\overline{v}$ function, denoting the ex ante value of being admitted to the set of colleges O_i as $\overline{v}(O_i)$. We do the same for the V function, denoting the value of an application set Y_i for student i as $V(Y_i)$. Finally, we drop the X_i argument and denote the probability that student i is admitted to the set of colleges O_i as $P(O_i)$.

Under Assumption 1, the value of application set Y_i is given by

$$\begin{aligned} V(Y_i) &= \sum_{O_i \subseteq Y_i} P(O_i)\,\overline{v}(O_i) + \left(1 - \sum_{O_i \subseteq Y_i} P(O_i)\right) E(u_{i0}) \\ &= \sum_{O_i \subseteq Y_i} P(O_i)\,\overline{v}(O_i). \end{aligned} \tag{9}$$

Pick any school in Y_i, say, y^*. $O_i' \subseteq Y_i \backslash y^*$ are sets that do not include y^*. (9) can be written as

$$V(Y_i) = p_{iy^*} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', y^*\}) + (1 - p_{iy^*}) \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(O_i')$$

Consider the set Y_i and Y_i' where y^* is replaced by k.

$$V(Y_i') = p_{ik} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', k\}) + (1 - p_{ik}) \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(O_i')$$

$$\begin{aligned} V(Y_i) - V(Y_i') &= p_{iy^*} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', y^*\}) - p_{ik} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', k\}) \\ &\quad - (p_{iy^*} - p_{ik}) \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(O_i'). \end{aligned}$$

$V(Y_i) - V(Y_i') \geq 0$ implies

$$\begin{aligned} & p_{iy^*} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', y^*\}) - p_{ik} \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(\{O_i', k\}) \\ \geq\ & (p_{iy^*} - p_{ik}) \sum_{\{O_i'\} \subseteq Y_i \backslash y^*} P(O_i')\,\overline{v}(O_i') \end{aligned}$$

B Choice sets

For each student, the choice set always includes (1) colleges in the actual application set and (2) public colleges with the highest median SAT in each state that has a reciprocity agreement with

the student's state of residency, or the public college with the highest median SAT in the nearest neighboring state if the student's state of residency has no reciprocity agreement with any other state. The rest of the choice set are drawn randomly from the remaining colleges.

Specifically, for each student, all colleges are divided into ten groups that are exhaustive and mutually exclusive: (i) top public in state, (ii) other public in state, (iii) top private in state, (iv) other private in state, (v) top public out of state, (vi) middle public out of state, (vii) other public out of state, (viii) top private out of state, (ix) middle private out of state, and (x) other private out of state. An in-state public (private) college is classified as top if it meets at least one of three criteria: (1) median SAT ranks among the top 10% of all public (private) colleges in the country, (2) median SAT ranks first among all public (private) colleges in state, (3) enrollment ranks first among all public (private) colleges in state. An in-state public (private) college is classified as other if it does not meet any of the three criteria. An out-of-state public (private) college is classified to be: (1) top if its median SAT ranks among the top 10% of all public (private) colleges in the country, (2) middle if its median SAT ranks among the top 10-30% of all public (private) colleges in the country, and (3) other if its median SAT ranks among the bottom 70% of all public (private) colleges in the country.

The number of colleges drawn from each of the ten groups is proportional to the weighted group size, subject to the following modifications:

1. Colleges in groups (vii) and (x) receive a weight of 0.5. All other colleges receive a weight of 1.
2. There must be at least one college drawn from each of the four in-state groups (i)-(iv), unless the group is empty. The numbers for other groups are adjusted so that they are still proportional to the weighted group size.
3. If the number of colleges for a group is not large enough to include the colleges in the actual application set and the best public colleges in reciprocity (or the nearest neighboring) states, it is increased to the number of the two types of colleges in the group. The numbers for other groups are adjusted so that they are still proportional to the weighted group size.
4. If necessary, a random number is used to make sure the resulting numbers are all integers. As an example, suppose steps 1-3 imply that the numbers of colleges drawn from the first two groups are 1.6 and 3.4, respectively, while the numbers for the other groups are all integers. We would draw a number from the uniform distribution [0,1]. If the number drawn is less than 0.6, we set the numbers for the first two groups to 2 and 3, respectively. Otherwise, they are set to 1 and 4, respectively.

The colleges in the actual application set and the best public colleges in reciprocity (or the nearest neighboring) states are drawn first. If the number of these two types of colleges in a group is smaller than the number to be drawn, the rest are drawn randomly from the remaining colleges in the group.

While this procedure almost always ensures that the flagship university in a student's home state is included in her choice set, theoretically it may not be included if (1) the size of group (i) is larger than one and (2) the flagship is not in the student's application set.

C Estimation algorithm

We use the following algorithm to construct the quasi-likelihood for student i's observed choices (application set and enrollment decision):

1. Simulate M copies of student i with the same characteristics X_i but different preference "shocks" $\epsilon_{k,i}$, so that each simulated student $m \in \{1, .., M\}$ is characterized by $\left(X_i, \epsilon_{k,i}^m\right)$, which are fixed throughout. Also draw a set of S post-application shocks $\{\eta_{ij}, \epsilon_{i0}\}$, which are fixed throughout.

2. For each simulated student, compare each college j in the observed application set Y_i with every college k in $\mathcal{J}_i \backslash Y_i$, checking the necessary condition from Proposition 1. For Y_i to be optimal among the set of alternatives involving one-for-one swaps, it must be that $V(Y_i, X_i, \beta_i^m) \geq V(Y_i^{k\backslash j}, X_i, \beta_i^m)$ for all possible swaps (j, k), where $Y_i^{k\backslash j}$ denotes i's application set with college k replacing college j. The probability is calculated using a kernel smoothed frequency simulator (McFadden (1989)), which converges to the frequency simulator as the smoothing parameter ι goes to zero, so that the likelihood of the application decision for copy m of student i is

$$L_{i,m}^{app} = \frac{\exp\left(\frac{V(Y_i, X_i, \beta_i^m)}{\iota}\right)}{\exp\left(\frac{V(Y_i, X_i, \beta_i^m)}{\iota}\right) + \sum_{j \in Y_i} \sum_{k \in \mathcal{J}_i \backslash Y_i} \exp\left(\frac{V(Y_i^{k\backslash j}, X_i, \beta_i^m)}{\iota}\right)}.$$

3. For each college o in $O_i \cup \{0\}$ (the set O_i of colleges that admitted student i augmented to include the outside option of not enrolling in any college $j = 0$), compute the ex post utility U_{imso} for the random coefficient draw m and post-application shocks s. Letting o^* be the observed choice and $r_{imso}^e \equiv U_{imso^*} - U_{imso}$, we compute the smoothed likelihood of the enrollment decision for copy m of student i as

$$L_{i,m}^{enr} = \frac{1}{S} \sum_s \frac{1}{\sum_{o \in O_i \cup \{0\}} e^{-\frac{r_{imso}^e}{\iota}}} \tag{10}$$

4. The smoothed log-likelihood function of the sample is then calculated as

$$L = \frac{1}{N} \sum_i \ln\left(\frac{1}{M} \sum_m L_{i,m}^{app} L_{i,m}^{enr}\right). \tag{11}$$

The model parameters are then chosen to minimize this log-likelihood function using a standard numerical optimization algorithm.

D Additional Tables for Model Fit

Table 12 reports model fit statistics for three groups of family income, and Table 13 reports the same statistics for three groups of student SAT.

Table 12: Model fit by family income

	Data			Model		
Family income ($1,000)	≤ 35	$(35, 100]$	> 100	≤ 35	$(35, 100]$	> 100
Panel A: College characteristics - application						
Admission probability	0.68	0.76	0.77	0.71	0.78	0.74
Tuition ($1,000)	9.28	11.03	14.17	8.69	10.13	13.31
Aid probability	0.61	0.50	0.36	0.63	0.54	0.37
Aid amount ($1,000)	9.04	7.79	6.39	9.41	8.29	6.56
$(SAT_i - SAT_j)_+^2$	0.37	0.73	0.94	0.34	0.68	0.75
$(SAT_i - SAT_j)_-^2$	4.88	2.40	1.41	4.21	2.08	1.59
Median SAT (100)	10.66	10.99	11.52	10.59	10.96	11.66
Private	0.26	0.31	0.40	0.32	0.38	0.44
Distance (100 km)	2.70	3.26	4.79	1.64	2.33	4.86
Out of state	0.18	0.25	0.41	0.07	0.14	0.33
NCAA Division I sports	0.30	0.34	0.39	0.27	0.32	0.39
Panel B: College characteristics - enrollment						
Admission probability	0.76	0.82	0.79	0.74	0.80	0.74
Tuition ($1,000)	9.68	10.91	14.12	8.31	9.36	12.75
Aid probability	0.64	0.52	0.37	0.64	0.53	0.36
Aid amount ($1,000)	9.32	7.74	6.39	9.30	7.92	6.39
$(SAT_i - SAT_j)_+^2$	0.47	0.88	0.88	0.39	0.70	0.65
$(SAT_i - SAT_j)_-^2$	2.36	1.23	0.92	3.32	1.69	1.45
Median SAT (100)	10.67	10.98	11.58	10.70	11.02	11.76
Private	0.30	0.32	0.40	0.28	0.33	0.40
Distance (100 km)	2.51	3.12	4.57	1.42	2.03	4.45
Out of state	0.16	0.23	0.40	0.06	0.11	0.29
NCAA Division I sports	0.27	0.34	0.42	0.30	0.35	0.44
Panel C: Admission and enrollment rate as a share of students						
Admission rate	0.81	0.91	0.96	0.86	0.92	0.96
Enrollment rate	0.77	0.85	0.91	0.83	0.90	0.92

The admission rate is the fraction of students who were admitted to at least one of the colleges they applied to, and the enrollment rate is conditional on being admitted to at least one college.

E Simulations

For baseline simulations, we proceed as follows:

1. For each individual, we keep the consideration set (with size J) and the draws of random coefficients and shocks (to financial aid and the outside option) used in estimation.

2. For each individual, we draw J random numbers to determine the admission outcome of each college in the consideration set.

3. For each draw of the random coefficients, we calculate the expected value of each possible combination of n colleges, where n is the number of applications in data. The combination with the largest expected value will be the application set.

Table 13: Model fit by student SAT

	Data			Model		
Student SAT	≤ 950	$(950, 1130)$	≥ 1130	≤ 950	$(950, 1130)$	≥ 1130
Panel A: College characteristics - application						
Admission probability	0.63	0.80	0.81	0.68	0.80	0.80
Tuition ($1,000)	8.90	10.46	14.85	8.15	9.96	13.65
Aid probability	0.46	0.48	0.54	0.49	0.52	0.56
Aid amount ($1,000)	7.21	7.19	8.89	7.35	7.87	9.30
$(SAT_i - SAT_j)_+^2$	0.01	0.20	1.93	0.01	0.17	1.71
$(SAT_i - SAT_j)_-^2$	6.49	1.07	0.24	5.36	1.36	0.33
Median SAT (100)	10.41	10.93	11.81	10.24	11.02	11.93
Private	0.25	0.29	0.43	0.32	0.36	0.45
Distance (100 km)	2.52	3.19	4.79	2.12	2.45	3.70
Out of state	0.19	0.24	0.38	0.12	0.14	0.24
NCAA Division I sports	0.27	0.34	0.41	0.23	0.33	0.41
Panel B: College characteristics - enrollment						
Admission probability	0.72	0.83	0.82	0.70	0.80	0.80
Tuition ($1,000)	8.54	10.33	14.43	7.72	9.19	12.57
Aid probability	0.47	0.48	0.54	0.48	0.51	0.54
Aid amount ($1,000)	6.70	7.10	8.82	7.06	7.43	8.81
$(SAT_i - SAT_j)_+^2$	0.01	0.21	1.82	0.01	0.16	1.56
$(SAT_i - SAT_j)_-^2$	3.86	0.88	0.17	4.71	1.28	0.27
Median SAT (100)	10.20	10.87	11.82	10.28	11.02	11.96
Private	0.26	0.29	0.42	0.29	0.31	0.40
Distance (100 km)	2.31	3.06	4.35	2.00	2.20	3.17
Out of state	0.18	0.23	0.34	0.11	0.12	0.20
NCAA Division I sports	0.22	0.33	0.44	0.25	0.36	0.46
Panel C: Admission and enrollment rate as a share of students						
Admission rate	0.77	0.95	0.99	0.84	0.95	0.97
Enrollment rate	0.70	0.87	0.94	0.82	0.91	0.93

The admission rate is the fraction of students who were admitted to at least one of the colleges they applied to, and the enrollment rate is conditional on being admitted to at least one college.

4. Given the application set and the random numbers that determine the admission outcome of each college, we have the admission set. We can then calculate the enrollment outcome for each draw of the shocks to financial aid and outside options.