Journal of Monetary Economics 129 (2022) 49–64

Contents lists available at ScienceDirect

Journal of Monetary Economics

journal homepage: www.elsevier.com/locate/jmoneco

Students' heterogeneous preferences and the uneven spatial distribution of colleges☆

Chao Fu[a], Junjie Guo[b], Adam J. Smith[b], Alan Sorensen[a,*]

[a] University of Wisconsin & NBER, United States
[b] University of Wisconsin, United States

ARTICLE INFO

Article history:
Received 25 February 2022
Accepted 1 March 2022
Available online 4 March 2022

Keywords:
Higher education
College competition
Geography of opportunity

ABSTRACT

The uneven geographic distribution of colleges in the United States endows students with uneven access to colleges depending on where they live. To examine the implication of this for student welfare, 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, and find that place indeed matters: the expected value of applying to college differs dramatically across states and across counties within a state.

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 the flagship college in the student's home state, as proxied by the median SAT score of admitted freshmen.[4] At the lower end, 5% of the students

☆ We are grateful for helpful comments from Emily Cook, Adam Kapor, Chris Sleet, Jeff Smith, and seminar participants at the University of Oregon, the NBER Labor Studies Meeting, the North American and China meetings of the Econometric Society, and the Carnegie-Rochester-NYU Conference on Public Policy.

* Corresponding author.
E-mail addresses: cfu@ssc.wisc.edu (C. Fu), jguo27@wisc.edu (J. Guo), ajsmith26@wisc.edu (A. Smith), sorensen@ssc.wisc.edu (A. Sorensen).

[1] See, for example, Brewer et al. (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.

https://doi.org/10.1016/j.jmoneco.2022.03.001
0304-3932/

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.

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 km 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 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, raising 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 the 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 combinatorially 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. (2014a) and several other recent papers that have emphasized the geography of opportunity,[8] we use our estimates to calculate ex ante

[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.

[6] Some studies have attempted to quantify which factors are influential in students' college choice decisions (e.g., Avery and Hoxby, 2004; Dillon and Smith, 2017; Long, 2004; Manski and Wise, 1983); while another set of studies has focused specifically on the impact of tuition or financial aid on college choices (e.g., Avery and Hoxby, 2004; Curs and Singell, 2002; Dynarski, 2003, Kane 2007 and Deming and Walters, 2017).

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

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 equivalent to more than 2200 tuition dollars, compared to about 1600 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, 60% 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 this 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 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 et al. (2006), Fu (2014), Bodoh-Creed et al. (2018), Fillmore (2016), Cook (2020) and Kapor (2020) 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 et al. (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 et al. (2004) to construct a revealed preference ranking of U.S. universities.[9] 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 7409 students, whose characteristics are summarized in Table 2.

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

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

[9] Data on students' application and admission sets have been used elsewhere to study heterogeneous effects of colleges on student outcomes, e.g., Dale and Krueger (2002), Arcidiacono et al. (2016), Bleemer (2021), and Mountjoy and Hickman (2021).

Table 2
Summary of student characteristics ($N = 7,409$).

			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				

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.

academies and colleges whose Carnegie classification is "Special Focus Institution".[10] The resulting sample includes 1337 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 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]

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 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.

[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."
[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.

	1	2	3	4	5+
All students	0.30	0.24	0.17	0.11	0.18
Low income, Low SAT	0.39	0.30	0.16	0.07	0.08
Low income, Middle SAT	0.34	0.26	0.16	0.10	0.14
Low income, High SAT	0.27	0.15	0.15	0.14	0.28
Middle income, Low SAT	0.41	0.27	0.18	0.07	0.07
Middle income, Middle SAT	0.32	0.25	0.19	0.11	0.13
Middle income, High SAT	0.24	0.22	0.17	0.13	0.24
High income, Low SAT	0.27	0.30	0.19	0.13	0.10
High income, Middle SAT	0.24	0.22	0.15	0.12	0.27
High income, High SAT	0.15	0.15	0.17	0.14	0.39

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).

Table 5
Examples of application overlaps.

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

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.

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 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, a 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.,

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

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

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

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

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 (2020)) or by grouping colleges into a small number of types (e.g., Epple et al. (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, a student's application problem (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.,

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

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.[15]

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.

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

[15] Note that we have assumed that the cost of application $C(|Y|)$ depends only on the size of the application set $|Y|$ rather than the components of Y. In reality, some colleges may have higher (pecuniary and/or non-pecuniary) application costs than others, and when we estimate our model these differences will be absorbed into students' preferences for colleges. For example, if higher-quality colleges are more likely to require supplemental essays, our estimates of students' preferences for quality will absorb these higher application costs.

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 Assumption 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 $k \notin Y_i$,*

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

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. In a setting similar to ours, Larroucau and Rios (2018) prove that a condition analogous to ours (relating to what they call "one-shot swaps") is sufficient for optimality. However, this is not the case in our model: because the post-application shocks can create complementarities between certain pairs of colleges, the condition from Proposition 1 is necessary for optimality, but not sufficient.

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 briefly describe our estimation of these two components, and then 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.[16] 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.

[16] 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.

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,[17] 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 students will tend to be higher than the aid amounts offered to admitted students, so our model may slightly overpredict aid amounts.[18] 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

$$\begin{aligned} u_{ij} = &-(\gamma_1 LowInc_i + MidInc_i + \gamma_2 HighInc_i)t_{ij} \\ &+ \alpha_0 + \alpha_1 (SAT_i - SAT_j^c)_+^2 + \alpha_2 (SAT_i - SAT_j^c)_-^2 + \alpha_3 Black_i + \alpha_4 Hispanic_i \\ &+ \exp(\beta_{1,i})\left[q_j + \delta_1 \left(q_j - q^{85th}\right)_+\right] + \beta_{2,i}\left[\ln\left(Dist_{ij}\right) + \delta_2 OutState_{ij}\right] \\ &+ \beta_{3,i} Private_j + \beta_{4,i} NCAAI_j. \end{aligned} \tag{7}$$

The first component of this function reflects the student's sensitivity to net tuition $\left(t_{ij}\right)$, 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. We normalize the tuition coefficient for middle-income students to 1, so student preferences for various college attributes are measured in tuition dollars. Parameters γ_1 and γ_2 measure the price sensitivity of low- and high-income students, respectively.

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 quality q_j. Following the method in Dillon and Smith (2017), we construct q_j as an index that combines three characteristics of each college j : the median SAT, the average faculty salary, and the faculty-student ratio.[19] Since quality differences are most likely to be meaningful for colleges toward the upper end of the distribution, we allow the slope of college quality to depend on whether or not q_j is above the 85th percentile of the q_j distribution across our sample of four-year colleges ($q^{85th} = 0.73$).[20]

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. Estimates of students' preferences for proximity depend on the extent to which students in our data apply to (and enroll in) colleges close to home. In reality these "preferences" can arise from factors beyond the cost of physical distance and the attachment to one's home state. For example, a student may know less about colleges farther away and may therefore be less likely to apply. Also, applications to multiple nearby colleges may be driven by application costs rather than preferences for poximity, since some colleges (e.g. in the University of California and Pennsylvania State systems) can be applied to in bundles. Thus, what we estimate and label as a preference for proximity should be interpreted broadly, since it may also reflect unmodeled application costs and differences in the information students have about close vs. distant colleges.

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 the student experience at college j.

[17] 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.

[18] To check whether this selection effect is likely to be important, we examined data from the National Longitudinal Survey of Youth, 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.

[19] Dillon and Smith (2017) use the percent of applicants rejected as a fourth measure. We do not use this measure because we are interested in application and admissions. See Appendix C for details.

[20] The results are similar if we use $SAT_j^c + \delta_1\left(SAT_j^c - SAT^{c,85th}\right)_+$ instead of $q_j + \delta_1\left(q_j - q^{85th}\right)_+$ as the quality measure. The correlation between SAT_j^c and q_j is 0.91.

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 quality $(\beta_{1,i})$ and private colleges $(\beta_{3,i})$ are allowed to vary with family income. The mean taste for distance $(\beta_{2,i})$ is allowed to vary with family income and the average daily travel distance in one's home state ($AveD_i$), which captures the possibility that students may "benchmark" the home-college distance against their everyday experience.[21] The mean taste for Division I sports $(\beta_{4,i})$ is assumed to be common across students. Formally, the mean tastes for these attributes are given by

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

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.

Identification

As in a standard discrete choice model, students' overall propensities to choose different colleges identify their average preferences for the college characteristics included in the model. What may be less obvious is how the data can identify *heterogeneity* in students' preferences—i.e., the variances σ_k^2 of preference coefficients $\beta_{k,i}$. For that, we rely on the richness of our data, which goes beyond a typical dataset on consumer's choices in two aspects. First, we observe students' preference-revealing choices at both the application stage (which colleges to apply to) and the enrollment stage (which college to enroll in among the set of colleges the student was admitted to). The second and more important aspect is that we observe students' full application sets. The strength of a particular student's preference for a given college characteristic is reflected in the similarity of the applied-to colleges in that characteristic. For example, if a student applied mostly to nearby colleges (e.g., the first two examples in Table 5), the model will infer that the student had a strong preference for proximity to home (a large negative $\beta_{2,i}$); if a student applied mostly to academically competitive colleges (e.g., the last example in Table 5), the model will infer that the student had a strong preference for college quality (a large $\beta_{1,i}$); similarly, a student who applied only to universities with NCAA Division 1 sports programs will be inferred to have a high $\beta_{4,i}$. Data on application sets thus provide information about an *individual* student's preferences, and heterogeneity in preferences can then be naturally identified from variation across students.[22]

Since we rely on similarities between colleges in students' application sets to learn about their preferences, the choices of students who applied to many colleges are the most informative.[23] However, data from students in college-rich locations are not necessarily more informative than from students in locations with few colleges. The consideration set varies across students, but its size is large and the same across all students. This means even students in regions with few colleges are considering a large number of options with different characteristics, which helps identify the preferences for those characteristics.

Note that there are multiple reasons why students face different net prices, which lead to rich variation we use for identifying their tuition sensitivity. The sources of price variation include tuition differences across colleges, in-state versus out-of-state tuition differences in the case of public colleges, and financial aid differences across college-student pairs.

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.

[21] Data for $AveD_i$ are obtained from the National Household Travel Survey (NHTS) 2009 Transferability Statistics. It first sums the travel distance of all members in a household on a weekday, and then takes the average across households in a state. The variable ranges from about 68.4 km in DC to about 116.7 km in South Dakota.

[22] Our use of data on application sets is related to the use of "second-choice data" in demand estimation, as in Berry et al. (2004).

[23] Some students in our data applied to only one college. Data on these students are more like typical datasets where consumers choose one product from a menu of available products. While these students' college choices are still informative about average preferences, they are less helpful for identifying heterogeneity in preferences.

Table 6
Utility parameter estimates.

Variable	Parameter	Estimate	Std. Error
Tuition×Low Income	γ_1	1.283	0.097
Tuition×High Income	γ_2	0.228	0.033
Constant	α_0	33.461	0.393
$\left(SAT_i - SAT_j^c\right)_+^2$	α_1	–1.028	0.059
$\left(SAT_i - SAT_j^c\right)_-^2$	α_2	0.152	0.039
Black	α_3	–0.077	1.166
Hispanic	α_4	–2.645	1.009
$(q_j - q^{85th})_+$	δ_1	2.650	0.461
Quality index: q_j	$\mu_{1,0}$	0.927	0.088
q_j×Low Income	$\mu_{1,1}$	–0.085	0.087
q_j×High Income	$\mu_{1,2}$	–0.454	0.086
Out of state	δ_2	1.891	0.148
Distance	$\mu_{2,0}$	–1.435	0.113
Distance×Low Income	$\mu_{2,1}$	–0.897	0.118
Distance×High Income	$\mu_{2,2}$	0.841	0.077
Distance×State Average Travel Distance	$\mu_{2,3}$	–0.011	0.001
Private	$\mu_{3,0}$	–2.366	0.275
Private×Low Income	$\mu_{3,1}$	–0.274	0.528
Private×High Income	$\mu_{3,2}$	0.002	0.356
NCAA Division 1	$\mu_{4,0}$	0.675	0.225
Std dev. of quality (q_j) preference	σ_1	0.525	0.060
Std dev. of distance preference	σ_2	1.188	0.058
Std dev. of Private preference	σ_3	5.193	0.588
Std dev. of NCAA Div. 1 preference	σ_4	4.532	0.632
Std dev. of shocks to the outside option	λ_0	2.677	0.030
Std dev. of shocks to the outside option (low income)	λ_1	0.290	0.054
Std dev. of shocks to the outside option (high income)	λ_2	0.314	0.038

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. The set $\mathcal{J}_i$ 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 D.

5. Results

5.1. Parameter estimates

Table 6 reports the parameter estimates and associated standard errors.[24] The estimated tuition sensitivity declines with family income, which is in line with findings from the previous literature. The coefficient on $(SAT_i - SAT_j^c)_+^2$ is negative, suggesting that students dislike being undermatched (own SAT being above college SAT). However, we do not find similar distastes for being overmatched (own SAT being below college SAT); the coefficient on $(SAT_i - SAT_j^c)_-^2$ is actually slightly positive.

The main parameters of interest are the ones related to the heterogeneity in preferences—i.e., the distributions of student-specific coefficients for various college attributes (academic quality, distance, 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

[24] 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 7
Preference heterogeneity.

	Preference at:		
	$\hat{\mu}-\hat{\sigma}$	$\hat{\mu}$	$\hat{\mu}+\hat{\sigma}$
Increase q_j from -0.5 to 0 (27th to 59th percentile)			
Low income	0.54	0.90	1.53
Middle income	0.75	1.26	2.14
High income	2.08	3.52	5.95
Increase q_j from 1 to 1.5 (90th to 95th percentile)			
Low income	1.95	3.30	5.58
Middle income	2.73	4.61	7.80
High income	7.60	12.85	21.71
Increase distance from 16km to 160km (10 to 100 miles)			
Low income	−8.42	−6.29	−4.15
Middle income	−8.74	−6.00	−3.26
High income	−29.76	−17.76	−5.76
Out of state vs. in state			
Low income	−6.91	−5.16	−3.41
Middle income	−7.17	−4.93	−2.68
High income	−24.44	−14.59	−4.73
Private vs. public			
Low income	−6.11	−2.06	1.99
Middle income	−7.56	−2.37	2.83
High income	−33.14	−10.37	12.41
NCAA Division I sports			
Low income	−3.01	0.53	4.06
Middle income	−3.86	0.68	5.21
High income	−16.92	2.96	22.84

To calculate the numbers related to distance reported in the third and fourth panels, we set the value for the state average travel distance at the median across all states.

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 ($\hat{\gamma}_2 = 0.228$), so the heterogeneity in their preferences for non-tuition college characteristics is amplified when expressed in terms of tuition dollars.

Several points stand out from Table 7. First, there is considerable heterogeneity in how much students value academic quality. A middle-income student at the higher end of the quality preference ($\beta_{1,i}$) distribution would be willing to pay $7,800 more in tuition to attend a college with a quality index of 1.5 vs. 1 (rough examples would be University of Michigan-Ann Arbor vs. UT Austin, or College of William and Mary vs. Bard College), whereas a student at the lower end of the $\beta_{1,i}$ distribution would be willing to pay only $2,730. Second, most students have strong preferences for attending colleges close to home. For example, for a middle-income student with an average distance preference—i.e., with the mean value of $\beta_{2,i}$—an increase in distance from 10 to 100 miles is equivalent to a nearly $6,000 increase in tuition. However, 0.2 percent of low-income students and 6.9 percent of high-income students have a positive preference for distance. Similarly, while most students exhibit strong home-state biases (for reasons beyond tuition and distance), a small fraction of students prefer to study out of their home states. Third, all else being equal, an average student values public colleges over private colleges, but over 30% of students prefer private colleges. Finally, our estimates suggest that 56 percent of students favor a college with an NCAA Division I football team over an otherwise equivalent college without such a team, while the rest have the opposite preference.

5.2. Model fit

As discussed in Section 3.2.1, all of our simulations take as given the observed number of colleges $|Y_i|$ a student applied to. To evaluate how well our model fits the data, we simulate each student i's optimal application set given size $|Y_i|$ 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 quality, 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 col-

Table 8
Model fit.

Panel A: College characteristics	Data		Model	
	Application	Enrollment	Application	Enrollment
Admission probability	0.74	0.80	0.76	0.78
Tuition ($1,000)	11.34	11.51	10.71	10.03
Aid probability	0.49	0.50	0.52	0.52
Aid amount ($1,000)	7.76	7.69	8.20	7.78
$\left(SAT_i - SAT_j\right)_+^2$	0.70	0.81	0.64	0.65
$\left(SAT_i - SAT_j\right)_-^2$	2.74	1.36	2.53	1.98
Median SAT (100)	11.03	11.08	10.99	11.06
Quality index q_j	0.24	0.26	0.28	0.32
Private	0.32	0.33	0.39	0.34
Distance (100 km)	3.47	3.38	3.07	2.77
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	0.89		

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.

lege, we first take the average of (college-student-specific) characteristics across the colleges a student applied to, and then average across students.[25] 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. Overall, the model fits the data well. However, it underpredicts the tuition and home-to-college distance (physical distance and out-of-state status) for both applied colleges and enrolled colleges. 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.

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.[26] 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.

Fig. 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 to the bottom) and student SAT tercile (low, middle, high shown from the left to the right).[27] 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.[28] This preference for

[25] 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., $\frac{1}{I}\sum_{i=1}^{I}\overline{p}_i$.

[26] To construct the group averages, we use means for continuous variables and medians for categorical variables.

[27] 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.

[28] In the discussion here we interpret our estimates as literal preferences for proximity. As we noted above, however, the estimates may also reflect unmodeled differences in application costs, or gaps in the amount of information students have about near vs. distant colleges.

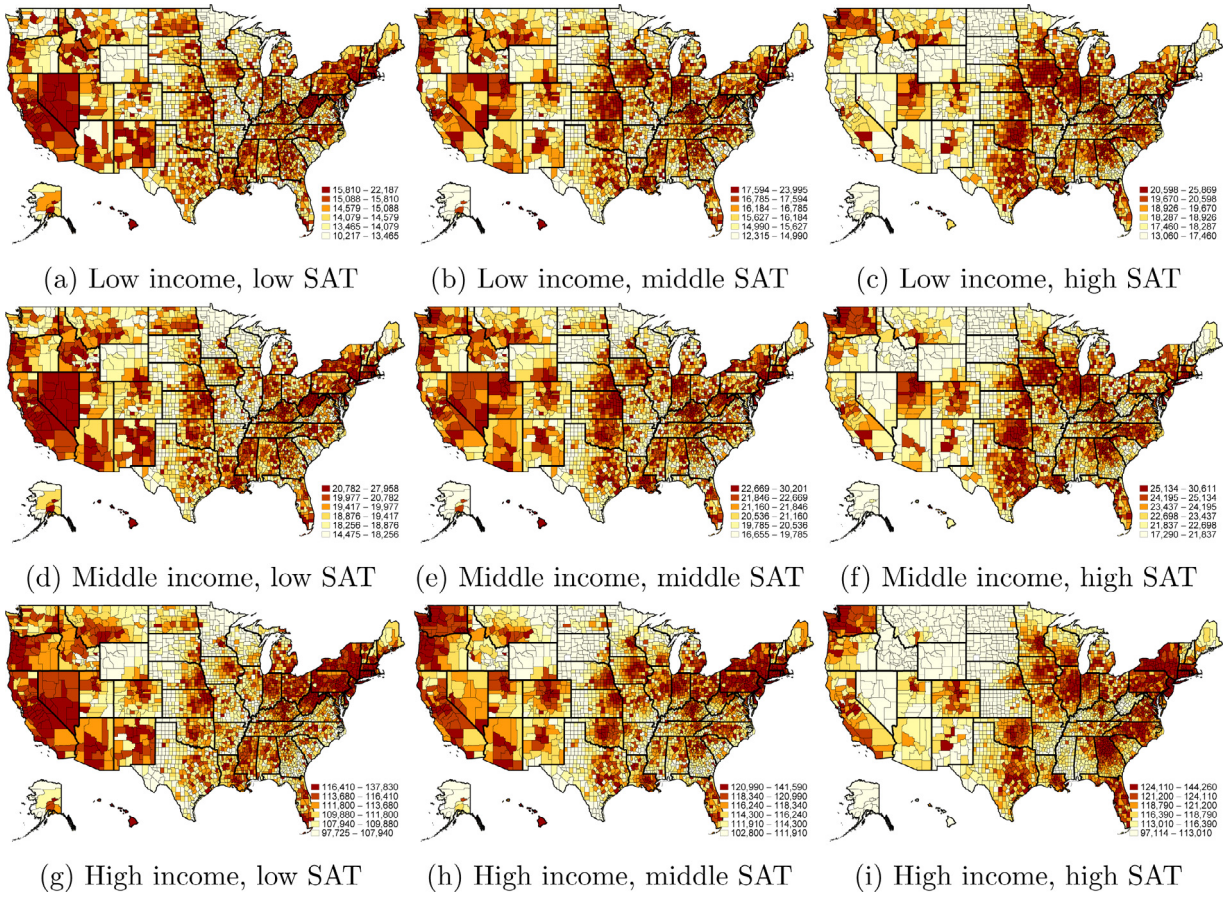

Fig. 1. Expected utility of college education across U.S. counties. Expected utility is measured in tuition dollars. From top to bottom, the family income for the three rows are $\leq 35,000$, (35,000,100,000], and $> 100,000$, respectively. From left to right, the student SAT for the three columns are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

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 Fig. 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 South Dakota, 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 Fig. 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 five predicted outcomes across counties. For instance, the enrollment probability of a low-income student with a middle SAT score varies by 5.88 percentage points between the 25th and 75th percentile counties, and the median SAT score of the enrolled college varies by 53 points. Comparing across rows in the table, we find that the higher a student's family income, the less her home location matters for her enrollment probability and the quality of her enrolled college, and the more her location matters for home-college distance and net tuition. This suggests that students from richer families are more able to offset their location disadvantages (in terms of college access) by traveling further and/or paying higher net tuition.

There are also meaningful correlations between the same student's welfare and her enrollment outcomes across counties, as shown in Table 10. Each row in the table refers to one of the 9 representative students; the five columns show how each representative student's expected utility is correlated with her college enrollment probability and the characteristics of the enrolled college (q_j, SAT, 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,

Table 9
Interquartile ranges of simulated outcomes.

	Expected	Enrollment	Enrolled college			
Student group	utility ($)	Prob (%)	q_j	SAT	Dist	Net Tui
Low income, Low SAT	1593	6.11	0.32	52	199	1,545
Low income, Middle SAT	1775	5.88	0.35	53	155	1,540
Low income, High SAT	2212	5.35	0.37	58	174	2,118
Middle income, Low SAT	1708	4.77	0.31	46	240	1,337
Middle income, Middle SAT	1987	4.43	0.34	47	196	1,443
Middle income, High SAT	2293	4.15	0.32	49	214	2,193
High income, Low SAT	5930	2.56	0.28	36	291	1,975
High income, Middle SAT	6340	1.68	0.24	37	246	1,952
High income, High SAT	7604	1.95	0.26	36	302	2,922

Expected utility values are divided by the relevant tuition coefficient in order to express utility in terms of tuition dollars. Dist is distance in kilometers, and NT is net tuition in dollars.

Table 10
Correlations with the expected utility.

	Enrollment	Enrolled college				
Student group	Probability	q_j	SAT	Dist	NT	R^2
Low income, Low SAT	0.92	−0.14	−0.26	−0.65	−0.36	0.29
Low income, Middle SAT	0.87	0.03	−0.08	−0.68	−0.12	0.21
Low income, High SAT	0.86	0.27	0.27	−0.67	−0.04	0.31
Middle income, Low SAT	0.91	0.02	−0.18	−0.52	−0.28	0.33
Middle income, Middle SAT	0.88	0.14	−0.02	−0.57	−0.14	0.25
Middle income, High SAT	0.87	0.33	0.29	−0.64	0.03	0.36
High income, Low SAT	0.88	0.18	−0.13	−0.36	−0.01	0.38
High income, Middle SAT	0.92	0.22	−0.01	−0.45	0.14	0.35
High income, High SAT	0.90	0.43	0.41	−0.62	0.09	0.60

Cells in columns 2–6 report correlations of the indicated outcome with ex ante expected utility across counties. Dist is distance in kilometers, and NT is net tuition in dollars. The last column reports the R^2 from a regression of county expected utilities on state fixed effects.

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 both the quality index q_j and the median SAT of her enrolled college, while this correlation is much weaker for low- and middle-SAT students.

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 20 to 60 percent 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 former, 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, 60% 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.[29] Table 11 reports the differences in average outcomes between the two simulations (counterfactual minus baseline).

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

Table 11
Simulated changes when In-state subsidies are removed.

	Characteristics of enrolled college				
Student group	Quality q_j	SAT	% out of state	Distance (km)	% Private
Low income, Low SAT	−0.04	−7.17	6.28	78.74	25.52
Low income, Middle SAT	−0.05	−11.20	5.15	75.69	32.91
Low income, High SAT	−0.05	−10.18	6.02	90.38	34.91
Middle income, Low SAT	−0.03	−8.79	10.08	129.29	25.55
Middle income, Middle SAT	−0.04	−11.05	8.92	117.73	29.99
Middle income, High SAT	−0.02	−5.99	9.53	130.08	30.54
High income, Low SAT	0.01	−0.91	3.50	49.74	6.85
High income, Middle SAT	0.01	0.04	4.12	62.98	8.19
High income, High SAT	0.04	4.35	4.16	66.32	8.07
All students	−0.01	−5.10	7.55	103.03	23.58

Naturally, we find that eliminating subsidies leads students to substitute away from their home states' universities. Overall, students are 7.55 percentage points more likely to attend a college outside of their home states, and the average distance to the enrolled college increases by 103 km. 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 academic quality as measured by both the index q_j and the median SAT score. 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.[30] [31]

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 tuition would be that the increased tuition revenue could be used to improve the quality of public universities, which—as shown in the previous section—would be especially valuable in states with relatively low-quality flagships.

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

[30] 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 $2,134 to $7,374, compared to $5,170 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.

[31] The predicted enrollment rate (not shown in the table) also drops by 3.7 percentage points, but our simulation may underpredict the drop in enrollment because we hold the number of applications fixed at the baseline. An increase in tuition levels reduces the net benefit of applying to colleges; this may lead some students to apply to fewer colleges and discourage some others from applying to colleges at all. This would imply further reductions in education quality that could more than offset any gains from resolving Peltzman's mismatch problem.

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.

Supplementary material

Supplementary material associated with this article, including all Appendixes referred to in the text, can be found in the online version at 10.1016/j.jmoneco.2022.03.001

References

Abbott, B., Gallipoli, G., 2017. Human capital spill-overs and the geography of intergenerational mobility. Rev Econ Dyn 25, 208–233.
Arcidiacono, P., 2005. Affirmative action in higher education: how do admission and financial aid rules affect future earnings? Econometrica 73 (5), 1477–1524.
Arcidiacono, P., Aucejo, E.M., Hotz, V.J., 2016. University differences in the graduation of minorities in STEM fields: evidence from California. American Economic Review 106 (3), 525–562.
Avery, C., Glickman, M., Hoxby, C., Metrick, A., 2004. A revealed preference ranking of US colleges and universities. In: NBER Working Paper 10803.
Avery, C., Hoxby, C.M., 2004. 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. University of Chicago Press, pp. 239–302.
Berger, T., 2018. Places of persistence: slavery and the geography of intergenerational mobility in the United States. Demography 55 (4), 1547–1565.
Berger, T., Engzell, P., 2019. American geography of opportunity reveals european origins. Proceedings of the National Academy of Sciences 116 (13), 6045–6050.
Berry, S., Levinsohn, J., Pakes, A., 2004. Differentiated products demand systems from a combination of micro and macro data: the new car market. Journal of Political Economy 112 (1), 68–105.
Black, D.A., Smith, J.A., 2006. Estimating the returns to college quality with multiple proxies for quality. J Labor Econ 24 (3), 701–728.
Bleemer, Z., 2021. Affirmative action, mismatch, and economic mobility after California' s proposition 209. Q J Econ.
Bodoh-Creed, Aaron, L., Hickman, B.R., 2018. College assignment as a large contest. J Econ Theory 175, 88–126.
Brewer, D.J., Eide, E.R., Ehrenberg, R.G., 1999. Does it pay to attend an elite private college? J Hum Resour 34 (1), 104–123.
Chade, H., Lewis, G., Smith, L., 2014. Student portfolios and the college admissions problem. Review of Economic Studies 81 (3), 971–1002.
Chade, H., Smith, L., 2006. Simultaneous search. Econometrica 74 (5), 1293–1307.
Chetty, R., Hendren, N., Kline, P., Saez, E., 2014. Where is the land of opportunity? the geography of intergenerational mobility in the United States. Q J Econ 129 (4), 1553–1623.
Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N., 2014. Is the United States still a land of opportunity? recent trends in intergenerational mobility. American Economic Review 104 (5), 141–147.
Cook, E., 2020. Competing Campuses: An Equilibrium Model of the U.S. Higher Education Market. Working paper. University of Virginia.
Corak, M., 2018. The Canadian geography of intergenerational income mobility. The Economic Journal.
Curs, B., Singell Jr, L.D., 2002. An analysis of the application and enrollment processes for in-state and out-of-state students at a large public university. Econ Educ Rev 21 (2), 111–124.
Dale, S.B., Krueger, A.B., 2002. Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. Q J Econ 117 (4), 1491–1527.
Deming, D.J., Walters, C.R., 2017. The impact of price caps and spending cuts on US postsecondary attainment. In: NBER Working Paper 23736.
Dillon, E.W., Smith, J.A., 2017. Determinants of the match between student ability and college quality. J Labor Econ 35 (1), 45–66.
Dynarski, S.M., 2003. Does aid matter? measuring the effect of student aid on college attendance and completion. American Economic Review 93 (1), 279–288.
Ellickson, P.B., Houghton, S., Timmins, C., 2013. Estimating network economies in retail chains: a revealed preference approach. Rand J Econ 44 (2), 169–193.
Epple, D., Romano, R., Sieg, H., 2006. Admission, tuition, and financial aid policies in the market for higher education. Econometrica 74 (4), 885–928.
Fillmore, I., 2016. Price discrimination and public policy in the US college market. Employment Research Newsletter 23 (2), 2.
Fu, C., 2014. Equilibrium tuition, applications, admissions, and enrollment in the college market. Journal of Political Economy 122 (2), 225–281.
Hillman, N.W., 2016. Geography of college opportunity: the case of education deserts. Am Educ Res J 53 (4), 987–1021.
Howell, J., 2010. Assessing the impact of eliminating affirmative action in higher education. J Labor Econ 28 (1), 113–166.
Kapor, A., 2020. Distributional effects of race-blind affirmative action. Working paper. Princeton University.
Larroucau, T., Rios, I., 2018. Do 'short-list' students report truthfully? strategic behavior in the Chilean college admissions problem. Working paper.
Long, B.T., 2004. How have college decisions changed over time? an application of the conditional logistic choice model. J Econom 121 (1–2), 271–296.
Manski, C.F., Wise, D.A., 1983. College choice in America. Harvard University Press.
Monge-Naranjo, A., Lochner, L., 2012. Credit constraints in education. Annu Rev Econom 4, 225–256.
Mountjoy, J., Hickman, B.R., 2021. The returns to college(s): relative value-added and match effects in higher education. In: NBER Working Paper 29276.
Peltzman, S., 1973. The effect of government subsidies-in-kind on private expenditures: the case of higher education. Journal of Political Economy 81 (1), 1–27.
Turley, R.N.L., 2009. College proximity: mapping access to opportunity. Sociol Educ 82 (2), 126–146.