College Quality and Student Education Choices

Su Wang*

October 26, 2026

Click here for the latest version

Abstract

Students' labor market outcomes vary across different colleges, reflecting differences in both student composition and the potential effects of college attendance. In this paper, I utilize Texas administrative data to estimate how students' enrollment decisions are influenced by their heterogeneous preferences and quality of the colleges. Contrary to the literature that defines college quality by how selective the schools are, I define quality as the school's average causal impact on students' future outcomes. I estimate the impact of schools on student potential earnings using a generalized Roy model and account for selection bias with control functions that utilize exclusion restrictions on distance and tuition. I further measure how these earnings impacts, at the average level and also the student-college matching effects, affect students' enrollment. After controlling for student characteristics and selection on unobservables, I find nontrivial but modest variation in quality across four-year public universities, consistent with previous results from the literature. Students' choices are influenced heavily by peer composition, but do not reflect the causal earning impacts of the colleges.

*PhD Candidate, Department of Economics, Johns Hopkins University. Email: swang307@jhu.edu. Website: suwang-sam.github.io. I thank my advisors, Professor Matthew Wiswall, Professor Robert Moffitt, and Professor Yingyao Hu, as well as Professor Jonathan Elliott and Professor Yujung Hwang, for their guidance and support on this project and others during my PhD years. I thank the JHU applied micro group and participants of the 2025 MidSouth Education Policy Workshop for their helpful comments and suggestions. I also thank my peers, John Green, Yongzheng Liang, Yingfei Mu, and Junhyeok Shin for helping me throughout the PhD. All errors are my own.

Disclaimer: The conclusions of this research do not necessarily reflect the opinion or official position of the University of Houston Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas.

1 Introduction

Choosing which college to attend is a major decision for many students, since it could have long-lasting impacts on their future life paths. However, students face widely different options when making this decision, due to "the geography of opportunity" — research finds that people's childhood locations have profound effects on their lifetime earnings (Chetty and Hendren 2018). One potential contributing factor is the uneven spatial distribution of educational opportunities, in particular higher education institutions (Hillman 2016, Fu et al. 2022, Fabre 2023, Ishimaru 2025). The differences are not just limited to the availability of colleges. Different college educations provide varying benefits to students in terms of earning gains (Kirkeboen et al. 2016, Hoxby 2018, Chetty, Friedman, et al. 2020, Sekhri 2020, Jia and Li 2021), where those varying benefits can be viewed as or partly due to differences in education quality across schools.

Previous literature frequently equates "quality" with how selective the colleges are (Arcidiacono 2005, Fu et al. 2022, Ishimaru 2025). Graduates from elite colleges tend to have better labor market outcomes, but the admitted students have higher academic ability on average in the first place. Hence, selectivity measures seem less like a description of how good the schools are in benefitting individual students. Figure 1 below shows the positive correlation between average SAT scores of admitted students and median income of graduates across different post-secondary institutions in the US.[1] In light of this, one might ask: how much of the higher income is brought to the student through the education experiences in that specific university, and how much is from the student's innate ability and own background?

In this paper, I disentangle the causal effect on students' future outcomes of enrolling at

[1]Data from the US Department of Education College Scorecard project: https://collegescorecard.ed.gov/. Institutions are limited to those that offer at least a Bachelor's degree. The income measures 10 years after entry are pulled from the most recent cohort csv file (award year 2021-2022), but are likely from a pooled cohort of AY2010-11 and AY2011-12, adjusted to 2023 dollars. The average SAT score is pulled from the 2011-2012 cohort csv, which shows the average score for Fall 2011 enrollees.

a particular college from the effect of students' own characteristics.[2] I define the school's average impact on student earnings as the "value-added," "treatment effect," or "quality" of that given institution. These measurements have important policy implications, since the effectiveness of higher education policies depends on the quality of institutions that students attend. By showing which schools are particularly successful in improving their student outcomes and where they might need attention and improvement, this paper can guide the better allocation of limited policy resources.

Many papers examine the differential returns of higher education institutions on student earnings (Dale and Krueger 2002, Dale and Krueger 2014; Kirkeboen et al. 2016; Hoxby 2015, Hoxby 2018; Mountjoy and Hickman 2021; Mountjoy 2026). I complement these papers by not only estimating heterogeneous returns to individual colleges and controlling for selection bias, but also exploring how students value these potentially different returns to college education when making enrollment decisions. I develop a discrete choice model with heterogeneous preferences, with a nested logit structure to allow for correlated unobserved preferences for different colleges, reflecting the possibility that students have common unobserved factors affecting the attractiveness of different schools. Understanding student preferences is key to measuring the students' response to education access policies, such as financial aid programs, when they modify or offer more choices to the students. This is integral to measuring the real impact of these policies.

The central empirical challenge is that college attendance is endogenous: students who choose a particular college may differ systematically from those who choose other colleges in ways that affect their potential earnings. I address this selection problem using control function methods, using the estimated choice probabilities and, in some specifications, distance and tuition directly to flexibly control for the component of unobserved student outcomes that is related to the enrollment decision. This relies on the exclusion restriction of these variables. I consider both the parametric correction of Lee (1983) and a family of semi-

[2]The peer effects on student earnings will be absorbed by the college quality and differential returns to student characteristics, if those peer effects depend on student characteristics as well.

parametric control functions, generalizing from Abdulkadiroğlu et al. (2020)[3] while following Dahl (2002) and related approaches.

I then examine the relationship between these estimates of college-specific returns and students' enrollment decisions. I distinguish between three components: the peer composition, the average causal effect on student earnings, and the match between students and colleges. The first captures how selective the schools are, while the second item is my primary measure of college quality. The third captures heterogeneous returns to attending different colleges. This three-component analysis framework is partly inspired by APSW, who examined a similar question but in the setting of high school choices. Comparing the relationship between students' estimated college preferences and the three components provides evidence on whether students' enrollment decisions are primarily associated with the selectivity of the school, the college's quality, or the student-college match.

The results reveal sizable overlap in estimated quality across institutions, although complete equality of college quality is strongly rejected. This finding is robust across outcome measures and control function specifications and is consistent with the findings of Mountjoy and Hickman (2021).[4] Students' college preferences are more consistently associated with peer composition, a measure of selectivity, than with college quality or student-college matching effects. This pattern suggests potential information frictions in the college choice process and raises questions about the extent to which expanding access alone can improve student-college matches and student outcomes.

The rest of the document is organized as follows. Section 1.1 ties the project to previous literature and discusses my contribution compared to several relevant papers. The next section summarizes the data used to answer the research questions. Section 3 describes my theoretical model, followed by section 4 discussing several empirical strategies and detailed estimation procedures. Section 5 shows the estimation results and discusses their implications. The last section concludes and discusses the potential policy impact and future

[3]Hereafter, APSW.
[4]Hereafter, MH21.

extensions of this project.

1.1 Previous Literature

This project is grounded in the literature studying inequality of educational resources and opportunities (Chetty and Hendren 2018, Agostinelli et al. 2024). Although my quantitative methods are inspired by economic theory, the research question is not limited to the field of economics and receives much attention from sociologists and education researchers. For example, Hillman (2016) quantifies the uneven distribution of college institutions by measuring the number, type, and admission selectivity of colleges within 709 commuting zones in the US. I add to this discussion by examining the spatial distribution of the objective quality of college education, highlighting a potentially even more asymmetric distribution of educational opportunities.

My paper connects to three other strands of literature: one on students' college choices, one on the effects of college "quality" on student outcomes, and one measuring the causal returns to attending post-secondary institutions.

A large body of literature studies college choices with heterogeneous college quality (Brewer et al. 1999, Arcidiacono 2005, Dillon and Smith 2017, Fu 2014, Fu et al. 2022). One subset of this school choice literature examines education and migration decisions simultaneously (Fabre 2023, Kennan 2020, Kim 2022, Ishimaru 2025). In these papers, however, returns to higher education only vary across a few different types of institutions, for instance, along the dimensions of public vs. private and elite vs. non-elite. This is partly due to the complexity of some models, but also partly due to data limitations — small sample sizes from survey data, such as NLSY or ELS2002, will not be able to support rich institution-specific college returns. My project complements these papers by examining the wider range of heterogeneous returns to college education with the help of large-scale administrative datasets.

Previous papers estimating the returns to college quality tend to define "quality" using

the selectivity of colleges, such as the median SAT scores of admitted students, or some predefined labels of "elite schools." They arrive at the causal effect of colleges after controlling for selection bias in college attendance, using careful methods to compare only students with similar ability (Dale and Krueger 2002, Dale and Krueger 2014; Anstreicher and Miller 2023, Kirkeboen et al. 2016, Sekhri 2020, Jia and Li 2021). As opposed to these papers, I define college quality using the differences in student outcomes directly, instead of using the ability of admitted students or other predefined labels.

This idea brings me closer to papers measuring the causal impact of institutions on student outcomes. Hoxby (2015, 2018) calculate the earnings value added for a wide range of US colleges using administrative data from the College Board, National Student Clearinghouse, and tax return documents. She compares the outcomes of marginal students in a certain range of test scores where admission might arguably be random. One shortcoming of this work is that she does not have admission data and has to impute admission probabilities and enrollment rates.

Closely related is the paper by MH21. They have detailed data on students' application and admission sets in the Texas public university system. They compare students with the same admission results but enrolled in different colleges, arguing that admission results control for the majority of information on the unobserved ability, ambition, and potential of the student. The authors show that differences in value-added for universities are quite small but statistically different from 0. For earnings after graduation, the causal estimates are only very weakly related to the average outcome without correcting for selection, suggesting that selectivity drives much of the difference in the outcomes across colleges. They also show that the estimates have a modestly positive relationship with non-peer inputs in each university.[5]

I extend their work on a few aspects, exploring a few other related questions. The first is to examine the students' enrollment decisions and whether the differences in college quality matter for their choices. This was not the focus of MH21, yet if we want to evaluate the

[5]This includes instructional expenditures, full-time faculty share, tenure-track faculty share, and faculty-student ratio.

impact of policies that only provide incentives and offer more choices to the students, how individuals choose where to attend college matters for the end result. MH21 also limit their earnings outcome to the average from 8 to 10 years after college entry. I expand the earnings data to gain a more comprehensive view of the impact on earnings from the schools.

The only other similar work to my paper is APSW, who examined high school choices in New York City. They estimate the schools' impact on student outcomes and recover agents' preferences for that impact, assuming that students select into schools only through observable characteristics, distances, and some distribution of their unobserved preferences. Their results showed that high school choices are only weakly influenced by the treatment effects of the schools, but more strongly by the peer composition of the schools. This presented a rather interesting puzzle for the researchers. Do people not care about the real effect of schooling when making the decision? Is it even possible for a policy to improve student outcomes? I explore these questions in the post-secondary setting to see if the story is the same. Does college quality matter for college choices?

2 Data

Identifying causal impacts of individual schools requires a large amount of data that is only available from administrative sources. Survey data, limited by budget and their goal to be nationally representative, cannot cover a group of schools with enough individual observations for each institution. Therefore, this project utilizes and combines administrative datasets from three separate state institutions in Texas: Texas Education Agency (TEA) for K-12 student and school information, Texas Higher Education Coordinating Board (THECB) for state public university (PU) data, and Texas Workforce Commission (TWC) for worker earnings measurements.[6]

Studying Texas also has its advantages. Being the second-largest state in the US in terms of population and area, Texas provides large geographic variations and greater sample sizes,

[6]The data are accessed through the Education Research Center at the University of Houston.

which is much needed for this project. The state has a very low out-migration rate (Aisch et al. 2014). For example, less than 4% of Texas public high school graduates in 2008 and 2009 cohorts attend out-of-state colleges. This highly self-contained population means that I am able to track more individuals over time and gather more information from their choices and outcomes.

2.1 Data Description

College Admission and Enrollment

The central piece of data for my project is the admission results and enrollment decisions of Texas university students. Collected by THECB, this information is only available for 4-year public universities, which enroll over 80% of 4-year college students in Texas.[7] These datasets also collect information on students' gender, ethnicity, residence location, family income group, and parental education levels.

I merge the admission data with fall enrollment reports of the same fiscal year to determine students' initial enrollments. I assign the enrollment from the corresponding fall semester as the "treatment" school of the students. This means that, similar to MH21, I estimate the effect of "initial school enrollment" on students. The impact of transferring or dropping out is all included as possible future paths following the initial enrollment at this school, and thus part of the initial enrollment treatment effect.

TEA Data

The TEA data contain rich individual data on student demographics, family backgrounds, and test scores of high school students. I will mainly use two types of information: the geographic location of high schools and standardized test scores as measures of students' academic abilities.

[7]There are currently 38 individual institutions, most of which belong to the 6 public university systems of Texas: Texas A&M, University of Texas, Texas State, Texas Tech, University of Houston, and University of North Texas.

I treat the students' high school of graduation as their home location when constructing their distances to the universities. This provides the most granular information, since information from other datasets only goes down to the county level. Details of the distance calculations are provided in appendix A.1.

TEA reports statewide high school exit-level standardized test scores. I use the scores from TAAS, whose exit-level test was administered until 2003 to students in grade 10 or above.[8] The test has three subjects—math, reading, and writing—with scaled scores. For math and reading, there is also a 0-to-100 measure, Texas Learning Index, which provides a consistent measure of academic ability across different TAAS tests.

In some cases where I reduce the dimensions of the test score information of the students, reading and writing will be combined into one "verbal" measure using principal component analysis on the scores after standardization.

Net Tuition

FADs data from the THECB contain information on all sources of student financial aid, including grants, scholarships, loans, and work-study programs, as long as these forms of aid require some financial-need analysis. I abstract away from the dynamic labor supply decisions after graduation, since this is not the current focus of my project. Therefore, I chose to exclude work-study programs and student loans when calculating the total amount of financial aid. If the individuals are not liquidity-constrained and make labor supply decisions optimally, these are options that they will choose even without these programs, thus not providing external funds to their education that they do not need to pay back.[9]

The "sticker prices" of the universities are also needed to calculate the net tuition paid by students. These are obtained from the Integrated Postsecondary Education Data System (IPEDS), which reports the total price for a student living on campus, a reasonable baseline for an incoming freshman.

[8] The measure can be expanded to later tests like TAKS and STAAR.

[9] An exception might be subsidized loans that are otherwise unavailable.

To make the individual observations across years comparable, all dollar amounts are deflated, using the CPI-U series for the South, to the price levels in the second half of 2000, the starting point of the first cohort in college.

For each student, aid details are only reported at some schools, most likely the school in which they enrolled.[10] Therefore, some imputation is needed to recover the counterfactual financial aid for alternative schools.

The imputation is carried out in two steps. The first step is to determine the "eligibility" of students for need-based financial aid. Although the financial aid data only report students who received some amount of aid, I can expand my data to students who enrolled in public universities but have no FADs information at all: they must not be receiving any need-based aid at their enrolled school. Using a logistic regression, I predict the student's probability of receiving aid from their observable characteristics. I then estimate the amount of aid received, conditional on receiving any aid, with a linear regression with flexible interactions trained on the FADs data only. The net tuition used in the analysis is the expected net tuition combining both cases. More details of the imputation process can be found in appendix A.2.

Earnings

The outcome variable for this project is the earnings of college graduates. I have quarterly earnings figures from the TWC from the year 2003 to 2024. These are aggregated into annual income and also deflated to the second half of 2000 to make all variables consistent. Instead of using one earnings measurement like MH21, I collect student earnings 6 to 10 years after their initial enrollment, but also take average earnings over the 5-year period as well as over the 8-to-10 year period.

[10] There is a small amount of data on the financial aid of students from non-enrolled schools. These observations have much lower average aid than the enrolled-school observations.

2.2 Descriptive Statistics

For this estimation, I use data from students in the high school classes of 2000 to 2004 who have at least two offers from Texas public universities.[11] This data consists of 101,671 offers from 41,943 students, about 58% of them female, and 36% of them non-White. I use the last year's income reported in the application dataset to define family income groups.[12] Around 25% of the students are from the lowest income group, under $40,000, whereas about 42% have family income over $80,000.

Test Scores

Figure 2 shows the patterns of the math and reading test scores of the students in the sample. Average reading scores are always higher than math scores, with male and white students performing better. Test scores also increase with family income as well.

Admission and Enrollment

Among the students in the sample, 37.9% are admitted to UT Austin and 40.4% to Texas A&M. Only 29.9% have more than two offers. Figure 3 shows the detailed breakdown of the proportion of students admitted to two flagship universities.[13] Male students and those with higher math scores have slightly higher offer rates. Offer rates also increase with family income.

Comparing the schools in each student's admission set with the one in which they eventually enroll, we can see that they tend to choose schools that are slightly closer—by about 5–10%—to their home location. This pattern is fairly consistent, with a few exceptions in the racial minority groups, as shown in Figure 4. Racial majorities and male students, on

[11] I selected students with at least 2 offers to make sure that the students have comparable outside options. More details of the sample selection and merging process are available in appendix A.3.

[12] The dataset reports 5 family income categories, which I further collapsed into 3 groups: $0-40k, $40k-80k, and $80k+. There is also itemized income information in the financial aid data that could be more precise, but it is available for much fewer people.

[13] This is different from the admission probability, since it is also a result of the application behavior of the students.

average, get admitted to schools that are further away. Students who do not enroll in public univeristies are generally further away too.

Patterns regarding net tuition are closely related to family income levels. Low-income students tend to enroll when net tuition is lower and choose schools that are cheaper for them. This trend is reversed as we move up the income ladder.

Female and racial minority students generally enroll with lower net tuition, but the racial gap is more significant. Some of the gender and racial gap is driven by the application behavior and admission results, as similar trends can be found for posted tuition (Figure 6). The amount of financial aid, which is the gap between the posted and net tuition, is highly negatively correlated with family income, as expected (Figure 7).

On average, 24.9% of the students in the sample do not enroll in PUs. Over time, these figures do not show any clear trend (Figure 8). This rate is generally higher for low-income students and racial minorities, as shown in Figure 9.

Student Earnings

Figure 10 shows the real earnings of students 6 to 10 years after their first college entry. White, male, high-income, and top math performers tend to have higher earnings. All average earnings increase over time, but top-performing high-income males seem to have the highest growth rate in earnings.

3 Model

To understand the weight that students place on college effectiveness, my model needs two components — one describing the factors affecting students' outcomes, the other outlining the decision process of high school graduates on college enrollment.

3.1 Student Outcomes

A student, denoted by i, with a vector of observable characteristics X_i, can choose school $s_i = j$ to attend from a set of schools $\mathcal{J} = \{1, \ldots, J\}$. Their potential outcome, Y_{ij}, is

$$Y_{ij} = \alpha_j + \beta_j X_i + \epsilon_{ij} \tag{1}$$

This is a very general way of stating that the student outcomes are influenced by both the college education and the students' characteristics. Note that

$$\mathbb{E}[Y_{ij}|X_i = 0] = \alpha_j + \mathbb{E}[\epsilon_{ij}|X_i = 0] = \alpha_j$$

$X_i = 0$ can be specifically constructed to be meaningful. For all continuous variables, I normalize them to have mean 0, so $X_i = 0$ represents the average student in that variable. On the other hand, $X_i = 0$ for categorical variables naturally stands for the base category of students. Therefore, α_j can be interpreted as the mean potential outcome at school j for an average student in the base category.

The school's effect on individual outcomes also depends on the personal characteristics X_i, because colleges can have different returns β_j to observable characteristics. Some colleges may provide a larger benefit to certain subgroups of the population because of their instructional environment or curricula. This is the "matching effect" between students and schools due to observables. Any other variation in student potential outcomes uncontrolled by X_i is included in the unobserved ϵ_{ij}.

Similar to APSW, I can also rewrite equation 1 by separating out the average potential outcome of student i, denoted as $\bar{Y}_i$. This object does not depend on the school j and can be interpreted as a benchmark of the student's potential or a proxy measure of their inherent ability. Define $\bar{Y}_i = \frac{1}{J}\sum_{s=1}^{J} Y_{is} = \bar{\alpha} + \bar{\beta} X_i + \bar{\epsilon}_i$, with $\bar{\alpha} = (1/J)\sum_j \alpha_j$, $\bar{\beta} = (1/J)\sum_j \beta_j$, and

$\bar{\epsilon}_i = (1/J)\sum_j \epsilon_{ij}$. Then,

$$Y_{ij} = \bar{Y}_i + [\alpha_j - \bar{\alpha}] + \underbrace{[(\beta_j - \bar{\beta})X_i + (\epsilon_{ij} - \bar{\epsilon}_i)]}_{M_{ij}} \tag{2}$$

$$ATE_j(0) = \mathbb{E}[Y_{ij} - \bar{Y}_i | X_i = 0] = [\alpha_j - \bar{\alpha}] + \mathbb{E}[\epsilon_{ij} - \bar{\epsilon}_i] \tag{3}$$

One way to think about the impact of individual schools, in this multiple "treatment" case, is to look at the resulting student outcome benchmarked against this $\bar{Y}_i$. The average of the individual treatment effects of school j is shown in equation 3. When $X_i = 0$, and we have $\mathbb{E}[\epsilon_{ij} - \bar{\epsilon}_i] = 0$, the average treatment effect is $\alpha_j - \bar{\alpha}$. This is the main measurement I use for the education quality of each school.[14]

The last component in equation 2 is the individual matching effect M_{ij} from attending that particular college, arising from either differential returns of observables or some unobservable shocks of student outcome.

In reality, only the post-graduation outcomes from the students' enrolled schools are observed. For any particular school j, the average observed outcome of its enrolled students is also decomposed into three similar terms:

$$\mathbb{E}[Y_{ij}|s_i = j] = \underbrace{\mathbb{E}[\bar{Y}_i|s_i = j]}_{Q_j} + ATE_j + \mathbb{E}[M_{ij}|s_i = j] \tag{4}$$

Q_j is thus the average of the enrolled students' potential $\bar{Y}_i$ at school j, a measure of peer composition and selectivity of the school.[15] ATE_j is the quality measure of school j. $\mathbb{E}[M_{ij}|s_i = j]$ is the average matching effect of enrolled students at school j. We would expect this term to be positive, since students are likely to choose schools that are a good

[14] Continuing forward, I will always refer to this average treatment effect when $X_i = 0$ and will only write ATE_j for simplicity. With categorical variables, normalizing them to have mean 0 makes $X_i = 0$ and thus α_j lose their economic meaning, but will make the average treatment effect truly averaged over all X_i's.

[15] This measure includes the common effect of attending public universities on student outcomes, but I am primarily interested in the variations and student choices among the public universities, instead of the extensive margin effect of attending any public university.

match for them.

3.2 Student Choices

I construct a model of students choices generalized from the Roy model of self-selection. Students from different backgrounds might have different preferences for colleges. I address this heterogeneity by grouping students into several cells c using their observable characteristics. All students have an outside option $U_{i0} = V_{i0} + \eta_{i0}$.

If the student i in cell c chooses $s_i = j \neq 0$, their utility is:

$$U_{ij} = w(Y_{ij}, Q_j) - \tau_c T_{ij} - \delta_c D_{ij} + \nu_{ij} \tag{5}$$

where $w(\cdot)$ is some flexible function of the peer composition Q_j at school j and the student's individual outcome Y_{ij}, representing the value of attending school j. This could either mean that students' value is not just a linear function of their outcome, or it could mean that students do not have perfect information about their potential outcome when making the choices.

There are two sources of cost associated with attending j. One is D_{ij}, the distance between school j and student i's home. Compared to APSW, I also added the net tuition T_{ij} paid by student i at school j, since college tuition is an important determinant of students' choices. ν_{ij} is unobserved preferences only known to the decision-makers.

Rewrite the individual potential outcome in equation 1 as a sum of the average potential outcome of students in cell c, denoted as Y_{cj}, and the individual deviation from that average, denoted as y_{ij}.

$$Y_{ij} = Y_{cj} + y_{ij}, \quad \text{where } Y_{cj} = \frac{\sum_{i \in c} Y_{ij}}{\sum_{i \in c} 1}$$

Assuming $w(Y_{ij}, Q_j) = w_1(Y_{cj}, Q_j) + \omega_{ij}$ for some function w_1 and unobservable individual

shock ω_{ij}, we can rewrite the utility function as:

$$U_{ij} = \overbrace{w_1(Y_{cj}, Q_j) - \tau_c T_{ij} - \delta_c D_{ij}}^{V_{ij}} + \underbrace{\omega_{ij} + \nu_{ij}}_{\eta_{ij}} \tag{6}$$

$$= \mu_{cj} - \tau_c T_{ij} - \delta_c D_{ij} + \eta_{ij} \tag{7}$$

where $\mu_{cj} = w_1(Y_{cj}, Q_j)$ is the average utility of students in cell c for school j, since w_1 only varies at the cell-school level. I will be estimating equation 7 directly, treating $(\mu_{cj}, \tau_c, \delta_c)$ as the parameters to be estimated.

Due to the existence of ω_{ij} in η's, there is a potential correlation between η_{ij}'s for all $j \neq 0$. For any $j \neq s$,

$$\begin{aligned} \text{Cov}(\eta_{ij}, \eta_{is}) &= \text{Cov}[\omega_{ij} + \nu_{ij}, \omega_{is} + \nu_{is}] \\ &= \text{Cov}[\omega_{ij}, \omega_{is}] \end{aligned}$$

$\text{Cov}[\omega_{ij}, \omega_{is}] \neq 0$ for general ω's. To allow for this, the vector of $\{\eta_{ij}\}$ is assumed to follow a Generalized Extreme Value distribution (nested logit), with all $j \neq 0$ in one nest, and the outside option in another. The student chooses from their admission set $A_i = \{0, 1, \ldots, J\}$, including the outside option, to maximize their utility:

$$s_i = \arg\max_{s \in A_i} U_{is}$$

$$s_i = j \iff U_{ij} \geq \max_s U_{is}$$

The resulting choice probabilities have the following closed-form expressions:

$$\mathbb{P}_{i,j\neq 0} \equiv \mathbb{P}_i(s_i = j) = \frac{\exp(\iota_c IV_i)}{\exp V_{i0} + \exp(\iota_c IV_i)} \cdot \frac{\exp\left(V_{ij}/\iota_c\right)}{\sum_{s\neq 0} \exp\left(V_{is}/\iota_c\right)} \tag{8}$$

$$\mathbb{P}_{i0} \equiv \mathbb{P}_i(s_i = 0) = \frac{\exp V_{i0}}{\exp V_{i0} + \exp(\iota_c IV_i)} \tag{9}$$

$$\text{where } IV_i = \ln\left(\sum_{s\neq 0} \exp\left(V_{is}/\iota_c\right)\right)$$

$\iota_c \in (0, 1]$ is the nesting parameter for cell c, which measures the strength of correlation of η between schools. The closer ι_c is to 0, the stronger the correlation of η's between schools. When $\iota_c = 1$, the model reduces to a standard logit model with independent η's.

4 Empirical Strategy

The key issue for the estimation is the existence of selection bias in the student outcome regression, equation 1. Researchers only observe Y_i, the potential outcome for chosen school s_i. Denote $D_i = \{D_{is}\}$ the vector of distances, $T_i = \{T_{is}\}$ the vector of net tuition, and $V_i = \{V_{is}\}$ the vector of conditional value functions for individual i. The feasible version of the outcome regression with OLS, given V_i, is:

$$\mathbb{E}[Y_{ij}|s_i = j, X_i, V_i] = \alpha_j + \beta_j X_i + \underbrace{\mathbb{E}[\epsilon_{ij}|s_i = j, X_i, V_i]}_{\equiv \lambda_j(X_i, V_i)} \tag{10}$$

One approach to estimate this equation is to maintain the "selection on observables" assumption, which assumes that the unobserved components of student outcomes ϵ_{ij} are independent of the school choice conditional on a rich set of observables X_i. This would mean that in the choice model above, ω_{ij} does not contain any information about ϵ_{ij}. Under this strong assumption, $\lambda_j(X_i, V_i) = 0$, leading to consistent OLS estimates of α_j and β_j in equation 1.

More realistically, there might be selection on unobservables, so $\lambda_j(X_i, V_i) \neq 0$. This is

because the decision s_i is affected by η_{ij}, which contains ω_{ij}, perhaps some private information about the student's future prospects, and correlated with unobserved components ϵ_{ij} in student outcome. If the student knows that they will get a higher ϵ_{ij} at school j, they are more likely to choose that school, and therefore ϵ_{ij} are usually "selected" to be higher at the attended school. In this case, the conditional mean independence of OLS fails. Elements in the selection bias $\lambda(\cdot)$ will be captured by $\widehat{\alpha}_j$ and $\widehat{\beta}_j$, leading to inconsistent estimates. This selection bias needs to be accounted for in the estimation.

4.1 Control Function Approach

There are a few ways to address the selection issue. My baseline method takes the parametric control function approach (Lee 1983). This method is a generalization of the Heckman selection models, specifying a particular joint distribution of $(\epsilon_{ij}, \{\eta_{is}\}_s)$. In my setting, $s_i = j$ when

$$V_{ij} + \eta_{ij} \geq \max_s \{V_{is} + \eta_{is}\} \iff V_{ij} \geq \underbrace{\max_s \{V_{is} + \eta_{is}\} - \eta_{ij}}_{\equiv \eta^*_{ij}}$$

The vector V_i's is known when the choice model is estimated. Under some distribution of η_{is}'s, we can derive the distribution of $\eta^*_{ij} \sim G(\eta^*)$ for some G. Lee (1983) showed the results under independent Gumbel distributions of η's, but this can be easily generalized to the nested logit GEV case.

The distribution $G(\cdot)$ can be transformed into a standard normal distribution through probability integral transformation. Now define $z^\eta_{ij} = \Phi^{-1}(G(\eta^*_{ij})) \sim \mathcal{N}(0,1)$, and assume $(z^\eta_{ij}, \epsilon_{ij}) \sim \mathcal{N}\left((0,0), \begin{pmatrix} 1 & \rho\sigma_\epsilon \\ \rho\sigma_\epsilon & \sigma^2_\epsilon \end{pmatrix}\right)$. Under the joint normality assumption, $\epsilon_{ij} = \rho\sigma_\epsilon z^\eta_{ij} +$

w_{ij} for some independent normal white noise w_{ij}. Therefore, the selection bias term $\lambda(\cdot)$ is

$$\begin{aligned}
\mathbb{E}[\epsilon_{ij}|s_i = j, X_i, V_i] &= \mathbb{E}[\epsilon_{ij}|\eta^*_{ij} \leq V_{ij}, X_i, V_i] \\
&= \mathbb{E}[\rho\sigma_\epsilon z^\eta_{ij}|\eta^*_{ij} \leq V_{ij}, X_i, V_i] + \mathbb{E}[w_{ij}|\eta^*_{ij} \leq V_{ij}, X_i, V_i] \\
&= \rho\sigma_\epsilon \cdot \mathbb{E}[z^\eta_{ij}|z^\eta_{ij} \leq \Phi^{-1}(G(V_{ij})), X_i, V_i] \\
&= -\rho\sigma_\epsilon \cdot \frac{\phi(\Phi^{-1}(G(V_{ij})))}{G(V_{ij})} \qquad (11)
\end{aligned}$$

Note that $G(V_{ij}) = \mathbb{P}(\eta^*_{ij} \leq V_{ij}) = \mathbb{P}(s_i = j) = \mathbb{P}_{ij}$, so the selection bias is only a function of the choice probability of the chosen alternative. ρ and σ_ϵ will not be separately identified.

Lee (1983)'s method relies on the specific distributional assumptions of the vectors ϵ_{ij} and η_{ij}, which may or may not be true. Dahl (2002) relaxes the Gaussian copula assumption to allow for selection correction under more general cases. Under some formulations of the choice model and selection process, the selection bias can be reduced to a single-index problem — $\lambda_j(V_i)$ is just a function of the probability of the chosen option $\mathbb{P}_{ij}$.[16] As we just derived above, Lee (1983) is a special case of that. Under the generalization, λ is approximated by the polynomial expansions of $\mathbb{P}_{ij}$.

This method can also be extended to multiple-index problems. For the general selection bias function $\lambda_j(V_i)$, other variables in the choice value vector V_i, such as the distance and net tuition can also affect the conditional mean. As long as the number of dimensions is controlled, the selection bias can be approximated by multivariate polynomial expansions of those indices. APSW can be viewed as one particular implementation of this, with some additional functional form assumptions.[17] In their application, they assumed the selection

[16] This is the main assumption in Dahl (2002), called "index sufficiency:" the selection bias is only a function of a limited amount of indices of the choice problem. A sufficient condition for the assumption is that the bivariate distributions of the unobserved ability ϵ_{ij} and the unobserved utility η_{ij} are the same as the joint distribution of ϵ_{ij} and the indices, such as the η^* in Lee (1983). One particular example Dahl (2002) gave was a Roy model selection only on potential outcomes, which is a sum of individual fixed ability and independent random shocks, with no correlated unobserved tastes in choice. There can be other models that also satisfy this condition.

[17] In the paper, the authors seem to indicate this is the exact functional form of $\lambda(\cdot)$, derived not from a choice model but only from a structural assumption on the conditional mean.

bias is a function of distances D_i and unobserved utility shocks $\{\eta_{is}\}$:

$$\mathbb{E}[\epsilon_{ij}|s_i = j, X_i, D_i, \{\eta_{ij}\}] = \sum_{s=1}^{J} \left[\kappa_s^d D_{is} + \psi_s(\eta_{is} - \gamma)\right] + \phi(\eta_{ij} - \gamma)$$

where γ is the Euler constant, mean of T1EV(0,1) distribution. Therefore,

$$\lambda_j(X_i, D_i) = \sum_{s=1}^{J} \left[\kappa_s^d D_{is} + \psi_s \mathbb{E}[\eta_{is} - \gamma|s_i = j, D_i]\right] + \phi\mathbb{E}[\eta_{ij} - \gamma|s_i = j, D_i]$$

I can expand the APSW specification with the addition of net tuition T_{ij} in the control function and the polynomial expansions of (D, T) for more flexibility.

In general, these control function approaches rely on the exclusion restriction on distance and net tuition. Conditional on the included observable characteristics X_i, (D, T) can only affect the student's observed outcome through the school choice. They cannot be connected to other unobserved factors that also impact student potential outcomes. In section 4.3, I will go into more details about the empirical specifications of the model and what is included in X_i.

4.2 Matched Applicant Approach

The control function methods above rely on some assumptions on the choice process. The correlation between unobserved choice shock η and unobserved ability ϵ, explicitly modeled or implicitly assumed, help us derive a function of $\mathbb{E}[\epsilon_{ij}|s_i = j, V_i]$ using information from the value functions V_i in the choice model.

An alternative method is using the matched applicant approach by MH21. This is a different style of approach to address selection bias. MH21's method relies on the assumption that students with the same admission sets A_i have similar unobserved ability, regardless of

their school choice. If $\epsilon_{ij} = \epsilon_i + e_{ij}$, then

$$\mathbb{E}[\epsilon_{ij}|s_i = j, A_i] = \mathbb{E}[\epsilon_{ij}|s_i \neq j, A_i] = \mathbb{E}[\epsilon_{ij}|A_i] \tag{12}$$

Consider a simple case where $X_i = 0$ (so we ignore X's for now), and two schools j and j' are in the admission set A_i. When controlling for the admission set, we have

$$\mathbb{E}[Y_{ij}|s_i = j, A_i] = \alpha_j + \mathbb{E}[\epsilon_{ij}|s_i = j, A_i] \tag{13}$$

$$\mathbb{E}[Y_{ij'}|s_i = j', A_i] = \alpha_{j'} + \mathbb{E}[\epsilon_{ij'}|s_i = j', A_i] \tag{14}$$

$$\mathbb{E}[Y_{ij}|s_i = j, A_i] - \mathbb{E}[Y_{ij'}|s_i = j', A_i] = \alpha_j - \alpha_{j'} + \mathbb{E}[\epsilon_{ij} - \epsilon_{ij'}|A_i] \tag{15}$$

Thus, the difference in mean outcomes identifies the difference in school effects if the conditional mean of the unobserved component is also equal across the two schools:

$$\mathbb{E}[\epsilon_{ij}|A_i] = \mathbb{E}[\epsilon_{ij'}|A_i]$$

with some additional assumptions on sorting.[18]

Thus, controlling for admission sets allows for comparison and computing "relative value-added" of schools. In application, adding controls for admission sets makes it tricky to represent the intercepts α_j in the outcome regression – a baseline admission set needs to be chosen if I wish to keep all α_j without the "relative" interpretation. Therefore, I will only use the matched applicant approach to compare with the control function approach when using a baseline school in estimation.

[18] The exact MH21 formulation is slightly different than mine. Admission sets controls for vertical selection – the selection from unobserved student ability that is invariant across schools j. In their formulation, the variation across j in ϵ_{ij} will be included as part of the school value-added. Under the strictest homogeneous value-added assumption, that variation will go away.

4.3 Empirical Specifications

Students' earnings outcome is the main focus of this project. Y_i is their log annual earnings 6 to 10 years from initial college enrollment. I focus on students who are enrolled in a public university, so they can be assigned to a particular "treatment" school j.

For the X variables in the outcome equation, (1), I include a set of student characteristics, including gender, racial minority status, 3 levels of family income, high school GPA (whether they are top 10% of their high school class), economic disadvantaged status (receiving reduced-price or free lunch), TAAS test scores (reading, writing, and math), and standardized measure of population density in their home county to control for urban/rural differences. The test scores are normalized to have mean 0 and standard deviation 1, and a white middle-income male student, not in the top 10%, not economically disadvantaged, is chosen as the base category. Cohort fixed effects are included to control for macroeconomic shocks and other time-varying factors. After controlling for these observable characteristics, distance D_{ij} and net tuition T_{ij} are not correlated with unobserved factors of student potential earnings and only affect the observed student outcome through their enrollment choices.

For the choice model, the outside option value V_{i0} is assumed to be dependent on the admission results of the students. It is likely that students with "better" offers, particularly those with offers from the 2 state flagship universities (UT Austin and TAMU), have better outside alternatives. Let

$$V_{i0} = \mu_c^+ \cdot \mathbb{1}(\text{Admitted to UT Austin or TAMU}) \tag{16}$$

μ_c^+ varies across student groups c, so it accounts for different values of outside options for different types of students.

4.4 Detailed Procedure

The main estimation process is carried out in three steps, as shown in Figure 11. The first step is to estimate the utility parameters $(\mu_{cj}, \delta_c, \tau_c)$ in equation 7, as well as the correlation parameter ι_c and the outside option value V_{i0}. I use data on student school choices (admission and enrollment), net tuition T_{ij}, and distance matrix D_{ij} to search for the parameters with maximum likelihood estimation, matching the students' choices with their theoretical choice probabilities shown in equations 8 and 9.

After estimating the choice model, I can construct the conditional value functions, $\widehat{V}_{ij} = \widehat{\mu}_{cj} - \widehat{\delta}_c D_{ij} - \widehat{\tau}_c T_{ij}$. The conditional value functions are then used to construct choice probabilities used in the control functions.

The second step is to estimate the outcome equation, addressing the selection bias with control function methods. The control function specifications are as follows:

$$Y_{ij} = \alpha_j + \beta_j X_i + \psi^l \cdot \frac{\phi(\Phi^{-1}(\mathbb{P}_{ij}))}{\mathbb{P}_{ij}} + \zeta_{1i} \tag{Lee}$$

$$Y_{ij} = \alpha_j + \beta_j X_i + \psi^l \cdot \frac{\phi(\Phi^{-1}(\mathbb{P}_{ij}))}{\mathbb{P}_{ij}} + \sum_{s=1}^{J} \left[p^3(D_{is}) + p^3(T_{is})\right] + \zeta_{2i} \tag{Lee+DT}$$

$$Y_{ij} = \alpha_j + \beta_j X_i + \sum_s \left[p^3(D_{is}) + p^3(T_{is}) + \psi_s \mathbb{E}[\eta_{is} - \gamma | s_i = j, V_i]\right] + \psi^0 \mathbb{E}[\eta_{ij} - \gamma | s_i = j, V_i] + \zeta_{3i} \tag{APSW}$$

$$Y_{ij} = \alpha_j + \beta_j X_i + p^5(\mathbb{P}_{ij}) + \zeta_{4i} \tag{Dahl}$$

where $p^n(\cdot)$ is a n^{th} degree polynomial expansion of the variables.

The first specification is the baseline Lee (1983) method. I then expand this into "Lee + DT," by adding polynomial functions of distance and net tuition. Another way to generalize "Lee" is to follow Dahl (2002) and replace the Lee control function with a flexible polynomial expansion of choice probability. I also follow the APSW idea, but added third degree polynomials of distance and net tuition as control functions.[19]

[19] Control function variables are only included if the school is inside the student's admission set.

From this step I obtain estimates $\widehat{\alpha}_j$ and $\widehat{\beta}_j$. This allows us to decompose the student outcomes, following the definitions in section 3.1. Firstly, the $\widehat{ATE}_j$ is obtained by simply demeaning $\widehat{\alpha}_j$. Then,

$$\widehat{\bar{Y}}_i = \bar{\widehat{\alpha}} + \bar{\widehat{\beta}} X_i \tag{17}$$

$$\widehat{Q}_j = \mathbb{E}[\bar{Y}_i | s_i = j] = \bar{\widehat{\alpha}} + \bar{\widehat{\beta}} \mathbb{E}[X_i | s_i = j] \tag{18}$$

The student composition in each school $\widehat{Q}_j$ is the average ability of students in that school given the observed choice. Finally, I calculate the average matching effect of observable characteristics for each cell:

$$\widehat{M}_{cj} = \frac{1}{|c|} \sum_{i \in c} (\widehat{\beta}_j - \bar{\widehat{\beta}}) X_i \tag{19}$$

where $|c|$ is the number of students in cell c.

Once we have the estimates above, I link the average "value" of schools μ_{cj} with three different objects related to the student outcome:

$$\widehat{\mu}_{cj} = \rho_{0c} + \rho_1 \widehat{Q}_j + \rho_2 \widehat{ATE}_j + \rho_3 \widehat{M}_{cj} + \xi_{cj} \tag{20}$$

The regression is weighted by the number of students in each cell c. The three components of the student outcome — peer composition $\widehat{Q}_j$, the treatment effect $\widehat{ATE}_j$, and the average observed matching effect $\widehat{M}_{cj}$ — are also standardized to have unit variance, so the resulting estimates $\{\widehat{\rho}_1, \widehat{\rho}_2, \widehat{\rho}_3\}$ are directly comparable. These tell us how strongly each factor explains the preferences of the agents and which factor plays a more important role in their enrollment decisions. I compare ρ_2 and ρ_3 against ρ_1, to see whether students put more weight on the quality of colleges, their match with the college, or just the selectivity of colleges.

5 Results

In this section, I present the estimation results from the baseline specification. Population cells are defined using gender, race, family income, and test score status – whether the student has top-quartile test scores in math, reading, and writing tests.[20] All standard errors are bootstrapped across the three-step estimation procedure, with 100 replications.

5.1 Utility Function Estimates

I obtain utility function parameters $(\tau_c, \delta_c, \iota_c, \mu_c^+)$ for each cell c. Full estimates of the utility function parameters are reported in Table B1. Consistent with the patterns in Figure 5, the tuition sensitivity τ_c decreases significantly with family income. Gender differences are minimal, whereas racial minorities and students with higher test scores are less sensitive to tuition. The distance sensitivity δ_c also decreases with family income and test performance. The estimates of the outside-option value, $\widehat{\mu}_c^+$, vary across demographic and test-score groups, but are mostly insignificant and close to zero. They do not exhibit a clear monotonic relationship with family income, gender, or race.

To gain a better sense of the quantitative relationship between students' choices and distance and tuition, let us look at a specific case of students choosing between two flagship universities, UT Austin and TAMU. This is the most commonly observed admission set in my data — over 3000 students are in this scenario. For each cell c, I compute the predicted choice probabilities of those two schools at the average distances and net tuition levels.[21] I then look at the percentage change[22] in enrollment probabilities when the distance or expected net tuition of TAMU is increased by 1%, while holding all other factors the same. The choice elasticities are reported in Table 1.

Elasticity estimates confirm the pattern in the parameter estimates. Low-income students

[20] With reading and writing scores combined into one "verbal" skills index using principal component analysis.

[21] Assuming both schools have the same distance and net tuition, which is the average level of the mean distance and net tuition within the admission set for the students in the corresponding cell.

[22] Not percentage points.

are substantially more responsive to a 1% change in net tuition at TAMU, decreasing their enrollment probability by 4.4% to 10.4%. The tuition elasticity drops to between -2.8 and -1.2 for middle-income students and becomes close to zero for most high-income students.

Distance elasticities are smaller than tuition elasticities but also exhibit a generally decreasing pattern with family income. A 1% increase in distance to TAMU decreases the enrollment probability by 0.06% to 1.8%, 0.1% to 0.4% for middle-income students, and from 0.0% to 0.5% for high-income students.

The estimates in Table B1 imply large and heterogeneous distance–tuition tradeoffs. Tuition is measured in units of $10,000 and distance in units of 1,000 kilometers, so a $100 increase in tuition (about 1% of the average net tuition within the admission set for all students) is equivalent to $10\tau_c/\delta_c$ additional kilometers of distance. Using the corresponding estimates, this tradeoff is about 103 kilometers for low-income students, 29 kilometers for middle-income students, and 9 kilometers for high-income students. These comparisons reinforce that tuition is an important determinant of enrollment decisions, especially for low-income students.[23]

5.1.1 School Values

The estimated school values μ_{cj} in the choice equation 6 also show substantial heterogeneity across institutions and student groups. Table B2 reports the school-level mean, median, minimum, maximum, and standard deviation of the μ_{cj}. UH Downtown has the highest mean value, followed by PVAMU, UT Dallas, and UT San Antonio. UT Austin and TAMU remain toward the upper part of the distribution. Negative values indicate that schools are unattractive relative to the estimated outside option for at least some student groups.

With the same distance and tuition, different levels of μ_{cj} across schools give different

[23]The tuition sensitivity for low-income students seems to be very large. Fu et al. (2022) estimate that a ten-fold increase in distances is equivalent to a $6000 increase in tuition for low-income (below $35,000) and middle-income students ($ 35,000 - 100,000).

enrollment probabilities. Within the public university system,

$$\ln \frac{\mathbb{P}_{ij}}{\mathbb{P}_{ij'}} = \frac{V_{ij} - V_{ij'}}{\iota_c} = \frac{\mu_{cj} - \mu_{cj'}}{\iota_c} \tag{21}$$

Consider the low-income non-white female students with top-quartile math scores, for which Table B1 reports $\widehat{\iota}_c = 0.470$. Holding distance, tuition, and the choice set fixed, if one school has a 0.1 higher value than another, its conditional enrollment odds are multiplied by $\exp(0.1/0.470) \approx 1.24$, a 24% increase. A 0.5 value difference would imply odds about 2.90 times as large. These odds ratios only apply to within-cell comparisons.

5.1.2 Choice Model Fit

How well does the estimated model describe the observed choice behavior? I compute the conditional choice probabilities for each student in the sample and compare them to their enrolled school. I look at two measures: 1) the average predicted probability for the student's observed choice, both overall and separately for students who do or do not enroll, and 2) the proportion of students whose choice is the one with the maximum predicted probability.

Overall, the model fits the data reasonably well. About 69.3% of the students have their chosen option predicted with the highest probability. The number is 67.2% for enrolled students and 75.3% for non-enrolled students. The average predicted probability of the chosen option is 0.334 for enrolled students and 0.500 for non-enrolled students.

Figure 12 shows the proportion of correct predictions for each cell. Only two cells have a prediction rate lower than 50%. The model struggles a little for some smaller cells due to less data, but for larger cells the model captures choice patterns well.

5.2 Earnings Regression

Figure 13 shows the estimated α_j for different schools in the outcome regression under a few different specifications and using the average annual earnings 6 to 10 years after initial

college enrollment. Here, α_j can be interpreted as the average log earnings for an average-test-score white middle-income male student who is not economically disadvantaged and not in the top 10% of their high school graduating class.[24] For control function specifications, 25455 enrolled students are included in the regression.

The point estimates vary across schools, but the confidence intervals overlap substantially for many schools. A few schools stand out: TAMU, UT Dallas, Tarleton State University, Angelo State University, and Midwestern State Universitiy are toward the upper end of the distribution. Estimates from different specifications are fairly close to one another, with control function estimates generally lower than those from OLS. However, the overlapping confidence intervals do not indicate strong positive selection on unobservables, if any. The standard deviation of $\widehat{\alpha}_j$ is 0.13 to 0.15 log points, depending on the specification. The simple summary measure used by MH21 to describe the dispersion of school-level effects, the variance adjusted for sampling errors,

$$\widehat{\sigma}_{\alpha}^2 = \text{Var}_j(\hat{\alpha}_j) - \frac{1}{J}\sum_{j=1}^{J} SE_{\alpha_j}^2$$

is estimated to be negative or 0 for almost all specifications but Dahl (2002), implying that the dispersion of value-added across colleges is small relative to the sampling noise.

For every specification, I tested if α_j are jointly equal to each other with a Wald χ^2 test. The null hypothesis is rejected at the 1% level universally, suggesting that there are indeed differences in earnings value added across schools.[25] Combined with the figure and the results from the simple dispersion measure, this seems to sugggest that only a handful of schools can distinguish themselves from many others than have similar qualities. This is broadly consistent with the findings in MH21, where the authors found that the differences

[24] Earnings are measured in 2000H2 dollars, so 10 on the graph is $\exp(10) \approx \$22,026 \approx \$40,125$ 2024 USD.

[25] I test if the 20 schools in regression have the same α_j, with the chi-square distribution with 19 degrees of freedom. Using the current average earnings measurement, test statistics range from 50.45 with APSW model to 90.71 with OLS, leading to p values at magnitudes equal or lower than 10^{-4}. This is also conducted for all other outcome measures. The only failed rejection at the 1% level is the OLS model with earnings 8 years after enrollment. The test statistic is 35.04 and the p value is 0.014.

in value added across schools are "relatively tight, though non-degenerate."

To have a better comparison, I also run the same regressions but benchmark earnings against UT Austin. The results are shown in Figure **??**. I replicated MH21's matched applicant method as well, adding admission set as controls. The estimates are mostly insignificant, with the exception of TAMU and Angelo State University, showing about 20% and 30% higher earnings than UT Austin, respectively. This pattern is consistent with the main results. Estimates from the matched applicant approach follow similar patterns with those from other models.

5.2.1 Returns to Student Characteristics

α_j only shows part of the picture of the college returns. There is also the student-college matching effect – students have different returns to their own characteristics X_i at different schools, captured by the β_j coefficients in equation 1. A school might not be the best for the average student in general, but it might be greatly beneficial for a particular group of students. Figures 15–24 show the estimated β_j for different student characteristics and test scores.

The dispersion of estimated school-specific returns remains broadly similar in shape, with substantial overlap across institutions and many estimates that are insignificant. This pattern could be consistent with the main result that dispersion of college quality is nontrivial but modest. The estimates between different specifications, in particular those from control functions compared to those of OLS, do not show much difference. This could suggest that the selection bias is not systematically related to some of the observable characteristics.

Specifically, relative returns of female students (to their male counterparts) are generally less than 0, reflecting the gender gap in earnings. Some schools might be relatively better for female students, closing the gap, such as UT Austin and University of Houston. Returns of high-income students relative to middle-income students are mostly positive but not significant, except for UT Austin and Texas Tech. For low-income students, Sam Houston State

University seems to be better for them—all else being equal, low-income students have more than 20% higher earnings than middle-income students. The estimates are not consistently significant, however.

The patterns for the racial gap in earnings and reading scores are less conclusive. UT Dallas is better for students with higher writing test scores, but the two flagship universities, UT Austin and TAMU, are more valuable for students with better math skills—possibly highlighting differences in the focus of their curricula.

Top 10% GPA students benefit more, about 10-20% in earnings, from UT Austin, TAMU, Texas State, and Texas Tech. Economically disadvantaged students benefit more from UT Austin.

Effects for student hometown population density are mostly insignificant, but students from more urban areas seem to benefit slightly from UT Austin and TAMU, and much more from Tarleton State – about 4% higher earnings.

5.3 Utility Weights

The final step of the analysis is to understand the relation between students' value and the peer composition, quality, and match of the school. Figure 25 shows the estimated utility weights (ρ_1, ρ_2, ρ_3) for those three components of student outcomes in equation 20, when the measures were constructed using different models in the outcome regression. The coefficients on the OLS estimates do not appear to be much different from those of the control function estimates. This naturally follows from the fact that the α_j and β_j estimates are not substantially different between OLS and control function methods.

ρ_1 is consistently larger than 0, ranging from 0.15 to 0.2, indicating that students care about the peer composition of their chosen schools, even though Q_j does not directly impact their future earnings. One standard deviation increase in the average enrolled student ability of a school increases the utility of students by 0.15 to 0.2, which translates into an increase in conditional enrollment odds of at least $\exp(0.15/0.470) - 1 \approx 0.38 = 38\%$ for low-income non-

white female students with top-quartile math scores, using our calculation in section 5.1.1. ρ_1/τ_c tells us that students' willingness to pay for a one standard deviation increase in peer composition is \$200 to \$700 tuition for low-income students.

The estimates of ρ_2 and ρ_3 are negative and not significant, suggesting value-added and matching effects are less related to students' college choice decisions. This finding is robust across different specifications and also consistent with the conclusions of APSW in the high school choice context.

This result, however, does not necessarily mean that students' choices are completely unrelated to the quality of schools. Some weak positive correlation exists between school quality and student composition, ranging from 0.20 to 0.35 depending on the regression specification, as shown in Figure 26.[26] This is lower than the APSW numbers of around 0.6. The standard deviation of peer composition across schools is around 0.03, much smaller than the variation in school quality. An imperfect relationship between student choices and school quality could be due to a number of reasons, most likely information frictions, since it is difficult for students to gather enough information to fully understand the causal effects of schools.

6 Conclusion

This paper develops a framework for studying both the causal effects of colleges on student earnings and the role of those effects in enrollment decisions. Using Texas administrative data and control-function methods to address the endogeneity of college attendance, I find meaningful but relatively modest differences in earnings value added across public universities. The estimates vary across institutions, but their confidence intervals overlap substantially.

The choice estimates show that students respond strongly to tuition and peer composition when selecting among colleges. Peer composition is the most consistent of the three outcome-

[26] The same results using other outcome variables can be seen in Figure B13.

related factors in explaining enrollment choices, whereas college quality and student-school matching effects do not seem to matter.

These findings have several policy implications. Because students respond strongly to tuition, financial aid policies that reduce the cost of attendance or improve access to nearby institutions may have larger effects on enrollment than policies that simply provide information about average earnings outcomes. This requires carefully designed interventions to move students toward higher-value-added colleges, but also target students with better match with the school as well. Nevertheless, even if it is difficult to distinguish the role of information frictions and individual preferences with the currently available data, the modest and indirect relationship between school quality and student choices suggests that information interventions might still improve choices.

Policies that expand access should also account for peer composition: changes in enrollment may affect the characteristics of a college's student body, which students value directly, even when peer composition is not itself a source of value added in the earnings regressions.

This project leads to future work on several fronts. Is there a way to break down the treatment effect from colleges? How much of it is due to better education and possibly higher skill levels, as opposed to better social networks and thus benefits when searching for jobs in the labor market? To distinguish these two factors, we would need detailed individual data on those pieces of information. Another dimension to explore is the field of study of the students. This aspect is already explored by some papers, such as Kirkeboen et al. (2016). The treatment effects could potentially vary significantly by college majors.

Another avenue to advance this project is to incorporate job choices after graduation into the choice model. This might allow us to decompose the value added of schools into local labor market benefits and mobility benefits. The challenge is to modify this framework to allow for multi-period choices because the income of graduates is now not simply derived from school choices, but also a result of the work location choice.

References

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters (2020). "Do Parents Value School Effectiveness?" *American Economic Review* 110.5, pp. 1502–1539. ISSN: 0002-8282. DOI: 10.1257/aer.20172040. URL: https://pubs.aeaweb.org/doi/10.1257/aer.20172040 (visited on 04/08/2025).

Agostinelli, Francesco, Margaux Luflade, and Paolo Martellini (2024). "On the Spatial Determinants of Educational Access". Working Paper Series 32246. DOI: 10.3386/w32246. URL: http://www.nber.org/papers/w32246.

Aisch, Gregor, Robert Gebeloff, and Kevin Quealy (2014). "Where We Came From and Where We Went, State by State". *The New York Times*. URL: https://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html#Texas.

Anstreicher, Garrett and Lois Miller (2023). "Who Scars the Easiest? College Quality and the Effects of Graduating into a Recession". *EdWorkingPaper* 23–761.

Arcidiacono, Peter (2005). "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings?" *Econometrica* 73.5, pp. 1477–1524.

Brewer, Dominic J., Eric R. Eide, and Ronald G. Ehrenberg (1999). "Does It Pay to Attend an Elite Private College? Cross-cohort Evidence on the Effects of College Type on Earnings". *The Journal of Human Resources* 34.1, pp. 104–123.

Chetty, Raj, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan (2020). "Income Segregation and Intergenerational Mobility across Colleges in the United States". *The Quarterly Journal of Economics* 135.3, pp. 1567–1633. DOI: 10.1093/qje/qjaa005.

Chetty, Raj and Nathaniel Hendren (2018). "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects". *The Quarterly Journal of Economics* 133.3, pp. 1107–1162.

Dahl, Gordon (2002). "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets". *Econometrica* 70, pp. 2367–2420. DOI: 10.1111/j.1468-0262.2002.00443.x.

Dale, Stacy Berg and Alan B. Krueger (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables". *The Quarterly Journal of Economics* 117.4, pp. 1491–1527.

— (2014). "Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data". *Journal of Human Resources* 49.2, pp. 323–358.

Dillon, Eleanor Wiske and Jeffrey Andrew Smith (2017). "Determinants of the Match between Student Ability and College Quality". *Journal of Labor Economics* 35.1, pp. 45–66.

Fabre, Anaïs (2023). "Education and Migration Choices: Evidence from France". URL: https://drive.google.com/file/d/1Y1rDnk5ptVtynmGGmo9iYC0S1comYJgc/view.

Fu, Chao (2014). "Equilibrium Tuition, Applications, Admissions, and Enrollment in the College Market". *Journal of Political Economy* 122.2, pp. 225–281.

Fu, Chao, Junjie Guo, Adam J. Smith, and Alan Sorensen (2022). "Students' Heterogeneous Preferences and the Uneven Spatial Distribution of Colleges". *Journal of Monetary Economics* 129, pp. 49–64.

Hillman, Nicholas W. (2016). "Geography of College Opportunity: The Case of Education Deserts". *American Educational Research Journal* 53.4, pp. 987–1021.

Hoxby, Caroline (2015). *Estimating the Value-Added of U.S. Postsecondary Institutions*. Working Paper. Internal Revenue Service Statistics of Income Division.

— (2018). "The Productivity of U.S. Postsecondary Institutions". *Productivity in Higher Education*. Ed. by Caroline Hoxby and K. Stange. University of Chicago Press, pp. 31–66.

Ishimaru, Shoya (2025). "Geographic Mobility of Youth and Spatial Gaps in Local College and Labor Market Opportunities". *Journal of Labor Economics* 43.4, pp. 1251–1294. DOI: 10.1086/730542. eprint: https://www.journals.uchicago.edu/doi/pdf/10.1086/730542. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/730542.

Jia, Ruixue and Hongbin Li (2021). "Just above the Exam Cutoff Score: Elite College Admission and Wages in China". *Journal of Public Economics* 196, p. 104371.

Kapor, Adam (2025). "Transparency and Percent Plans". *Econometrica* 93.6, pp. 2123–2157. DOI: https://doi.org/10.3982/ECTA18385. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18385. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18385.

Kennan, John (2020). *Spatial Variation in Higher Education Financing and the Supply of College Graduates*. Working Paper 21065. NBER.

Kim, Jiwon (2022). "The Effect of State Tuition Subsidies for Public Colleges on Student Migration and State Income Tax Revenue".

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad (2016). "Field of Study, Earnings, and Self-Selection". *The Quarterly Journal of Economics* 131.3, pp. 1057–1111.

Lee, Lung-Fei (1983). "Generalized Econometric Models with Selectivity". *Econometrica* 51.2, pp. 507–512. DOI: 10.2307/1912003.

Mountjoy, Jack (2026). "Marginal Returns to Public Universities". *The Quarterly Journal of Economics* 141.1, pp. 429–497. ISSN: 0033-5533. DOI: 10.1093/qje/qjaf055. eprint:

https://academic.oup.com/qje/article-pdf/141/1/429/65830274/qjaf055.pdf. URL: https://doi.org/10.1093/qje/qjaf055.

Mountjoy, Jack and Brent R. Hickman (2021). *The Returns to College(s): Relative Value-Added and Match Effects in Higher Education*. Working Paper 29276. NBER.

Sekhri, Sheetal (2020). "Prestige Matters: Wage Premium and Value Addition in Elite Colleges". *American Economic Journal: Applied Economics* 12.3, pp. 207–225.

Tables and Figures

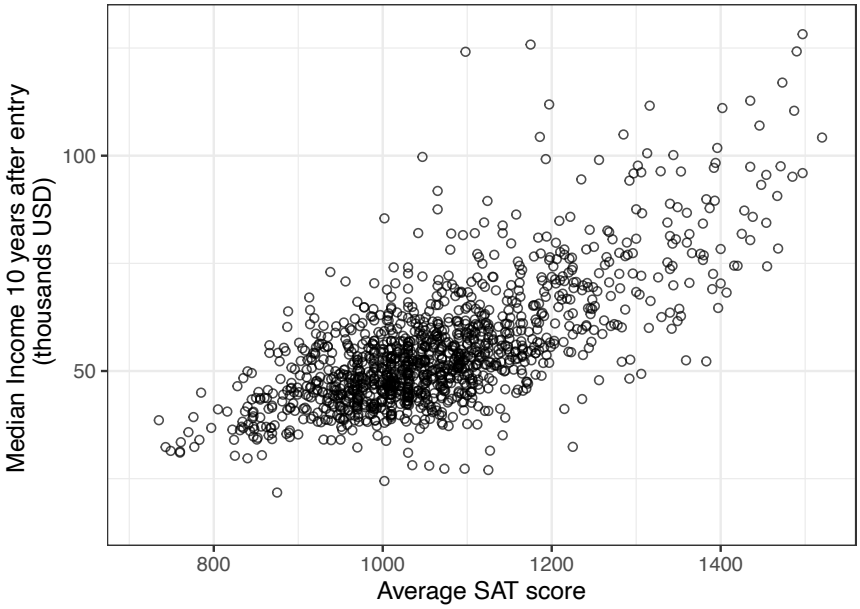


Figure 1: Median Graduate Income across Post-secondary Institutions

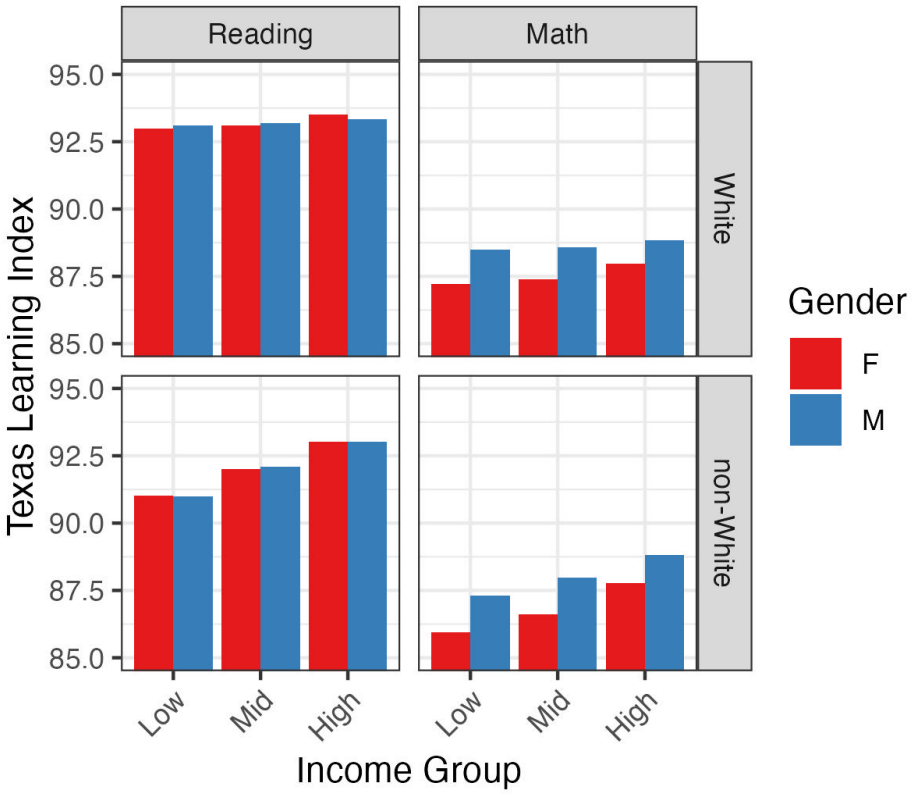


Figure 2: Average Reading and Math TLI across Groups

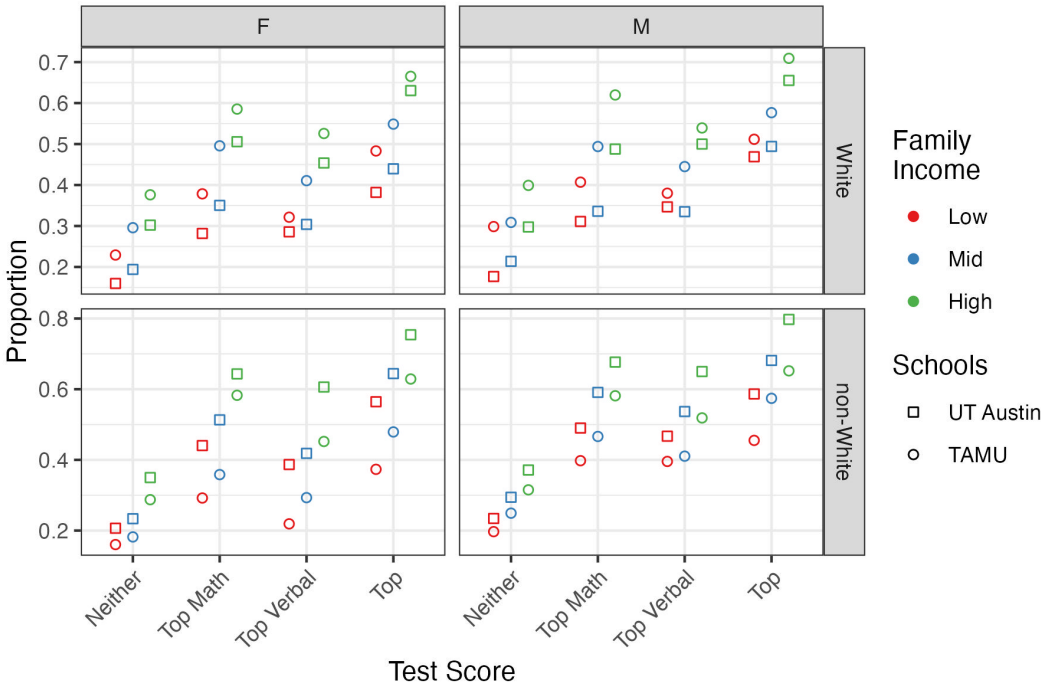


Figure 3: Proportion of Students Admitted to Flagship Universities

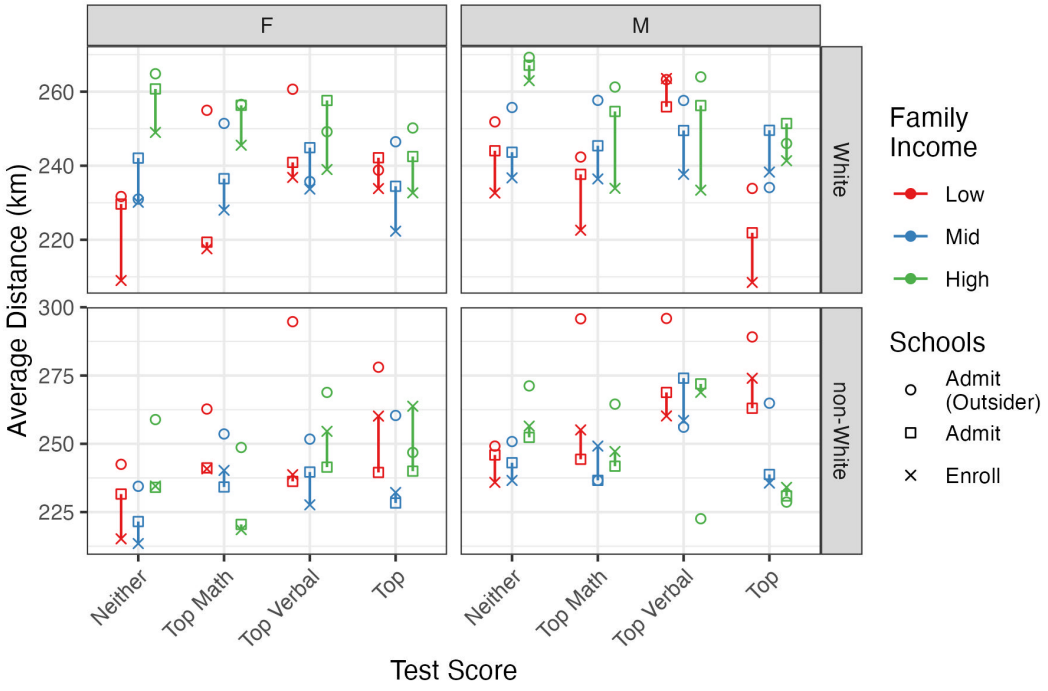


Figure 4: College Distances in Admission Sets

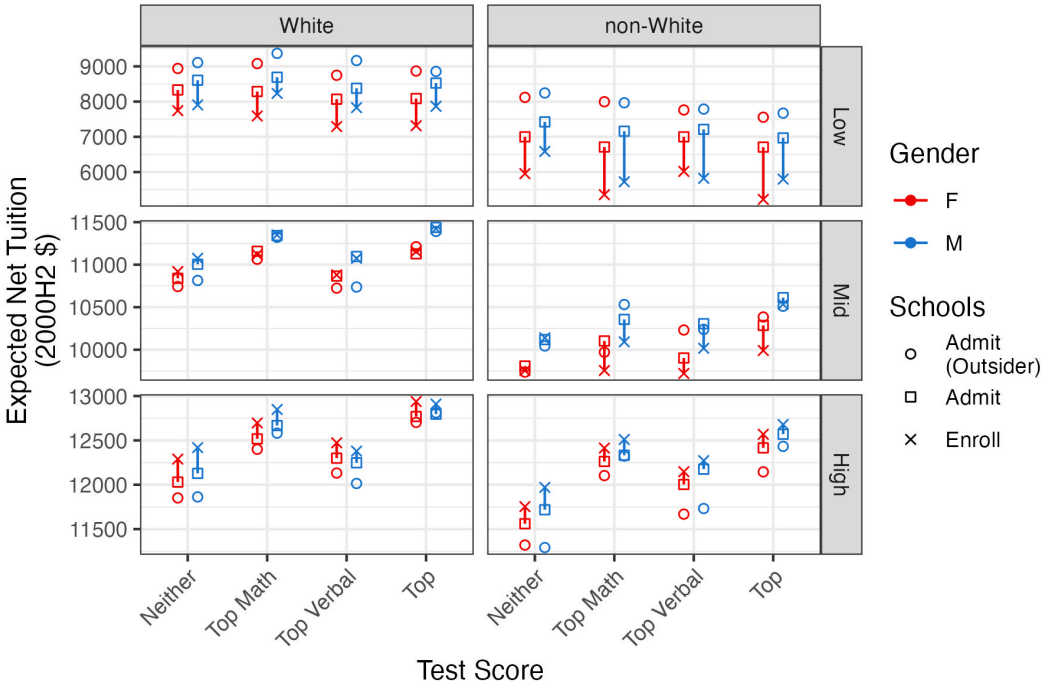


Figure 5: Net Tuition in Admission Sets

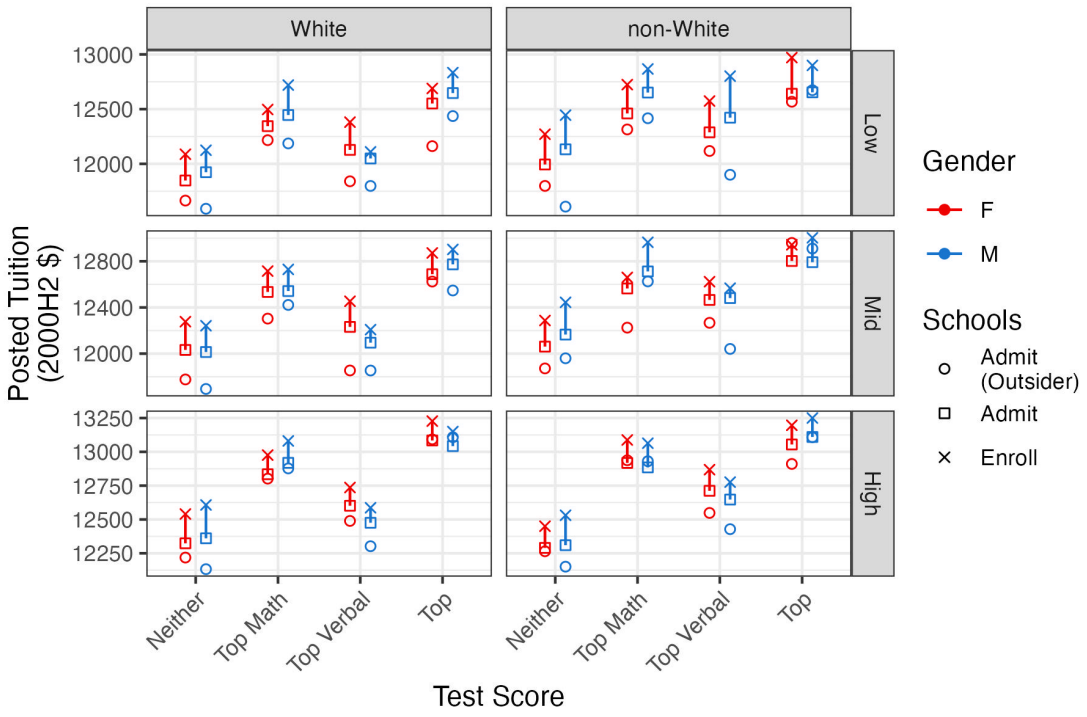


Figure 6: Posted Tuition in Admission Sets

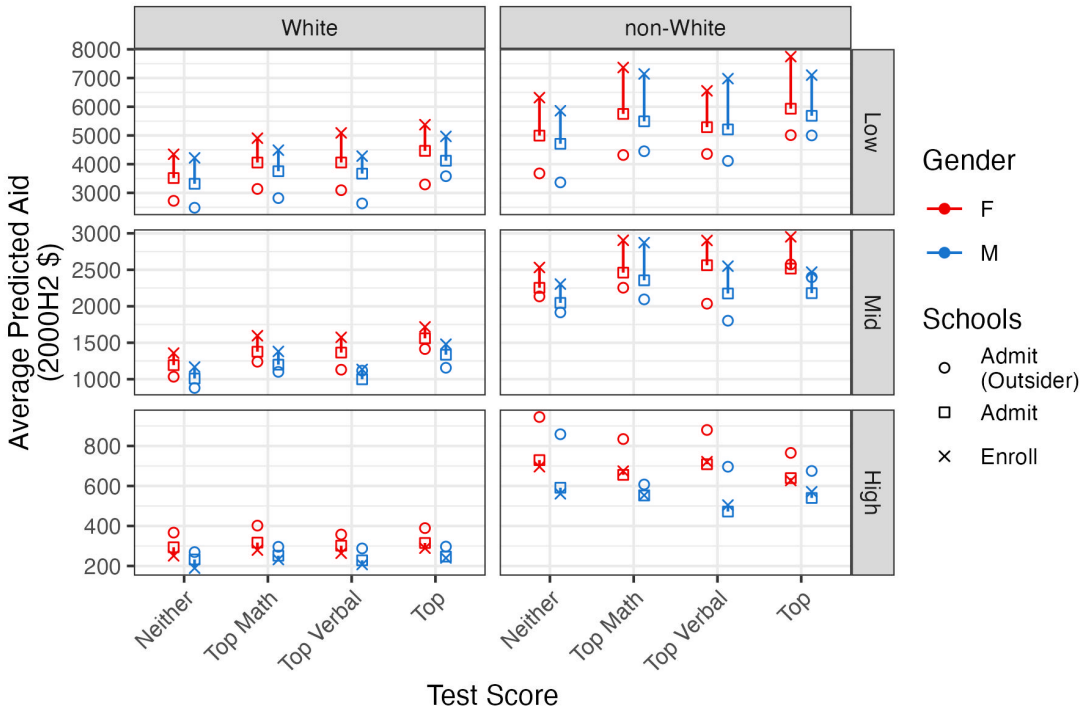


Figure 7: Aid Received in Admission Sets

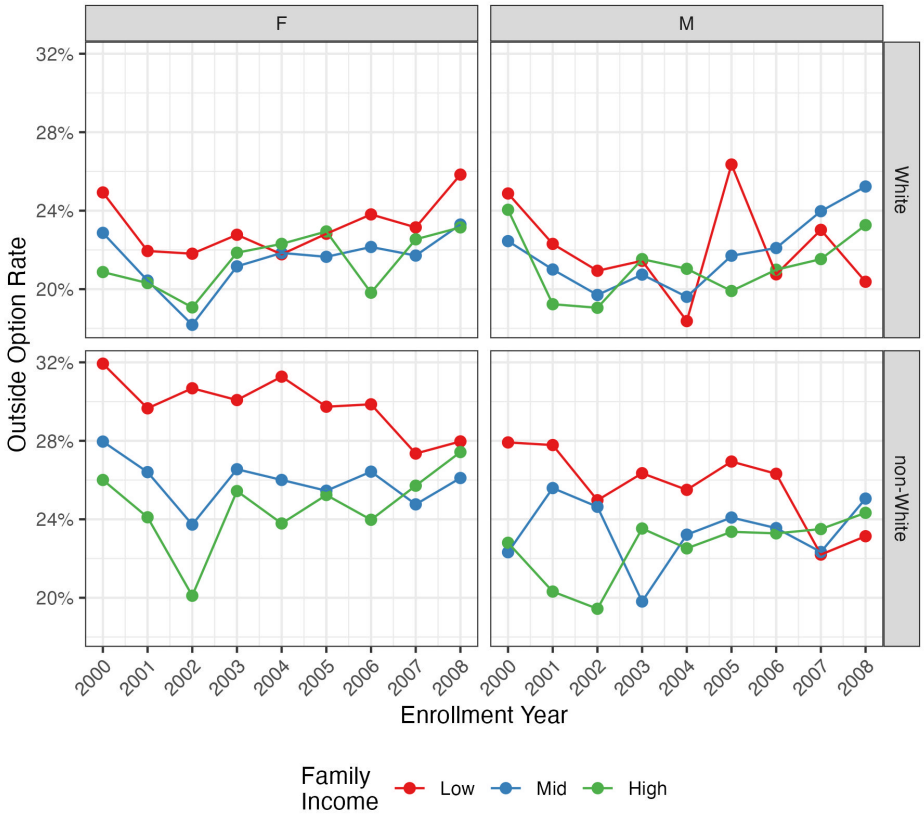


Figure 8: Proportion of Outside Option over the Years

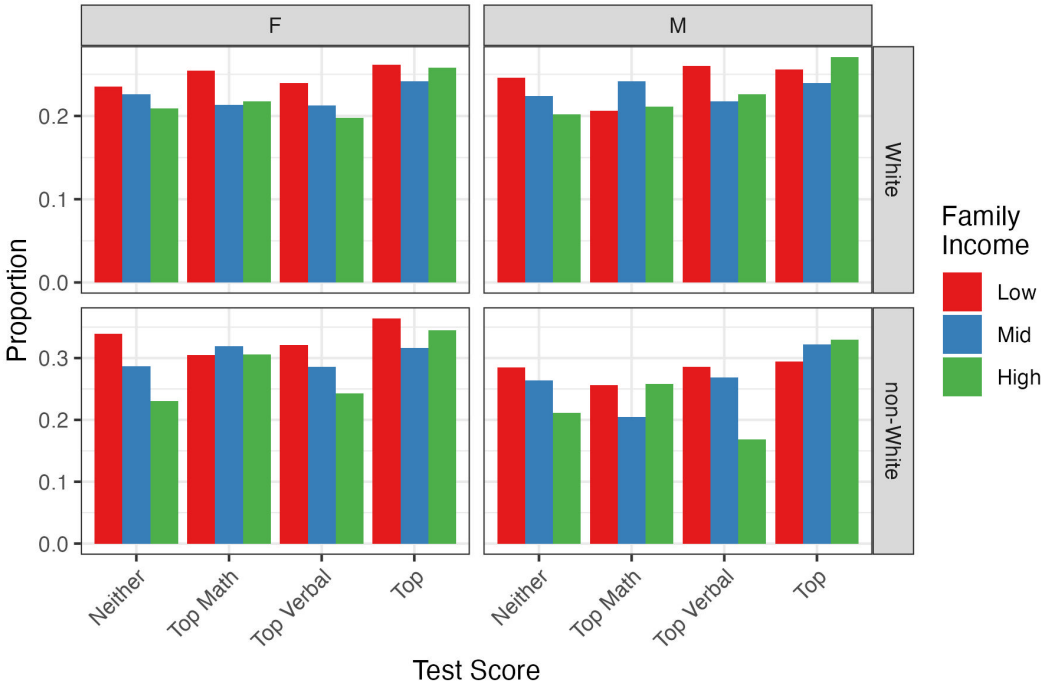


Figure 9: Proportion of Students Choosing Outside Option

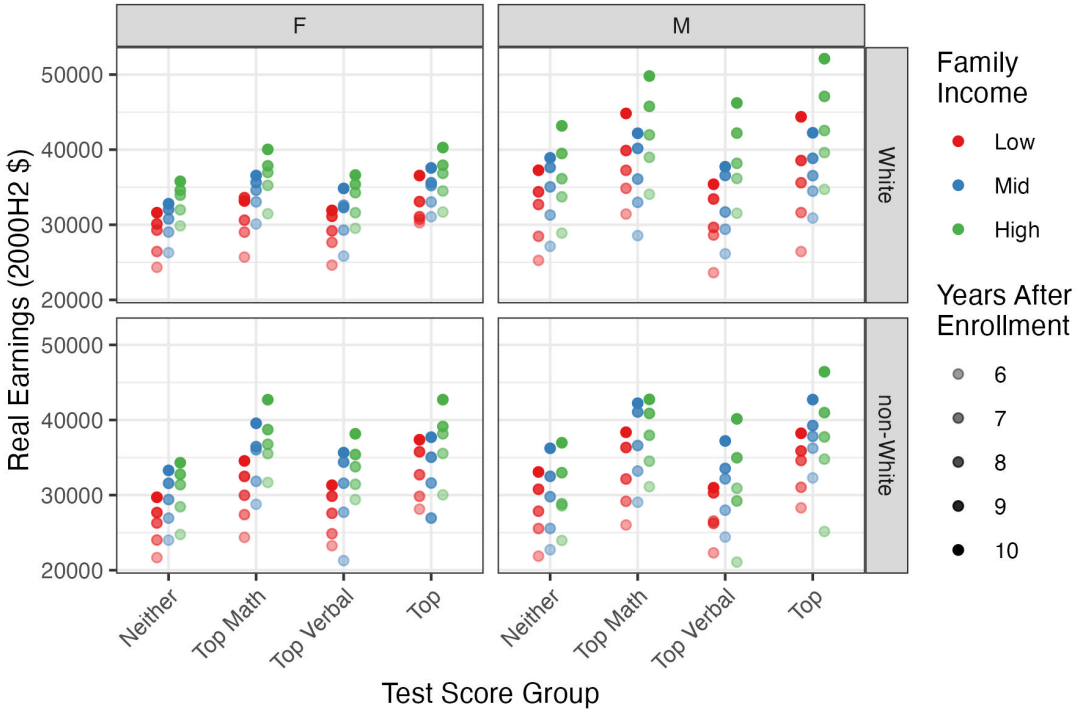


Figure 10: Average Earnings over time across Groups

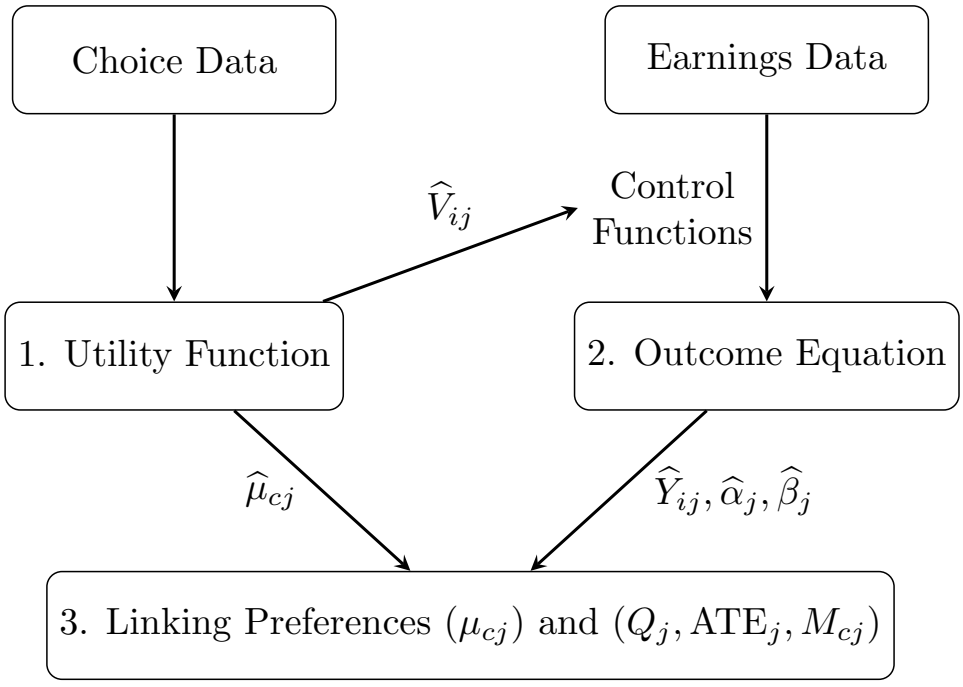


Figure 11: The Main Estimation Procedure

Table 1: Estimated Distance and Tuition Elasticities by Cells

Gender	Race	Family Income	Top Test Scores	$\mathcal{E}_\delta$	$\mathcal{E}_\tau$
F	White	Low	Both	-0.061	-5.669
F	White	Low	Math	-0.190	-7.410
F	White	Low	Verbal	-0.055	-7.681
F	White	Low	Neither	-0.297	-6.586
F	non-White	Low	Both	-0.556	-7.695
F	non-White	Low	Math	-0.607	-6.245
F	non-White	Low	Verbal	-0.788	-5.976
F	non-White	Low	Neither	-0.555	-5.422
M	White	Low	Both	-0.541	-9.718
M	White	Low	Math	-0.308	-4.794
M	White	Low	Verbal	-0.064	-4.443
M	White	Low	Neither	-0.118	-6.527
M	non-White	Low	Both	-0.524	-6.448
M	non-White	Low	Math	-0.345	-5.850
M	non-White	Low	Verbal	-1.838	-10.440
M	non-White	Low	Neither	-0.390	-5.091
F	White	Mid	Both	-0.187	-1.306
F	White	Mid	Math	-0.187	-2.160
F	White	Mid	Verbal	-0.178	-2.372
F	White	Mid	Neither	-0.131	-2.221
F	non-White	Mid	Both	-0.271	-2.273
F	non-White	Mid	Math	-0.114	-1.916
F	non-White	Mid	Verbal	-0.200	-2.507
F	non-White	Mid	Neither	-0.280	-1.237
M	White	Mid	Both	-0.139	-2.763
M	White	Mid	Math	-0.107	-2.072
M	White	Mid	Verbal	-0.091	-2.477
M	White	Mid	Neither	-0.159	-1.721
M	non-White	Mid	Both	-0.300	-2.279
M	non-White	Mid	Math	-0.179	-2.833
M	non-White	Mid	Verbal	-0.375	-2.510
M	non-White	Mid	Neither	-0.259	-1.748
F	White	High	Both	-0.110	-0.028
F	White	High	Math	-0.075	-0.001
F	White	High	Verbal	-0.204	-0.025
F	White	High	Neither	-0.182	-0.000
F	non-White	High	Both	-0.006	-0.117
F	non-White	High	Math	-0.360	-0.000
F	non-White	High	Verbal	-0.022	-1.012
F	non-White	High	Neither	-0.182	-0.002
M	White	High	Both	-0.137	-0.181
M	White	High	Math	-0.198	-0.014
M	White	High	Verbal	-0.051	-0.025
M	White	High	Neither	-0.091	-0.000
M	non-White	High	Both	-0.129	-0.013
M	non-White	High	Math	-0.317	-0.147
M	non-White	High	Verbal	-0.526	-1.298
M	non-White	High	Neither	-0.136	-0.157

Figure 12: Proportion of Correct Predictions for Each Cell

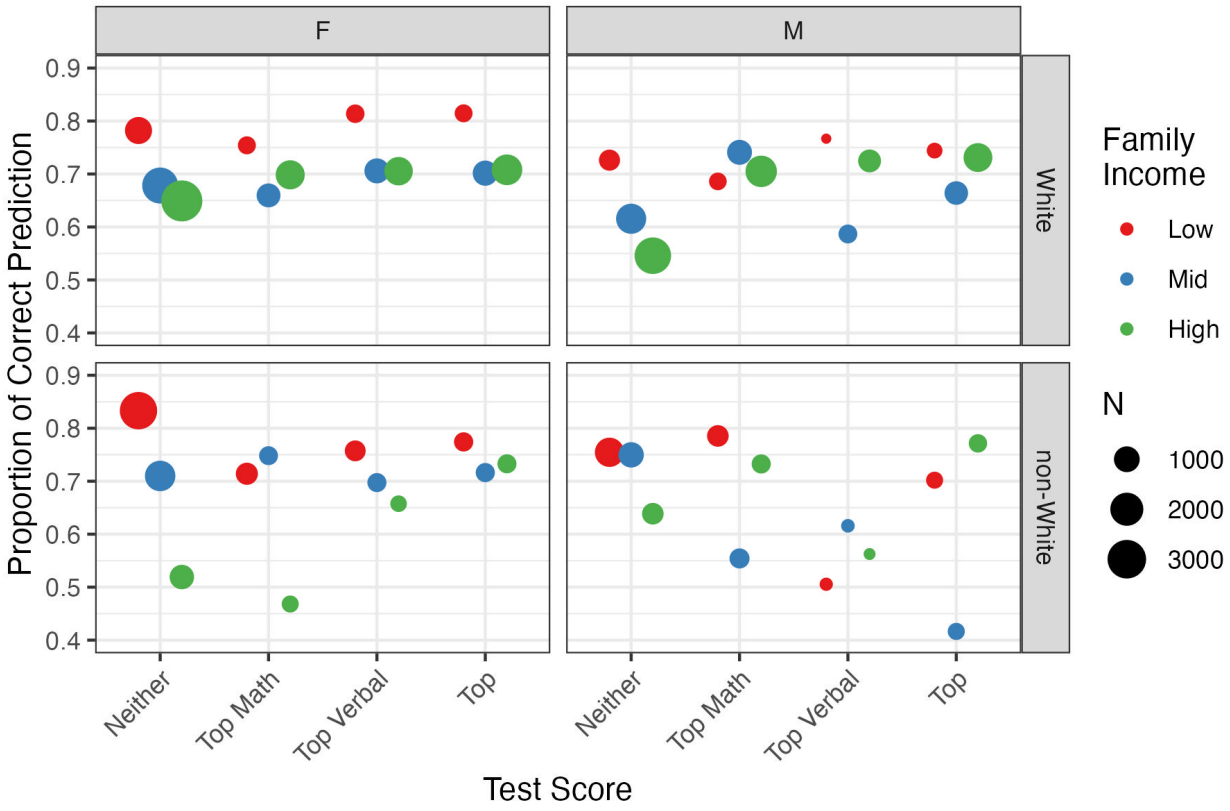


Figure 13: Estimated $\widehat{\alpha}_j$ with Average Earnings 6-10 Years after Enrollment

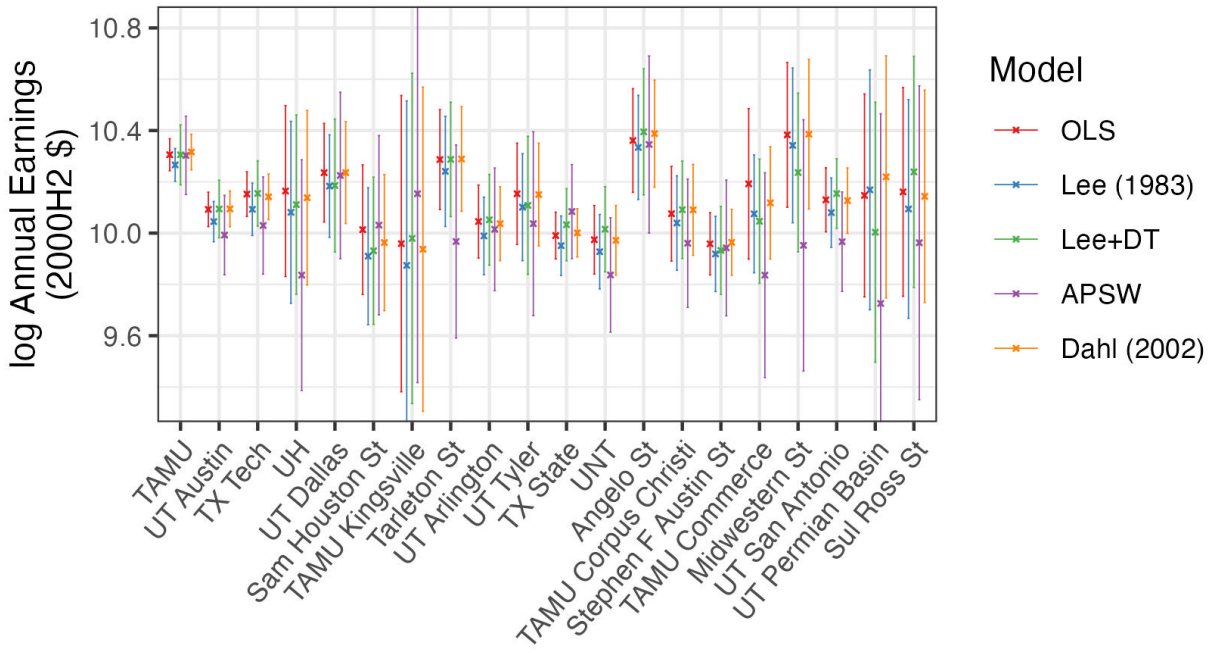


Figure 14: Estimated $\widehat{\alpha}_j$ with Average Earnings 6-10 Years after Enrollment (Baseline: UT Austin)

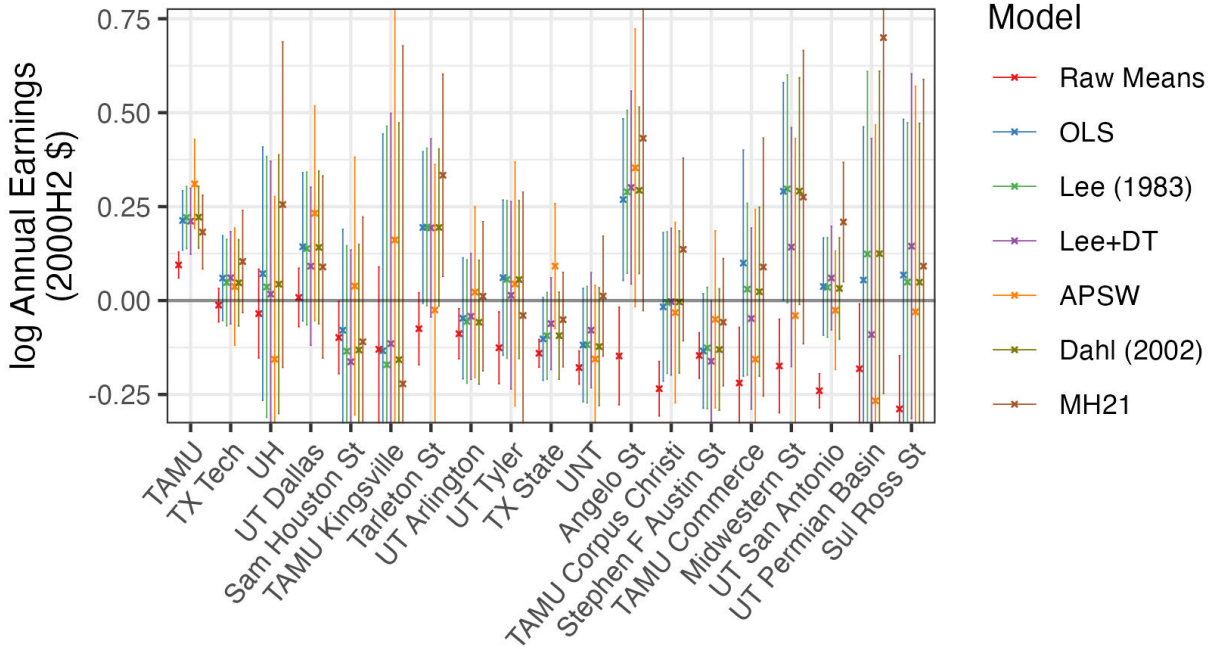


Figure 15: Estimated Relative Returns for Female Students

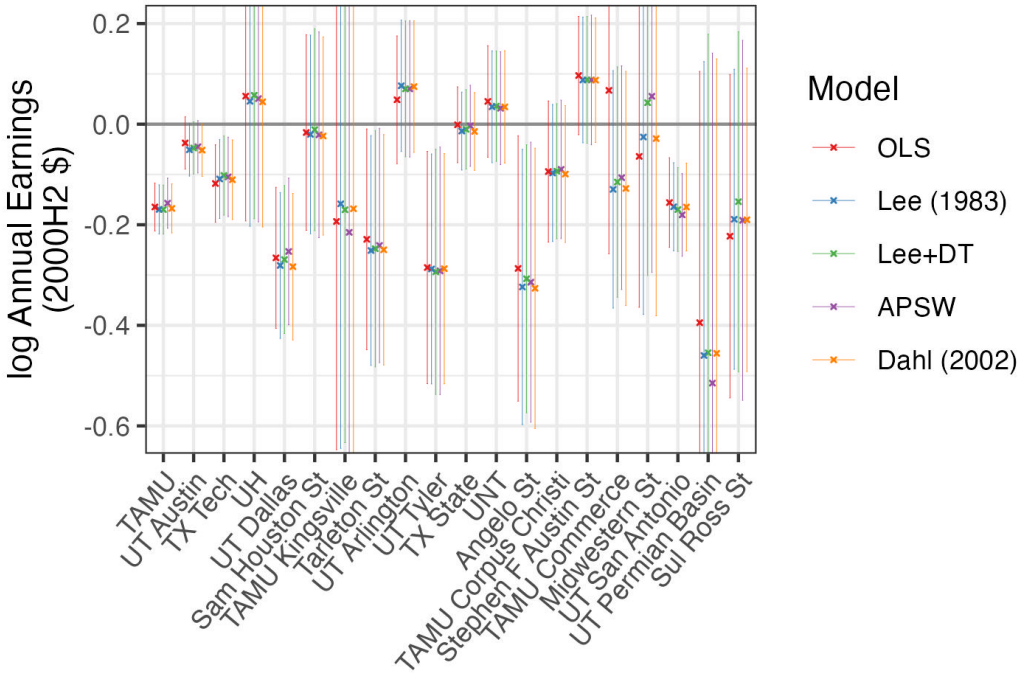


Figure 16: Estimated Returns for High-Income Students (Relative to Mid-Income)

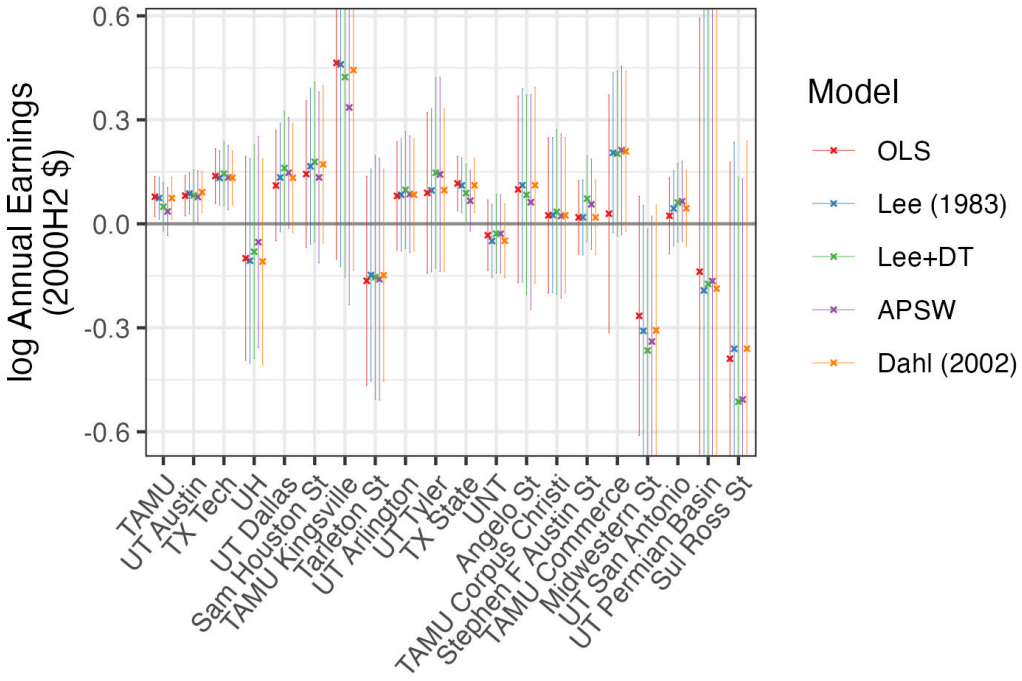


Figure 17: Estimated Returns for Low-Income Students (Relative to Mid-Income)

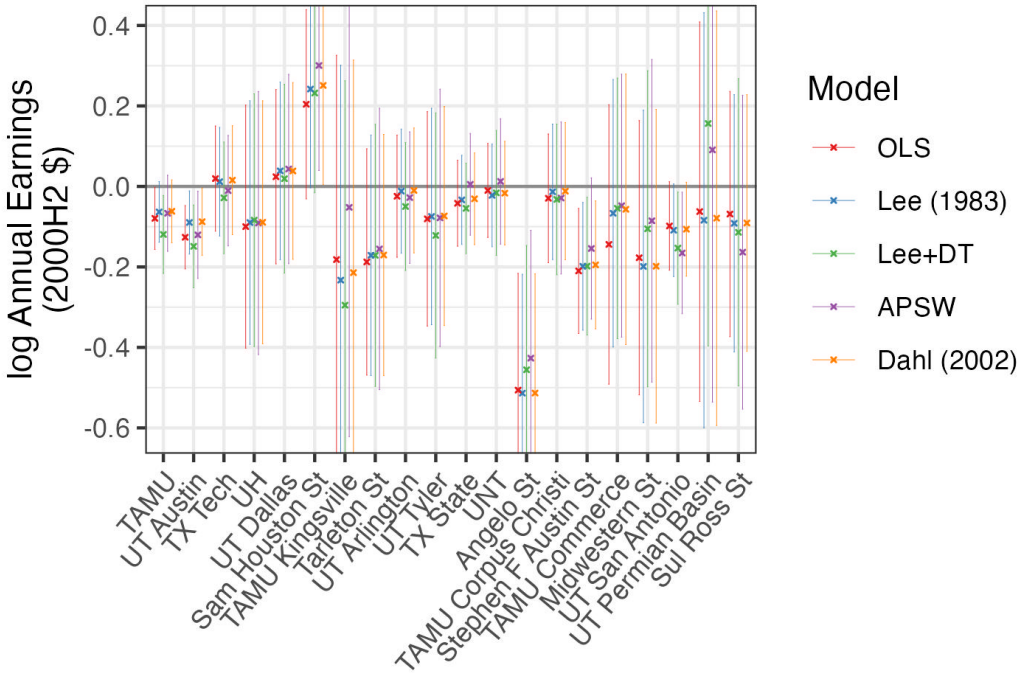


Figure 18: Estimated Relative Returns for Non-White Students

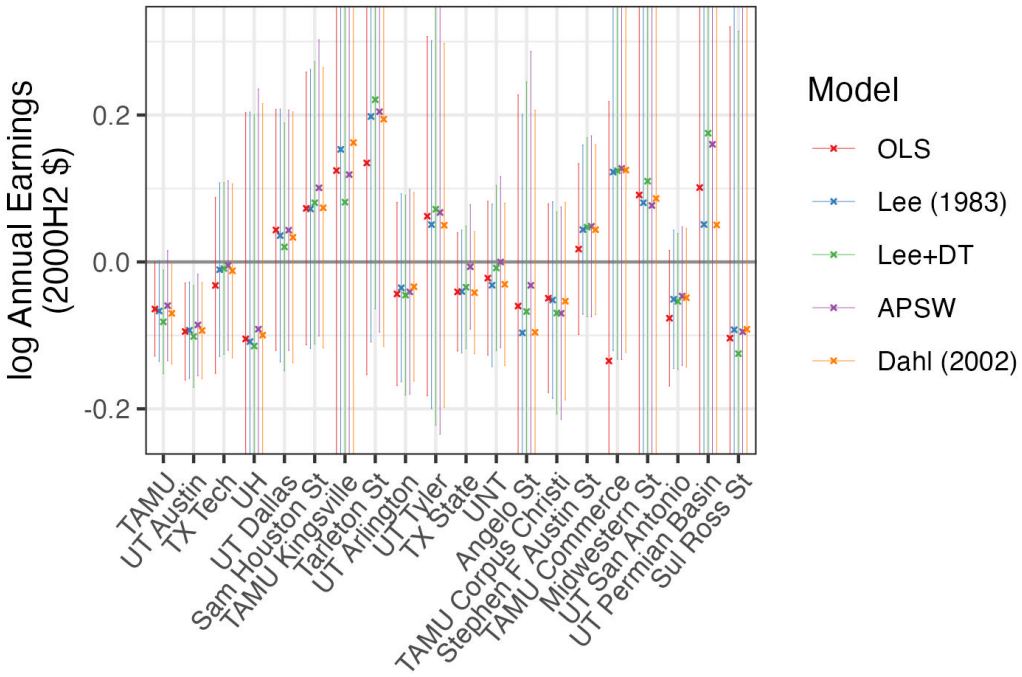


Figure 19: Estimated Relative Returns for Top 10% Students

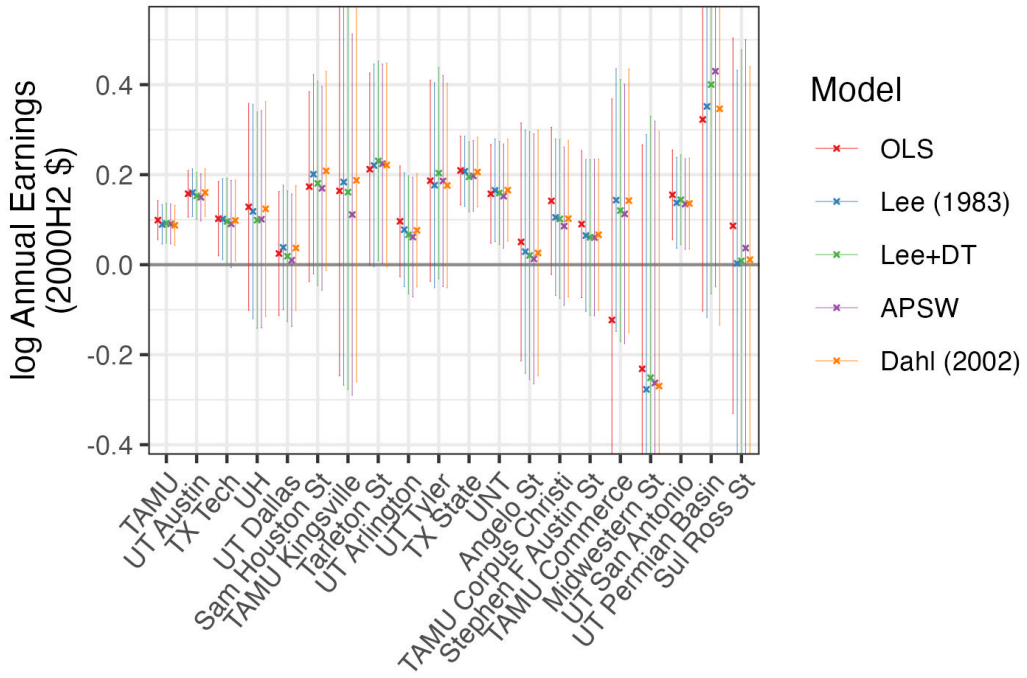


Figure 20: Estimated Relative Returns for Economically Disadvantaged Students

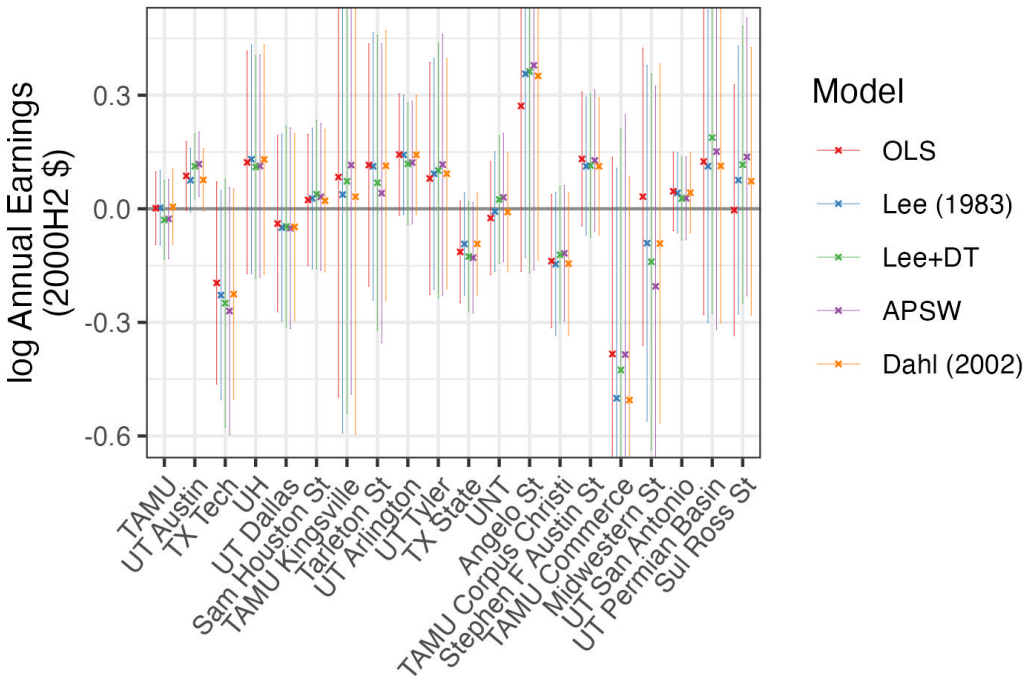


Figure 21: Estimated Returns for Reading Scores

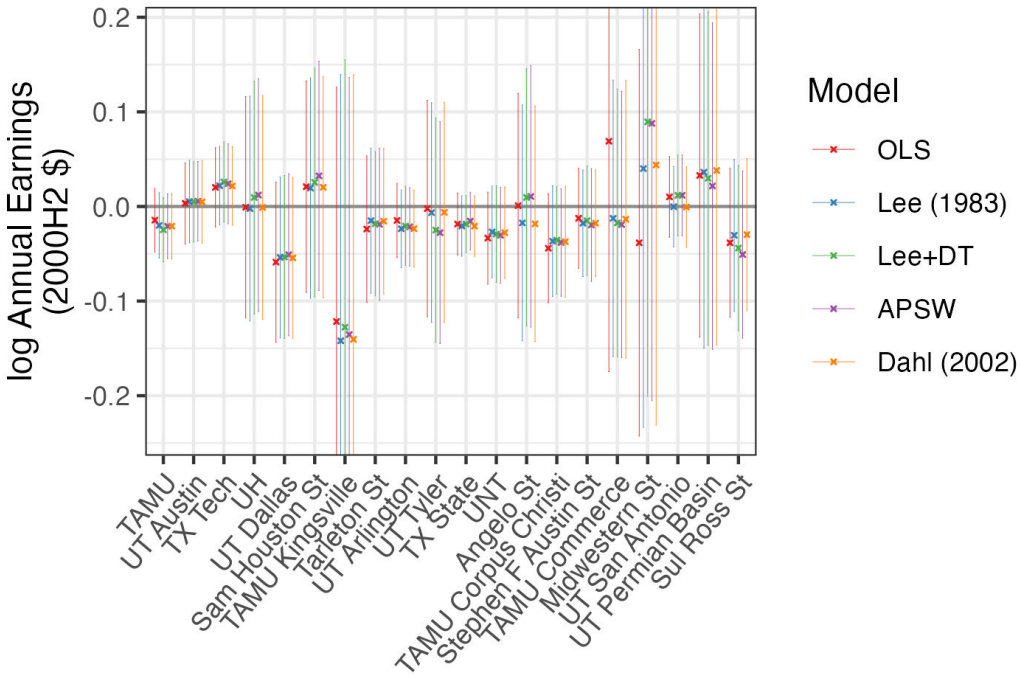


Figure 22: Estimated Returns for Writing Scores

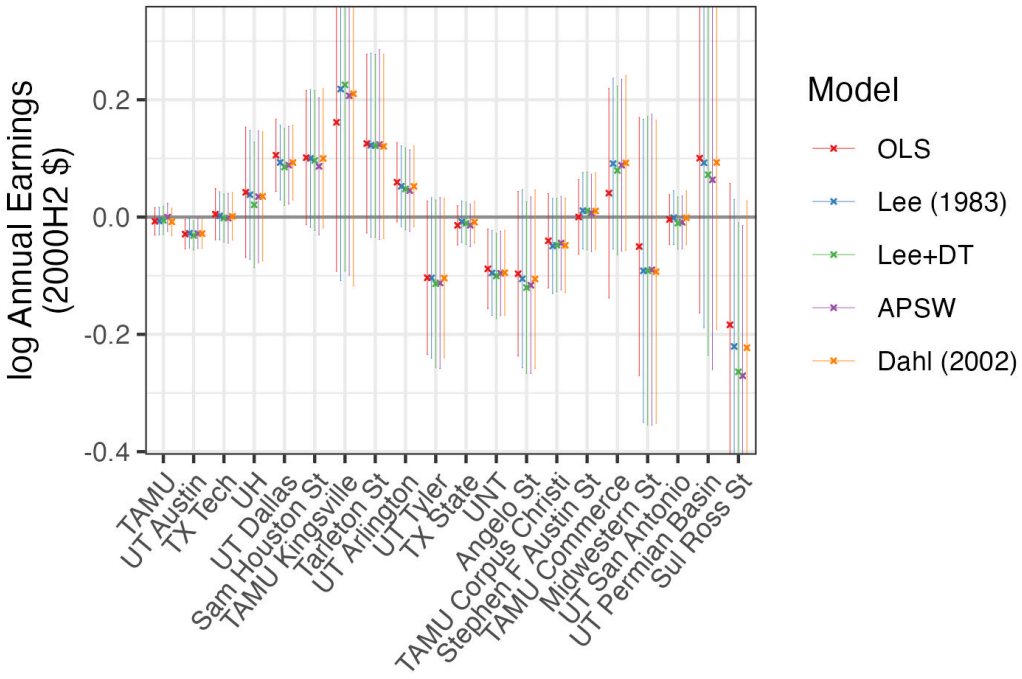


Figure 23: Estimated Returns for Math Scores

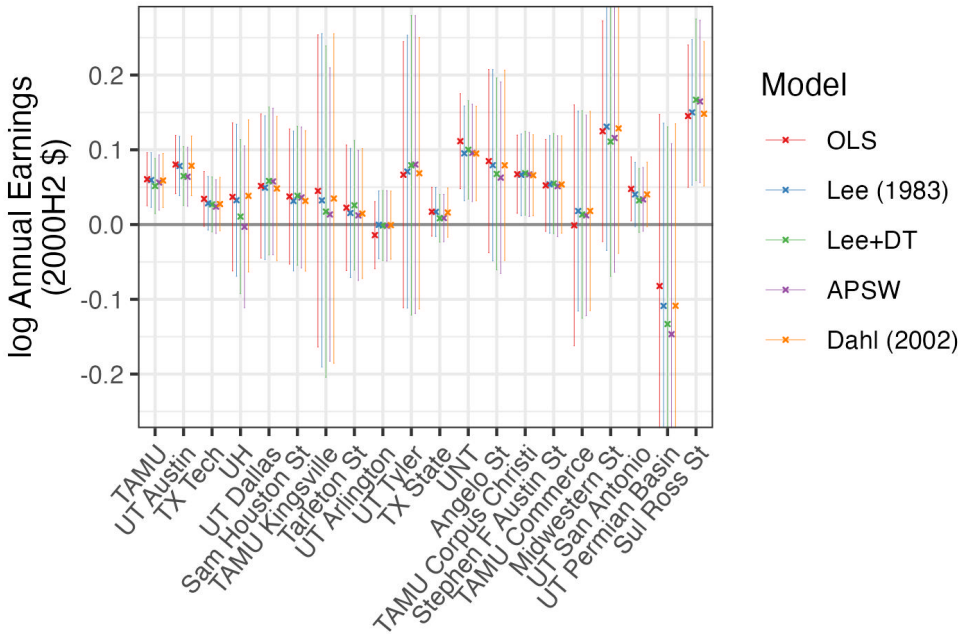
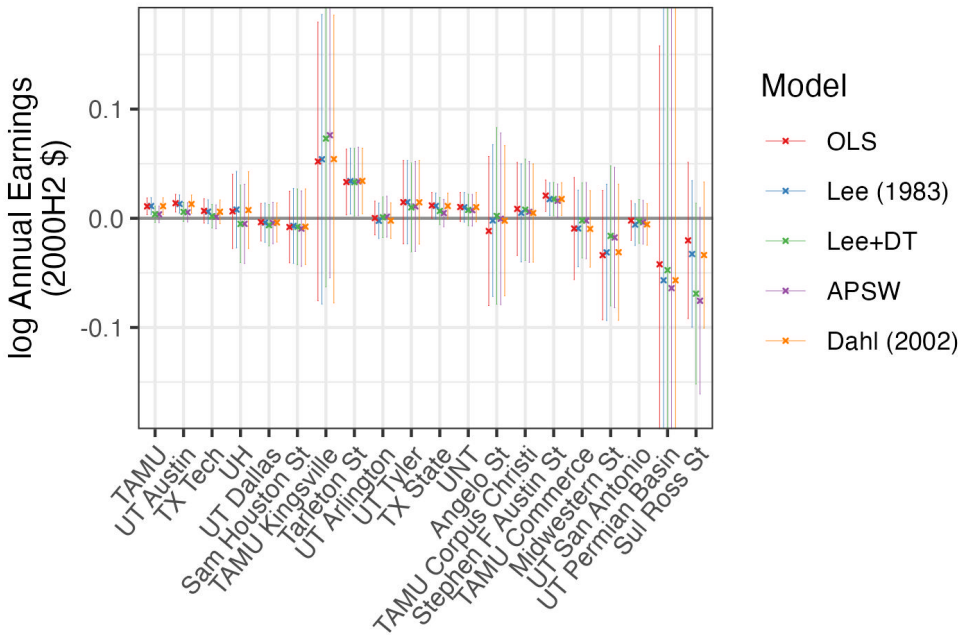


Figure 24: Estimated Returns for Different Hometown Population Densities



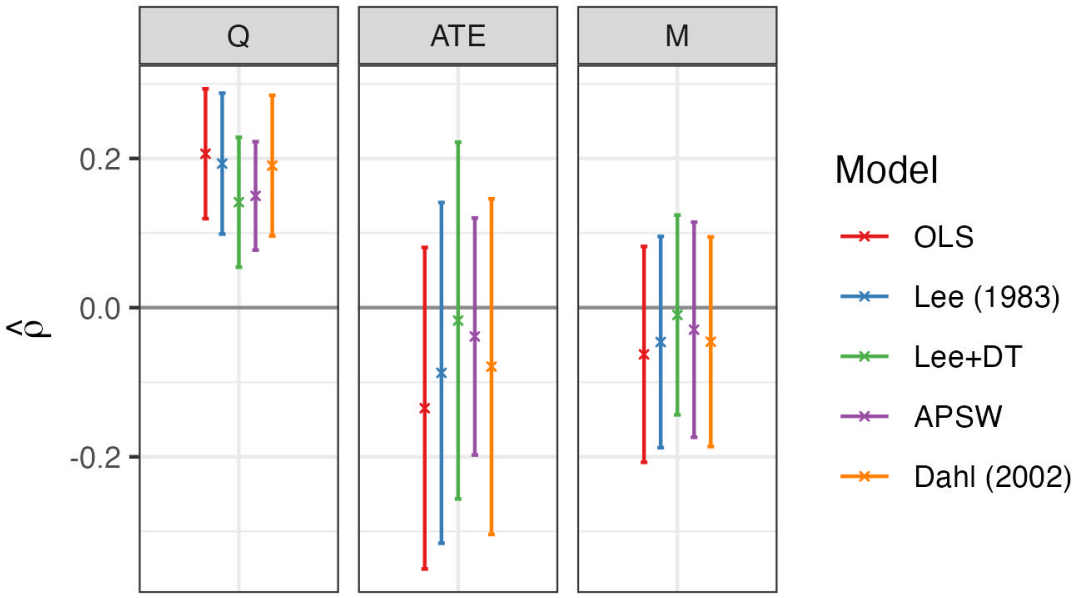


Figure 25: Estimated Utility Weights

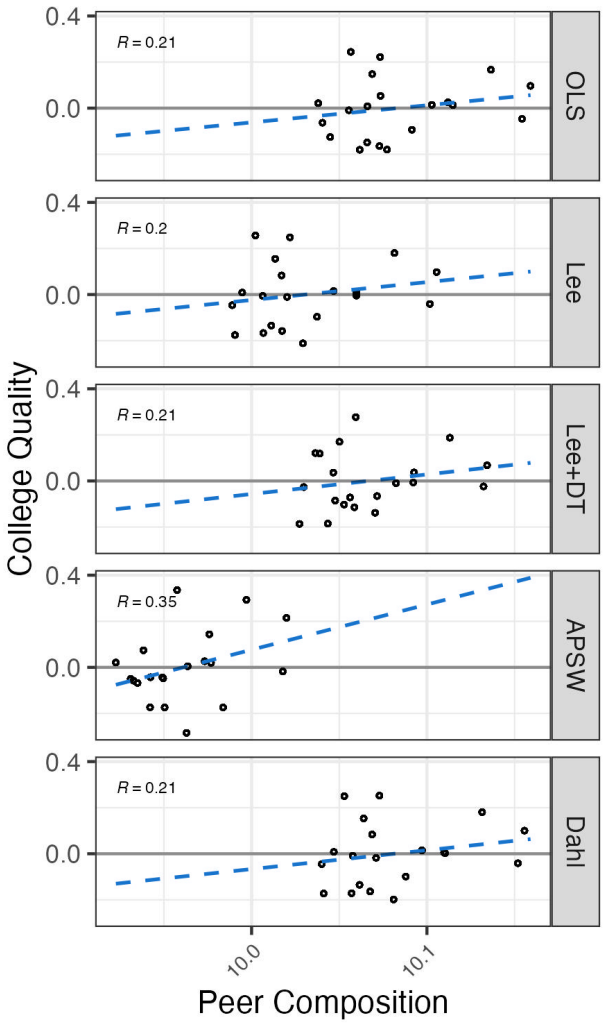


Figure 26: Estimated Correlation between Peer Composition and College Quality

Appendix

A Data Appendix

A.1 Distance Calculation

Among a few geographical information about the student's home, the location of their high school of graduation is possibly the most granular, compared to student residential county codes in the application, enrollment and financial aid datasets from THECB. Most of the geographic information for the high schools are obtained using the TEA public open data site, which records the current geographic coordinates of TEA-registered schools.[27] For these current schools with valid addresses in the TEA ArcGIS data, I assumed they never moved throughout the years. This might not be true, but the workload of data cleaning would explode otherwise.

For high schools that were in the administrative data but have since been closed and thus not showing up in the TEA ArcGIS data, I used the AskTED online directory to obtain historical records about their location and address. For these observations, I have a panel of their locations.

In some cases, when the site address of the schools is missing, I replaced it with a valid mailing address that is not a P.O. box, or the central coordinates of the city, school district, or county if the previous step fails. A few remaining schools required some manual correction, and I tried to pinpoint their location to the best of my ability using any information I could gather from the internet and the maps.

Distances are calculated as the great circle distance between the coordinates of the high schools and universities. The Earth's radius used in the calculation is 6373 kilometers or 3960 miles, about the radius plus elevation at the geographic center of Texas.

[27] Link: https://schoolsdata2-tea-texas.opendata.arcgis.com/. Data accessed in April 2025.

A.2 Net Tuition Imputation

To impute the net tuition of alternative schools, I use a 2-step process based on the extensive and intensive margins of financial aid. All dollar figures are deflated to the price levels in the second half of 2000, using the CPI-U series for the South, ID CUUR0300SA0 and CUUS0300SA0.

Total Cost of Attendance

Before dealing with the financial aid, I obtain the total cost of attendance for each school each year from the IPEDS data, including tuition, fees, and other living expenses. As stated in section 2, I use the reported total price for students living on campus as the baseline for the incoming freshmen, with some exceptions.

Because the University of Houston - Downtown does not have on-campus housing, I use the reported total price for students living off-campus without family. There are a few schools that do not have tuition information in IPEDS, such as Texas A&M at Galveston and University of Texas - Brownsville (closed in 2015), and a few other schools with sporadically missing data. It is not possible to impute the net tuition under the current process. I intend to collect more information to rectify this issue.

Net tuition

Many students do not pay the full cost, of course. The true price of college for students after taking into account the amount financial aid received is their "net tuition." Regarding the financial aid of the students, we have two main pieces of information:

1. Students with financial aid information in the FADs data, for selected schools (mostly the schools they enroll in but a small amount of non-enrolled schools),

2. Students who enrolled in public university but have no FADs entry = no financial aid at the enrolled school

A key piece of missing information is the net tuition levels for students who did not enroll in any public university. It is important to impute the counterfactual net tuition for those students, since otherwise many students choosing the outside option will be dropped from the estimation sample.

Extensive Margin: Eligibility The first step of the imputation is to learn the extensive margin of the financial aid. Does the student receive any financial aid at all? I use both pieces of information outlined above. A logistic regression is then used to predict the aid-receiving status of the student with their observable characteristics X_i^E. Say the probability for student i to receive any need-based aid at school j is $\mathbb{P}_{ij}^{\text{Aid}}$, then

$$\ln \frac{\mathbb{P}_{ij}^{\text{Aid}}}{1 - \mathbb{P}_{ij}^{\text{Aid}}} = f^E(X_i^E, j) + u_{ij}^E \tag{A1}$$

The f^E function is a linear function of the students' characteristics X_i^E, which include gender, ethnicity and race, parental education levels, an indicator for single-parent households, Texas resident status (based on residence codes in the application data), whether the student received offers from UT Austin or TAMU, whether the student was ever economically disadvantaged during high school, and their family income levels interacted with the posted tuition of the schools.

Intensive Margin: Amount of Aid The second step is to estimate the amount of aid received, conditional on receiving any aid, with the following linear regression:

$$\text{Aid}_{ij} = f_j^I(X_i^I) + u_{ij}^I \tag{A2}$$

The regressors in equation A2 are fully interacted with school indicators j, and include those listed in X_i^E along with some extra variables — the inverse of the number of people in the household and whether the financial aid data is from an enrolled school or not —

interacted with family income levels. The predicted aid amount is then bounded by 0 and the posted tuition.

Expected Net Tuition After predicting $\widehat{\mathbb{P}}_{ij}^{\text{Aid}}$ and Aid_{ij} from the model, I calculate the expected net tuition:

$$T_{ij} = T_j - \widehat{\mathbb{P}}_{ij}^{\text{Aid}} \cdot \text{Aid}_{ij} \tag{A3}$$

where T_j is the posted tuition of school j.

There are alternative ways to impute the net tuition. Kapor (2025) drew a few realizations of expected family contribution (EFC) from income information in the CPS and then estimated the rate of "generosity" of each school in fulfilling the students' financial need. I have a predicted probability of receiving aid for each school, so it might be possible to also simulate a few different realizations of net tuition levels. However, the tricky part for simulation is figuring out the correlation between those probabilities.

A.3 Sample Selection

I have 20 years of admission data – from the fiscal year of 2001 (starting 2000 fall) to 2020. I selected first-time college students (not transfers) who applied to Bachelor's degrees. I study the students' enrollment decisions given their admission results, so I dropped observations of withdrawn applications and only kept students with offers from at least 2 Texas public universities. In the corresponding fall semester enrollment report by these universities, I selected only students classified as "college freshmen."

It is possible to study students with only one offer, but there are some challenges associated with the available data. Since the admission and enrollment data are only available for 4-year public universities, I cannot determine the destination of those who do not enroll—they could be out of state, at in-state private universities, or at other 2-year institutions. However, data patterns indicate that those students' choices are very different from those of students

with multiple offers—they have a much higher rate of non-enrollment in a public university. Since they might have widely different outside options from those who received at least two offers, I excluded them from the sample. This is about 78% of the students who received any offer from Texas PUs in the early 2000s.

Because I use high school location to calculate students' distances to universities, I further narrowed down the sample to those who were enrolled in a Texas high school in the TEA dataset. This could include nonresident students attending a Texas high school. Geographic information for non-Texas high school students is much less precise – it is only at the state level for US residents and country level for international students. Larger distances from home for these students also mean that the difference in distances between options within Texas will be largely insignificant.

The final dataset is combined from multiple sources, so naturally, it only includes people who can be tracked through all the datasets. They need to have a valid and unique student identifier 'ID2' to be successfully matched.[28] They also need to have observed outcome – valid earnings data. Therefore, only those who work in Texas after graduation and in the state UI dataset are included.

[28]There is an extra layer of complication - these identifiers are student SSN replacements for privacy purposes. However, for students without a valid SSN replacement, the school assigns a school-specific identifier instead. These IDs cannot be used in matching across schools and are flagged with 'stuinvssn' in the enrollment data. For students like this, since their ID is not consistent across schools, we cannot confidently observe their admission results, and they will never "receive" more than 1 offer. I checked the ratio in some enrollment data, and it is around 1.75% in the fall of 2000.

B Results Appendix

In this appendix, I show detailed results from the baseline model reported in the main text.

B.1 Baseline Results

Table B1 shows the estimated utility parameters for each group. The estimates are obtained from the baseline model with 48 cells. Standard errors are reported in parentheses below each estimate. Under this specifications, the estimated tuition sensitivity is close to the boundary of0 for high-income students. The choice patterns in Figure 5 foreshadow such results.

Table B2 and Table B3 show the estimated school value for each school and each group, respectively.

Apart from using the average earnings 6-10 years after initial enrollment, I also conducted regression with each of the years separately, and with two other averages — 8-10 years, matching up with MH21's measurement and 6-10 years with students without missing earnings. Figure B1 to B11 show the outcome regression coefficients $(\widehat{\alpha}_j, \widehat{\beta}_j)$. All standard errors used to construct the confidence intervals are bootstrapped with 100 replications. "APSW" and "Lee+DT" models are temporarily dropped from these figures due to extremely noisy estimates. Patterns are largely consistent across different outcome measures. Figure B12 reports the estimated $\widehat{\alpha}_j$ when benchmarking against UT Austin.

Finally, Figure B13 shows the correlation between peer composition and college quality for all regression specifications and outcome measures. Correlations are all positive, ranging from 0.11 to 0.64.

Table B1: Estimated Utility Parameters by Group

Gender	Race	Family Income	Top Test Scores	$\widehat{\tau}_c$	$\widehat{\delta}_c$	$\widehat{\iota}_c$	$\widehat{\mu}_c^+$
F	White	Low	Both	2.104	0.077	0.149	-0.370
				(2.450)	(0.200)	(0.181)	(0.297)
F	White	Low	Math	6.587	0.616	0.451	-0.278
				(1.533)	(0.504)	(0.164)	(0.347)
F	White	Low	Verbal	7.401	0.175	0.415	-0.056
				(1.341)	(0.289)	(0.106)	(0.263)
F	White	Low	Neither	6.116	1.003	0.416	-0.402
				(0.874)	(0.252)	(0.064)	(0.199)
F	non-White	Low	Both	4.159	0.810	0.277	0.217
				(0.776)	(0.455)	(0.060)	(0.295)
F	non-White	Low	Math	5.526	1.507	0.470	0.117
				(0.798)	(0.330)	(0.085)	(0.248)
F	non-White	Low	Verbal	5.382	1.978	0.487	0.077
				(0.914)	(0.423)	(0.096)	(0.278)
F	non-White	Low	Neither	5.001	1.576	0.516	-0.189
				(0.445)	(0.209)	(0.047)	(0.143)
M	White	Low	Both	5.529	1.167	0.243	-0.220
				(1.582)	(0.638)	(0.076)	(0.355)
M	White	Low	Math	8.101	1.917	0.811	0.387
				(1.665)	(0.631)	(0.189)	(0.360)
M	White	Low	Verbal	4.346	0.204	0.532	-0.970
				(3.125)	(0.513)	(0.339)	(0.548)
M	White	Low	Neither	5.284	0.337	0.331	-0.298
				(1.749)	(0.300)	(0.117)	(0.261)
M	non-White	Low	Both	5.921	1.250	0.470	0.154
				(1.115)	(0.529)	(0.122)	(0.335)
M	non-White	Low	Math	3.807	0.634	0.298	-0.032
				(0.985)	(0.355)	(0.088)	(0.258)
M	non-White	Low	Verbal	4.228	1.935	0.197	1.201
				(1.625)	(1.165)	(0.112)	(0.501)
M	non-White	Low	Neither	5.517	1.289	0.593	0.101
				(0.642)	(0.238)	(0.066)	(0.194)
F	White	Mid	Both	0.035	0.024	0.015	-0.408
				(0.008)	(0.005)	(0.003)	(0.114)
F	White	Mid	Math	1.500	0.601	0.426	-0.201
				(0.884)	(0.426)	(0.254)	(0.354)
F	White	Mid	Verbal	0.995	0.332	0.247	-0.485
				(0.686)	(0.223)	(0.179)	(0.227)
F	White	Mid	Neither	1.589	0.422	0.400	-0.174
				(0.479)	(0.147)	(0.128)	(0.149)
F	non-White	Mid	Both	0.671	0.343	0.227	-0.231
				(0.819)	(0.602)	(0.258)	(0.356)
F	non-White	Mid	Math	0.023	0.006	0.007	-0.164
				(0.003)	(0.001)	(0.003)	(0.161)
F	non-White	Mid	Verbal	0.896	0.292	0.219	-0.662
				(1.020)	(0.398)	(0.225)	(0.256)
F	non-White	Mid	Neither	0.237	0.233	0.101	-0.443
				(0.199)	(0.202)	(0.076)	(0.148)

Estimated Utility Parameters by Group (continued)

Gender	Race	Family Income	Top Test Scores	$\widehat{\tau}_c$	$\widehat{\delta}_c$	$\widehat{\iota}_c$	$\widehat{\mu}_c^+$
M	White	Mid	Both	0.074	0.017	0.015	-0.226
				(0.015)	(0.004)	(0.003)	(0.126)
M	White	Mid	Math	0.064	0.015	0.015	-0.507
				(0.012)	(0.003)	(0.003)	(0.139)
M	White	Mid	Verbal	0.050	0.008	0.011	-0.187
				(0.015)	(0.003)	(0.068)	(0.199)
M	White	Mid	Neither	1.891	0.780	0.538	0.220
				(0.605)	(0.295)	(0.153)	(0.204)
M	non-White	Mid	Both	0.060	0.034	0.019	0.849
				(0.031)	(0.016)	(0.008)	(0.245)
M	non-White	Mid	Math	1.496	0.412	0.348	0.201
				(0.897)	(0.354)	(0.222)	(0.365)
M	non-White	Mid	Verbal	0.029	0.017	0.008	-0.406
				(0.010)	(0.005)	(0.003)	(0.288)
M	non-White	Mid	Neither	0.497	0.303	0.169	-0.101
				(0.407)	(0.247)	(0.126)	(0.220)
F	White	High	Both	0.000	0.003	0.004	-0.168
				(0.000)	(0.000)	(0.000)	(0.063)
F	White	High	Math	0.000	0.084	0.166	-0.255
				(0.000)	(0.126)	(0.211)	(0.288)
F	White	High	Verbal	0.000	0.009	0.007	-0.348
				(0.000)	(0.001)	(0.000)	(0.094)
F	White	High	Neither	0.000	0.200	0.151	-0.177
				(0.000)	(0.149)	(0.104)	(0.151)
F	non-White	High	Both	0.002	0.001	0.016	-0.094
				(0.000)	(0.000)	(0.002)	(0.283)
F	non-White	High	Math	0.000	0.309	0.151	0.760
				(0.000)	(0.481)	(0.239)	(0.545)
F	non-White	High	Verbal	0.012	0.001	0.008	-0.339
				(0.004)	(0.000)	(0.032)	(0.254)
F	non-White	High	Neither	0.000	0.110	0.087	0.145
				(0.000)	(0.221)	(0.137)	(0.254)
M	White	High	Both	0.001	0.005	0.006	-0.006
				(0.000)	(0.000)	(0.000)	(0.067)
M	White	High	Math	0.000	0.021	0.017	-0.209
				(0.000)	(0.003)	(0.002)	(0.089)
M	White	High	Verbal	0.001	0.006	0.015	-0.586
				(0.000)	(0.001)	(0.004)	(0.182)
M	White	High	Neither	0.000	0.338	0.478	0.564
				(0.000)	(0.159)	(0.144)	(0.224)
M	non-White	High	Both	0.000	0.006	0.008	-0.128
				(0.000)	(0.001)	(0.003)	(0.170)
M	non-White	High	Math	0.001	0.010	0.006	-0.160
				(0.000)	(0.001)	(0.001)	(0.162)
M	non-White	High	Verbal	0.019	0.035	0.013	0.752
				(0.014)	(0.016)	(0.125)	(0.483)
M	non-White	High	Neither	0.004	0.015	0.018	-0.221
				(0.000)	(0.005)	(0.095)	(0.230)

Table B2: Distribution of μ_{cj} across Schools

School	Min	Median	Mean	Max	Std. Dev.
TX Southern	-0.383	-0.145	-0.154	0.056	0.214
West TAMU	-1.097	0.290	0.641	3.109	0.980
TAMU Kingsville	-0.636	0.307	0.692	5.968	1.213
Sam Houston St	-0.567	0.331	1.182	6.507	1.715
Sul Ross St	-0.380	0.336	1.458	6.226	1.851
UT Rio Grande Valley	-0.292	0.337	1.105	6.241	2.011
Stephen F Austin St	-0.262	0.388	1.477	6.308	1.807
TAMU Int'l	-0.249	0.404	1.151	3.615	1.223
UT El Paso	-0.263	0.413	0.753	3.355	0.922
TAMU Commerce	-0.245	0.389	1.368	5.455	1.681
UH	-0.482	0.429	1.485	7.045	1.903
TAMU Corpus Christi	-0.260	0.439	1.719	8.053	2.104
Midwestern St	-0.354	0.442	1.627	7.314	1.971
UNT	-0.254	0.455	1.620	7.097	1.986
Lamar	-0.164	0.459	1.101	7.786	1.894
TX Woman	-1.178	0.491	1.246	4.389	1.496
UT Arlington	-0.262	0.562	1.715	7.608	2.060
TX Tech	-0.256	0.589	1.950	8.978	2.335
TX State	-0.250	0.607	1.932	8.670	2.334
TAMU	-0.248	0.640	1.625	7.437	1.837
UT San Antonio	-0.251	0.645	2.100	9.513	2.559
Tarleton St	-0.018	0.658	1.811	7.319	2.045
UT Dallas	-0.244	0.679	2.127	7.523	2.499
UT Austin	-0.244	0.705	1.627	6.972	1.782
UT Permian Basin	-0.229	0.742	1.276	4.133	1.289
UT Tyler	-0.234	0.905	1.462	7.753	1.814
Angelo St	-0.220	1.380	1.862	7.488	1.962
PVAMU	-1.207	1.411	2.261	8.832	2.908
UH Downtown	0.312	5.355	5.000	7.673	2.501

Table B3: Distribution of μ_{cj} across cells

Gender	Race	Family Income	Top Test Scores	Min	Median	Mean	Max	Std. Dev.
F	White	Low	Neither	1.530	4.673	4.502	7.389	1.373
M	White	Low	Neither	2.461	4.270	4.299	5.791	0.806
F	non-White	Low	Neither	-0.383	2.590	2.478	4.392	1.166
M	non-White	Low	Neither	0.946	3.418	3.499	6.336	1.284
F	White	Low	Math	1.305	4.976	5.003	7.523	1.348
M	White	Low	Math	5.298	7.314	7.213	9.513	1.120
F	non-White	Low	Math	-0.287	3.049	3.047	5.652	1.232
M	non-White	Low	Math	0.198	2.569	2.384	3.745	1.024
F	White	Low	Verbal	2.420	5.493	5.320	7.766	1.408
M	White	Low	Verbal	1.500	3.057	3.064	4.193	0.693
F	non-White	Low	Verbal	0.413	3.514	3.363	7.673	1.472
M	non-White	Low	Verbal	3.303	3.771	4.119	5.495	0.744
F	White	Low	Both	0.224	1.465	1.436	2.155	0.450
M	White	Low	Both	2.838	4.733	4.770	6.722	0.981
F	non-White	Low	Both	-1.207	2.201	2.027	3.880	1.155
M	non-White	Low	Both	1.238	3.749	3.908	5.855	1.173
F	White	Mid	Neither	-0.046	1.531	1.398	2.393	0.677
M	White	Mid	Neither	0.206	1.943	1.823	3.024	0.799
F	non-White	Mid	Neither	-0.292	0.090	0.050	0.312	0.142
M	non-White	Mid	Neither	-0.179	0.364	0.329	0.704	0.272
F	White	Mid	Math	0.478	1.495	1.516	2.259	0.505
M	White	Mid	Math	-0.007	0.041	0.037	0.115	0.026
F	non-White	Mid	Math	-0.271	-0.251	0.195	5.587	1.488
M	non-White	Mid	Math	0.737	1.903	1.800	2.458	0.475
F	White	Mid	Verbal	0.307	0.974	0.967	1.510	0.324
M	White	Mid	Verbal	0.400	0.422	0.423	0.440	0.012
F	non-White	Mid	Verbal	-0.279	0.482	0.428	0.974	0.330
M	non-White	Mid	Verbal	-0.079	-0.052	0.170	3.995	0.955
F	White	Mid	Both	-0.017	0.027	0.250	4.968	1.054
M	White	Mid	Both	0.182	0.228	0.222	0.262	0.021
F	non-White	Mid	Both	-0.175	0.145	0.190	0.593	0.199
M	non-White	Mid	Both	0.496	0.650	1.540	8.832	2.549
F	White	High	Neither	-0.164	0.236	0.221	0.475	0.172
M	White	High	Neither	-1.097	0.207	0.190	2.050	0.674
F	non-White	High	Neither	0.109	0.350	0.556	5.123	0.999
M	non-White	High	Neither	0.270	0.378	0.392	0.968	0.127
F	White	High	Math	-0.181	0.082	0.051	0.305	0.147
M	White	High	Math	0.292	0.319	0.319	0.345	0.016
F	non-White	High	Math	-1.178	0.435	0.319	0.738	0.433
M	non-White	High	Math	0.095	0.106	0.283	3.238	0.738
F	White	High	Verbal	0.274	0.290	0.538	6.241	1.215
M	White	High	Verbal	-0.057	-0.025	0.171	3.906	0.879
F	non-White	High	Verbal	0.042	0.058	0.057	0.073	0.009
M	non-White	High	Verbal	1.352	1.387	1.384	1.413	0.016
F	White	High	Both	0.072	0.078	0.079	0.084	0.003
M	White	High	Both	0.154	0.170	0.170	0.181	0.006
F	non-White	High	Both	-0.253	-0.223	-0.208	-0.065	0.045
M	non-White	High	Both	-0.249	-0.239	-0.235	-0.175	0.017

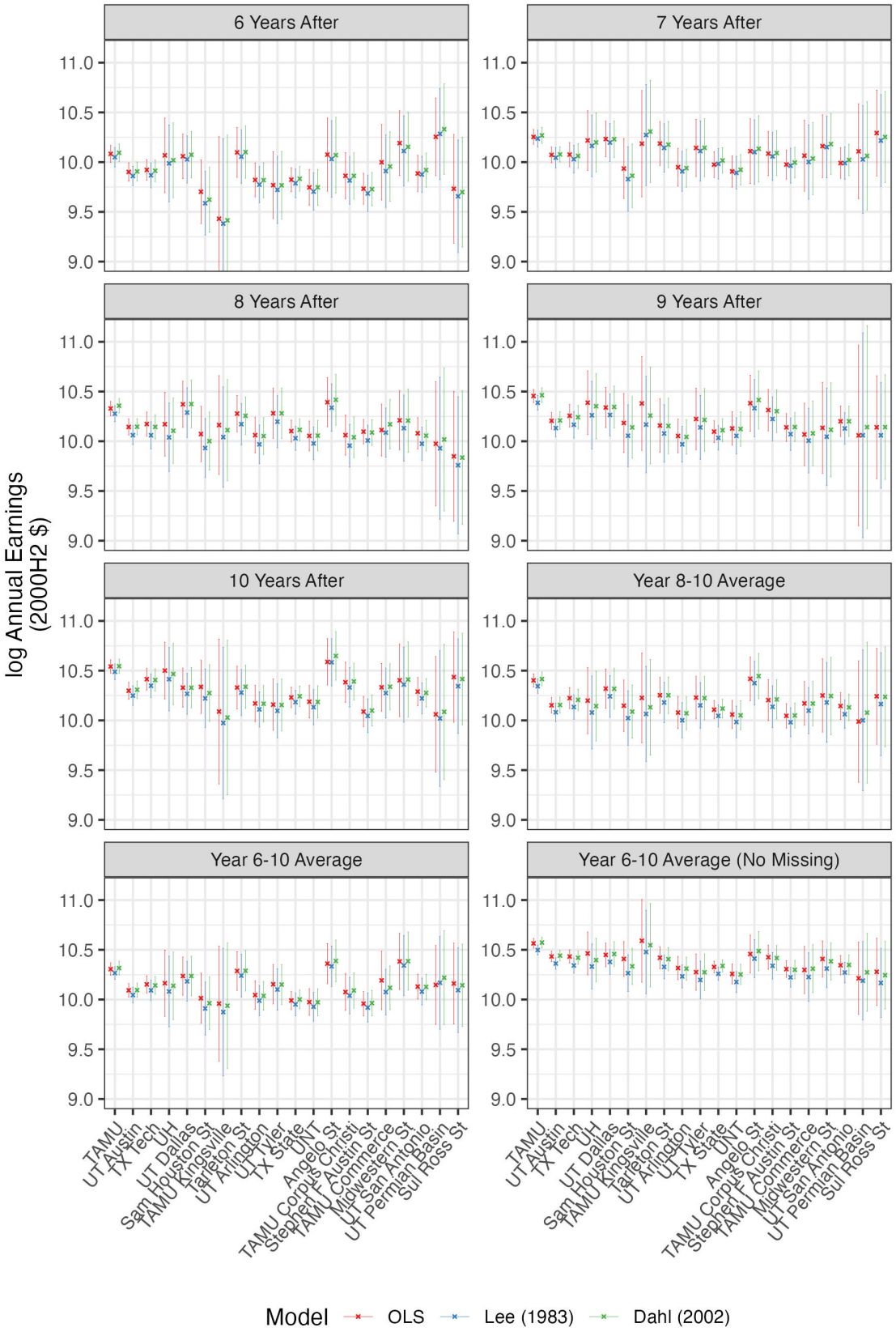


Figure B1: Estimated $\widehat{\alpha}_j$ with different Outcome Variables

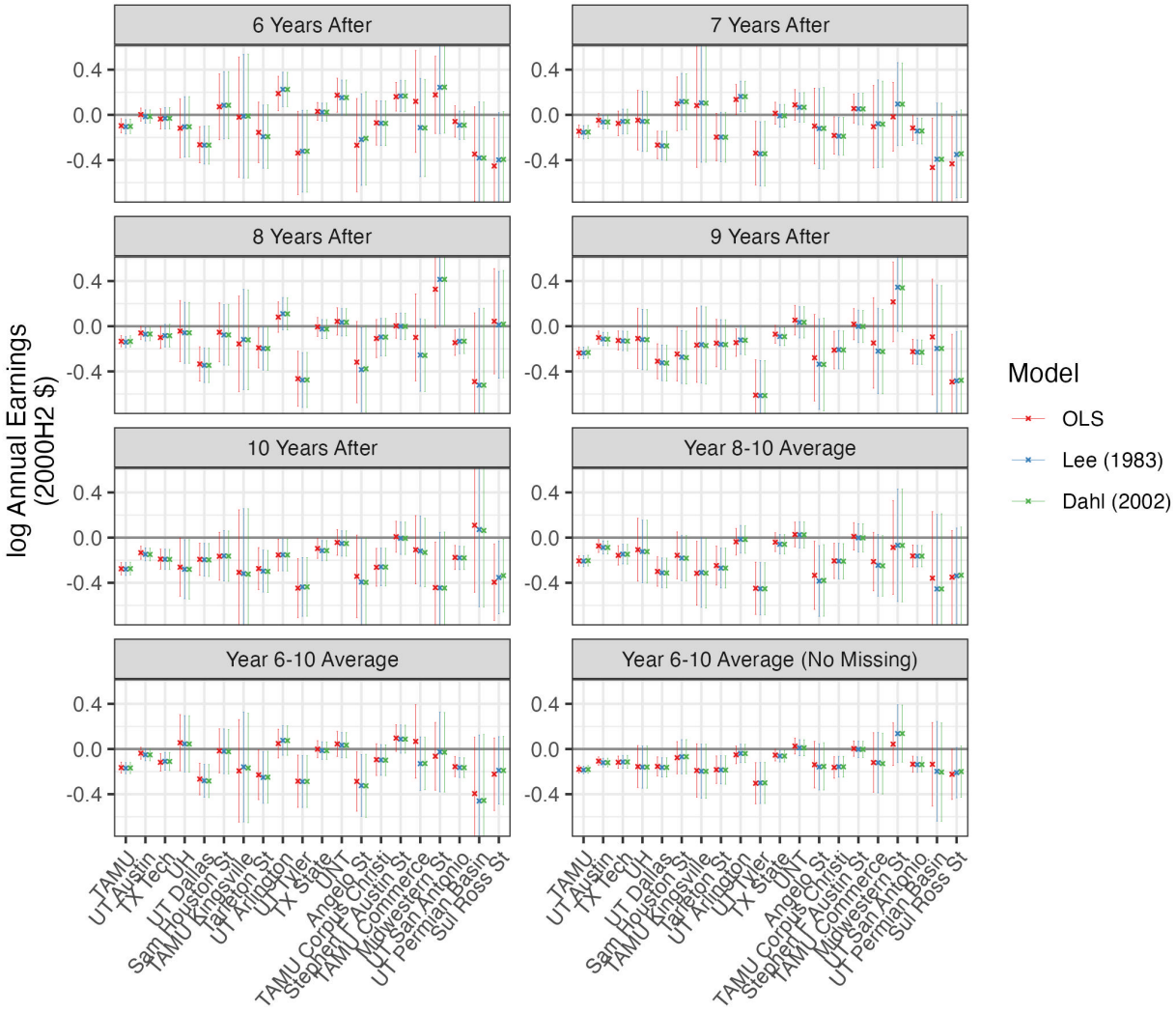


Figure B2: Estimated Returns for Female Students (Relative to Male)

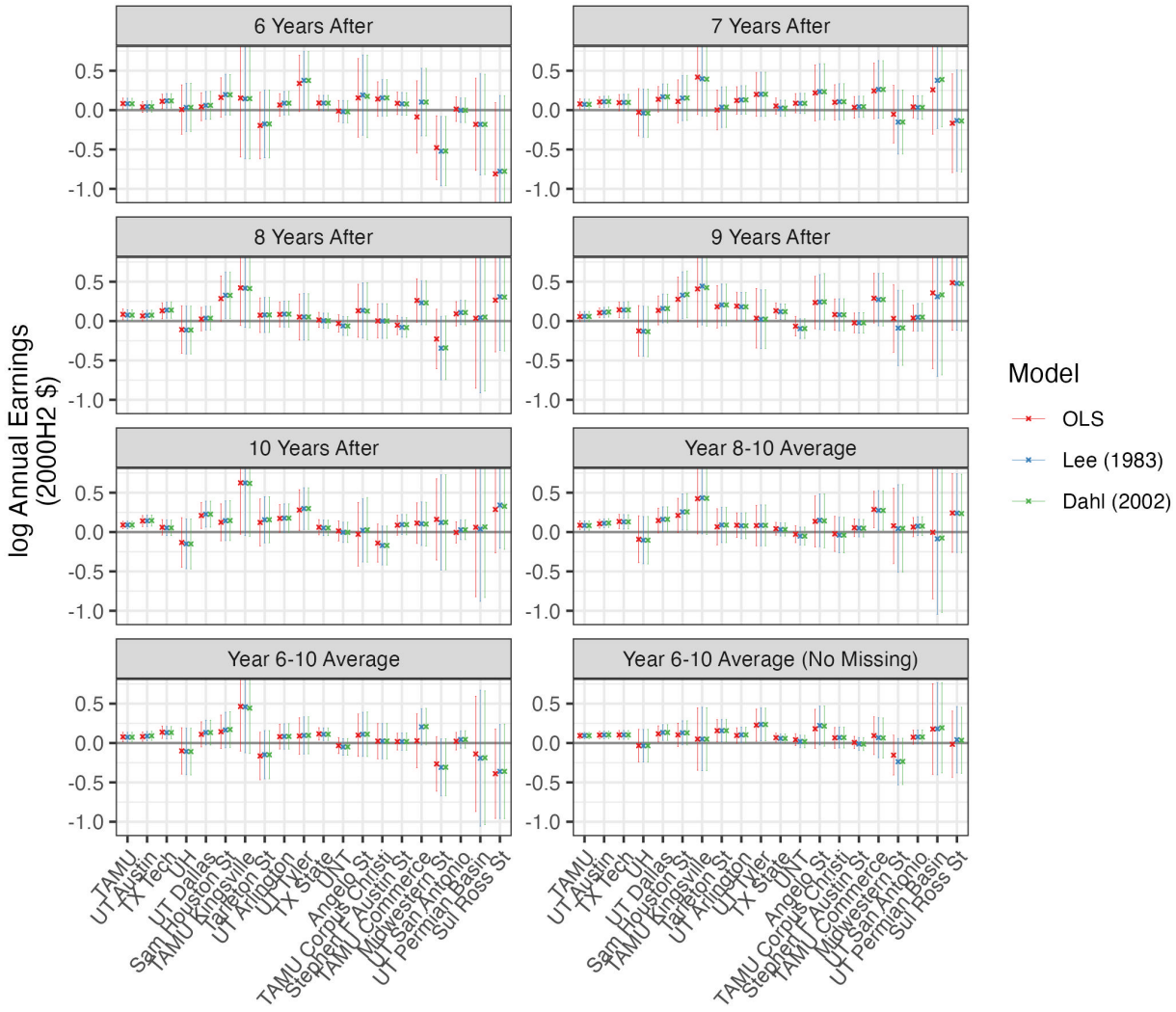


Figure B3: Estimated Returns for High-Income Students (Relative to Mid-Income)

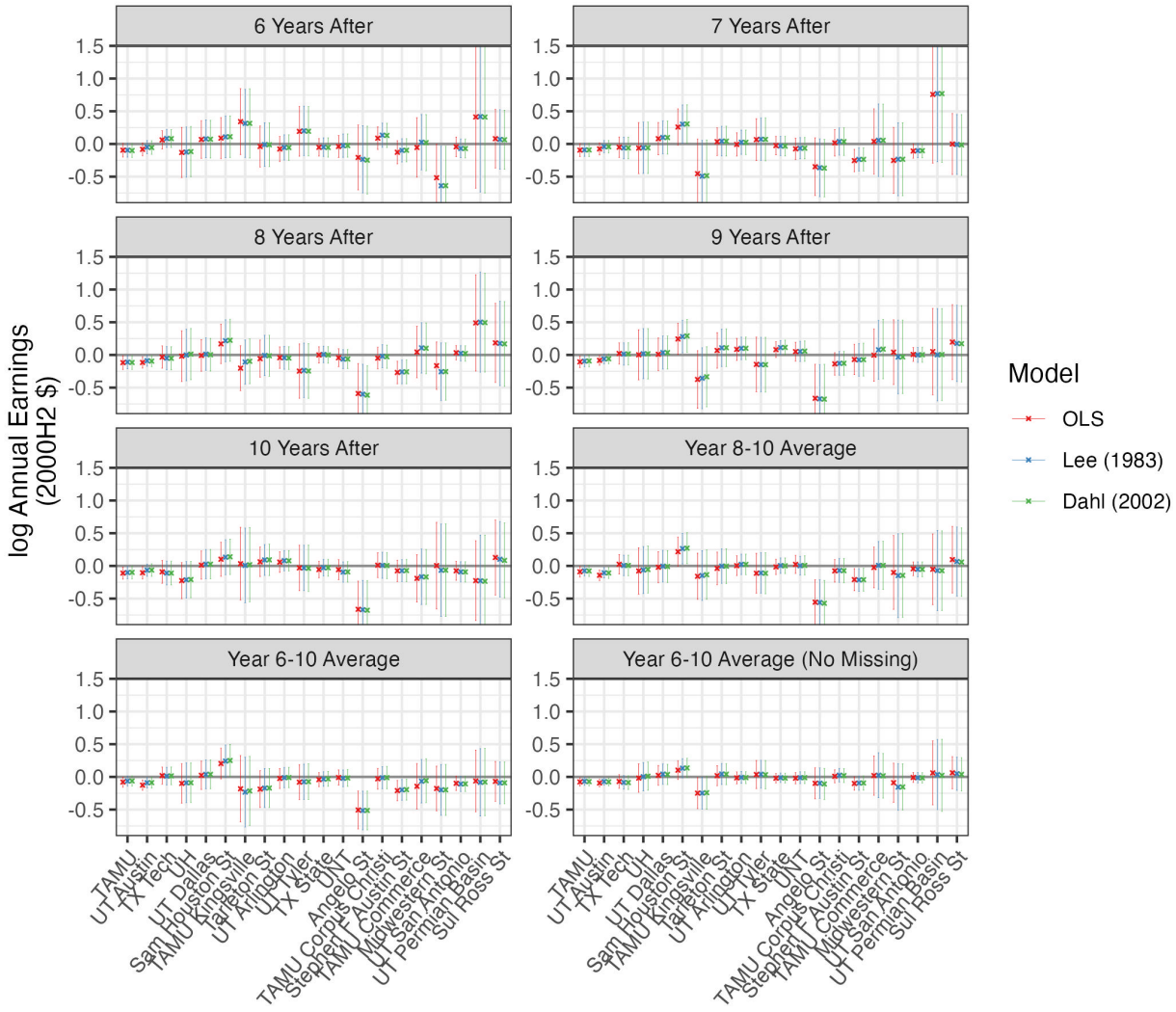


Figure B4: Estimated Returns for Low-Income Students (Relative to Mid-Income)

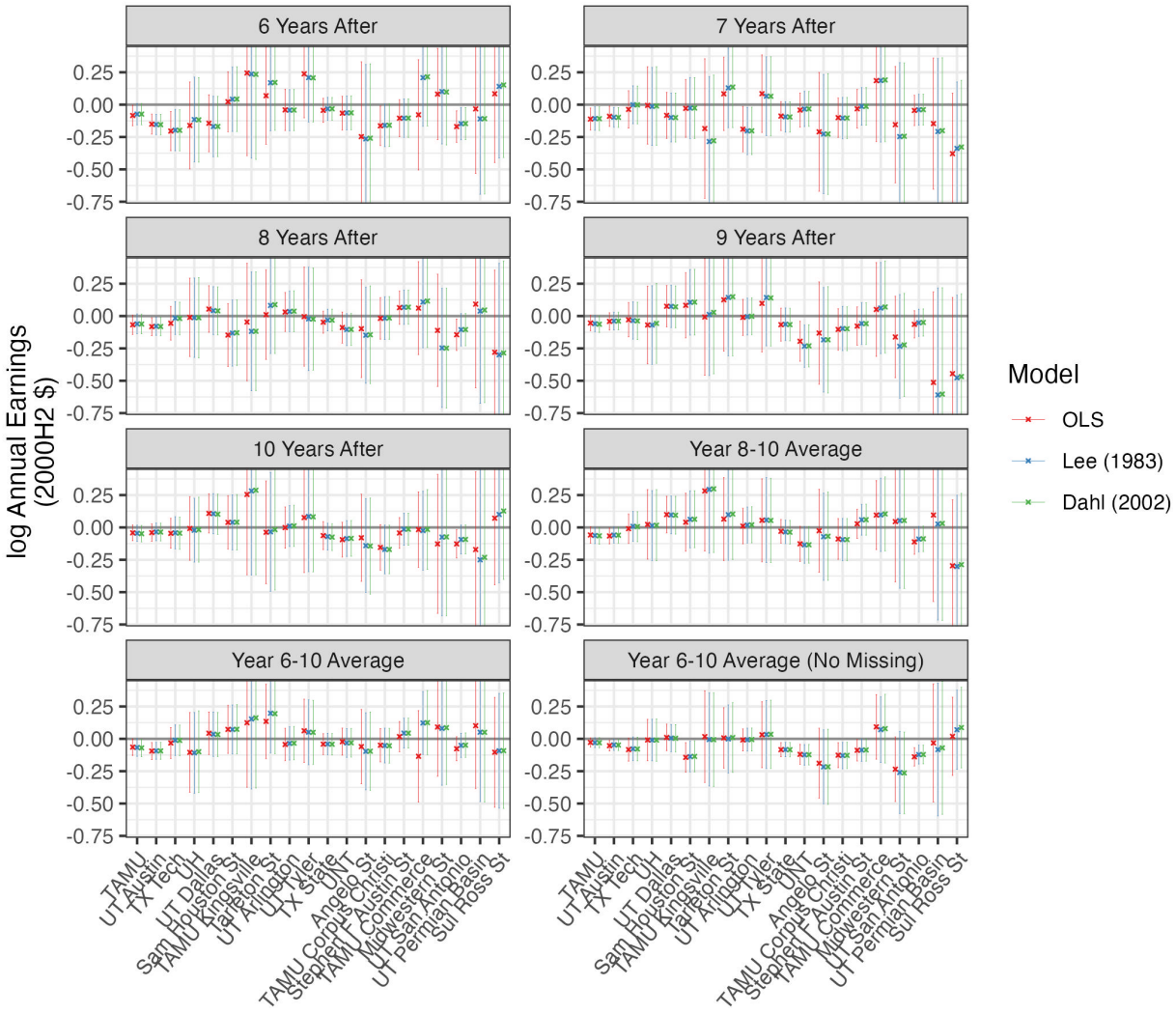


Figure B5: Estimated Returns for Non-White Students (Relative to White)

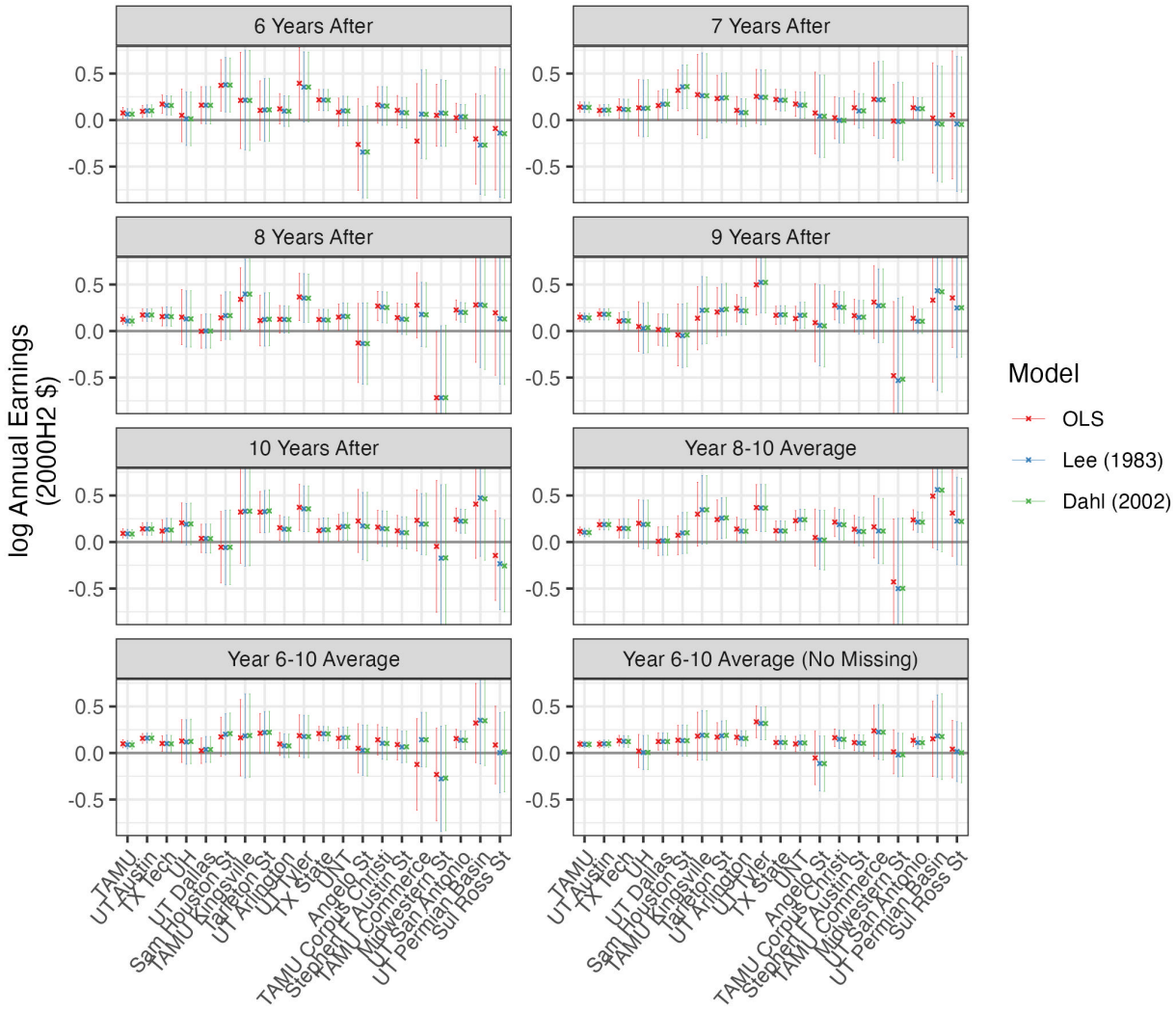


Figure B6: Estimated Returns for Top 10% Students

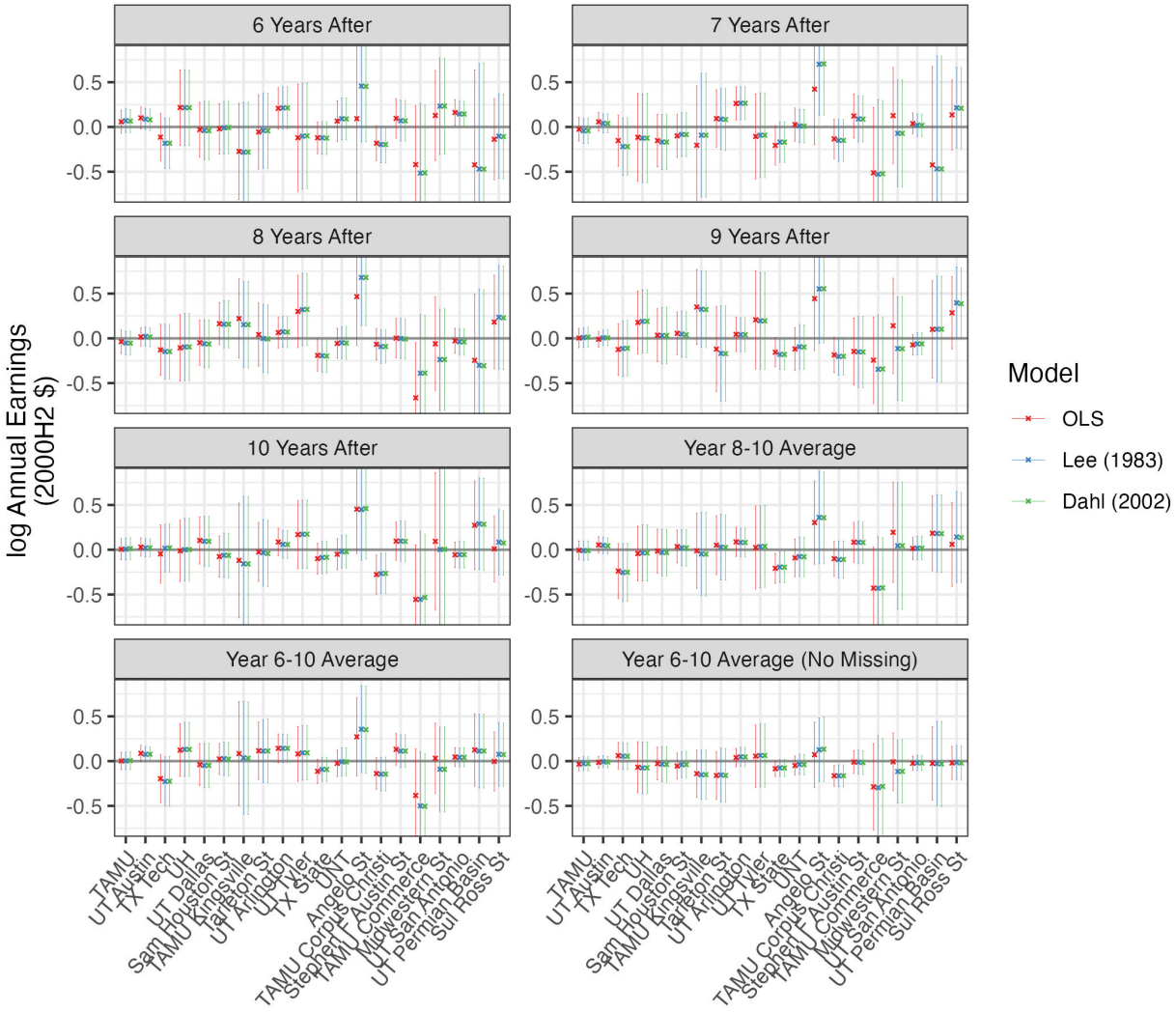


Figure B7: Estimated Returns for Economically Disadvantaged Students

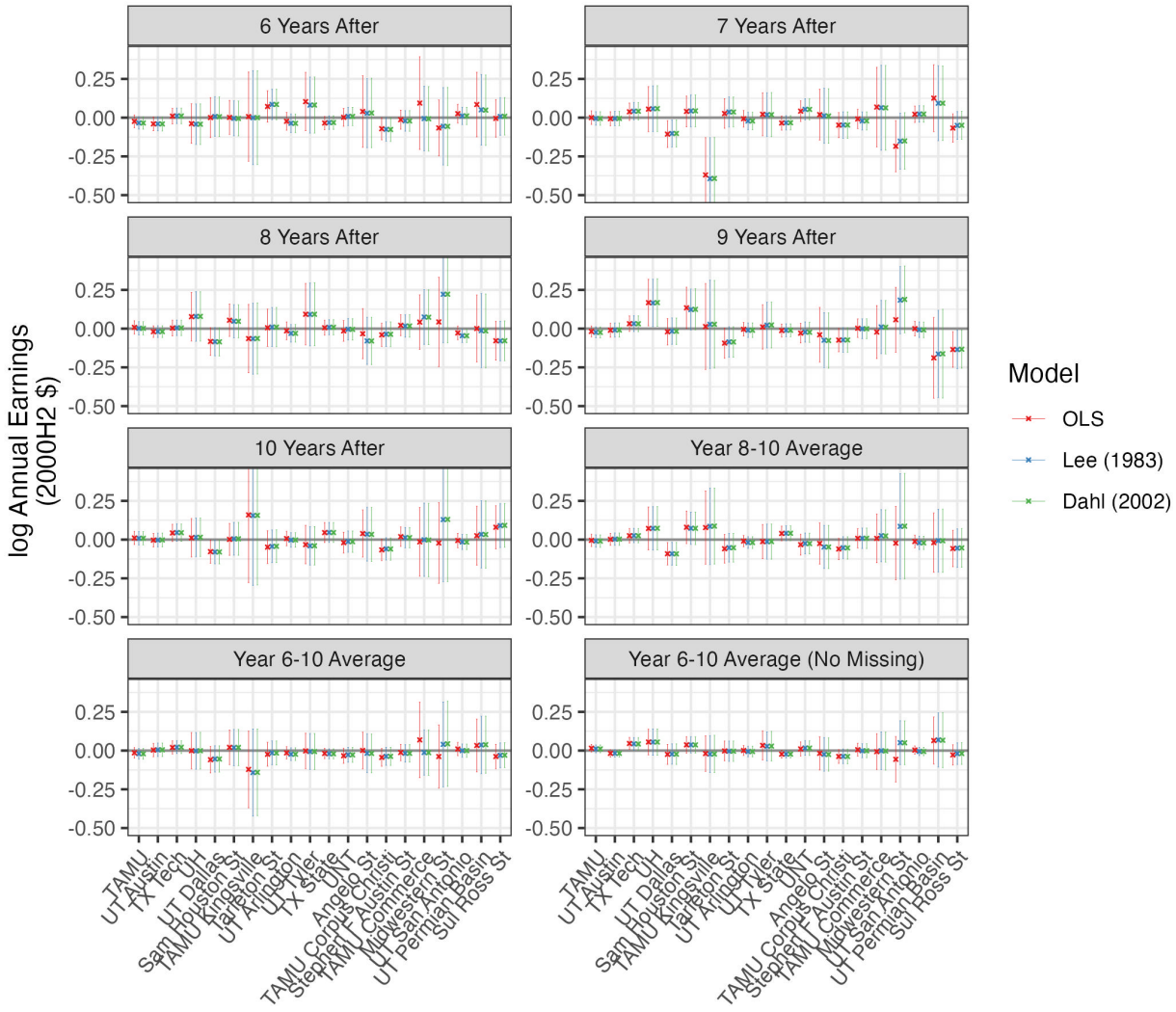


Figure B8: Estimated Returns for Reading Scores

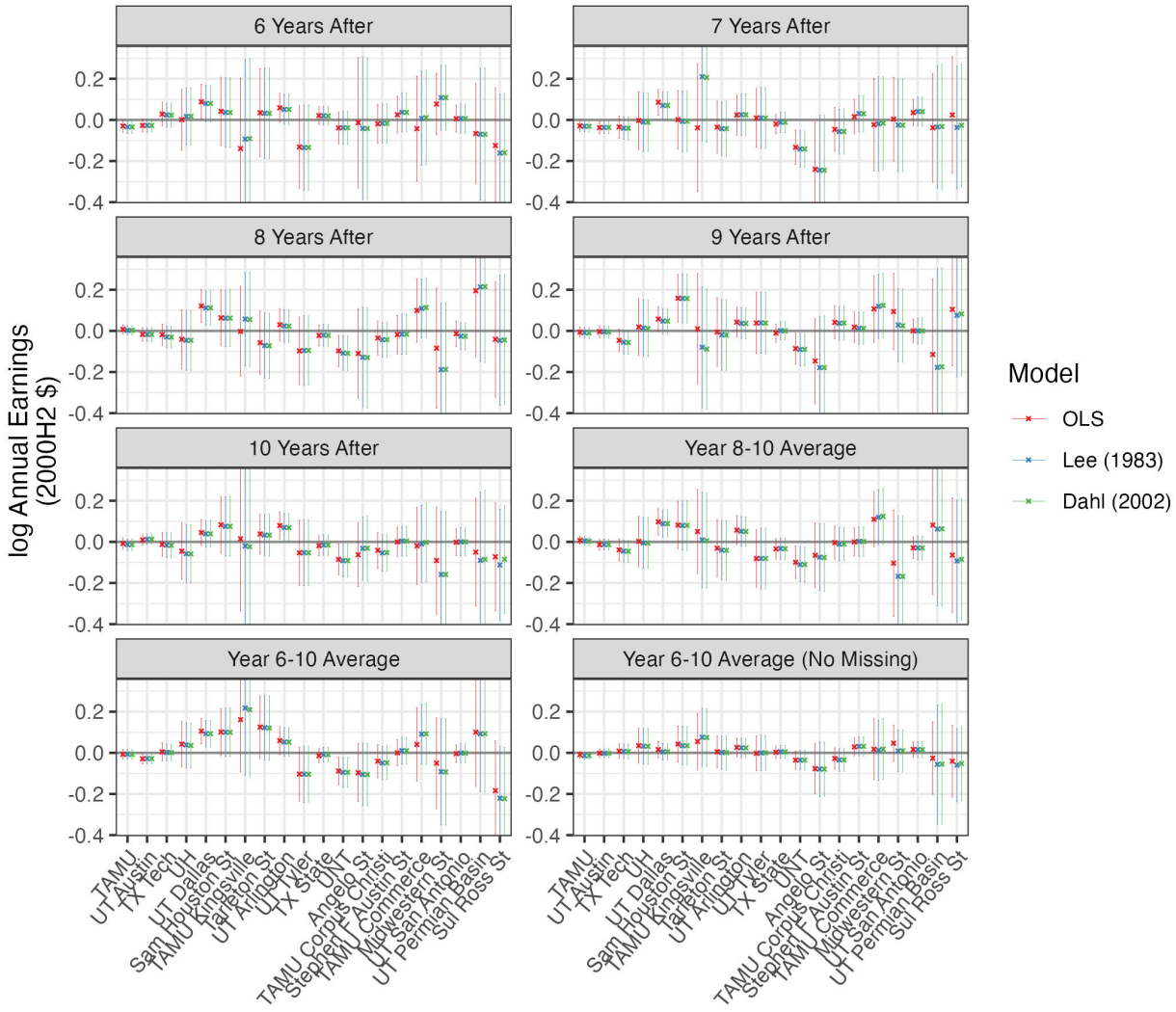


Figure B9: Estimated Returns for Writing Scores

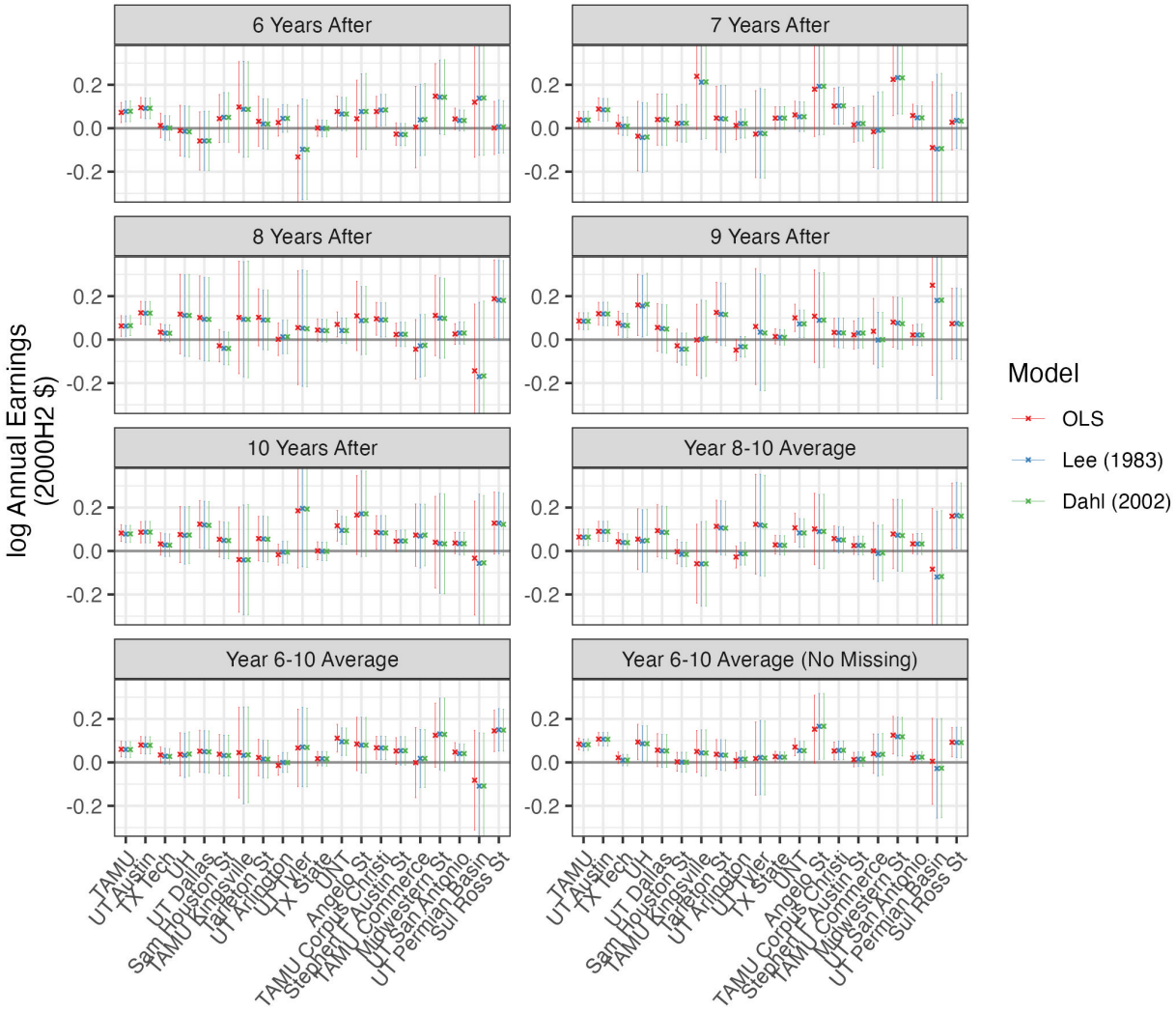


Figure B10: Estimated Returns for Math Scores

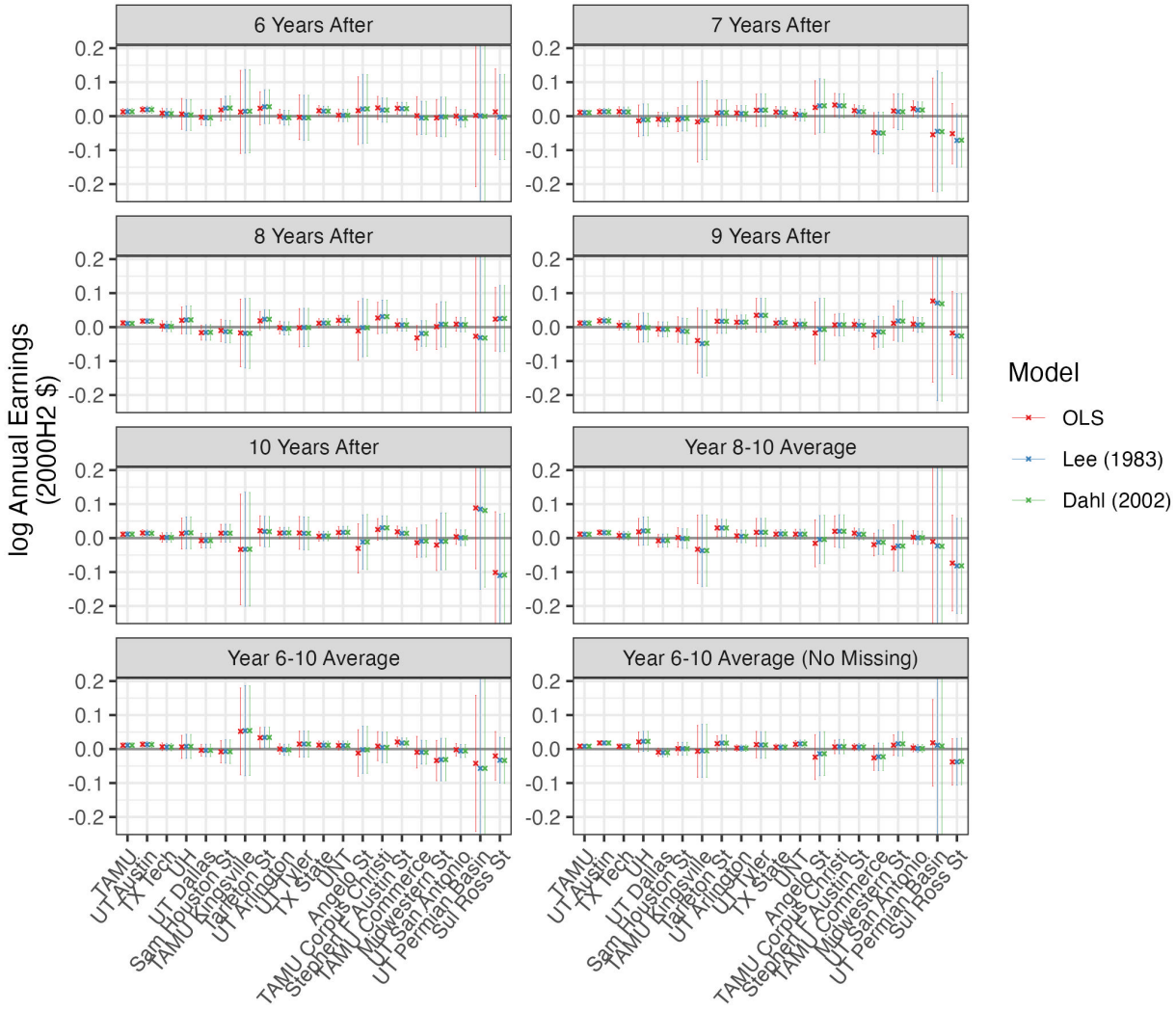


Figure B11: Estimated Returns for Different Hometown Population Densities

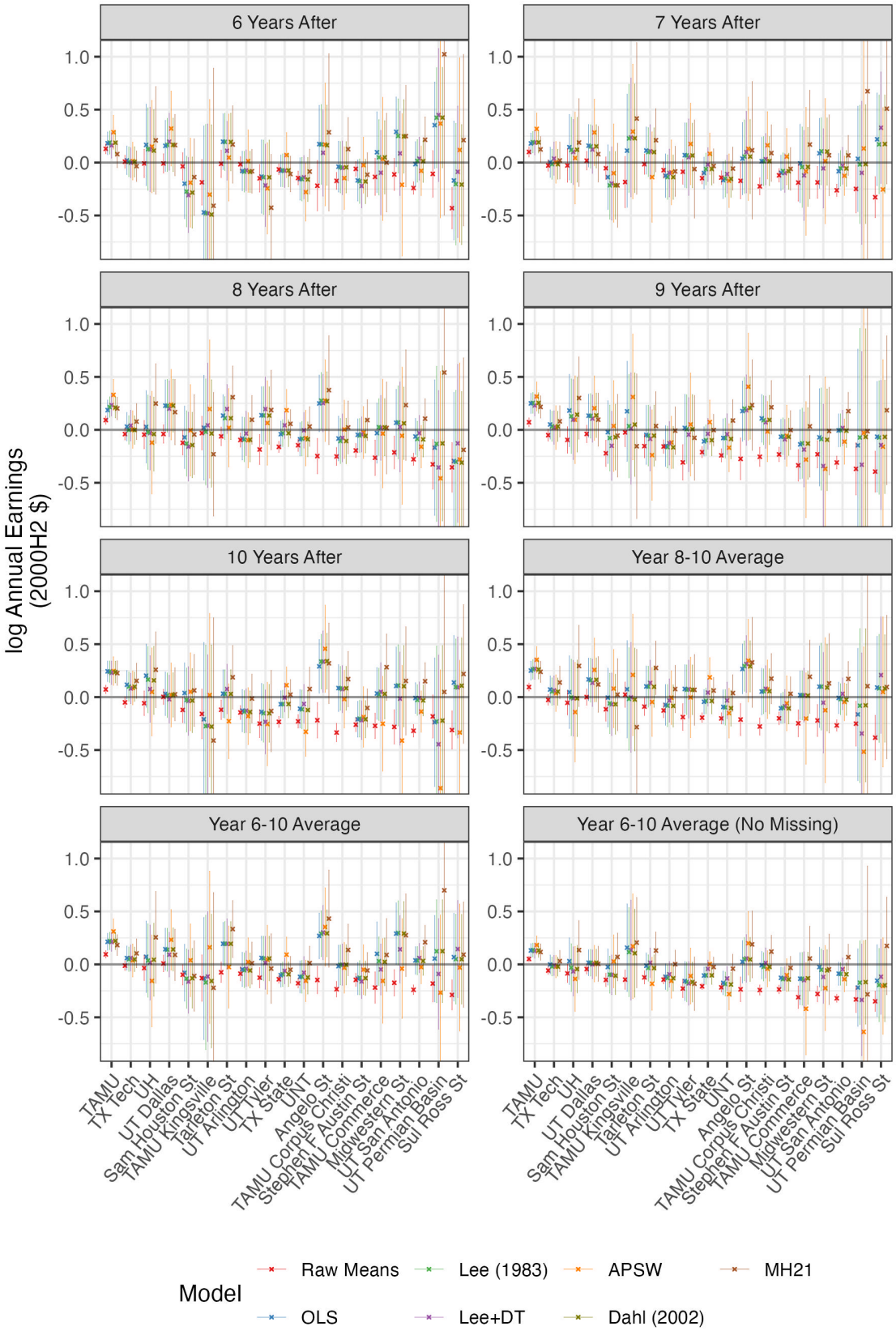


Figure B12: Estimated $\widehat{\alpha}_j$ with UT Austin as the Base School

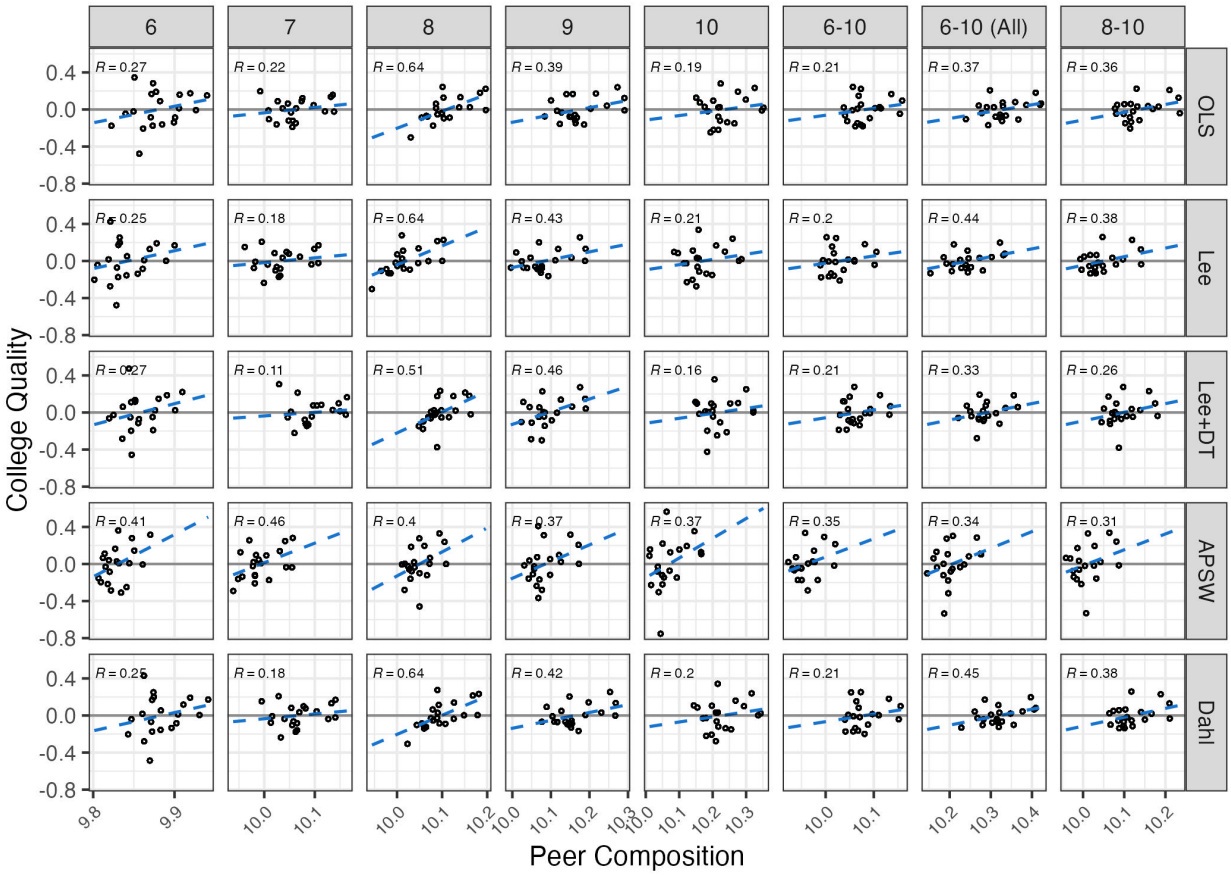


Figure B13: Estimated Correlation between Peer Composition and College Quality