THE ROLE OF DISTANCE IN COLLEGE UNDERMATCHING

by

Lois Miller

and

Humberto Barreto

DePauw University

Comments Welcome
Contact Author: hbarreto@depauw.edu
DePauw University Economics Working Papers Series, 2017-01

16 June 2017

Keywords: college application, Mechanical Turk, matching, sorting, education, inequality, social capital

JEL Classification Codes: I21, I23, I24

Acknowledgements

We thank the J. William and Dorothy A. Asher Funds for providing financial support. We also thank Frank M. Howland and Allison Roehling for comments.

Abstract

This paper explores factors explaining why so many high-achieving, low-income students apply to and enroll at universities with relatively low academic standards, despite generous financial aid packages and evidence that these students would be successful at colleges that are more selective. Amazon's Mechanical Turk was used to gather data, and the entire file is freely available at academic.depauw.edu/hbarreto_web/working. A probit analysis confirms an established result that low-income students are more likely to undermatch. The key result is that as the distance between a student's home and the university they attend increases, the probability that the student will undermatch decreases. At a distance of 500 miles between a student's home and college, the difference in the probability of undermatching between low-income students and high-income students is 25.5 percentage points. At 3,000 miles, the gap is only 8.7 percentage points.

I. Introduction

The college application process is a complicated, multi-step game played by students and schools. The stakes are high, impacting life prospects and driving societal mobility and inequality trends. When a student attends a college that is much less academically rigorous than they could handle, it is called *undermatching*. This phenomenon is not randomly distributed across applicants.

Hoxby and Avery (2013, 2, footnote omitted) found that "a large number—probably the vast majority—of very high-achieving students from low-income families do not apply to a selective college or university." In addition, they provide evidence that this is not a small, inconsequential issue: "We estimate that there are at least 25,000 and probably about 35,000 low-income high achievers in each cohort in the United States." (Hoxby and Avery, 2013, 14-15, footnote omitted.) Perhaps somewhat counter-intuitively, most of these students are not densely concentrated, inner-city residents, instead they are isolated and in non-urban areas. They are, like dust, thinly spread out and invisible, but once swept up and aggregated, they form a strikingly large group.

Increasing attention is being paid to the behavior and application strategies of these students. While there are many factors that affect college application and enrollment decisions, we focus on the student's family income and the distance between a student's home and the college they attend. In addition to verifying previous findings that low-income students have a higher probability of undermatching using a novel data source, our work focuses on the role of distance between a student's home and the college they attend. We find that the farther from home a student attends or considers attending college, the less likely they are to undermatch. The magnitude of this effect varies with income, so that

at farther distances from home, the effect of income on undermatching is smaller than at distances close to home. This result has implications for public policy and may provide a way for admission offices at selective colleges to tap into the high-achievement, low-income talent pool, which is much deeper than we thought.

The remainder of this paper is organized as follows. Section II presents a brief review of previous literature on the student college choice process and how it is affected by socioeconomic status, as well as literature on undermatching. Section III provides an overview of the methodology of this study, and describes Amazon's Mechanical Turk, which was used to collect survey data. Section IV gives empirical results and Section V concludes with implications from the results and areas for further research.

II. Literature Review

There are two main steps in the college decision process—whether to attend college and where to attend college. At first, research focused mainly on whether students attended college or not, without giving much consideration to the specific universities that students were choosing. Chapman (1981) turned the conversation away from exclusively focusing on whether students continued their education beyond high school or not by presenting a non-mathematical model of student college choice in which the student is choosing among various schools.

In the last decade, a body of literature has emerged surrounding the more specific issue of undermatching. Different measures of academic achievement and rigor of college can be used, and different thresholds can be used to define high-achieving, which affect the rates of undermatching observed (Winston and Hill, 2005, 19.7). The most common

yardstick of academic achievement in high school is a student's SAT or ACT score, since they are standardized and the average SAT or ACT scores for the student bodies of most colleges are readily available.

Winston and Hill (2005, 19.1) use the national population of SAT and ACT test takers to analyze the issue of a very low proportion of low-income students being represented at the United States' most selective colleges. They see two possible explanations for this discrepancy—either low-income students are not high-achieving enough to attend selective private schools ("the COFHE schools"), or there exist low-income, high-achieving students who are being excluded in favor of higher-income students (Winston and Hill, 2005, 19.1). Results depend on the threshold used, but they find clear evidence that there exist enough high-ability, low-income students for the COFHE schools to be able to mirror the national low-income distribution of high-achieving students.

A natural question that follows is *why* these students are not attending selective colleges. Hoxby and Avery (2013) find that a large portion of these students do not even apply to selective schools. They define high-achieving students as those who scored in the top 10% of students on the SAT or ACT test (1300 on the combined Critical Reading and Mathematics sections of the SAT or 29 composite ACT score), and self-reported a grade point average of A- or higher in high school (Hoxby and Avery, 2013, 10).

Using the schools to which each student sends their SAT or ACT scores as a proxy for which schools that student applies to, they identify two distinct groups of low-income high achievers by their application patterns. Some low-income, high-achieving students apply in a very similar manner to high-income high achievers, they are labeled

achievement-typical. They follow the advice of expert counselors, applying mostly to peer schools, a few reach schools, and include safety schools, along with their state's flagship university (Hoxby and Avery, 2013, 23-24). Reach schools are defined as schools which have a median test score more than 5 percentiles above the student's own, peer schools are those where the school's mean test score is within 5 percentiles of the student's own, and safety schools are those which have median test scores between 5 and 15 percentiles below the student's own (Hoxby and Avery, 2013, 21). Eight percent of low-income high-achievers fall into this achievement-typical category, applying to "at least one peer college, at least one safety college with a median score not more than 15 percentiles lower than their own, and…no nonselective colleges" (Hoxby and Avery, 2013, 26).

Income-typical students are low-income, high-achieving students who apply using a different strategy. Comprising 53% of low-income high-achievers, they "apply to no school whose median score is within 15 percentiles of their own, and they do apply to at least one nonselective college." (Hoxby and Avery, 2013, 26). Finally, the remaining 39% of low-income, high-achieving students use a variety of strategies that do not fit either profile, and do not show a clear pattern (Hoxby and Avery, 2013, 27-28).

To assess factors that are associated with a student's choice of where to apply to college, Hoxby and Avery use a "conditional logit model in which a student can apply to all colleges in the United States but decides to apply only to some" (Hoxby and Avery, 2013, 28). Results show that high-income students strongly favor reach colleges, disfavor safety colleges, strongly disfavor nonselective institutions, and have a mild preference for in-state schools and their state's flagship university. They dislike high net costs but like high sticker prices, and like higher per-student resources. Finally, they dislike distance, but the

quadratic term of distance is associated with an increase in probability of applying, which implies that these students only dislike distance up to a point (Hoxby and Avery, 2013, 30). Low-income students strongly favor nonselective institutions. They disfavor high sticker prices but do not have a preference for net costs, and favor higher per-student resources, but less so than high-income students do. Low-income students disfavor distance within 100 miles, and are indifferent to distance for schools farther than 100 miles away (Hoxby and Avery, 2013, 31).

Two further conditional logit models demonstrate that, conditional on applying to a specific college, high-income and low-income students do not behave differently in their enrollment or progress towards a degree (Hoxby and Avery, 2013, 31). Thus, it is primarily in the application stage that low-income, high-achieving students who could attend selective colleges are being lost.

Geography plays a key role in determining income- versus achievement-typical behavior. Hoxby and Avery (2013, 38-39) show that "65 percent of achievement-typical students live in the main city of an urban area, whereas only 30 percent of income-typical students do" and only 21 percent of achievement-typical students live in a nonurban area, compared to 47 percent of income-typical students. The achievement-typical students are much more geographically concentrated, since "the radius needed to gather 50 high achievers is 37.3 miles for the average income-typical student, but only 12.2 miles for the average achievement-typical student" (Hoxby and Avery, 2013, 42).

Although Hoxby and Avery's study has the advantage of being nationwide, there have been several studies on undermatching restricted to certain areas of the United States. Bowen, Chingos, and McPherson (2009) focus on high school seniors in North Carolina in

1999, for whom the researchers have a large body of data including race/ethnicity, gender, and socioeconomic status. They aim to determine how many students have undermatched in their college choices, and if there are "disproportionate numbers of undermatches among certain groups of students—defined by race/ethnicity, family background, level of high school attended, academic qualifications, and rural or urban location" (Bowen, Chingos, and McPherson, 2009, 100). The authors measure a student's ability to gain access to selective schools using a combination of their SAT/ACT scores and self-reported high school GPA. Since NC State and UNC-Chapel Hill account for over 90 percent of enrollments in the top-tier selectivity institutions in North Carolina (SEL A), a student is assumed to be able to get into a SEL A institution if more than 90 percent of students with the same test score/GPA combination who applied to NC State or UNC-Chapel Hill were admitted (Bowen, Chingos, and McPherson, 2009, 101). Ninety percent was chosen as a cut-off to be conservative in eligibility criteria, so that the results are more likely to underestimate the number of undermatches than overestimate them (Bowen, Chingos, and McPherson, 2009, 102).

Results showed that of the 6,217 students who met the eligibility criteria, 40 percent undermatched by not attending a SEL A institution, enrolling instead in a SEL B, an HBCU (Historically Black Colleges and Universities), a two-year college, or no college (Bowen, Chingos, and McPherson, 2009, 102). Family income and parental education have strong effects on enrollment patterns, since students are more likely to undermatch the lower their family income, and the less education their parents have. These effects remained when controlling for quality of high school, high school GPA, and SAT scores.

Bowen, Chingos, and McPherson (2009, 105) find that of students who undermatch, 64% don't apply to any SEL A institutions, 28% are accepted but don't enroll, and 8% are rejected. Bowen, Chingos, and McPherson (2009, 104) hypothesize that "the primary forces leading to such high undermatch rates were a combination of inertia, lack of information, lack of forward planning for college, and lack of encouragement," noting that these are the factors emphasized by the Chicago Consortium in a report on undermatching (Roderick, et al., 2008).

Another study that focuses on a specific area of the United States takes advantage of an admissions policy in Texas to explore the impact of *a priori* knowledge on admissions behavior (Lincove and Cortes, 2016, 3). The Texas Top 10% plan allows students who rank in the top 10% of their class during their junior year to be automatically admitted to all Texas public universities (Lincove and Cortes, 2016, 5). The researchers compare public school students who qualify for the Texas Top 10% plan to those who graduate in the top 11-25% of class rank, who "have a high probability of admissions in a holistic process, but without certainty" (Lincove and Cortes, 2016, 6). The sample is limited to students who are either low income (family income less than $40,000) or high income (family income greater than $80,000) to allow for comparisons of how *a priori* knowledge of admission affects the income groups differently (Lincove and Cortes, 2016, 8).

Although they use the same terminology as Hoxby and Avery (2013), Lincove and Cortes (2016) use slightly different measures of safety, match, and reach schools. A safety school has a median SAT score more than 10 percentile points below the student's, a closely-matched school's median SAT is within 10 percentile points of the student's own, and a reach school has a median SAT score more than 10 percentile points higher than the

student's own. Using this definition, "34.4 percent of all Texas public high school graduates who enroll at Texas public universities are undermatched by at least 10 percentile points in enrollment" (Lincove and Cortes, 2016, 15-16).

Dividing students who have SAT scores in the top 25% into four subgroups defined by class rank (top 10% or top 11-15%) and family income, descriptive statistics show that top 11-25% students are more likely to apply to a safety school than top 10% students, regardless of income. High-income students of all class ranks are similarly likely to apply to at least one closely-matched school, but low-income students are more likely to apply to closely-matched schools if they have automatic admissions. Results are similar for enrollment rates (Lincove and Cortes, 2016, 16-17).

In their regression, Lincove and Cortes (2016, 18) control for "student demographics (race, ethnicity, gender, and whether the student's mother attended college), observable college readiness (percentile rank of SAT scores and Texas high school exit exam scores, and the number of AP or IB courses completed in high school), and high school fixed effects" in addition to including income, admissions status, and the interaction between income and admission status. Results show that students with automatic admissions were 21.3 percentage points less likely to undermatch and 15.4 percentage points more likely to apply to a closely-matched school. Low-income students were 4.4 percentage points more likely to apply to a safety school, 14.8 percentage points less likely to apply to a closely-matched school, and 20.6 percentage points less likely to apply to a flagship campus, compared to high-income students (Lincove and Cortes 2016, 18). Low-income students with automatic admissions were 8.7 percentage points more likely to apply to a closely-matched school and 6.5 percentage points more likely to apply to a

flagship campus than high-income top 11-25% students. Overall, results show that "top 10% eligibility reduces undermatch overall, and also appears to have a larger effect on low-income students than high-income students" (Lincove and Cortes, 2016, 19).

Lincove and Cortes (2016, 21) also find that low-income students are much more affected by proximity of the college than high-income students, across all class ranks. A low-income student is much more likely to apply and enroll in a college that is within commuting distance than a high-income student, but beyond 60 miles, the effects of distance from home are similar between low-income and high-income students (Lincove and Cortes, 2016, 21).

In summary, there is evidence that low-income students with high levels of academic high-school achievement perform well in college, especially at selective schools. However, there is a surprisingly large population of these students who undermatch by only applying to and attending institutions that are much less rigorous than they could handle. These students tend to be spread out geographically, where they are not around many other high-achieving students.

Proximity to home is an important factor in all students' college choices, but it affects low-income students more than high-income students. Low-income students may be more risk-averse than high-income students, evidenced by the equalizing effect that the safety of *a priori* admission had across incomes in Lincove and Cortes (2016). Moving far away for college implies taking more of a risk, which could explain why low-income students are more likely to stay close to home.

III. Methodology

After receiving IRB approval, our data were obtained through a survey distributed on Amazon's Mechanical Turk: "a crowdsourcing web service that coordinates the supply and the demand of tasks that require human intelligence to complete" (Paolacci, Chandler, and Ipeirotis, 2010, 411). It has many uses, but has become particularly popular among social scientists to collect experimental data through surveys. Although it has not gained much popularity in economics, it has been shown to be a reliable way to quickly obtain high-quality data at low cost (Buhrmester, Kwang, and Gosling, 2011, 3).

Mechanical Turk got its name from a chess-playing automaton hoax from the 18th century. This machine was actually operated by a hidden person, but was presented as pure machine (Paolacci, Chandler, and Ipeirotis, 2010, 411). Amazon has given MTurk the slogan, "Artificial Artificial Intelligence" based on the idea that "there are still many things that human beings can do much more effectively than computers" (www.mturk.com/mturk/help?helpPage=overview). It is an online labor market that can easily match workers (employees who will be paid to do tasks) to requesters (employers who pay per task completed). The Human Intelligence Tasks, or HITs, are posted by requesters to be completed by workers for a monetary reward. Workers decide which tasks they will complete from the online database, which they can sort based on the reward amount, maximum time allotted, and tags associated with the type of task. Each task is listed with a short description and requesters can also limit which workers are eligible to complete their tasks based on certain criteria such as country of residence or rate of accuracy in previous HITs. All workers and requesters are anonymous, and requesters can

only link responses to unique worker IDs assigned by Amazon (Paolacci, Chandler, and Ipeirotis, 2010, 411-412).

Rewards paid to workers are generally very low, between $0.01 and $1.00 per simple task. Workers typically make much less than a typical minimum wage, and are usually internally motivated, completing tasks for enjoyment rather than monetary gains (Buhrmester, Kwang, and Gosling, 2011, 3). Somewhat surprisingly, "even at low compensation rates, payment levels do not appear to affect data quality," although offering higher rewards on MTurk generally allows data to be collected faster (Buhrmester, Kwang, and Gosling, 2011, 4).

Further, using subjects from MTurk does not pose a threat to obtaining a representative sample. Poalacci, Chandler, and Ipeirotis (2010, 412) found their MTurk sample to be "slightly younger than the U.S. population as a whole and the population of Internet users," whereas Buhrmester, Kwang, and Gosling (2011, 4) found MTurk participants to be older than participants in a standard Internet sample. They also found similar gender splits among MTurk participants and Internet participants, roughly 55% female (Buhrmester, Kwang, and Gosling, 2011, 4). Poalacci, Chandler, and Ipeirotis (2010, 412) found MTurk users to have higher levels of education but lower income than the general United States population. Both studies found samples from Mechanical Turk to be more diverse than traditional American college samples (Buhrmester, Kwang, and Gosling, 2011, 4; Paolacci, Chandler, and Ipeirotis, 2010, 412).

One concern with conducting surveys on MTurk is that users will rush through and randomly click answers to questions without reading them, thus producing unreliable data. To combat this problem, requesters can implement attention checks into surveys to test

whether participants are thoughtfully replying. Attention checks are extremely easy questions, and if participants fail to answer correctly, requesters can reject their work and withhold payment. Paolacci, Chandler, and Ipeirotis (2010, 415) conducted a study comparing a Mechanical Turk sample to a traditional subject pool at a large Midwestern U.S. university and to an Internet sample obtained from visitors of online discussion boards. They included an attention check, "*While watching the television, have you ever had a fatal heart attack?*" embedded into a series of questions with responses ranging from "*Never*" to "*Often*" (Paolacci, Chandler, and Ipeirotis, 2010, 415). Results shows that MTurk users had the lowest proportion of participants fail the attention check by not selecting "*Never"*, although "the number of respondents who failed the catch trial is very low and not significantly different across subject pools" (Paolacci, Chandler, and Ipeirotis, 2010, 416).

Mechanical Turk provides a shockingly cheap and efficient way to collect data that is just as reliable as traditional surveys. The total paid for the 1,073 responses used in this project was $1,500, and data were collected in 9 days. Approximately 500,000 workers are part of the Mechanical Turk workforce, so even after placing several restrictions on participation, there were plenty of workers eligible and willing to complete the survey. For our project, participants were required to be in the United States, between the ages of 18 and 25, be currently attending or have attended college, and remember and be willing to report their SAT/ACT scores. Surveys were released on MTurk in batches of 50 with a few smaller batches at the beginning and end, and batches were usually complete within 2 to 4 hours.

The survey was designed and implemented with Qualtrics software. The "display logic" and "skip logic" features allowed certain questions to be asked based on respondents'

previous answers and certain questions to be skipped depending on answers to previous questions. The skip logic was particularly useful for implementing attention checks, so respondents who failed an attention check were immediately sent to the end of the survey. Miller (2017) provides full documentation for this research. Along with the final dataset, the entire survey questionnaire, annotated to include how questions were developed, is available at academic.depauw.edu/hbarreto_web/working.

IV. Results

The dummy dependent variable *undermatch* takes a value of 1 if the student has undermatched by attending a school with a median SAT that is 15 or more percentiles below their own. If a student's college has a median SAT score that is below the student's, but less than 15 percentiles below or the student's college has a higher SAT score, they have not undermatched.

Survey respondents are asked whether they took the SAT, the ACT, or both. If they had taken the SAT or both, they were asked to report their combined Critical Reading and Mathematics SAT scores. If they had only taken the ACT, they were asked to report their composite ACT scores, which were converted to SAT scores using a concordance table (ACT, 2009). SAT scores for most colleges were obtained through the Integrated Postsecondary Education Data System (IPEDS) (nces.ed.gov/ipeds). Colleges report SAT scores for their 25th and 75th percentile, for each section separately. The midpoint of the 25th and 75th percentile is taken as a proxy for the median SAT for each section, and then the Critical Reading and Mathematics scores are added together to give a final score to be used for each college's "median" SAT score. Some colleges did not report their SAT scores

to IPEDS, but reported ACT scores. Median ACT scores were obtained from IPEDS using the same method, and then converted to SAT scores using the concordance tables. Some colleges did not report SAT or ACT scores. Of these colleges, if the highest degree they grant is an Associate's degree, or if they are a non-degree granting institution, they were labeled as *nonselective* instead of being assigned a median SAT score. For the remaining schools that did not report SAT or ACT scores to IPEDS, we searched for their median SAT scores using a variety of online college-planning sources (PrepScholar 2017; College Simply 2017; College Factual 2017; Princeton Review 2017). These sources were able to either provide a median SAT/ACT score, or provide enough information about admissions policies to be able to label the school nonselective. Eight institutions were not classifiable.

Next, all SAT scores (for students and colleges) were converted to their percentile ranks among all students who took the SAT (SAT 2014). For each student, *diffattend* is given by the difference between the percentile rank of the median SAT for the college they attended and the student's percentile rank. By construction, students who undermatch will have highly negative values for *diffattend*, since their SAT scores will be much higher than the college they attend. There is no obvious threshold for determining if a student has undermatched, but the previous literature has typically used between -10 and -15. For this study, to keep a conservative definition of undermatching, a student has undermatched if their *diffattend* score is less than or equal to -15. This means that if they attend a college with a median SAT score that is more than 15 percentiles lower than their own, they have undermatched.

Following Hoxby and Avery (2013), we only include students who score at the 90th percentile or above to be sure the focus is on high-achievers. Excluding nonselective colleges, the average SAT percentile rank for colleges' median SAT score is 69. So, on average, a student would need to score at least above the 84th percentile to undermatch at a college. After using this cut-off, the sample size is 338 and 59% of these students undermatched (see Table 1).

Undermatch	Freq.	Percent	Cum.
0	138	40.83	40.83
1	200	59.17	100.00
Total	338	100.00	

Table 1. Frequency of Undermatching

The two independent variables of interest are *distattend* and *income*. *Distattend* gives the distance between the student's home at the time that they were applying to college and the college they attended. These distances were constructed using zip codes. Survey respondents answered a question about the zip code of their hometown when they were applying to college. The zip codes of the colleges were obtained from IPEDS for most colleges, and for the colleges that were missing from the IPEDS data, we used Google Maps. *Distattend* gives the fastest driving distance between these two zip codes for each observation, as given by Google Maps (Google Maps, 2017). Figure 1 shows summary statistics for *distattend* and its distribution.

	Obs	Mean	Std. Dev.	Min	Max
Distattend	338	241.5	487.9	0	3087.8

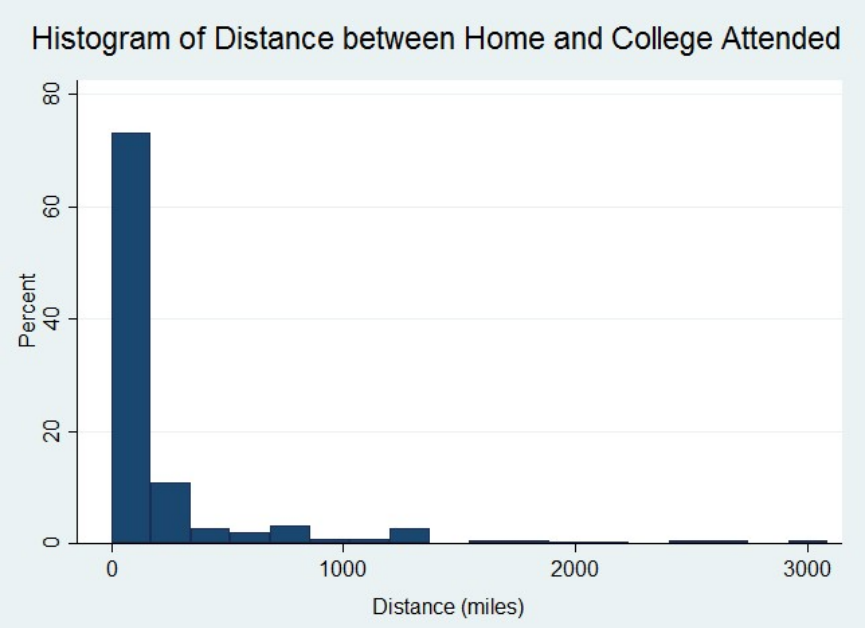

Figure 1. Summary statistics and distribution of Distattend

The *income* variable gives the income category of the student. Respondents were asked about their family's income at the time that they were applying to college. If their family income was less than $40,000, they are labeled as low income. If their family income was greater than $100,000, they are high income. The middle income category includes those in between ($40,000 to $100,000). Table 2 shows that roughly half of the sample is middle income with one-quarter above and below.

Income	Freq.	Percent	Cum.
Low	75	22.19	73.37
Middle	173	51.18	51.18
High	90	26.63	100.00
Total	338	100.00	

Table 2. Frequencies of Income Categories

Demographic variables are included in the regression as controls. These include dummy variables for race, gender, and whether the respondent is Hispanic (see Tables 3, 4, and 5 for frequencies). Although much work has been done on the role of race and gender in education and undermatching, Hoxby and Avery (2013, 18) point out that:

> A student's being an underrepresented minority is not a good proxy for his or her being low-income. Thus, if a college wants its student body to exhibit income diversity commensurate with the income diversity among high achievers, it cannot possibly attain this goal simply by recruiting students who are underrepresented minorities.

Race	Freq.	Percent	Cum.
White	254	75.15	75.15
Asian	52	15.38	90.53
Black	22	6.51	97.04
Native American	4	1.18	98.22
Asian and White	3	0.89	99.11
Black and White	3	0.89	100.00
Total	338	100.00	

Table 3. Frequencies of Race

Gender	Freq.	Percent	Cum.
Female	133	39.35	39.35
Male	205	60.65	100.00
Total	338	100.00	

Table 4. Frequency of Gender

Hispanic	Freq.	Percent	Cum.
Not Hispanic	308	91.12	91.12
Hispanic	30	8.88	100.00
Total	338	100.00	

Table 5. Frequency of Hispanic

Our theoretical model focuses on the effect of distance and income on the probability of undermatching and is straightforward:

$$undermatch = f(distance, income, demographic\ controls)$$

Because undermatch is a dummy dependent variable, we follow common practice and use probit regression. To get a rough handle on the effects of the *X* variables, we also estimate the model with ordinary least squares (OLS), i.e., the linear probability model. Table 6 shows our results. The OLS and probit regressions are in general agreement.

The coefficient on *distattend* is the effect of *distattend* on a student's probability of undermatching, holding their income category, race, gender, and if they are Hispanic constant. For OLS, the estimated coefficient is - 0.000203, so a student who goes to college 100 miles farther away decreases his or her probability of undermatching by 2 percentage points. The estimated standard error is 5×10^{-5} so an interval estimate of a 100-mile increase in *distattend* is a decrease of 2 percentage points ± 0.5 percentage point. A student who goes to college 500 miles farther away decreases his or her probability of undermatching by 10 ± 2.5 percentage points.

Middle income ($40,000 to $100,000) is treated as the base case. The coefficient on low income for OLS in Table 6 is 0.07, which implies that a low-income student is 7 percentage points more likely to undermatch than a middle-income student, holding all other included variables constant. The coefficient on high income is - 0.17, which means that a high-income student is 17 percentage points less likely to undermatch than a middle-income student, holding all other included variables constant. This result matches with findings from previous literature that low-income students are more likely to undermatch than rich students. The effect (24 percentage points from low to high) is quite large.

	Probit		OLS
	Coefficients	Percentage point Impact	Coefficients
distattend	-0.000619***	-9.8	-0.000203***
	(-3.59)		(-3.73)
low income	0.203	7.4	0.0700
	(1.07)		(1.05)
high income	-0.472**	-18.6	-0.171**
	(-2.77)		(-2.77)
male	-0.283	-11.0	-0.0930
	(-1.88)		(-1.75)
hispanic	-0.568*	-22.0	-0.193*
	(-2.13)		(-2.05)
_cons	1.369		0.971***
	(1.89)		(3.95)
6 Race Dummies Included	Yes		Yes
Pseudo R^2/R^2	.0861		.1091

N = 338 for all models
t statistic in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 6. Determinants of Undermatch Regression Results

OLS imposes the restrictive assumption that the effects of the independent variables are linear. So, in the OLS model, no matter the initial distance, each additional mile between a student's home and college gives the same decrease in the probability of undermatching. The probit model relaxes this assumption, but the coefficients cannot be interpreted directly. The second column of Table 6 gives the percentage point impact of a change in the variable. For categorical variables, we report the percentage point impact associated with having that characteristic. For example, males are 11 percentage points less likely to undermatch than females. Since income is a categorical variable with more than one

category, the percentage point impact is compared to the base case, middle income. Since *distattend* is a continuous variable, the given percentage point is the change in the probability of undermatching associated with a one standard deviation change from the mean. The standard deviation of *distattend* is roughly 480 miles, so if a student goes 480 miles farther away for college, instead of the average of 240 miles away, they are 9.8 percentage points less likely to undermatch.

Table 7 shows the probabilities of undermatching at varying levels of *distattend*, holding all other variables at their means. An individual with 0 miles between their home and college (practically speaking, this is an individual who attends college in the same zip code area as their home) has a 65.1% probability of undermatching, whereas an individual who goes to college 3,000 miles away from home has only a 7.1% probability of undermatching.

	Margin	Std. Err.	z	P>z	95% Conf. Interval	
Distance						
0	.651	.03	21.64	0.000	.592	.710
500	.531	.03	15.51	0.000	.464	.599
1000	.408	.06	6.91	0.000	.293	.524
1500	.294	.08	3.68	0.000	.138	.451
2000	.197	.09	2.26	0.024	.026	.369
2500	.123	.08	1.51	0.131	-.036	.282
3000	.071	.07	1.08	0.281	-.058	.200

Table 7. Predicted Probability of Undermatch at Varying Levels of Distattend

Returning to the effect of income, but this time focusing on the probit regression, the percentage point impact in Table 6 of 0.074 on low income means that if you have two otherwise similar individuals, but one is low-income and one is middle-income, the low-

income student is 7.4 percentage points more likely to undermatch. The marginal effect of high income can be interpreted similarly. Holding the other variables at their means, a high-income student is 18.6 percentage points less likely to undermatch than a middle-income student.

Thus, we have shown that both the OLS and probit regressions associate being low-income and going to college closer to home with an increase in a student's probability of undermatching. We can also determine if distance from home affects the probability of undermatching differently among the three income groups. That is, we can see if the effect of going to college farther from home is different for low-income students than it is for high-income students.

Table 8 uses the probit regression results to compute the predicted probabilities of undermatching at varying levels of distance from home, by income category. For example, a low-income student who goes to college 500 miles away from home has a 63.7% probability of undermatching, while a similar high-income student only has a 38.2% probability of undermatching.

However, at 3,000 miles, the story is quite different. The high-income student's chance of undermatching is low (only 3.6%), but so is the low-income student's (12.6%). This is an important result: distance can mitigate the effect of income on undermatching.

To emphasize the powerful effect of distance, Figure 2 plots the predicted probabilities in Table 8. As can be seen in Figure 2, for all income categories, the probability of undermatching decreases with distance. However, it does not decrease at the same rate for all income categories. Compared to richer students, low-income students see a faster drop in the probability of undermatching.

	Margin	Std. Err.	Z	P>z	95% Conf.	Interval
Distance						
0 miles						
Low income	.742	.05	14.36	0.000	.641	.843
Middle income	.675	.04	18.05	0.000	.602	.748
High income	.500	.05	9.22	0.000	.393	.605
500 miles						
Low income	.637	.06	10.57	0.000	.519	.755
Middle income	.561	.04	13.45	0.000	.480	.643
High income	.382	.05	6.99	0.000	.275	.489
1000 miles						
Low income	.521	.08	6.53	0.000	.365	.677
Middle income	.443	.06	7.01	0.000	.319	.566
High income	.274	.06	4.30	0.000	.150	.399
1500 miles						
Low income	.402	.10	3.98	0.000	.204	.600
Middle income	.329	.08	3.91	0.000	.164	.493
High income	.184	.07	2.68	0.007	.049	.318
2000 miles						
Low income	.292	.11	2.56	0.010	.069	.516
Middle income	.229	.10	2.43	0.015	.044	.413
High income	.115	.06	1.78	0.075	-.012	.241
2500 miles						
Low income	.199	.11	1.75	0.081	-.024	.422
Middle income	.148	.09	1.64	0.101	-.029	.326
High income	.067	.05	1.26	0.209	-.037	.171
3000 miles						
Low income	.126	.10	1.25	0.211	-.071	.323
Middle income	.090	.08	1.17	0.241	-.060	.240
High income	.036	.04	0.93	0.354	-.040	.112

Table 8. Predicted Probabilities of Undermatch by Distance and Income Category

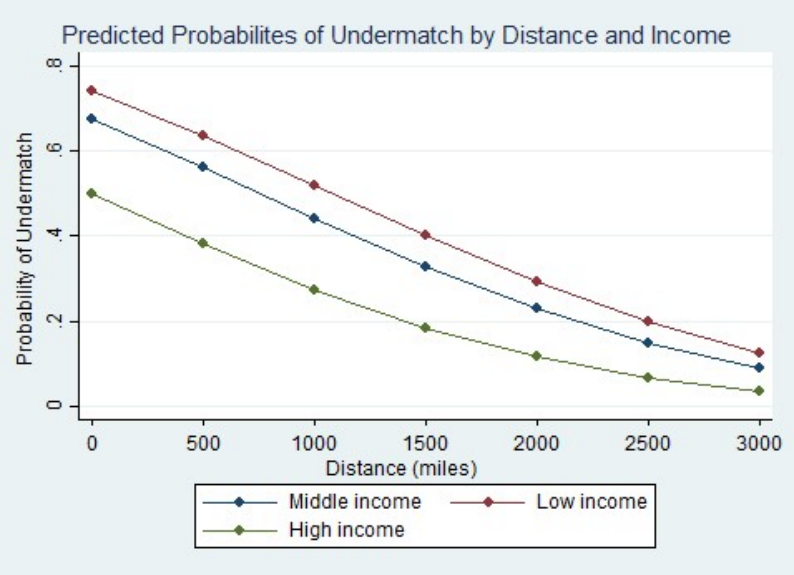

Figure 2. Predicted probabilities of undermatch falls as distattend increases for all income groups

To further analyze the relationship between distance and income, Table 9 shows the marginal effect of each income category at varying distances. Looking at the second row under "Low income," it shows that if a student attends college 500 miles away from home, holding other included variables constant, the effect of being low-income increases their probability of undermatching by 7.6 percentage points, compared to middle-income students. A high-income student who attends college 500 miles from home is 18.0 percentage points less likely to undermatch than a middle-income student who attends college 500 miles from home.

Notice that the undermatch gap falls as distance rises. At 3,000 miles, it has closed considerably. Once again, there is much less difference in the probability of undermatching when we compare students who attend colleges far away from home.

	dy/dx	Std. Err.	Z	P>z	[95% Conf.	Interval]
Low income						
0 miles	.067	.06	1.10	0.273	-.053	.187
500 miles	.076	.07	1.08	0.279	-.062	.213
1000 miles	.078	.07	1.07	0.285	-.065	.221
1500 miles	.074	.07	1.05	0.292	-.064	.211
2000 miles	.064	.06	1.02	0.307	-.059	.186
2500 miles	.050	.05	0.96	0.335	-.052	.153
3000 miles	.036	.04	0.87	0.382	-.050	.118
High income						
0 miles	-.176	.06	-2.78	0.005	-.299	-.052
500 miles	-.180	.06	-2.84	0.004	-.304	-.056
1000 miles	-.169	.06	-2.88	0.004	-.283	-.054
1500 miles	-.145	.05	-2.75	0.006	-.248	-.042
2000 miles	-.114	.05	-2.31	0.021	-.210	-.017
2500 miles	-.082	.05	-1.74	0.081	-.173	.010
3000 miles	-.054	.04	-1.28	0.202	-.136	.029

Table 9. Marginal effects of income on undermatch at various distances

Figure 3 plots the dy/dx results in Table 9. The blue (top) line shows the marginal effect of distance on a low-income student's probability of undermatching, whereas the red (bottom) line shows the marginal effect for high-income students. At every distance, low-income students are more likely to undermatch than high-income students. However, the gap between the low-income and high-income probability of undermatching shrinks as distance increases. This is a key result: high-income students have a much greater advantage over low-income students in terms of undermatching at 500 miles from home than at 3000 miles from home.

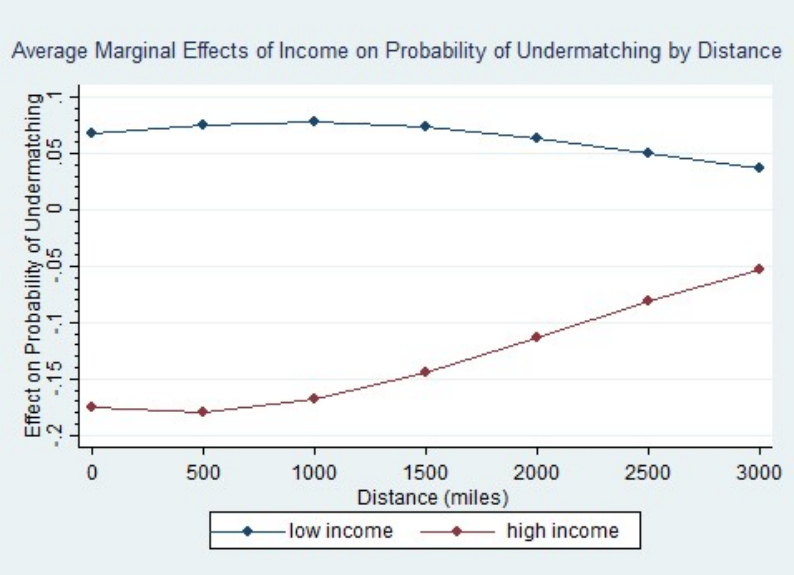

Figure 3. Income's effect on the probability of undermatch falls as distattend increases

We find that income has an effect on a student's probability of undermatching, and that high-income students are much less likely to undermatch than low-income students. This is the expected result and it has been demonstrated repeatedly in previous work. Additionally, our results show that increasing the distance between a student's home and college decreases their probability of undermatching. Finally, the magnitude of the effect of a student's income on their probability of undermatching decreases as distance between their home and college increases.

We conclude this section with a brief report on several questions from our survey. We asked respondents, "Did you apply to any colleges that you would consider prestigious or elite?" Figure 4 shows the percent of students who answered "Yes" to this question, by income category. Low-income students were the least likely to apply to an elite college, roughly 30%, and much less likely than high-income students (almost 80%).

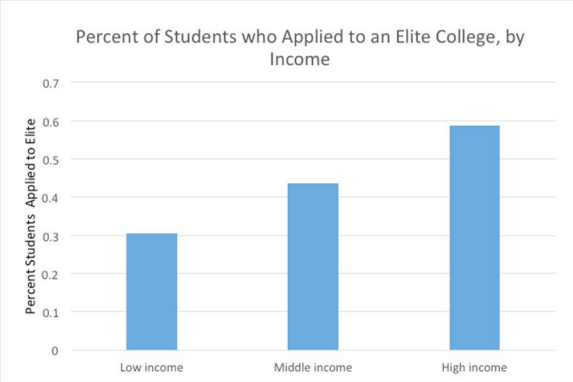

Figure 4. Applying to at least one "prestigious or elite" college rises as income increases

It logically follows that if students don't apply to selective colleges, they are more likely to undermatch with the college they attend. This is verified in Figure 5, which shows the percentage of students who undermatch by whether they applied to an elite college or not. Nearly 80% of students who did not apply to any colleges that they considered prestigious or elite undermatched, while only 37% of students who applied to at least one elite college undermatched.

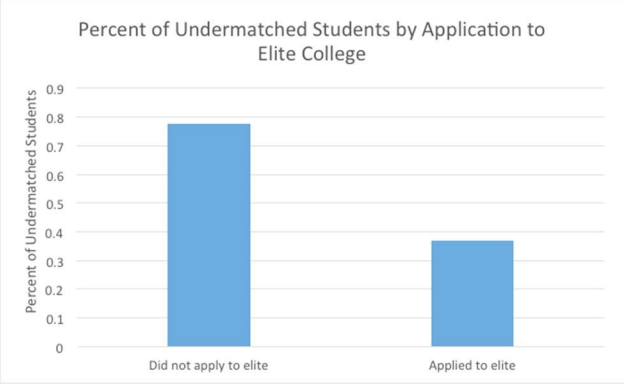

Figure 5. Undermatch less common if applied to a "prestigious or elite" college

Another survey question asked respondents, “Have you ever been eligible for the free- or reduced-price lunch program at school?” If the respondent selected “Yes,” they were additionally asked, “During the years when you were in Kindergarten through 12th grade, how many years were you eligible for the free- or reduced-price lunch program? If you're unsure, please approximate.” These questions were included as an alternative way of measuring financial resources, rather than simply asking the students about their household income. The follow-up question about how many years a student was eligible for free- or reduced-price lunch (FRPL) was inspired by findings from Michelmore and Dynarski (2016) that demonstrate that students who were persistently eligible for FRPL fared worse academically than those who were intermittently eligible. Figure 6 shows the percentage of undermatched students by general FRPL eligibility and the percentage of students who undermatched by the number of years they were eligible for FRPL.

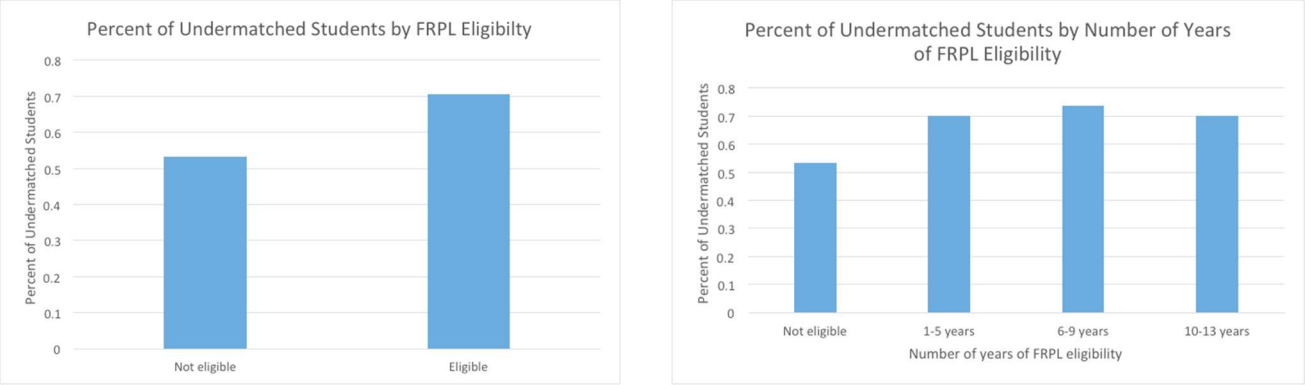

Figure 6. Undermatch and the free- or reduced-price lunch program

As the left panel in Figure 6 shows, students who are at some point eligible for the free- or reduced-price lunch program are less likely to undermatch than those students who were never eligible. However, in a departure from what one might assume given

Michelmore and Dynarski's (2016) results, the number of years that a student was eligible for FRPL does not seem to affect their probability of undermatching. This result should be interpreted with caution, since no other variables are being controlled for.

Another survey question was, "What were the most important factors in choosing your college? Explain." This question was asked before any questions about the specific influence of distance or other factors, so that respondents would not be primed before answering the open-ended question. We identified four main factors that were most often mentioned: 1) cost of attending college (including mentions of financial aid); 2) location of the college; 3) academic programs or reputation of the college; and 4) atmosphere or culture of the college. Some responses contained several of the categories, while some did not mention any. Figure 7 shows the percent of students who mentioned each of the four factors in their open-ended response, by income category.

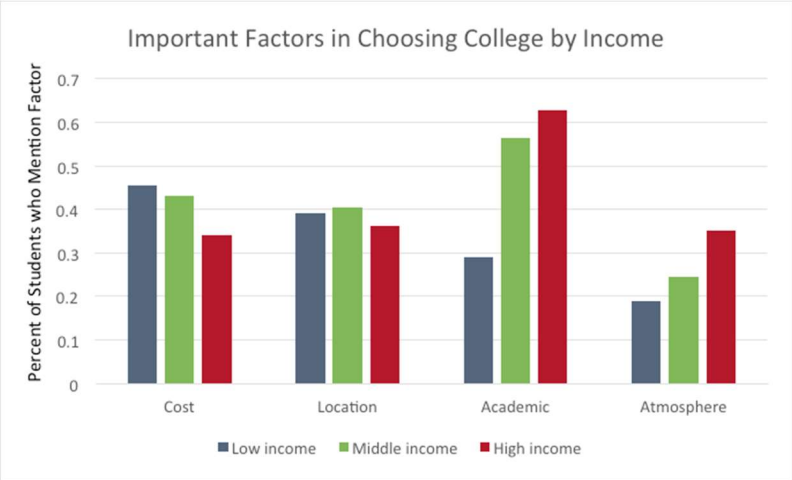

Figure 7. Percentage of students who mentioned each of four main factors as important in their college choice by income category

Clearly, low-income students were much more concerned with cost and location than academic programs or the atmosphere of the program. This supports the finding that low-income students are more likely to undermatch, since they are more likely to attend a college that is cheaper or closer to home than one that has a good academic reputation. High-income students have the privilege of being able to choose a college based on its atmosphere and academics, since cost is likely not as big of a concern for them. Figure 7 shows that high-income students are more likely to mention academics than any of the other categories when asked about the most important factors in their college decision, implying that this is the factor that they care about most. This also supports the regression findings that high-income students are least likely to undermatch, since they prioritize attending a college with rigorous academics.

In addition to asking respondents about the zip code of the area in which they lived when applying to college in order to calculate the actual distance between a student's home and the college they attended, the survey asked several questions about how distance from home played a role in the students' college decision process. One question asked, "When deciding colleges to apply to, what was the farthest distance you considered?" Figure 8 shows the percent of students who undermatched, by the farthest distance they considered going for college. The farther from home students consider going to college, the less likely they are to undermatch. Figure 8 uses a separate distance measure than the regression analyses, since it is based on the farthest distance students *considered,* rather than the actual distance of the college they attended. However, it gives the same general result as the regressions, which found that the farther from home a student attends college, the less likely they are to undermatch.

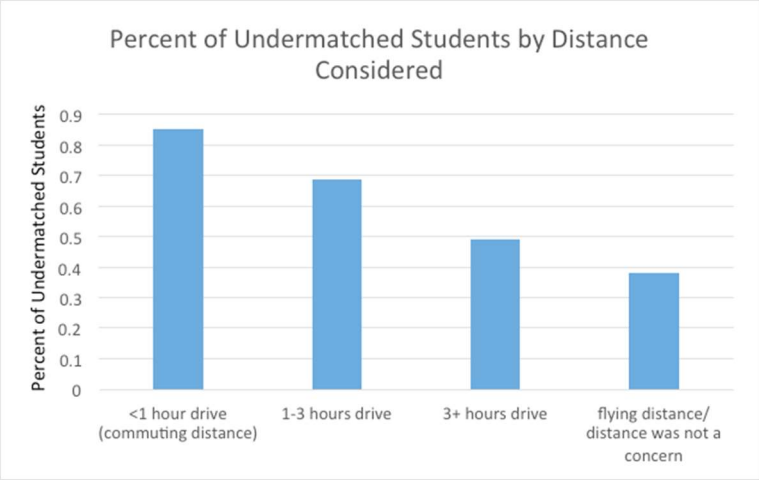

Figure 8. *Undermatch falls as the farthest distance from home they considered going to college increases*

These survey questions strengthen the regression results since they come to the same general conclusions. The high-achieving, low-income students in the sample are less likely to apply to elite colleges and more likely to undermatch. The farther away from home students consider going to college, the less likely they are to undermatch.

V. Conclusions

The novel dataset used in this study has confirmed an established result that among high-achieving students, those who are low-income are more likely to undermatch than their high-income peers. Controlling for gender, race, ethnicity, and the distance between a student's home and the college they attend, low-income students are 7.4 percentage points more likely to undermatch than middle-income students, and high-income students are 18.6 percentage points less likely to undermatch than middle-income students.

Our key finding is that increasing the distance between a student's home and the college they attend decreases the probability that they will undermatch. Furthermore, the

effect of income on the probability of undermatching decreases as distance increases. At a distance of 500 miles between a student's home and college, the difference in the probability of undermatching between low-income students and high-income students is 25.5 percentage points. At 3,000 miles, the gap is only 8.7 percentage points.

These results do not have a direct causal interpretation because students choose the college they attend and this includes how far away it is from home. Thus, both distance and undermatch are endogenous variables. In addition, a variety of factors are subsumed by distance as a measure. Students may be more likely to consider and visit colleges farther from home if they have more social capital. For example, having connections with family alumni at a college far away might make a student more likely to attend, or having mentors who are familiar with the college search process may lead a student to consider a broader range of options, including colleges that are farther away from home. Although increased social capital is often associated with higher income, some low-income students may have access to people and resources that allow them to consider a broader array of colleges. These students would be more likely to attend colleges farther from home and less likely to undermatch, which would be captured in the *distattend* variable of the regression.

The *distattend* variable could also be capturing the effect of a student's risk-taking behavior on their college choices. Low-income people are less likely to take financial risks, since they do not have the resources to recover from negative outcomes. They are less likely to take a risk on the first (and maybe only) chance they get. Lincove and Cortes (2016) showed that for high school students in Texas in the top 10% of their class, eliminating the uncertainty of admission had an equalizing effect among income groups. Going to college far away from home is certainly risky. It is easier for high-income students

to travel far from home because they know that they will have the money to come home, or the opportunity to try again if something goes wrong. This could explain why going to college far from home makes all students, but especially low-income students, less likely to undermatch. Those students who are willing to take the risk to move far away may be more likely to also take the risk of attending a rigorous college.

To help students reach their potential and for society to capitalize on its talent, efforts should be made to reduce undermatching. To give a specific example, elite colleges could pay for high-achieving, low-income students to come on campus visits, even if it involves paying for a flight because the student lives far away. These students, who do not have the resources to pay for their own visit to a campus far away, are the kind of students who are likely to undermatch by attending a college close to home. Paying for their campus visit and providing them with knowledge and encouragement on the application and financial aid process would increase the likelihood that these students would apply to and enroll in selective colleges.

This work has demonstrated that increasing the distance between a student's home and college is associated with a decrease in the probability of undermatching, and has offered some potential explanations of the relationship. However, further research is needed on the underlying reasons that distance affects undermatching, as well as investigations of additional determinants to undermatching, such as number of siblings in college, whether the student is first generation, and the median SAT of high school attended (lower median SAT scores may suggest something about availability of information about college). Research is also needed to find specific policies that can effectively reduce undermatching, especially for low-income students.

References

ACT. 2009. Concordance between ACT Composite Score and Sum of SAT Critical Reading and Mathematics Scores. Edited by ACT-SAT Concordance Tables: ACT Research and Policy.

Bowen, William G, Matthew M Chingos, and Michael S McPherson. 2009. *Crossing the Finish Line: Completing College at America's Public Universities*: Princeton University Press.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, 6 (1):3-5.

Chapman, David W. 1981. "A model of student college choice." *The Journal of Higher Education*, 52 (5): 490-505.

College Factual. 2017. "Applications and Admissions: Stats at a Glance," www.collegefactual.com, accessed March 20, 2017.

College Simply. 2017. "Admission Chances," www.collegesimply.com, accessed March 20, 2017.

Michelmore, Katherine and Susan Dynarski. 2016. "The Gap Within the Gap: Using Longitudinal Data to Understand Income Differences in Student Achievement" NBER Working Paper No. 22474, www.nber.org/papers/w22474.

Google Maps 2017. maps.google.com, accessed April 10, 2017.

Hoxby, Caroline and Christopher Avery, 2013. "The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students," *Brookings Papers on Economic Activity*, 2013(1), pp. 1-65. doi.org/10.1353/eca.2013.0000

Integrated Postsecondary Education Data System, nces.ed.gov/ipeds, accessed March 25, 2017.

Lincove, Jane Arnold, and Kalena E Cortes. 2016. "Match or Mismatch? Automatic Admissions and College Preferences of Low-and High-Income Students," NBER Working Paper No. 22559, www.nber.org/papers/w22559.

Miller, Lois. 2017. "Reaching Further: The Role of Distance in College Undermatching," Honor Scholar thesis, www.depauw.edu/learn/econexcel.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5):411-419.

PrepScholar. 2017. "Requirements for Admission," www.prepscholar.com, accessed March 20, 2017.

Princeton Review. 2017. "Find Your Dream School" www.princetonreview.com, accessed March 20, 2017.

Roderick, M, J Nagaoka, V Coca, E Moeller, K Roddie, J Gilliam, and D Patton. 2008. *From High School to the Future: Potholes on the Road to College*. Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

SAT. 2014. SAT Percentile Ranks for Males, Females, and Total Group, secure-media.collegeboard.org/digitalServices/pdf/sat/sat-percentile-ranks-composite-crit-reading-math-writing-2014.pdf, accessed March 20, 2017.

Winston, Gordon C, and Catharine B Hill. 2005. "Access to the most selective private colleges by high-ability, low-income students: are they out there?" Williams College, DP-69. ideas.repec.org/p/wil/wilehe/69.html