4-2017

Reaching Further: the Role of Distance in College Undermatching

Lois Miller
DePauw University

REACHING FURTHER:

THE ROLE OF DISTANCE IN COLLEGE UNDERMATCHING

Lois Miller

Honor Scholar Program Senior Project

2017

Sponsor: Humberto Barreto

1st Reader: Rebecca Bordt, 2nd Reader: Ted Bitner

Contents

Acknowledgements

I would like to thank Bert Barreto for being an invaluable mentor and devoting a great amount of time and effort to helping me throughout the project. I would also like to thank Rebecca Bordt and Ted Bitner for serving on my committee and providing advice and insights to guide my research. Thank you to J. William and Dorothy A. Asher for providing funding for me to collect data through Amazon's Mechanical Turk. Thank you to Amy Welch, Peg Lemly, and Kevin Moore for their work with the Honor Scholar program. Finally, thank you to my parents, Sheila and Rich Miller, to my sister, Kelly Miller, and to all of my friends who have helped me develop ideas, tested my survey, read sections of my thesis, and provided support throughout the year.

Abstract

This thesis 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 a probit analysis confirms an established result that low-income students are more likely to undermatch. The primary contribution of this work is the result that as the distance between a student's home and the university they attend increases, the probability that the student will undermatch decreases. The decrease in likelihood of undermatching from attending college an additional 500 miles from home ranges from 5 to 12 percentage points. Additionally, this study finds that as distance increases, the effect of income on the probability of undermatching decreases.

I. Introduction

College choice in the United States is a complex process, but not all students experience it in the same way. Even among students with similar high school academic achievement, there is much variation in the caliber of the colleges they attend. When a student attends a college that is much less academically rigorous than they could handle given their high school achievement, it is called undermatching. There are many factors that can affect college application and enrollment behavior and subsequent undermatching. This thesis will focus on two factors and their effect on a student's probability of undermatching: the student's family income, and the distance between a student's home and the college they attend. It will also explore how the distance between a student's home and college affects students' likelihood of undermatching differently depending on whether they are high-income and low-income.

The inspiration for this thesis came from a story about an extremely bright low-income high school boy in Los Angeles, told by Malcolm Gladwell in an episode of his podcast, *Revisionist History* (Gladwell 2016). The boy, given the pseudonym Carlos for the episode, transferred from a large public high school in which he was never challenged to an elite private school before his sophomore year. He was supported by the YES program, which identifies high-achieving low-income middle school students and helps them by providing tuition to private schools, along with tutoring and encouragement. Carlos and the other students in the YES program have a great chance of attending a college that matches their academic talents. But Carlos is an exception, and there are many smart, low-income students who don't attend elite colleges, or college at all. There are so many obstacles that can derail a low-income student from capitalizing on their talent, from lack of knowledge or

encouragement to family violence. Gladwell (2016) describes the difference between being privileged and being poor in America as, "how many chances you get". While wealthy high school students can usually overcome setbacks and still attend colleges that match their talents, often low-income students don't have to resources to do so.

Low-income students are more likely to undermatch than their wealthier peers. Even when financial aid packages enable these high-achieving students to attend elite schools for little to no cost, the vast majority of the approximately 25,000 to 35,000 such students each year do not apply to any selective colleges (Hoxby and Avery 2012, 2 and 15). When students undermatch, they don't receive an education that fully leverages their potential, and society's capitalization rate suffers. Gladwell (2016) describes a society's capitalization rate as "the percentage of people in any group who are able to reach their potential." It is important to address undermatching and its effects on disadvantaged members of our society, both to combat the inequality undermatching perpetuates and to improve society as a whole by increasing the productivity of its members.

In addition to verifying previous findings that low-income students have a higher probability of undermatching, this thesis examines the role of distance between a student's home and the college they attend in undermatching. The farther from home a student attends or considers attending college, the less likely they are to undermatch. This effect is greater for low-income students than for high-income students, meaning that at farther distances from home, the effect of income on undermatching is smaller than at distances close to home.

The remainder of this thesis is organized as follows. Section II presents a 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 reviews Amazon's Mechanical Turk, which was used to collect survey data. Section IV gives a theoretical framework of a dummy dependent variable model, which is used for analysis. Section V gives empirical results, and Section VI concludes with implications from the results and areas for further research.

II. Literature Review

A. Models of Student College Choice

For many students, choosing where to attend college is one of the most important decisions of their lives. There are two main steps of the college decision - whether to attend college and where to attend college. Prior to 1981, research focused mainly on whether students attended college or not, without giving much consideration to the specific colleges that students were choosing. Chapman (1981) turned the conversation away from exclusively focusing on whether students attend college or not by presenting a non-mathematical model of student college choice in which the student is choosing between various colleges. His goals were to provide a model to aid further research in the area of student college choice, as well as encourage college admissions offices in their recruitment efforts (Chapman 1981, 499).

According to the model, a student's college choice is determined by a combination of student characteristics and external influences. The four main student characteristics are socioeconomic status, aptitude, level of educational aspiration/expectation, and high school performance (Chapman 1981, 492). Socioeconomic status manifests itself both directly though the costs of various colleges and financial aid packages the student receives from these colleges, and indirectly through the student's attitudes and behavior. In this way, socioeconomic status limits the set of colleges that a student considers both through what the student believes they can realistically afford, but also through their aspirations or expectations for their education and future lives (Chapman 1981, 493). Aptitude drives students to self-select where they feel they will fit in, based on their high school achievement and performance on standardized tests such as the SAT and ACT. They tend to

choose colleges where the other students have similar aptitude to their own (Chapman 1981, 493). A student's educational expectations are what they believe they will be doing in the future, whereas their aspirations what they hope to be doing. Expectations and aspirations are correlated with GPA, but only moderately so, signaling that two students with the same GPA could have substantially different expectations or goals for their educational future (Chapman 1981, 494). Finally, high school performance influences college choice both directly and indirectly. It is a direct influence because measures of high school performance are used in college admissions, which in turn influence prospective students to apply based on their beliefs of where they will be accepted and how they will fit in with the other students. High school performance also indirectly affects student college choice in a myriad of ways. For example, students with higher high school achievement will receive more encouragement and guidance, which affects their application and enrollment behavior (Chapman 1981, 494).

There are also many external influences on students' college choice processes. They fall into three general categories: significant persons, relatively fixed college characteristics, and college efforts to communicate with students. While many individuals can affect a student's college decision, usually parents are the most influential, followed by friends (Chapman 1981, 495). Relatively fixed college characteristics include cost, although the social background and family income of the students who attend each college has been shown to be more stratifying than the actual cost of the college. Financial aid is ideally designed to increase students' college choices by making more colleges affordable. Location also affects college choice, and the distance that a student travels to college is positively correlated with academic ability and negatively correlated with income. The final relatively

fixed college characteristic is the availability of desired course programs, which is one of the most important deciding factors, especially for students interested in specialized areas of training (Chapman 1981, 495-497). Finally, college efforts to communicate with students affect their choices, although students who expect to go to college are more likely to seek out college information (Chapman 1981, 498).

All parts of this conceptual model are filtered through students' general expectations of college life, which are often idealized and not accurate. Information can be distorted by students who make their college decisions based on false expectations of what their experience will be like (Chapman 1981, 499).

Litten (1982) expands upon Chapman's model by focusing on the process students go through in making their college decisions (rather than just the outcomes) and by reviewing existing literature to determine how different groups of students behave differently in their college choice process, and which groups behave similarly. While Litten (1982, 383) divides these groups by race, sex, ability level, parents' education levels, and geographic location, this literature review will focus on the latter three since they most directly relate to the research question of how distance affects undermatching.

Higher ability students began their college search process earlier than lower ability students, and were more interested in information about academic programs at various institutions (Litten 1982, 392). The highest ability students were also better able to process information about varying costs and financial aid – they understood that "price" (tuition, fees, room and board) was less important than "net cost" (price minus aid) when deciding among colleges (Litten 1982, 393).

Students with parents who were more highly educated started applying to college earlier than those students with less educated parents. Parents with a college education were more influential to their child's college choice, whereas students without college-educated parents were more likely to rely on their guidance counselor (Litten 1982, 394). Students whose parents had relatively little education placed more importance on a college's cost, rules and regulations for students, and career information (Litten 1982, 395).

Aside from the timing of the application process as it relates to regional deadlines and a few college attributes, geographic location did not have much of an effect on students' application behavior (Litten 1982, 396).

Perna (2006, 105) proposes a conceptual model that combines aspects of the two main models used to model college choice: the economic model of human capital investment, and the sociological model of status attainment. Human capital theory predicts that students make college decisions based on weighing the costs of attending various colleges with the benefits from increased earnings that result from their increased productivity that comes from higher education (Perna 2006, 106). This theory does not assume that actors have perfect information, but assumes that students act rationally with regards to the information they have. This shows how some differences in college access between groups can arise due to differences in access to information. Low-income students are less likely to have accurate and complete information about the costs of attending college, so while they may be acting rationally within their knowledge, they are put at a disadvantage compared to high-income students with better information (Perna 2006, 109).

While the economic model captures a valuable piece of the college decision process, it is not complete. Sociological issues play an important role in setting the context that a student is making their college decision within. Students have varying levels of social capital, defined by Peter Bourdieu (1992, 119) as "the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition." James Coleman views social capital as one of many potential resources that an actor can use, similar to human capital and physical capital, and notes that social capital frames an individual's knowledge of everyday information, norms and sanctions (Gauntlett 2011, 4).

Perna (2006, 111-113) embeds this human capital model within four layers of context, which come from sociological models of status attainment, social capital, and habitus. The first layer of context is the individual's habitus, or their "internalized system of thoughts beliefs, and perceptions that are acquired from the immediate environment" (Perna 2006, 113). This can shape the student's expectations and aspirations, which Perna (2006, 113) claims in agreement with Chapman (1981), affect their college choices. The second layer is the school and community context, which is equivalent to the organizational habitus from McDonough (1997, 158), defined as "the impact of a cultural group or social class on an individual's behavior through an intermediate organization and family habitus that is reasonable or rational behavior in context." McDonough (1997, 153) shows how secondary schools shape students' idea of college through their differences in counseling and available resources. The third layer is the higher education context, which conveys how colleges shape students' decisions though the information they provide, their location and proximity to the student's home, and their ability to accept applicants (Perna 2006, 118).

The fourth and final layer is the broader social, economic, and policy context, which recognizes that changes in the nation's demographics and economic conditions, as well as policies that affect education and financial aid, have a bearing on individual students' college decisions (Perna 2006, 119).

Nurnberg, Schapiro, and Zimmerman (2012, 1) quantitatively analyze factors that affect whether students will attend Williams College or not, after being admitted. The data is limited to William College, but the methodology could be applied to other colleges if admissions data was available. In the admissions process at Williams, readers use a rubric to assign each student academic and non-academic ratings from 1 to 9 (1 is the best, 9 is the worst). Academic ratings are based on "SAT scores, high school grades, essays, class rank, high school academic program, support from the high school administration, AP test score – or IB test scores – and teacher recommendations," while non-academic ratings are based on "extra-curricular activities, non-academic awards, community service work, non-academic skills – i.e. special musical, athletic, acting, or other ability – and other non-academic activities" (Nurnberg, Schapiro, and Zimmerman 2012, 3-4).

Results showed that as both academic and non-academic ratings increase, so does the probability of matriculation, meaning that weaker accepted students were more likely to attend Williams than stronger students (Nurnberg, Schapiro, and Zimmerman 2012, 5). Results also showed that white students were substantially more likely to matriculate than minority students (Nurnberg, Schapiro, and Zimmerman 2012, 5-6), which may be related to Chapman's findings that students tend to prefer colleges where the student body has a similar social background and family income to their own. Nurnberg, Schapiro, and Zimmerman (2012, 6) found that the elasticity of demand for a William's education is small,

with a $5000 increase in net price only decreasing the probability of matriculation by 1.3 percentage points. The exception is that when students are denied financial aid entirely, they are 12 percentage points less likely to attend. The authors note that their results cannot be interpreted as causal, so should be interpreted with caution and not applied to college generally (Nurnberg, Schapiro, and Zimmerman 2012, 7).

B. Effects of Socioeconomic Status

While the general conceptual models from Chapman, Litten, and Perna are helpful for setting a framework of student college choice, specific empirical studies can give further insights into students' college choices, and how they are related to their income and socioeconomic status.

Aughinbaugh (2008, 33) analyzed a sample from the National Longitudinal Survey of Youth 1997 (NLSY97) to explore the question, "Who goes to college?" The NLSY97 is a national sample of individuals who were aged 12 to 16 on December 31, 1996, and the study included annual interviews beginning in 1997 (Aughinbaugh 2008, 33). Aughinbaugh (2008, 34) used a subsample of the NLSY97 to include individuals who were 21 or older at the time of data collection, and defined a college-going individual as someone who had entered college by the age of 20.

Overall, 49 percent of the sample had attended college by age 20, and the college attendees "had parents who attained more schooling, had higher levels of family income, had mothers who were older at the birth of their first child, and were more likely to have lived with both of their parents at age 12" (Aughinbaugh 2008, 35). Compared to those who attended 2-year colleges, 4-year college attendees were more likely to be female, less likely to be Black or Hispanic, and had higher income, parents with more education, and mothers

who were an average of 1.5 years older at the age of their first birth (Aughinbaugh 2008, 36).

Focusing on income, results showed that a 1-percent increase in level of family income increases the probability of an individual attending college by 5 percentage points, and increases the probability of an individual completing their first year once they have begun at either a 2-year or a 4-year college by approximately 4 percentage points. This regression produced these estimates of the effect of family income after controlling for gender, race, mother's highest grade completed, father's highest grade completed, mother's age at first birth, whether the respondent lived with both parents at age 12, high school grades, scores on the Armed Services Vocational Aptitude Battery, year of birth, whether the respondent lived in an urban area at age 12, and the region of the country in which the respondent lived at age 12 (Aughinbaugh 2008, 39-41).

Bowen et al. (2005) review existing literature to shed light on how family income and parental education affect college outcomes in general, and perform their own analysis on how socioeconomic status affects admission, enrollment, and academic outcomes at the most selective colleges in the country. They begin by noting that in general, the socioeconomic gaps in college attendance can be partially but not fully explained by the two most obvious explanations of academic preparedness and financing problems. "Differential access to the information and assistance necessary to navigate the admissions and financial aid processes (also products of background and family circumstances)" is likely more important than students not being able to finance college (Bowen et al. 2005, 74).

However, academic preparedness varies greatly by socioeconomic status, and is the primary factor that holds students of low SES back from going to college. “Poor families have great difficultly investing sufficient resources to develop in their children, in the time before high school graduation, the abilities and outlooks necessary to enable their children to attend college and graduate” (Bowen et al. 2005, 77). In fact, many students from disadvantaged backgrounds don’t even take the SAT – 34.2% of high school graduates in the bottom quartile of the family income distribution take it, as opposed to 70.1% in the top quartile. Even of students who score in the top decile on standardized test scores earlier in their education (and would presumably do well on the SAT), only 68% of students from the bottom income quartile took the SAT compared to 88% from the top income quartile. Doing well on the SAT further divides students by family income; only 7.4% from the bottom quartile scored 1200 or above, compared to 21.4% from the top quartile (Bowen et al. 2005, 81).

Despite the importance of academic preparedness, gaps in rates of college attendance between income groups exist even when controlling for academic achievement. Using a group of students in the eight grade in 1988, Thomas Kane found that a student from the top income quintile was 15 percentage points more likely to attend college than a student from the bottom income quintile, holding math and reading test scores and parental education constant (Bowen et al. 2005, 85). Another study by Caroline Hoxby found that of students with “medium-high preparedness”, that is, students with SAT section scores between 500 and 600 and who ranked in the top third of their class, 3% from the highest income quartile did not attend college compared to 13% from the bottom income quartile. Additionally, the highest income quartile saw 52% attend one of the most

expensive colleges, as opposed to only 20% of the bottom income quartile (Bowen et al. 2005, 87).

The cost of college is another factor that affects students of different income groups in different ways. Hill, Winston, and Boyd found that in 2001-02, after financial aid has been applied, students in the lowest income quartile spent 49% of their income on tuition, whereas students in the highest income quartile spent 21-28% (Bowen et al. 2005, 88). Yet, actual costs are not the only thing that stops low-income students from attending college. "Several studies suggest that students and their parents, in making decisions about college, are not always performing a clear-cut cost-benefit analysis, or are not necessarily acting on the conclusions of such an analysis" (Bowen et al. 2005, 90). McPherson and Shapiro find that high-achieving students overestimate the cost of selective colleges, and Hoxby and Avery show that students who are not high-income don't fully understand financial aid. They value loans and work-study as much as they value grants, and they prefer scholarships to grants based solely on the name (even though they mean the same thing) (Bowen et al. 2005, 90-91).

In their own analysis, Bowen et al. (2005) focus on 19 of the most selective colleges and universities in the United States. They find that of the 1995 entering cohort, 10 to 11 percent of students were from the bottom income quartile, just over 6 percent were first-generation college students, and 3 percent were both. These numbers show that these groups are underrepresented, since 25 percent of the national population is in the bottom income quartile, 38 percent of the national population of 16-year-olds have parents who never attended college, and 19 percent fall into both categories (Bowen et al. 2005, 98).

Much of this unequal representation occurs at the application stage. Low-income students account for 11.7% of applicants, 9.1% of admitted students, 10.8% of enrolled students, and 10.6% of graduates. The lack of variability in these percentages "suggests that socioeconomic status does not affect progression through the stages" (Bowen et al. 2005, 99-100). After examining the effect of being low SES on various stages of students' educational journeys at these selective institutions, Bowen et al. (2005, 135) conclude,

> For those applicants who took the SAT, did well on it, and applied to one of these selective institutions, family income and parental education, in and of themselves, had surprisingly little effect on admissions probabilities, on matriculation decisions, on choices of majors, on subsequent academic performance and graduation rates, and even on later-life outcomes such as earnings and civic participation…but the odds of getting into this highly competitive pool in the first place depend enormously on who you are and how you grew up.

Chetty et al. (2017) generally agree with the findings of Bowen et al. (2005) in their investigation of how colleges affect the intergenerational mobility of its students. They give every college in the United States a "mobility report card", based on students' earnings in their early thirties and their parents' income (Chetty et al. 2017, 1). A college's "mobility rate" is measured as "the fraction of its students who come from the bottom quintile of the income distribution and end up in the top quintile" (Chetty et al. 2017, 3). This rate is dependent upon "access rates" (percent of students who come from the bottom income quintile) and "success rates" (percent of those students who reach the top income quintile). Ivy-Plus colleges (Ivy League colleges, University of Chicago, Stanford, MIT, and Duke) have very high success rates but low access rates, consistent with Bowen et al.'s finding that low-income students who get past the initial hurdle of doing well on the SAT and applying to selective colleges are successful in the rest of the process. Interestingly, Chetty et al. (2017,3) also identify a group of less selective universities with comparable success rates

but much higher access rates, leading to significantly higher mobility rates than the Ivy-Plus colleges.

Chetty et al. (2017,15) first analyze different colleges' distributions of students' income backgrounds. They find that colleges are highly segregated by income, evidenced by "ten percent of colleges (like Harvard) draw fewer than 3.7% of their students from the bottom income quintile, while 10% of colleges have more than 21.0% of such students." Using a two-group Theil index, they find that "the degree of income separation across colleges is thus comparable to income segregation across census tracts in the average American city" (Chetty et al. 2017, 16).

Next, Chetty et al. turn to analyzing outcomes of students' earnings. They measure earnings when the former students are 32 years old, the point at which individuals' relative positions in the income distribution have stabilized (Chetty et al. 2017, 18). Results show that the relationships between children's earnings and their parents' incomes has a much flatter slope for each college than for the overall United States population, revealing that "parent income is no longer predictive of children's outcomes conditional on college attendance" (Chetty et al. 2017, 19). This means that low-income students who attend selective colleges are very likely to achieve upward mobility. Some non-Ivy colleges increase chances of mobility even more than the elite schools.

> Nationally, children from the highest-income families are 40 pp more likely to be in the top quintile than children from the poorest families. Conditional on attending an elite college, this gap shrinks to approximately 12 pp, and at certain colleges, such as UC-Berkeley, SUNY-Stony Brook, and Glendale Community College, the gap is even smaller, at 6-9 pp. (Chetty et al. 2017, 19)

The college with the highest overall mobility rate is California State University – Los Angeles, where 9.9% of students have parents with incomes from the bottom quintile of the

family income distribution, but reach the top quintile themselves. The top ranked colleges in mobility rates tend to be mid-tier public institutions, rather than Ivy-Plus colleges or flagship public universities (Chetty et al. 2017, 25). However, Ivy-Plus colleges perform well in the "upper-tail mobility rate" rankings, which measure the percent of students who move from the bottom quintile to the top 1% of the family income distribution (Chetty et al. 2017, 32).

The finding that earnings are similar among students of varying incomes conditional on attending the same college implies that low-income students are not generally overmatched at elite colleges, and suggest "that colleges do not pay a large cost, in terms of reduced earnings outcomes, for any affirmative action policies currently in place that favor the admission of low income students" (Chetty et al. 2017, 22). Both of these findings are consistent with Bowen et al. (2005) and point to the application stage as being the problem for low income students, because once they are admitted and attend the selective colleges, they perform well and have increased earnings after college.

Chetty et al. (2017, 37) note that their analysis does not include specific policy recommendations, but that efforts to expand access to high-mobility-colleges (mostly mid-tier public institutions) is important and may have a great effect on increasing the "overall contribution of higher education to upward mobility." They emphasize that "access rates" are what primarily determines whether a college will foster high levels of social mobility or not. However, they do not examine what determines these "access rates." There are many possible explanations, many of which may be due to student behavior, which is not discussed. Students could lack knowledge of their realistic options or choose not to go to schools with low "access rates" for various reasons.

Some insights are revealed in a qualitative study of 12 female high school seniors in California, particularly about the many factors that play into the complex college decision, and how they vary by socioeconomic status and habitus (McDonough 1997). In order to make meaningful comparisons within the small sample size, all interviewees are white females who are middle-range academic performers (McDonough 1997, 15). These students were from four different high schools, which varied in their methods of counseling students through the college choice process. The high schools with more students of high socioeconomic status (SES) had much more counseling starting early in the students' high school careers, whereas the school with mostly middle- and lower-class students didn't start college counseling until the senior year, partially because counselors were more concerned with getting students to graduation than to college (McDonough 1997, 56-86).

Students of high SES approached financing college differently from students of low SES. For high SES students, cost was of little to no influence on their college decisions (McDonough 1997, 140). Unlike high SES students who relied heavily on their parents, low SES students assumed that they would personally pay for their college and planned to have jobs in college. Some low SES students were influenced to attend colleges close to home so that they could keep their high school jobs. Community college was a more common option for low SES students, and one student expressed that if she didn't know what she wanted to major in yet, community college was the best option (McDonough 1997, 142-145).

McDonough also had several findings related to the geographical influences on students' college decision. All of the high SES students attended colleges out of state, and visited the schools at much higher rates than the low SES students. When they did visit colleges and have opportunities to have meetings and ask questions about a college, high

SES students were much more sophisticated in their search because of their previous opportunities and experiences. This included the experience of travelling alone and thus having more flexibility in visiting colleges, and knowing which kinds of questions to ask based on background that they had gained by being immersed in a network of college-minded people for most of their lives (McDonough 1997, 135-138). This fits with Coleman's definition of social capital functioning as a source of everyday information, since the high SES students with more social capital had more information to draw from when asking questions and making college decisions (Gauntlett 2011, 4).

When speaking of the distances from home of the various colleges they were considering, all students spoke in terms of travel time rather than number of miles. However, the high SES students spoke in terms of air travel time, whereas the low SES students spoke in terms of ground transportation time (McDonough 1997, 133).

The effect of socioeconomic status on college access is widespread, and not limited to the United States. A study by Jerrim, Chmielewski, and Parker (2015) compares the effect of family background on college access in the United States, England, and Australia. Academic achievement is separated into the indirect and direct effects of higher socioeconomic status on an educational transition. The indirect effect is higher academic achievement, and the direct effects include "financial resources, knowledge of the application process, information, and family connections" (Jerrim, Chmielewski, and Parker 2015, 20).

Family background is measured using international standards for parental education and the father's social class (Jerrim, Chmielewski, and Parker 2015, 24). Academic performance is measured using the Programme for International Student

Assessment (PISA) scores, as well as high school course grades and other academic achievement toward the end of high school. To determine which institutions qualify as "high-status", researchers use the Russell Group in England, the Group of Eight in Australia, and the Carnegie classification for "highly/more selective" in the United States (Jerrim, Chmielewski, and Parker 2015, 25).

To quantify the effects of SES, the authors define two groups of students: advantaged and disadvantaged. Students are advantaged if their father works in a professional occupation and they have at least one parent with a bachelor's degree or higher, and students are disadvantaged if their father is working class and neither of their parents have more education than upper secondary school (Jerrim, Chmielewski, and Parker 2015, 26).

Results show that in all three countries, advantaged students are approximately 1.8 times more likely to complete high school than disadvantaged students, after controlling for their high school performance (Jerrim, Chmielewski, and Parker 2015, 26). In England and Australia, advantaged students are approximately 1.6 times more likely than disadvantaged students to attend a non-high-status institution, compared to not attending college. In the United States, the increased likelihood from being advantaged of attending a non-high-status institution is 2.5 times (Jerrim, Chmielewski, and Parker 2015, 26). Finally, the increased likelihood from being advantaged of entering a high-status institution compared to a non-high-status institution is 1.9 times in Australia, 2.1 times in England, and 2.4 times in the United States (Jerrim, Chmielewski, and Parker 2015, 27).

C. Undermatching

Among the studies of college choice and effects of socioeconomic status, a body of literature has emerged surrounding the more specific issue of undermatching. In general, undermatching occurs when a student attends a college that is less academically rigorous than the student could handle, given their academic achievement in high school. 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/ACT scores for the student bodies of most colleges are readily available. Secondarily, high school G.P.A is often used to measure achievement. Several studies have tackled the issue of undermatching and how it is related to income and SES.

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 these selective private schools ("the COFHE schools"), or that there exist low-income, high-achieving students who are being excluded in favor of higher-income students (Winston and Hill 2005, 19.1). They set out to determine if those low-income, high-achieving students exist. Specifically, they attempt to answer the question, "Of those who meet various minimum SAT-ACT criteria – various potential specifications of high ability – how many of them come from families in each of the five income quintiles?" (Winston and Hill 2005, 19.6)

Winston and Hill (2005, 19.4) believe that a reasonable goal is for the income distribution of the COFHE schools to mirror the income distribution of highly able students in the national population. They use varying thresholds of SAT scores to define "high ability" (Winston and Hill 2005, 19.7). Results depend on the threshold used, but generally show 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. If a score of 1420 on the combined Math and Critical Reading sections of the SAT were used as a cutoff, "nearly 85 percent of the low-income, high-ability students in the United States would have to go to one of these COFHE schools in order for them to mirror national population shares" (Winston and Hill 2005, 19.11). If the threshold were reduced to 1300, the percent of low-income, high-achieving students needing to attend COFHE schools drops to 22% (Winston and Hill 2005, 19.12). The authors note shortcomings in the data related to students who do not report their family income and students who take both the SAT and ACT tests, thus, are doubly counted in the data. However, evidence shows that these effects are more likely to understate the amount of low-income, high-ability students in existence, since low-income students are more likely to not report income than high-income students, and separately analyzing the SAT and ACT data gives similar results to the combined analysis (Winston and Hill 2005, 19.22, 19.27).

So, we see that large numbers of high-achieving, low-income students exist. A natural question that follows is *why* these students are not attending selective colleges. A study by Hoxby and Avery (2012, 1) shows that there is a large portion of these students who don't ever 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 2012, 10). Hoxby and Avery estimate each student's family income using data on several predictive variables from the students being studied, as well as data on previous cohorts of College Board students, for whom they have access to their CSS Profile records (which are used to compute grants and loans by financial aid officers). Specifically, they "regress accurate administrative data on family income using all of our previous Census variables, the IRS income variables, the high school profile variables, and the student's own race and ethnicity" (Hoxby and Avery 2012, 13).

Using the schools to which each student sends their SAT or ACT scores as a measure of which schools that student applies to, Hoxby and Avery (2012, 26) 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, whom Hoxby and Avery call "achievement-typical". This behavior follows the advice of expert counselors. High-income, high-achieving students apply mostly to peer schools, to some reach schools when possible (some students score so highly that no reach schools exist), fairly frequently to safety schools, and often to their state's flagship university (Hoxby and Avery 2012, 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 2012, 21). Eight percent of low-income high-achievers fall into this "achievement-typical" category, by 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 2012, 26).

There is another group of low-income, high-achieving students who apply using a different strategy, named "income-typical" students. These students "apply to no school whose median score is within 15 percentiles of their own, and they do apply to at least one nonselective college." Fifty-three percent of low-income, high-achievers fall into this category (Hoxby and Avery 2012, 26). Finally, the remaining 39 percent 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 2012, 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 2012, 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, after which they are indifferent (Hoxby and Avery 2012, 30-31). 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 further than 100 miles away (Hoxby and Avery 2012, 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 2012, 31). Thus, it is primarily in the application stage that low-income, high-achieving students who could attend selective colleges are being lost.

Descriptive statistics show that geography is the most striking factor that separates income-typical students from achievement-typical students. Hoxby and Avery (2012, 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 2012, 42).

Although Hoxby and Avery's study has the advantage of being nation-wide, there have been several studies on undermatching restricted to certain areas of the United States. Bowen, Chingos, and McPherson (2009, 93-94) 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, 104).

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. This mostly agrees with Hoxby and Avery's (2012, 31) and Bowen et al. (2005) finding that most students are lost at the application stage.

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). The authors also noted that in some cases, students may have good reasons to undermatch (such as a desire to be near to home or family) and in fact are maximizing their utility by attending a school that is less selective than they could be accepted to. However, this should not be the norm and reasons of lack of information and planning are not good reasons for a student to undermatch (Bowen, Chingos, and McPherson 2009, 101).

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 (2012), Lincove and Cortes (2016, 9) use slightly different measures of safety, match, and reach schools. A safety school has a median SAT scores 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 (Lincove and Cortes 2016, 16-17). Results are similar for enrollment rates (Lincove and Cortes 2016, 17).

In their ordinary least squares regression analysis, 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 (Lincove and Cortes 2016, 18). 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).

Results from conditional logit regression show that "low-income students with automatic admissions are 15 percent less likely to apply to a campus where they will be undermatched, relative to a campus where their SAT scores are similar to or below the median," but "high-income students with automatic admissions…are 68 percent more likely to apply to a safety school" (Lincove and Cortes 2016, 20). There are not significant differences in matching behavior between low-income and high-income students in the top 11-25% of class rank (Lincove and Cortes 2016, 20). Enrollment results are similar to these application results, suggesting that "low-income students are less likely to apply to and enroll at undermatched campuses when they have perfect admissions information, where high-income students are more likely to apply to undermatched campuses at all class ranks" (Lincove and Cortes 2016, 21). So, automatic admissions may have an equalizing effect across income groups (Lincove and Cortes 2016, 22).

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).

D. Key Findings From Literature

This literature review has revealed that although student college choice is a complex process affected by many determinants and layers of context, socioeconomic status undoubtedly plays a role in the United States and other developed nations. Students with higher family income and parental education are more likely to attend college in general, and more likely to attend elite universities. Differences in academic preparedness accounts for part of this gap, but substantial differences still exist once researchers have controlled for academic achievement. Varying levels of social capital among students plays a role in their college search processes, from differences in college counseling to levels of understanding of the financial aid system.

Low-income students with high levels of academic high-school achievement perform well in college, especially at selective schools. However, there is a large population of students who undermatch with their college 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 in 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-adverse 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. This study attempts to investigate the relationship between distance from home and undermatching, and how the relationship differs with family income.

III. Methodology

A. Mechanical Turk

Data for this thesis was obtained through a survey distributed on Amazon's Mechanical Turk. Mechanical Turk (MTurk) is "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. It has been shown to be a reliable way to quickly obtain high-quality data at a 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" (Frequently Asked Questions 2015). 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, and are posted by requesters to be completed by workers for a monetary reward (Paolacci, Chandler, and Ipeirotis 2010, 411-412). 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 written by the requester (Paolacci, Chandler, and Ipeirotis 2010, 412). 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 (Paolacci, Chandler, and Ipeirotis 2010, 412). All workers and requesters are anonymous, and requesters can only link responses to unique worker IDs assigned by Amazon (Paolacci, Chandler, and Ipeirotis 2010, 412).

Rewards paid to workers are generally very low, between $0.01 and $1.00 per simple task (Paolacci, Chandler, and Ipeirotis 2010, 412). 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 a 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, at 55% female and 57% female, respectively (Buhrmester, Kwang, and Gosling 2011, 4). Poalacci, Chandler, and Iperirotis (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, 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 cost of the 1073 responses used in this project was $1500.00, and data was collected within 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 this 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. Mechanical Turk proved to be an ideal way to collect survey data quickly, with a limited budget.

B. Survey Design

Previous studies on undermatching informed survey questions about respondents' academic achievement to determine if they had undermatched. Questions were also asked about students' income to determine its effect on undermatching. Further, because distance from home is a focus in this study, questions were needed about the student's decision-making process and how distance played a role in it. These questions, as well as many other questions included in the survey, were influenced by literature and conversations with the members of the author's thesis committee. Appendix I includes the survey questionnaire that was distributed, annotated to include how questions were developed.

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 send to the end of the survey.

C. Data Cleaning

Not all survey responses could be used for analysis, and some responses needed to be manipulated prior to being analyzed. Survey responses were deleted or modified for one or more of the following reasons: the respondent completed the survey outside of the United States, multiple responses came from the same IP address, the respondent did not provide a valid SAT score, or the respondent provided an ambiguous name of a college. This subsection explains how these reasons for data deletion/manipulation were identified and handled.

On Mechanical Turk, requesters can set qualifications that their respondents must meet in order to be able to complete HITs. When collecting data, the qualification that respondents must be from the United States was set. However, Mechanical Turk's screening system is not perfect, and some respondents from outside the United States were still able to take the survey. Qualtrics automatically tracks the location of each survey respondent, and includes the longitude and latitude values as variables in the data set of survey responses. These longitude and latitude values were used to determine which respondents took the survey outside of the United States, and those responses were removed from the data set.

There could be legitimate reasons for a respondent who lives in and attended college in the United States to submit a response from outside the United States, such as a respondent who was travelling or moved to a foreign country after completing college in the United States. However, since the probability of Mechanical Turk workers lying about where they live in order to complete more HITs is non-trivial, all observations from outside of the United States were removed to make sure the data set contained only responses from individuals who lived in and attended college in the United States. The total number of observations removed was 32.

Knowing a respondent's SAT or ACT score is vital to this study, since it is the primary measure used to determine if a student undermatched. Following suit of previous studies on undermatching, this study uses students' combined critical reading and mathematics SAT scores, which is scored on a scale from 0 to 1600, in increments of 10. If a respondent entered an SAT score that was not a multiple of 10, it was considered invalid. Considering that respondents who entered invalid SAT scores were likely not answering

questions honestly and thoughtfully, all responses with invalid SAT scores were removed from the sample. 58 observations were deleted using this criterion.

In order to deter respondents from submitting multiple responses, a Qualtrics feature called "Prevent Ballot Box Stuffing" was used, which places a cookie in the respondent's browser when they take a survey for the first time, and does not permit them to take the survey again as long as the cookie remains in the browser. However, respondents who clear their browser cookies or use a second browser to complete the survey a second time can avoid this restriction (Survey Protection 2017). Fortunately, multiple responses submitted from the same IP address could be identified in the data. If one IP address was connected to more than one survey responses, the first response was kept, and all subsequent responses were deleted. This caused 35 responses to be removed from the sample.

Finally, many survey respondents answered the question about which colleges they applied to with ambiguous names. The author manually matched college names to a standardized list of colleges and their zip codes obtained from the Integrated Postsecondary Education Data System (IPEDS), using several rules to identify which school the respondent most likely meant when their response did not exactly match the name of a college on the list. For example, when a respondent indicated that they attended "Ohio State", that was taken to mean that they attended the main campus of Ohio State University, rather than one of the branch campuses. However, some responses were too ambiguous to be clearly matched to a college on the list, because the respondent used abbreviations that were not clear. These colleges, along with colleges outside of the United States, were marked as missing values.

IV. Theoretical Framework

This thesis seeks to model a student's probability of undermatching. Undermatching is modeled as a dummy variable rather than a continuous variable, meaning that a student either undermatches or does not undermatch. The definition of undermatch will be described in more detail in the Data subsection of the Results section below. This section describes the intuition behind a dummy dependent variable model, which will be used for analysis in this study.

With a dummy dependent variable model, as with continuous dependent variable models, there is an error term that includes the randomness associated with generating the observed value (Barreto 2006, 666). Whether a student undermatches or not cannot be entirely explained by measurable characteristics of the student. No matter how many variables are included in the regression, there will still be some randomness in who undermatches and who doesn't. Consider a distribution from which to draw the error term for each student. The distribution has a threshold, and if the error is drawn from above the threshold, the student will undermatch. If the error is drawn from below the threshold, the student will not undermatch.

Figure 1 shows one such distribution with a threshold represented by the red line. Note that the threshold was arbitrarily placed for purposes of demonstration. In this case, since the distribution is normal and the threshold is placed one standard deviation to the right of the center of the distribution, approximately 84% of the distribution lies below the threshold, and approximately 16% of the distribution lies above the threshold (these percentages come from the properties of the normal distribution). In this case, 16% of the errors drawn from this distribution would imply that a student undermatched.

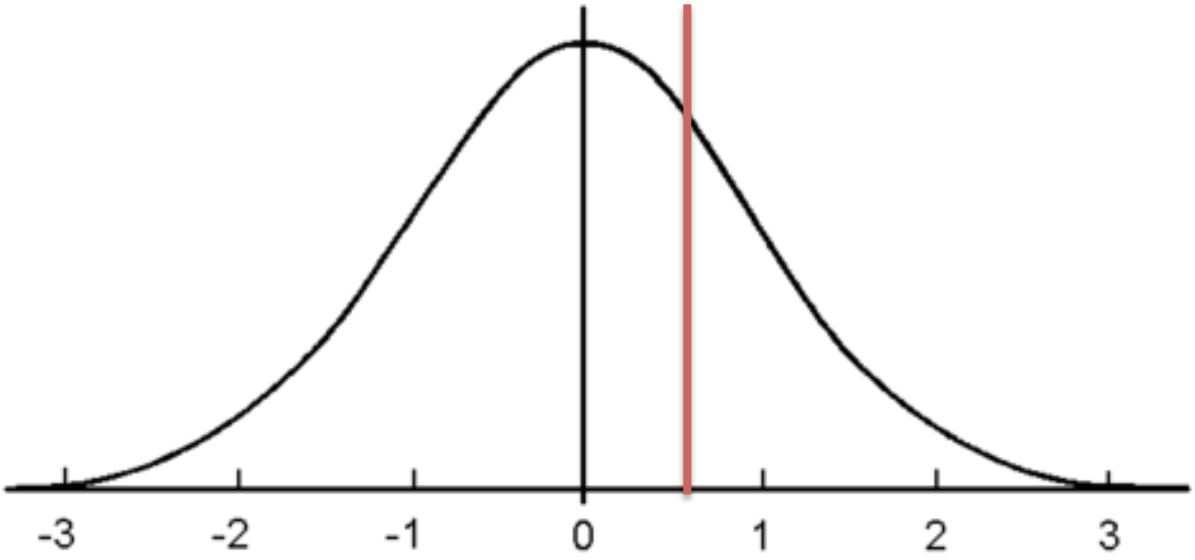

Figure 1. *Error distribution with threshold one standard deviation to the right of center*

Of course, whether a student undermatches is not entirely random, and there are variables that determine the threshold value. A student's income will have an effect on where the threshold value in placed. Figures 2 and 3 show two theoretical distributions, one for low-income students and one for high-income students. The figures have been constructed in a way so that low-income students are more likely to undermatch than high-income students. Focusing on figure 2 for low-income students, there is a still an element of randomness (and effects of other variables) that determines whether or not they undermatch. However, since the threshold is lower (father to the left) for low-income students, there is a greater chance that the error drawn from the distribution will be above the threshold, and thus that the student undermatches.

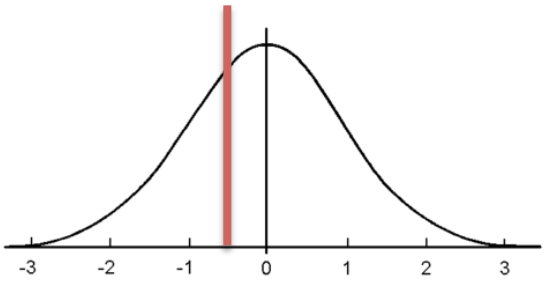

Figure 2. *Theoretical distribution of errors with threshold for low-income students*

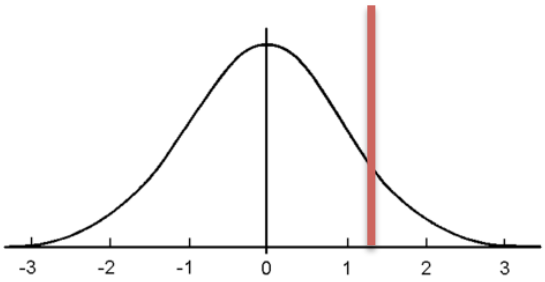

Figure 3. *Theoretical distribution of errors with threshold for high-income students*

However, with empirical data, the error distribution and threshold values are unknown. All that is observed is the outcome, that is, whether a student undermatches or not. Using a dummy dependent variable model allows us to estimate the threshold value for specific variables, which gives how these variables affect an individual's probability of undermatching. This study will be primarily focused on two variables: a student's household income, and the distance between a student's home and college. Using a probit regression, the effects of each of these two variables on a student's probability of undermatching will be estimated.

V. Results

A. Data

All results are calculated using the data obtained through the original study published on Amazon's Mechanical Turk. Data was cleaned in Microsoft Excel as described in the Methodology Section above. STATA was used for analysis leading to the following results. Microsoft Excel was used again for charts in the Understanding Fundamental Results subsection below.

The dependent variable, *undermatch*, indicates whether an individual has undermatched with the college they attended. It is constructed using the student's reported SAT/ACT score and the median SAT score for the college that they attended. 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 by the author using a concordance table. (ACT 2009) SAT scores for most colleges were obtained through the Integrated Postsecondary Education Data System (IPEDS) (U.S. Department of Education 2014). 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 median 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, the author manually 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 for the author to be able to label the school "nonselective". For the remaining 8 colleges in which case the author was not able to obtain an SAT score, their SAT score was reported as a missing value.

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, in order 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 percentage points lower than their own, they have undermatched. If they attend a college with a median SAT less than 15 percentage points below their own, or if they attend a college with a median SAT score higher than their own, they have not undermatched. The variable *undermatch* takes a value of 1 if the student has undermatched, and 0 otherwise.

For all analyses, the sample was restricted to students who scored in the top 10 percent of the SAT score distribution. Originally, analyses were performed with the entire sample, but many of the results were not statistically significant. This may be because students who do not score very high on the SAT don't have much of a chance of undermatching, since by definition, they would need to be scoring 15 percentage points higher than the median SAT of the college they attend. 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. Following suit of Hoxby and Avery (2012), this thesis only includes students who score at the 90th percentile or above to be sure the focus is on high-achievers. After using this cut-off, the sample size is 338. Table 1 shows the frequency of students who undermatched by this definition.

Table 1. *Frequency of Undermatching*

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

The 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, the author manually obtained the zip codes using Google Maps. *Distattend* gives the fastest driving distance

between these two zip codes for each observation, as given by Google Maps (Google Maps 2017). Table 2 shows summary statistics for *distattend*, and Figure 3 shows its distribution.

Table 2. *Distance between student's home and college*

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

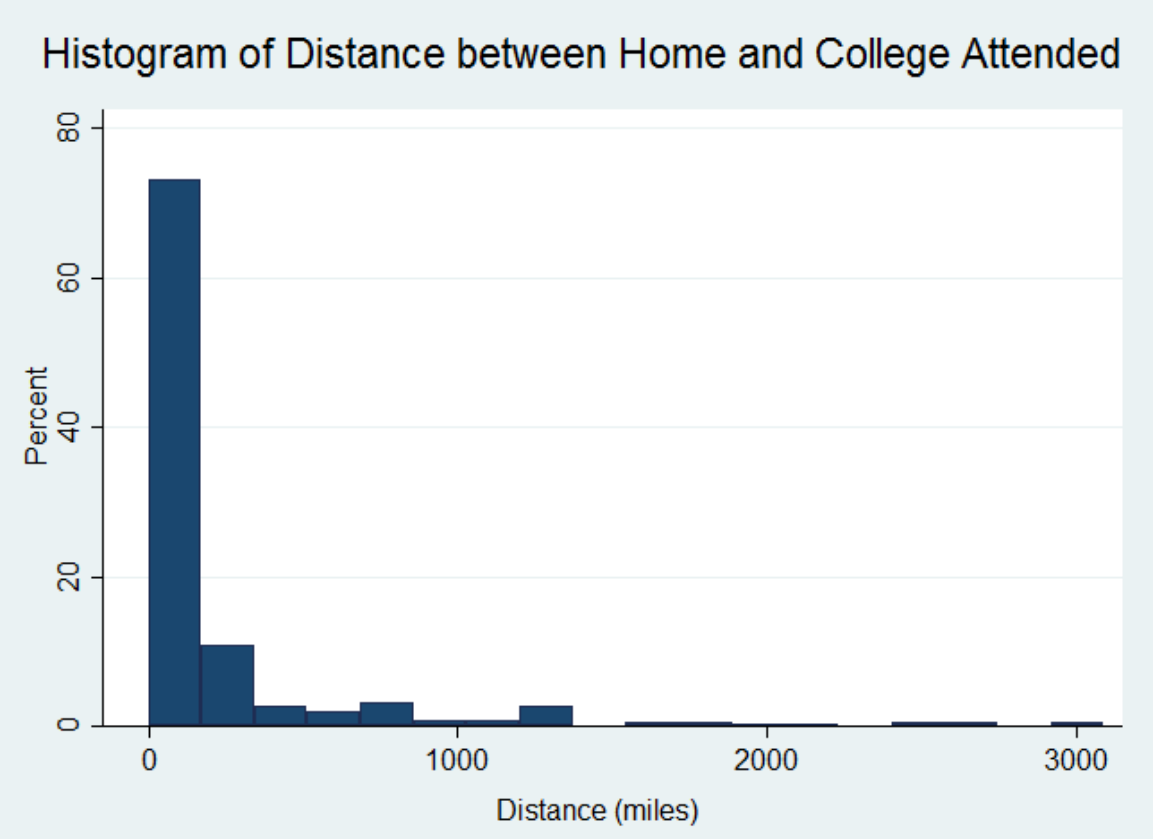

Figure 3. Distribution of *Distattend*

Income is an indicator variable that 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 labeled as high income. Students with family incomes between $40,000 and $80,000 are labeled as middle income.

Table 3. *Income Categories*

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	

Demographic variables are included in the regression as controlling variables. These include three dummy variables for race, gender, and whether the respondent is Hispanic. It is important to note that by including these controlling variables, when this thesis talks about the effect of income of undermatching, this is the effect of income *holding these demographic variables constant*. Although much work has been done on the role of race and gender in education and undermatching, Hoxby and Avery (2012, 18) show 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.

Since this study seeks to determine the effect of income of undermatching, regardless of race and gender, these variables are included in the regression. Tables 4, 5 and 6 show the frequencies of race, gender, and if the student is Hispanic.

Table 4. *Frequencies of Race*

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 4. *Frequency of Gender*

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

Table 5. *Frequency of Hispanic*

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

There are several limitations to the data used in this thesis. First, because survey data is being used, all answers are self-reported and may not be accurate due to respondents misremembering or not having full information to answer some survey questions. For example, a respondent may misremember the zip code of their hometown when applying to college, which would affect the value of *distattend*. Second, because survey respondents were being paid, there was some incentive for a respondent to complete the survey even if they did not fill the eligibility requirements. Also, some respondents could be just clicking randomly to get through the survey as quickly as possible, and not providing thoughtful answers. Much of this problem has been eradicated by including several attention checks in the survey as described in the Methodology section and in Appendix I, but there remains a possibility that some respondents were able to get through the attention checks without providing thoughtful answers to all questions.

Additionally, there may be determinants of undermatching that are not included as variables in the regression. Not all determinants of undermatching are known or able to be easily measured, which restricts how undermatching can be modeled. Nonetheless, the

regression provides insight into how income and distance from home affect undermatching, holding several key demographic variables constant.

B. Regression Results

Both ordinary least squares (OLS) and probit regression techniques were used to estimate the effects of distance from home and income on undermatching. The OLS model is included because of the ease of the interpretation of its coefficients, but has several disadvantages compared to the probit model. The OLS model forces a constant slope to a relationship that may be nonlinear and suffers from heteroskedasticity, which implies that interpretation of results may be unreliable. Probit regression models the true relationship better by allowing it to be nonlinear. However, the coefficients on the probit regression cannot be interpreted directly, and additional analysis of predicted probabilities is needed. Results for both models are shown in Table 7.

First, the OLS model is discussed. Primary interest is in the coefficient on *distattend*, which estimates the effect of the distance between a student's home and college (in miles) on their probability of undermatching. In multiple linear regression, when interpreting the coefficient on a single variable, it must be considered that all other included variables are being held constant. So, 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. The estimated coefficient is -0.000203, which means that an additional mile between a student's home and the college they attend gives a .02 ± .005 percentage point *decrease* in the probability of undermatching, holding all other included variables constant. When considering distances between students' homes and colleges across the United States, 1 mile is trivial, which explains why the coefficient is so small. To gain a

better understanding, multiply the coefficient by 100 to estimate the effect on the probability of undermatching of a student who attends college 100 miles further away. So, a student who goes to college 100 miles further away decreases his or her probability of undermatching by 2 ± .5 percentage points, or 1.5 to 2.5 percentage points. Similarly, a student who goes to college 500 miles further away decreases his or her probability of undermatching by 10 ± 2.5 percentage points.

Table 6. *Determinants of Undermatch*

	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	No	Yes	Yes
R^2/Pseudo R^2	.0861		.1091

N = 338 for all models

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

The coefficient on distattend is statistically significant with >99.9% confidence, since the hypothesis test against the null hypothesis that the coefficient equals zero is less than 0.001. This means that if we assume that the coefficient on distattend is equal to zero (and thus the distance between a student's home and college has no effect on his or her probability of undermatching), the chance of observing a value equal to the coefficient or more extreme is less than 0.1%.

Second, this thesis is interested in the coefficients on the income variables. Recall that students were grouped into three categories based on their reported family income: low income for less than $40,000/year, middle income for between $40,000/year and $100,000/year, and high income for greater than $100,000/year. Middle income is treated as the baseline in our regression, so when interpreting the coefficient on low income, we are estimating the effect of being low-income on a student's probability of undermatching, *compared to a middle-income student.* The coefficient on low income is .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 implies 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 lower income students are more likely to undermatch than higher income students.

Readers may notice that unlike distattend, the coefficients for low income have not been marked as statistically significant based on their p-values. This does not necessary mean that being low-income does not have a significant effect on the probability of undermatching. Because income is an indicator variable with multiple categories, in order

to determine if income has a significant effect on the probability of undermatching, all categories of income must be considered jointly. Single coefficients cannot be interpreted independently, since they are dependent on the coefficients of the other income categories. When considered jointly, income has a significant effect on the probability of undermatching.

Note that the R^2 value on the OLS model is .1091, which means that only about 11% of the variation in undermatching is explained by the displayed variables. Thus, there are additional determinants to the probability that a student undermatches that have not been accounted for in this regression.

Although there is some value in using the OLS model to determine the effects of distattend and income on undermatching, it is limited by making the assumption that the relationships between the variables are linear. So, in the OLS model, each additional mile between a student's home and college gives the same decrease in their probability of undermatching. The probit model relaxes this assumption, but requires further analysis.

Coefficients from the probit model cannot be interpreted directly. The second column of Table 7 gives the percentage point impact of a change in the variable. For categorical variables, the number in column 2 is the percentage point impact associated with having that characteristic. For example, males are 11 percentage points less likely to undermatch. Since income is a categorical variable with more than one category, the percentage point impact is compared to the baseline (in this case, middle income). The impact of income on the probability of undermatching is discussed in further detail below. 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. Since the

standard deviation of *distattend* is 479.8, the number in column 2 of table 7 indicates that if a student goes about 480 miles further away for college, they are 9.8 percentage points less likely to undermatch.

Table 8 shows the probabilities of undermatching at varying levels of *distattend*, holding all other variables at their means. Note that because other variables are being held at their means, this table does not take into account the difference in probability of undermatching between income categories. These differences will be discussed in Table 10 below. Table 8 shows that 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 3000 miles away from home has only a 7.1% probability of undermatching. The rightmost column shows the 95% confidence intervals for each distance, which has been constructed using the standard errors reported in the Std. Err. column. This means that 95% of intervals constructed this way will contain the true value for the probability of undermatching given a certain distance between a student's home and college. For example, if a student goes to college 500 miles away from home, it can be said with 95% confidence that they have a probability of undermatching that is between 46.4% and 59.9%.

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

	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

These results clearly show that as distance from home increases, a student's probability of undermatching decreases. Next, we will analyze the effect of income category on an individual's probability of undermatching, holding *distattend* and all other included variables at their means. Table 9 shows the marginal effect of a student's income category on their probability of undermatching. Recall that middle income is being used as the baseline, so the marginal effect of .074 on low income means that if you have two otherwise average individuals, but one is low-income and one is middle-income, the low-income student is 7.4 percentage points more likely to undermatch. This is similar to the results from the OLS regression, but results from the probit regression can be interpreted with more confidence since it does not suffer from heteroskedasticity or force a linear relationship. The marginal effect of high income can be interpreted similarly. Holding all other variables at their means, a high-income student is 18.6 percentage points less likely to undermatch than a middle-income student.

Table 8. *Marginal Effect of Income on Undermatch*

	dy/dx	Std. Err.	z	P>z	95% Conf. Interval	
Income						
low	.074	.069	1.09	0.275	-.059	.207
high	-.186	.063	-2.80	0.005	-.316	-.056

Table 9. *Predicted Probabilities by Distance and Income Category*

	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

So, we have shown that both being low-income and going to college closer to home increases a student's probability of undermatching. This thesis is also interested in determining if distance from home affects the probability of undermatching differently among the different income groups. That is, it seeks to determine if the effect of going to college further from home is different for low-income students than it is for high-income students. Table 10 shows 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, a middle-income student who goes to college 500 miles away has a 56.1% probability of undermatching, and a high-income student who goes to college 500 miles away has a 38.2% probability of undermatching. Figure 4 plots these probabilities.

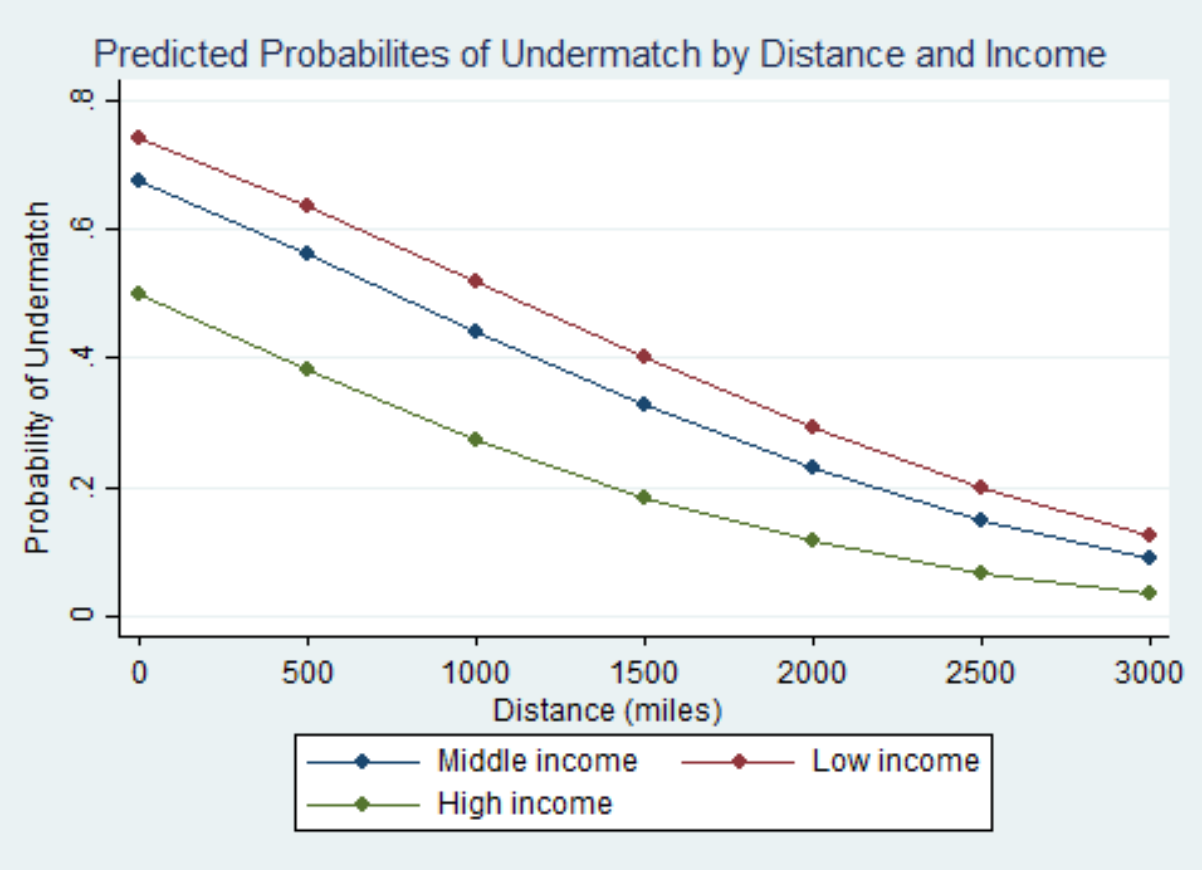

Figure 4. *Predicted probabilities of undermatch by distance between student's home and college, and income category*

Figure 4 shows how the relationship between distattend and undermatch varies by income category. As can be seen in the figure, for all income categories, the probability of undermatching decreases with distance. However, it does not decrease at the same rate for all income categories. In order to more clearly see how this effect varies between high-income and low-income students, Table 11 shows the marginal effect of each income

category at varying distances. Looking at the second row under "Low income", it is shown that if a student attends college 500 miles away from home, holding all other variables constant, the effect of being low-income increases their probability of undermatching by 7.6 percentage points, compared to middle-income students. To see the effect of being high-income at the same distance from home, look at the second row under "high income". The -.180 indicates that holding all other variable constant at their means, a high-income student who attends college 500 miles from home is 18 percentage points less likely to undermatch than a middle-income student who attends college 500 miles from home.

Table 10. *Marginal effects of income on undermatch at various distances*

	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

Figure 5 plots these effects. The blue line shows the marginal effect of being low-income on a student's probability of undermatching at various distances, whereas the red line shows the marginal effect of being high-income. 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. So, high-income students have a greater advantage over low-income students in terms of matching at 500 miles from home than at 3000 miles from home. This shows that the effect of distance on a student's probability of undermatching varies with income.

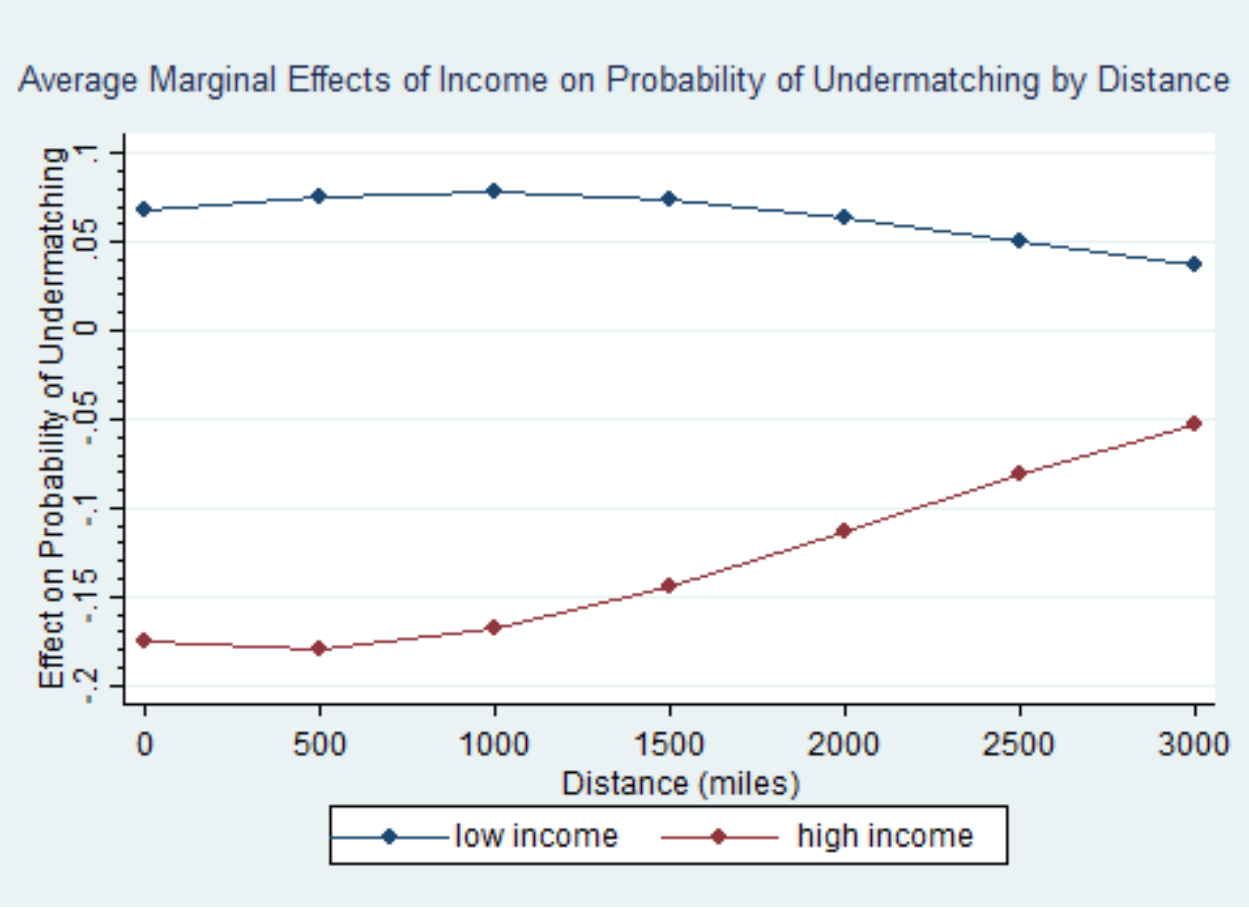

Figure 5. *Marginal effects of income on probability of undermatching at various distances between a student's home and college*

These regression results have shown 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, since it has been demonstrated in previous literature. Additionally, results show that increasing the distance between a student's home and college decreases their probability of undermatching,

regardless of income. 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.

Since probit regression does not minimize variance in the way OLS does, R-squared values cannot be calculated. However, a pseudo R-Squared statistics to assess the goodness-of-fit. In McFadden's pseudo R-Squared, "The log likelihood of the intercept model is treated as a total sum of squares, and the log likelihood of the full model is treated as the sum of squared errors… The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model. A likelihood falls between 0 and 1, so the log of a likelihood is less than or equal to zero. If a model has a very low likelihood, then the log of the likelihood will have a larger magnitude than the log of a more likely model. Thus, a small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model" (FAQ: What are pseudo R-Squareds? 2011). The McFadden pseudo R-squared for the probit model is .0861, which indicates that there are other factors that affect undermatching that are not included in the regression. Further research is needed to determine factors that affect undermatching.

C. Understanding Fundamental Results

Regression conveys results with the most statistical power, but is often not to best tool to make conclusions easy to understand. Graphical analyses do not control for variables as regression does, but can be useful in visualizing key results. Additionally, the survey distributed on Mechanical Turk for this study included many questions that were not transformed into variables to be used in the regression analysis. These variables can be used in graphical analyses to complement results from the regression findings, as well as conclusions from previous literature. Another way to foster understanding of key findings is to expand upon individual respondents, and tell their stories. This section provides additional insights into the role of income and distance on undermatching. Part a presents a series of charts using survey variables that were not included in the regression analysis to support key findings, and part b tells the stories of four individual students and their college search process.

a. Graphical Analyses

The survey asked respondents, "Did you apply to any colleges that you would consider prestigious or elite?" Figure 6 shows the percent of students who answered "Yes" to this question, by income category. The chart shows that low-income students were the least likely to apply to an elite college, high-income students were the most likely to apply to an elite college, and middle-income students fell in the middle. This supports the findings of Hoxby and Avery (2012), which show that of high-achieving students, low-income students are less likely to apply to selective colleges than their high-income peers.

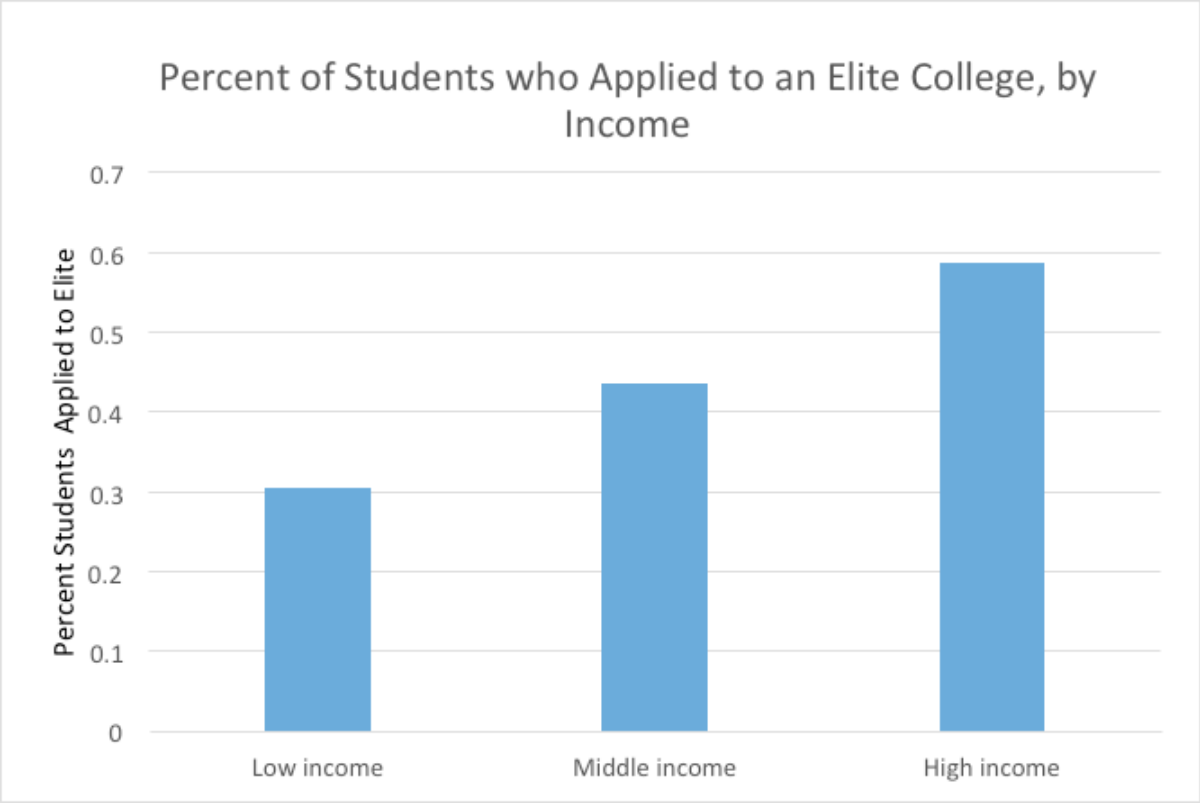

Figure 6. *Percent of students who indicated that they applied to at least one "prestigious or elite" college, by income category*

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 7, which shows the percent of students who undermatch by whether they applied to an elite college or not. The same definition of undermatch is used as described in the Data section above, that is, a student undermatches if they attend a college with a median SAT score at least 15 percentage points below the student's own SAT score. Nearly 78% 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.

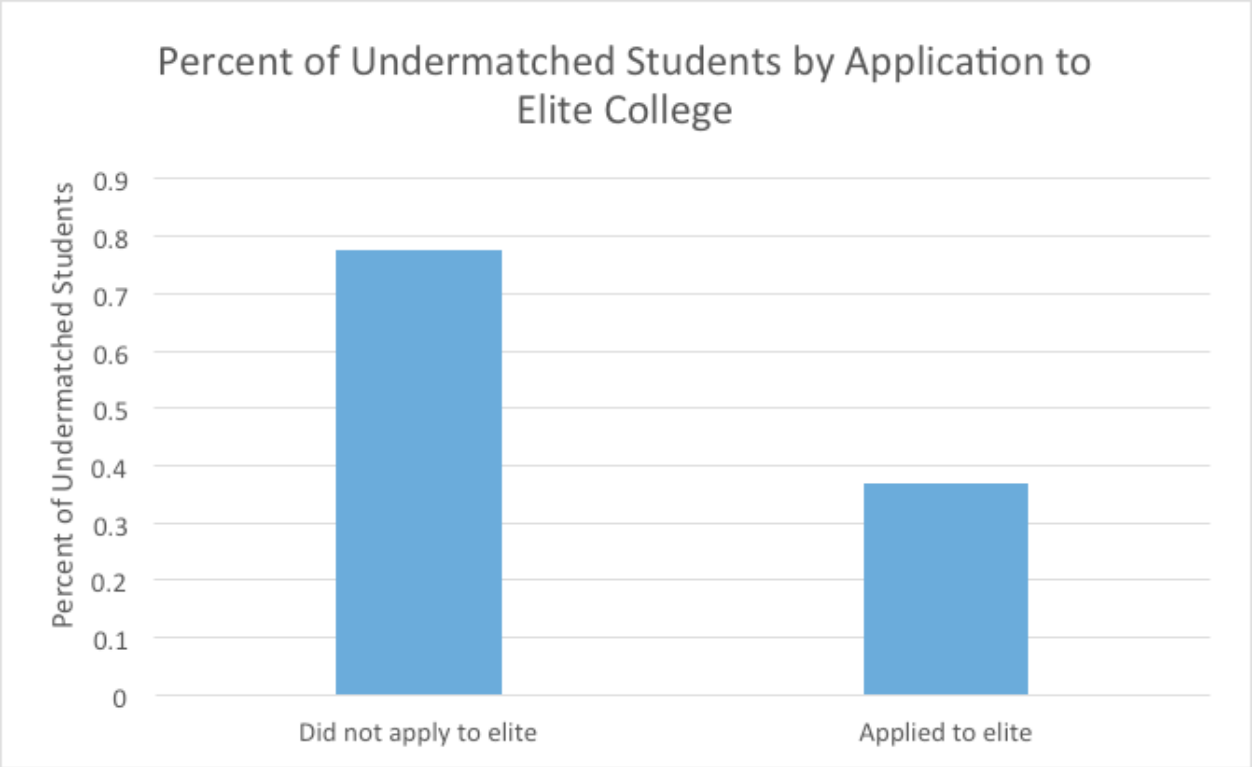

Figure 7. *Probability that a student will undermatch with the college they attend, by whether or not they applied to a "prestigious or elite" college*

Figures 6 and 7 have shown that low-income students are less likely to apply to elite colleges, and that applying to an elite college decreases the chance that a student will undermatch. This explains a portion of the difference in undermatching between low-income and high-income students, but figure 8 shows that it does not account for the entire gap. Because the results were very similar for low-income and middle-income students, the two categories have been collapsed into one for clarity.

Even when only considering students who applied to at least one elite college, low- and middle-income students are 11 percentage points more likely to undermatch than high-income students. Of students who did not apply to any elite colleges, low and middle-income students are 9 percentage points more likely to undermatch than high-income students. This is an additional insight that was not shown in the regression results, but

supports the finding from the regression that in general, high-income students are less likely to undermatch. This evidence is somewhat contrary to Hoxby and Avery (2012) and Bowen et al. (2005), whose findings indicated that if high-achieving low-income students apply to selective colleges, their socioeconomic status does not affect their progression through the stages of being admitted and enrolling in these selective institutions. However, it is important to note that this graphical analysis does not include any controlling variables as regression analysis would. So, although it suggests that high-income students are less likely to undermatch conditional on applying to an elite college, further research is needed to fully answer the question.

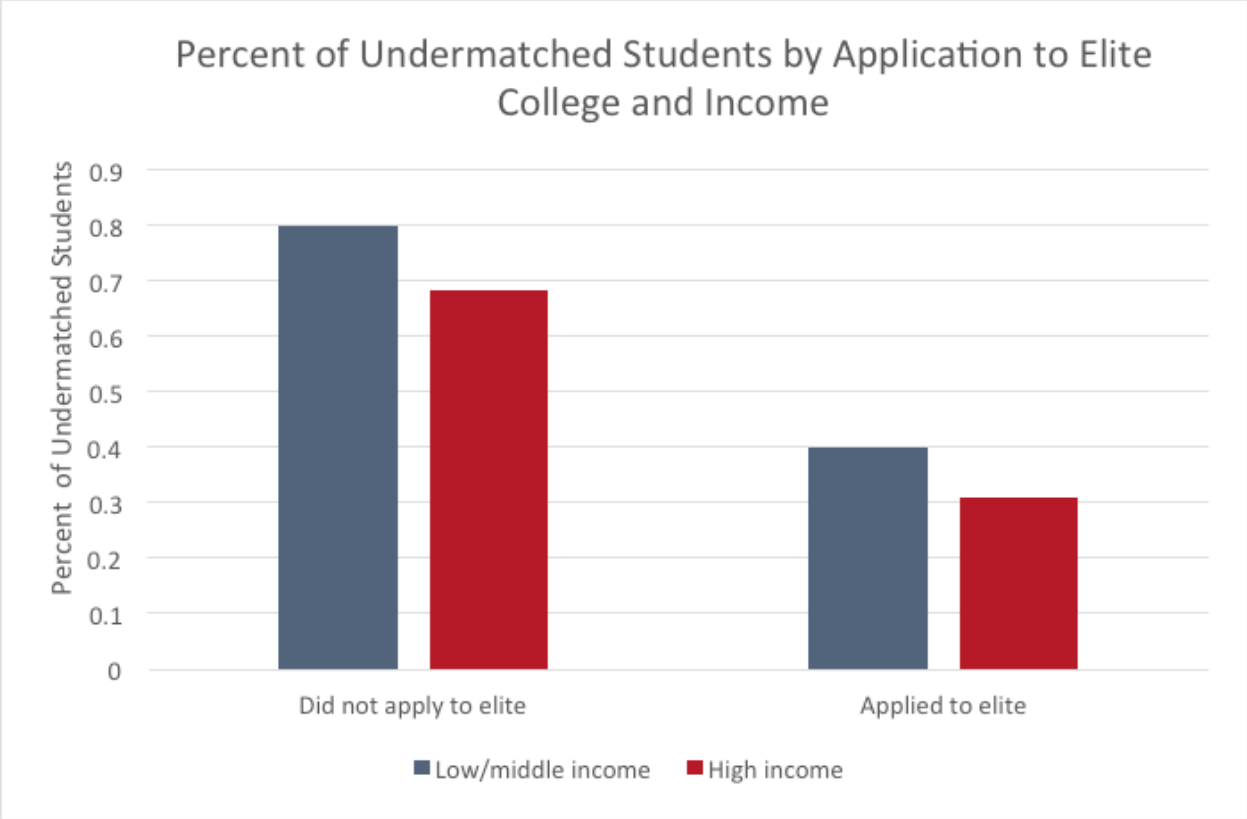

Figure 8. *Percent of students who undermatched, by whether the applied to "a prestigious or elite college" and income category*

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 poverty, 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 Dynarski (2016) that demonstrate that students who were persistently eligible for FRPL fared worse academically than those who were intermittently eligible. Figure 9 shows the percent of undermatched students by general FRPL eligibility, and figure 10 shows the percent of students who undermatched by the number of years they were eligible for FRPL.

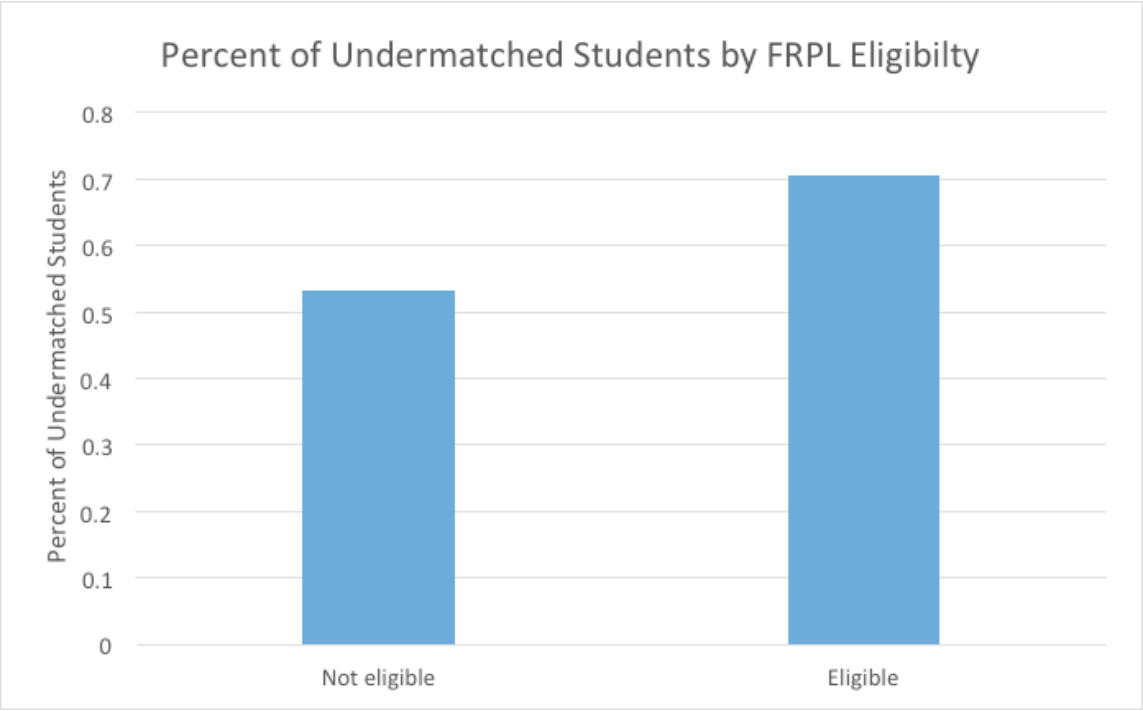

Figure 9. *Percent of students who undermatched by whether they were ever eligible for the free- or reduced-price lunch program*

As figure 9 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 Dynarski's (2016) results, the number of years that a student was eligible for FRPL does not seem to affect their probability of undermatching much. Once again, these results should be interpreted with caution, since no other variables are being controlled for.

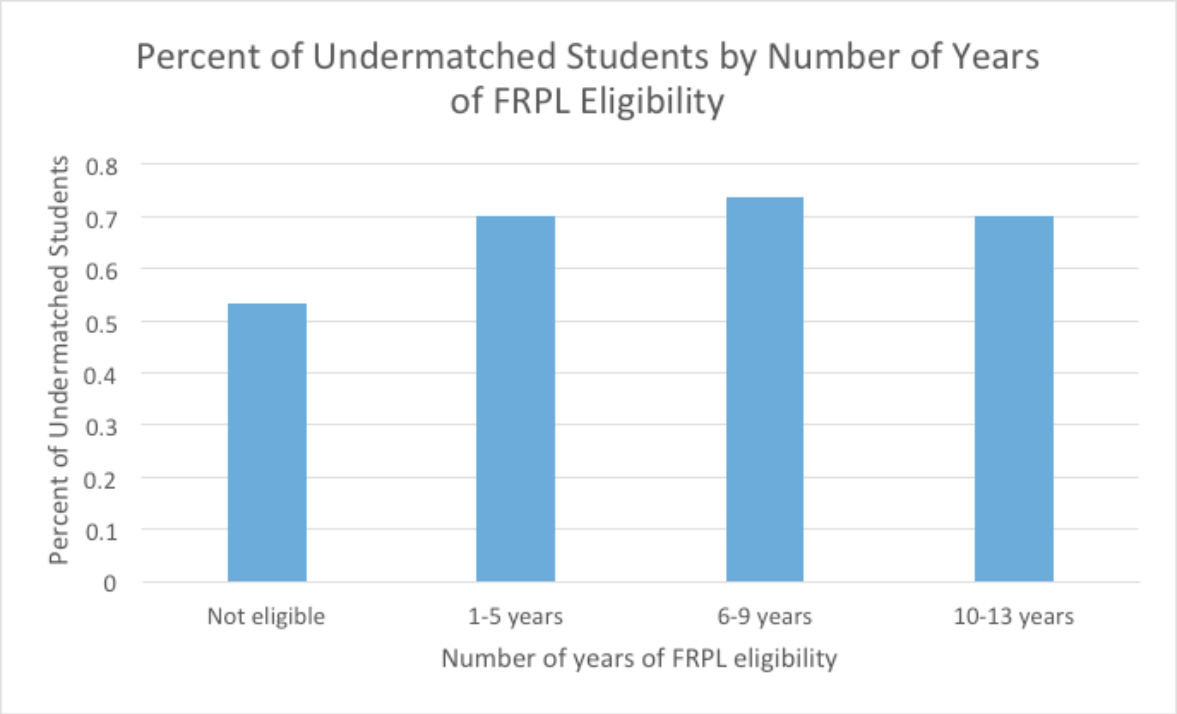

Figure 10. *Percent of students who undermatched by how many years they were eligible for the free- or reduced-price lunch program*

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. The author read through all responses to this question and identified four main factors that appeared throughout the responses. Each response was marked with whether it mentioned anything in the following four categories: 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 11 shows the percent of students who mentioned each of the four factors in their open-ended response, by income category.

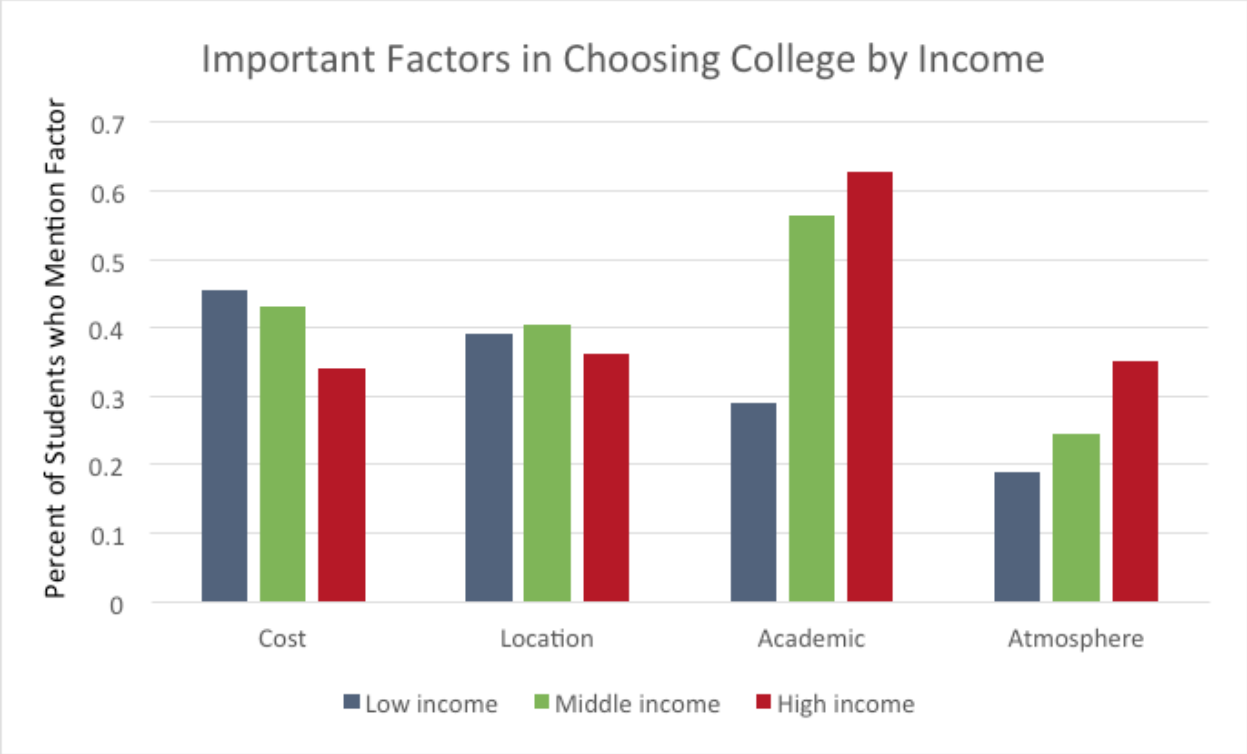

Figure 11. *Percent of students who mentioned each of four main factors as important in their college choice, by income category*

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 11

shows that high-income students are more likely to mention academics then 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 12 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 12 uses a separate measure than the regression, 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 regression, which found that the farther from home a student attends college, the less likely they are to undermatch.

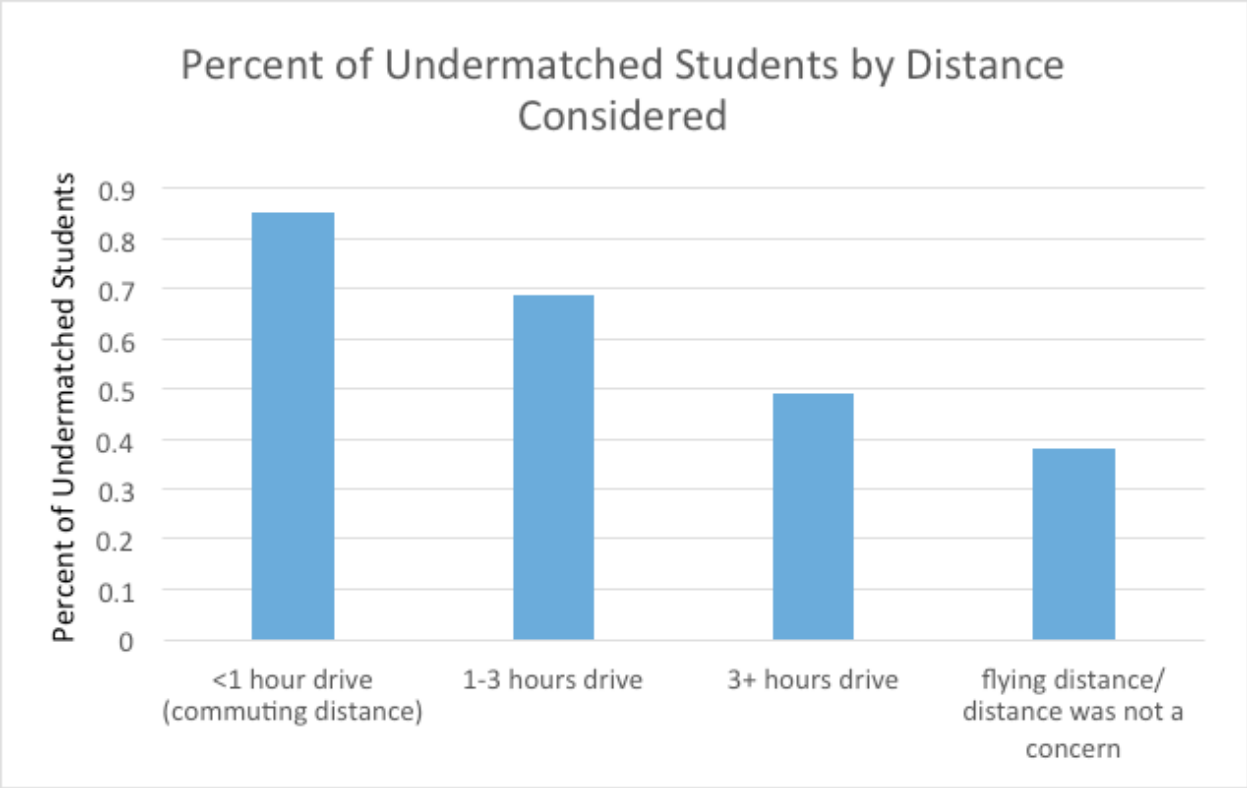

Figure 12. *Percent of students who undermatched by the farthest distance from home they considered going to college*

Results from this graphical analysis serve to strengthen the regression results, since they come to the same general conclusions using alternative survey questions as data. The charts in this section have shown that students are less likely to apply to elite colleges and more likely to undermatch if they are from families with lower income. The farther away from home students consider going to college, they less likely they are to undermatch.

b. Individual Cases

Often, specific stories can foster further understanding of statistical results. To complement the regression and graphical results, this section will present four individual cases of survey respondents, so that readers can get a sense of what a student goes through in their college search and how it contributes to undermatching. The respondents will not be identified.

Respondent A is a high-income student who scored in the 98th percentile on the SAT, and received mostly A's in high school. Both of his parents graduated from college. He applied to 11 colleges, but he was most interested in attending Massachusetts Institute of Technology, Northwestern University, Amherst College, Haverford College, or the University of Virginia. He was admitted to Northwestern University, Amherst College, and Haverford College, and decided to attend Northwestern University. Respondent A's standardized test scores were very similar to the median standardized test scores at Northwestern University, meaning Respondent A did not undermatch. Respondent A would not have undermatched at any of the five colleges he was most interested in attending. Northwestern University is nearly 800 miles from where he lived when he was applying to colleges. Respondent A considered both financial aid and distance from home "not at all important" in his college decision. He was most interested in attending college with a strong academic reputation, on a traditional campus with access to a big city. Respondent A is currently still attending Northwestern.

Respondent B also had a test score that placed him in the 98th percentile, and received mostly B's in high school. Respondent B comes from a low-income family - he was eligible for free- or reduced-price lunch every year from Kindergarten through 12th grade. His mother didn't graduate from college and he isn't sure whether his father did or not. Respondent B didn't seem to have any help with his college process – he indicated that his decision was not at all influenced by his parents, by other family members, by friends, by a high school counselor, or by a private counselor. Respondent B only applied to one college - San Diego State University. This was also the only college that Respondent B visited. The median SAT score of San Diego State is 34 percentage points below Respondent B's score,

signaling that he clearly undermatched. San Diego State is only 8 miles from Respondent B's home, which is the primary reason Respondent B chose to attend. When asked about the most important factors in his college decision, Respondent B simply answered, "location." He indicated in further questions that distance from home was extremely important to his decision, and that he only considered attending schools within commuting distance. Respondent B did not apply to any elite colleges, and he indicated that he strongly agrees that his reason for not doing so was to stay close to home. When asked why he wanted to stay close to home, he strongly agrees that it was to be near his family, and somewhat agrees that it to be near his friends. Financial aid was also extremely important to Respondent B, who is a Pell Grant recipient. Respondent B has not graduated, but is not currently attending college.

Respondent C was also a low-income student who was eligible for free- or reduced-price lunch for all 13 years Kindergarten-12th grade. He scored in the 98th percentile on the SAT, and his high school grades were mostly A's. Both of his parents graduated from college. Respondent C applied to 20 colleges, of which he was most interested in Johns Hopkins University, UC Berkeley, UCLA, UC Irvine, and UC San Diego. He was admitted to all five, and chose to attend Johns Hopkins University. Respondent C did not undermatch. When asked about the most important factors in his decision, he answered "financial aid, availability of desired major, research prospects, prestige, and met with alumni and local gathering." Johns Hopkins is over 2600 miles from respondent C's home, and he indicated that distance was not at all important in his decision. However, respondent C indicated that financial aid was extremely important. Respondent C has graduated from college.

Finally, Respondent D is a high-income student who achieved a perfect score of 1600 on the SAT, and received mostly A's in high school. His mother did not graduate from college, but his father did. Respondent C applied to 13 colleges of which he was most interested in University of Maryland – College Park, University of Ottawa, California Institute of Technology, Waterloo University, and University of Halifax. He was admitted to all five and attended University of Maryland, which was in the same zip code area as his home. He did not undermatch. Respondent D indicated that the programs for computer science, physics, and mathematics were the most important factors in his decision. He only considered distance "slightly important" to his decision, and he considered attending colleges that were flying distance from home. Respondent D is currently still attending University of Maryland – College Park.

As these four respondents illustrate, students with very similar high school achievement can experience vastly different outcomes. They were chosen to represent the key findings of the regression results. The high-income students, respondents A and D, did not undermatch and focused on academics when evaluating colleges. Respondent A traveled far from home for college, and although respondent D ended up going to college close to home, he considered colleges across the country. Respondent B is an example of a low-income student who did not capitalize on his talent. He only considered colleges very close to home, constrained by the desire to remain close to his family. Respondent C, on the other hand, was able to overcome the challenges presented by coming from a low-income background, and attended a prestigious college far from home. The purpose of this section has been to give detailed descriptions of students to personify the statistical results, but it is important to note that the sample contains a variety of students. Additionally, the types

of students represented by respondents A, B, C, and D are not equally common in the dataset. Regression results have shown that low-income students are likely to undermatch, so students similar to Respondent B may be more prevalent than students similar to Respondent C.

VI. Conclusions

This study has confirmed an established result that among high-achieving students, low-income students 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.

The new result that has been presented is that increasing the distance between a student's home and they college they attend has a negative impact on the probability that they will undermatch. Further, the effect of income on 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 3000 miles, the gap is only 8.7 percentage points.

These results may not have a direct casual interpretation, but rather include several potential factors that 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 further 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 vaster array of colleges. These students would be more likely to attend

colleges further from home and less likely to undermatch, which would be captured in the *distattend* variable of the regression. These could be the "achievement-typical" students from Hoxby and Avery's study (2012), or students like Carlos in Gladwell's podcast (2016).

The *distattend* variable could also be capturing the effect of risk-taking on choices about where to attend college. Low-income people are less likely to take risks, since they don't have the resources to do so. Like Gladwell (2016) said, poor students in America don't get second chances. So, they will be 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 (and thus reducing the risk of applying) was an equalizing effect among income groups. Going to college far away from home is certainly a risk. It's taking a step to leave your family and the comfort of your home behind, which is easier to do for high-income students who 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 to help the United States capitalize on its talent, efforts should be made to reduce undermatching. Broadly, this could be achieved by reducing the risk of applying to selective colleges for low-income students. 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 don't 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 greatly reduce the risk of applying, and increase the likelihood that these students would apply to and enroll in selective colleges.

This thesis has demonstrated that increasing the distance between a student's home and college decreases their 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. Research is also needed to find specific policies that can effectively reduce undermatching, especially for low-income students.

Appendix A. Survey Questionnaire

Consent Form

The research in which you are about to participate will ask you to answer questions about your high school academic achievement, family income, and college choice process.

The procedure in the study is very simple, and participation will not take long. The questions will take about 10 minutes to answer.

As a volunteer participant in research at DePauw University, you should understand that the following rights and conditions apply.

- Your participation is voluntary, and you may withdraw your participation at any time without penalty.
- To qualify for participation, you must be 18-25 years old, have attended or currently be attending college, and remember and be willing to report your SAT/ACT scores
- If you qualify for this survey and pay sufficient attention when responding to questions, you will be compensated $1.00 for completing the survey. You will not be compensated if you don't qualify or if you don't complete the survey.
- There are no foreseeable risks associated with completing this survey.
- The researchers have provided their phone numbers and email addresses below. This is to enable you to contact someone should questions or complaints arise. After April 10, 2017 you may contact one of the following to receive a full description of the nature, purpose and results of this study:

Lois Miller
loismiller_2017@depauw.edu
937-750-3186

Humberto Barreto
hbarreto@depauw.edu
765-658-4531

The results of the study will be confidential and anonymous. The data will not be recorded or reported in any manner that could reveal individual identity. No one, not even the researchers will be able to link your name with your responses. This study has been approved by the DePauw Institutional Review Board to insure that the study conforms to ethical principles in the conduct of research with human subjects. You may contact the IRB with questions or concerns at IRB@depauw.edu.

By consenting to participate in this survey, I verify that I am 18 years or over and have read and understood and agree to the conditions and rights listed above.

Q2 Do you consent to participate in the survey?

- ❍ Yes (1)
- ❍ No (2)

If No Is Selected, Then Skip To End of Survey

This consent form was designed based on a template provided by DePauw University's Institutional Review Board (Sample Informed Consent Form). It was submitted along with a proposal of the research project to DePauw's IRB, and then adjusted based on comments from a member of the board.

Q3 Are you between the ages of 18 and 25?

- ❍ Yes (1)
- ❍ No (2)

If No Is Selected, Then Skip To End of Survey

The sample was limited to individuals aged 18-25 so that they had been through the college search process recently enough to remember it. Another reason for restricting the age range was to ensure that effects observed were happening in the same time period. Non-traditional students (those who enter college when they are older than 25) are not included in the sample, which is consistent with how previous literature has analyzed high school graduates (Lincove and Cortes 2016, 7).

Q4 Are you currently or have you ever attended college?

- ❍ Yes (1)
- ❍ No (2)

If No Is Selected, Then Skip To End of Survey

The sample must be limited to those who attended college in order to study their college choice process.

Q5 Did you take the SAT or the ACT tests?

- ❍ SAT (1)
- ❍ ACT (2)
- ❍ Both (3)
- ❍ Neither (4)

If Neither Is Selected, Then Skip To End of Survey

The SAT and ACT standardized tests are the primary measure of high school achievement used to determine if a student has undermatched. To participate in the survey, individuals must have taken and remember their scores from one or both of the tests.

Display This Question:
If Did you take the SAT or the ACT tests?; SAT Is Selected
Or Did you take the SAT or the ACT tests?; Both Is Selected

Q6 What was your combined Critical Reading and Math SAT score (out of 1600)?

If the student has taken the SAT, regardless of whether they have also taken the ACT, they are asked to report their SAT score, since this is the measure used in analysis.

Display This Question:
If Did you take the SAT or the ACT tests? ACT Is Selected

Q7 What was your composite ACT score (out of 36)?

If the student only took the ACT, they are asked to report this score, so that it can be converted it to an SAT score, as is consistent with the literature.

Q8 What grades did you receive in high school?
- ❍ Mostly A's (1)
- ❍ Half A's and half B's (2)
- ❍ Mostly B's (3)
- ❍ Half B's and half C's (4)
- ❍ Mostly C's (5)
- ❍ Half C's and half D's (6)
- ❍ Mostly D's and below (7)

As a secondary measure of high school achievement, students are asked to report their grades. The style of this question was based on the question from the National Longitudinal Survey of Youth 1997, used in Aughinbaugh's (2008, 38) analysis.

Q9 What is your gender?
- ❍ Male (1)
- ❍ Female (2)

This question in included so that gender can be used as a controlling variable in the analysis.

Q10 What is your date of birth?

Month (number) (1)

Day (2)

Full Year (3)

Condition: Full Year Is Less Than 1991. Skip To: End of Survey.

Condition: Full Year Is Greater Than 2000. Skip To: End of Survey.

This question serves as an attention check, and a way to ensure that the respondent is between 18 and 25 years old. If they enter a birth year that ensure they are younger than 18 or older than 25, they will be excluded from the survey.

Q11 Are you of Hispanic, Latino, or Spanish origin?

- ❍ Yes (1)
- ❍ No (2)

Q12 Regardless of your answer to the prior question, please select one or more of the following that best describe you.

- ❑ American Indian or Alaska Native (including all Original Peoples of the Americas) (1)
- ❑ Asian (including Indian subcontinent and Philippines) (2)
- ❑ Black or African American (including Africa and Caribbean) (3)
- ❑ Native Hawaiian or Other Pacific Islander (Original Peoples) (4)
- ❑ White (including Middle Eastern) (5)

Similar to Q9, Q11 and Q12 are included so that race and ethnicity can be included as controlling variables. The wording of the questions is taken from the College Board Admitted Student Questionnaire (2015).

Q13 Have you ever been eligible for the free- or reduced-price lunch program at school?
- ❍ Yes (1)
- ❍ No (2)

Display This Question:
If Have you ever been eligible for the free- or reduced-price lunch program at school? Yes Is Selected

Q14 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.
- ❍ 1 (1)
- ❍ 2 (2)
- ❍ 3 (3)
- ❍ 4 (4)
- ❍ 5 (5)
- ❍ 6 (6)
- ❍ 7 (7)
- ❍ 8 (8)
- ❍ 9 (9)
- ❍ 10 (10)
- ❍ 11 (11)
- ❍ 12 (12)
- ❍ 13 (always eligible) (13)

Since self-reported household income may be unreliable for low-income students, Q13 and Q14 about free- and reduced-price lunch have been included since they have been used as a proxy for poverty. It has been shown that the number of years that a student has been eligible for free- or reduced-price lunch is a way to separate out low-income students into levels of poverty (Dynarski 2016).

Q15 Do you have knowledge of your family's total combined income (household income) to the nearest $10,000, for the year before you applied to college?
- ❍ Yes (1)
- ❍ No (2)

Display This Question:
If Do you have knowledge of your family's total combined income (household income) to the nearest thousand dollars, for the year before you applied to college? Yes Is Selected

Q16 The year before you applied to college, what was your family's total combined income (household income) before taxes? Please round to the nearest $10,000, and enter only the number (ex: 40,000)

Display This Question:

If Do you have knowledge of your family's total combined income (household income) to the nearest thousand dollars, for the year before you applied to college? No Is Selected

Q17 Although you do not know your family's income, please select from the given ranges your best estimate of your family's income (household income) before taxes, for the year before you applied to college.

- ❍ Less than $30,000 (1)
- ❍ $30,000 to $39,999 (2)
- ❍ $40,000 to $59,999 (3)
- ❍ $60,000 to $79,999 (4)
- ❍ $80,000 to $99,999 (5)
- ❍ $100,000 to $149,999 (6)
- ❍ $150,000 to $199,999 (7)
- ❍ $200,000 or higher (8)

Originally, Q16 and Q17 were separated so that a more precise measure of income could be used for respondents with knowledge of their household income to the nearest $10,000, and wider brackets could be used for individuals who were unsure (so that they would not guess incorrectly). However, in analysis, the two questions were collapsed into using the brackets so that all data could be analyzed together. Q17 was taken from the College Board Admitted Student Questionnaire (2015). Although the wording of Q17 was changed to account for the differing timeline of this survey (to ask about the year before applying to college), the responses were unchanged.

Q18 Did your mother graduate from college?

- ❍ Yes (1)
- ❍ No (2)
- ❍ Unsure (3)

Q18 was asked to measure family background.

Q19 What is the value of 2 + 2?

- ❍ 2 (1)
- ❍ 4 (2)
- ❍ 5 (3)

If 4 Is Not Selected, Then Skip To End of Survey

This was an attention check included to ensure that Mechanical Turk respondents were reading the questions and not just randomly selecting answers. If the respondent failed to answer this question correctly, they were immediately send to the end of the survey and not compensated.

Q20 Did your father graduate from college?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Unsure (3)

Q21 Are you currently in college?
- ❍ Yes (1)
- ❍ No (2)

Display This Question:
If Are you currently in college? No Is Selected

Q22 Did you graduate from college?
- ❍ Yes (1)
- ❍ No (2)

Q22 is used to determine of the students who are not currently in college, how many graduated and how many have dropped out or stopped attending college for the time being.

Q23 Have you ever transferred from one college to another?
- ❍ Yes (1)
- ❍ No (2)

Q24 Please answer the following questions in this survey in regards to the first time you applied to and enrolled in college.

The thesis seeks to determine how students made their college decision immediately after high school, so if they have transferred institutions, they are asked to answer with regards to their first college search. If a student has not transferred, the Q24 instructions are not displayed.

Q25 In what year did you first attend college?

Q25 serves as another attention check. Assuming no respondents have started college before the age of 16, if a respondent enters a year less than 2007, they must be over 25 years old and ineligible for the survey.

Q26 What was the zip code where you lived when you were applying to college?

Q26 is used to determine the distance between the student's hometown and the colleges they applied to.

Q27 How many colleges did you apply to?

Display This Question:

If How many colleges did you apply to? Text Response Is Equal to 5

Q31 Please list the full names (e.g. enter "Ohio State University", NOT "OSU") of all colleges to which you applied.

College 1 (1)
College 2 (2)
College 3 (3)
College 4 (4)
College 5 (5)

Respondents are asked to provide names of all colleges they have applied to, to determine how their SAT scores compared with the median SAT scores of colleges applied to (to measure undermatching) and to determine the proximity of the colleges to which the student applied to their home. Note: Q28-Q30 are omitted because they ask the same information, of students who applied to less than 5 colleges.

Display This Question:

If How many colleges did you apply to? Text Response Is Greater Than 5

Q32 Of the colleges to which you applied, please list the full names (e.g. enter "Ohio State University", NOT "OSU") of the top 5 that you were most interested in attending.

College 1 (1)
College 2 (2)
College 3 (3)
College 4 (4)
College 5 (5)

If a student applied to more than 5 colleges, they are asked to report the top 5 that they were most interested in. This follows the format of the College Board Admitted Student Questionnaire (2015).

Q33 Did you apply early decision? Early decision is a binding admission plan. When you apply early decision, you sign a statement agreeing to enroll in the college if you're accepted. Because of this binding agreement to enroll, you can only apply to one school early decision.

- ❍ Yes (1)
- ❍ No (2)

Q33 is asked to determine which students applied using early decision. Early decision has been shown to disadvantage low- and middle-income families (Bruni 2016).

Display This Question:
If How many colleges did you apply to? Text Response Is Greater Than 5

Carry Forward Entered Choices - Entered Text from "Of the colleges to which you applied, please list the full names (e.g. enter "Ohio State University", NOT "OSU") of the top 5 that you were most interested in attending."

Q38 Please select all colleges to which you were admitted.

Display This Question:
If How many colleges did you apply to? Text Response Is Greater Than 5

Carry Forward Selected Choices from "Please select all colleges to which you were admitted."

Q43 What college do/did you attend?

Q38 and Q43 give insight into which colleges the student was admitted to, and to which college they ultimately enrolled. Answer choices are carried forward, so that they can only select colleges to which they were admitted from the list of colleges to which they applied, and can only select which college they attended from the list of colleges to which they were admitted. Note: Q34-Q37 and Q39-Q43 have been omitted, since they provided the same information for students who applied to less than 5 colleges.

Display This Question:
If How many colleges did you apply to? Text Response Is Equal to 1

Q44 What college do/did you attend?

If a student indicated that they only applied to one college, they are not asked about which colleges they were admitted to.

Q45 How often did you suffer from a fatal heart attack when making your college decision?

- ❍ Daily (1)
- ❍ 4-6 times a week (2)
- ❍ 2-3 times a week (3)
- ❍ Once a week (4)
- ❍ Never (5)

If Never Is Not Selected, Then Skip To End of Survey

Q47 is the final attention check to ensure respondents are reading the questions.

Q46 During the search process, how many colleges did you visit?
- ❍ 1 (1)
- ❍ 2 (2)
- ❍ 3 (3)
- ❍ 4 (4)
- ❍ 5 (5)
- ❍ 6 (6)
- ❍ 7 (7)
- ❍ 8 (8)
- ❍ 9 (9)
- ❍ 10+ (10)

Visiting colleges may be easier and more common for high-income students, which could lead them to consider colleges further from home (McDonough 1997).

Q47 What were the most important factors in choosing your college? Explain.

This open-ended question is designed to determine what primarily drove the student's college decision. It is placed before the specific questions about influential factors, so respondents could answer this before being primed with specific factors.

Q48 Did you apply to any colleges that you would consider prestigious or elite?
- ❍ Yes (1)
- ❍ No (2)

Display This Question:

If Did you apply to any colleges that you would consider prestigious or elite? No Is Selected

Q49 Please rate the amount to which you agree or disagree with the following statements. I did not apply to any prestigious or elite colleges because...

	Strongly agree (1)	Somewhat agree (2)	Neither agree nor disagree (3)	Somewhat disagree (4)	Strongly disagree (5)
I did not think I would be admitted (1)	❍	❍	❍	❍	❍
I did not think I would be able to afford elite colleges (2)	❍	❍	❍	❍	❍
I wanted to go to college close to home (3)	❍	❍	❍	❍	❍
I did not think I would be able to handle the academic rigor (4)	❍	❍	❍	❍	❍

Display This Question:

If Did you apply to any colleges that you would consider prestigious or elite? No Is Selected

Q50 Are there any other reasons you did not apply to any prestigious or elite colleges?

Hoxby and Avery (2012) and Bowen et al. (2005) have shown that many students who could be admitted to selective colleges do not apply, but less attention has been given to why these students don't apply. Q49 and Q50 are designed to determine what has stopped a student from applying to elite colleges, if they did not apply to any.

Display This Question:

If Please rate the amount to which you agree or disagree with the following statements.I did not apply to any prestigious or elite colleges because... I wanted to go to college close to home - Strongly agree Is Selected

Or Please rate the amount to which you agree or disagree with the following statements.I did not apply to any prestigious or elite colleges because... I wanted to go to college close to home - Somewhat agree Is Selected

Q51 Please rate the degree to which you agree or disagree with the following statements. I wanted to go to college close to home because...

	Strongly agree (1)	Somewhat agree (2)	Neither agree nor disagree (3)	Somewhat disagree (4)	Strongly disagree (5)
I wanted to be near my family (1)	❍	❍	❍	❍	❍
I wanted to be near my friends (2)	❍	❍	❍	❍	❍
I wanted to keep my high school job (3)	❍	❍	❍	❍	❍
I wanted to be near my boyfriend/girlfriend (4)	❍	❍	❍	❍	❍

Display This Question:

If Please rate the amount to which you agree or disagree with the following statements. I did not... I wanted to go to college close to home - Strongly agree Is Selected

Or Please rate the amount to which you agree or disagree with the following statements. I did not... I wanted to go to college close to home - Somewhat agree Is Selected

Q52 Are there any other reasons you wanted to go to college close to home?

Q51 and Q52 are displayed only if the respondent indicates that they "strongly agree" or "somewhat agree" with the statement that they did not apply to any prestigious/elite colleges because they wanted to go to college close to home. It aims to determine students' primary reasons for wanting to be close to home.

Q53 To what extent were the following people influential to your college decision?

	No influence (1)	Minimal influence (2)	Moderate influence (3)	Strong influence (4)	Very strong influence (5)	N/A (6)
Parents (1)	❍	❍	❍	❍	❍	❍
Other family members (not parents) (2)	❍	❍	❍	❍	❍	❍
Friends (3)	❍	❍	❍	❍	❍	❍
High School Guidance Counselor (4)	❍	❍	❍	❍	❍	❍
Private Counselor (5)	❍	❍	❍	❍	❍	❍

Q30 is designed to measure a student's social capital, and how they used that social capital in their college search. The higher of an influence the various significant persons had on the student, the higher their social capital in making a college decision.

Q54 Thinking of your time in college, did/do you live on/near campus or did/do you commute from your family home? If both options apply, choose the one that was true for the majority of your time in college.

❍ Live on/near campus (1)

❍ Commute (2)

Q55 How important was the distance from home in your decision to attend your college?

❍ Extremely important (1)

❍ Very important (2)

❍ Moderately important (3)

❍ Slightly important (4)

❍ Not at all important (5)

Q56 When deciding colleges to apply to, what was the farthest distance you considered?
- ❍ <1 hour drive (commuting distance) (1)
- ❍ 1-3 hours drive (2)
- ❍ 3+ hours drive (3)
- ❍ flying distance/distance was not a concern (4)

Q54-Q56 are designed to measure the influence of distance from home on the student's college decision. The responses to question 56 were given in driving/flying time rather than miles because interviewees in McDonough's (1997) talked about distance from home in time rather than miles.

Q57 Did you receive a Federal Pell Grant?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Unsure (3)

Q58 Did you receive need-based financial aid?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Unsure (3)

Q59 Did you receive non-need-based financial aid by your college in recognition of your athletic, musical, artistic, or academic talent?
- ❍ Yes (1)
- ❍ No (2)
- ❍ Unsure (3)

Display This Question:
If Did you receive need-based financial aid? No Is Not Selected
Or Did you receive non-need-based financial aid by your college in recognition of your athletic, mus... No Is Not Selected

Q60 Did your financial aid package include

	Yes (1)	No (2)	Unsure (3)
Grants or scholarships? (1)	❍	❍	❍
One or more student loans? (2)	❍	❍	❍
A work package or campus job? (3)	❍	❍	❍

Q61 How important was the cost of your college and your financial aid package in your decision to attend your college?

- ❍ Extremely important (1)
- ❍ Very important (2)
- ❍ Moderately important (3)
- ❍ Slightly important (4)
- ❍ Not at all important (5)

Access to college across income groups is related to financial aid, so Q57-Q61 were designed to determine what kind of financial aid the student received and how it affected their decision. Q58-Q60 were taken from the College Board Admitted Student Questionnaire (2015).

Q62 Do you think you made a good college choice?

- ❍ Yes (1)
- ❍ No (2)

Q63 Do you think you would have been as satisfied at the other college(s) you considered but did not attend?

- ❍ Yes (1)
- ❍ No (2)

Undermatching may not be a negative consequence for all students, and some students may be maximizing utility by attending a college that has much lower median SAT scores than their own. Q62-Q63 are designed to measure if, in retrospect, the student believes they made a good college decision.

Display This Question:

If Have you ever transferred from one college to another? Yes Is Selected

Q64 Other than the first college you attended, please list all other colleges you have attended.

College 1 (1)
College 2 (2)
College 3 (3)
College 4 (4)

If the student has transferred colleges, after they have answered all questions with respect to their first college search, they are asked to provide which college(s) they have transferred to.

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.

"Admission Chances" 2017. College Simply, accessed March 20, 2017.

“Admitted Student Questionnaire/Admitted Student Questionnaire Plus User Manual” 2015. College Board. Accessed April 10, 2017. https://secure-media.collegeboard.org/digitalServices/pdf/professionals/admitted-student-questionnaire-user-manual.pdf

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

Aughinbaugh, Alison. 2008. "Who Goes to College-Evidence from the NLSY97." *Monthly Lab. Rev.* 131:33.

Barreto, Humberto, and Frank . Howland. 2006. *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel:* Cambridge University Press.

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.

Bowen, William G, Martin A Kurzweil, Eugene M Tobin, and Susanne C Pichler. 2005. Equity and excellence in American higher education (Thomas Jefferson Foundation distinguished lecture series). University of Virginia Press.

Bruni, Frank. 2016. "The Plague of 'Early Decision'." *The New York Times*, December 21, 2016, Opinion. Accessed January 13, 2017. https://www.nytimes.com/2016/12/21/opinion/the-plague-of-early-decision.html?_r=1.

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*:490-505.

Chetty, Raj, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility."

Dynarski, Susan. 2016. “The Gap Within the Gap: Using Longitudinal Data to Understand Income Differences in Student Achievement.”

“FAQ: What Are Pseudo R-Squareds?” 2011. UCLA Institute for Digital Research and Education. Accessed April 10, 2017. http://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

"Find Your Dream School" 2017. The Princeton Review, accessed March 20, 2017.

"Frequently Asked Questions" 2015. Amazon Mechanical Turk, accessed January 12. https://www.mturk.com/mturk/help?helpPage=overview.

Gauntlett, David. 2011. "Three approaches to social capital." *supplementary text to Making is Connecting: The Social Meaning of Creativity, from DIY and Knitting to Youtube and Web* 2.

Gladwell, Malcolm. 2016. Carlos Doesn't Remember. In *Revisionist History*.

Google Maps. 2017. Google. Accessed April 10, 2017. https://maps.google.com/

Hoxby, Caroline M, and Christopher Avery. 2012. The Missing" One-Offs": The Hidden Supply of High-Achieving, Low Income Students. National Bureau of Economic Research.

Jerrim, John, Anna K Chmielewski, and Phil Parker. 2015. "Socioeconomic inequality in access to high-status colleges: A cross-country comparison." *Research in Social Stratification and Mobility* 42:20-32.

Lincove, Jane Arnold, and Kalena E Cortes. 2016. Match or Mismatch? Automatic Admissions and College Preferences of Low-and High-Income Students. National Bureau of Economic Research.

Litten, L. 1982. "Different strokes in the applicant pool: Some refinements in a model of student choice." *Journal of Higher Education* 4:383-402.

McDonough, Patricia M. 1997. *Choosing colleges: How social class and schools structure opportunity*: Suny Press.

Nurnberg, Peter, Morton Schapiro, and David Zimmerman. 2012. "Students choosing colleges: Understanding the matriculation decision at a highly selective private institution." *Economics of Education Review* 31 (1):1-8.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running experiments on amazon mechanical turk." *Judgment and Decision making* 5 (5):411-419.

Perna, Laura W. 2006. "Studying college access and choice: A proposed conceptual model." In *HIGHER EDUCATION:*, 99-157. Springer.

"Requirements for Admission" 2017. PrepScholar, 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.

"Sample Informed Consent Form." accessed January 13. http://www.depauw.edu/offices/academic-affairs/grants-and-research/irb/topics/informed-consent/sample/.

SAT. 2014. SAT Percentile Ranks for Males, Females, and Total Group.

“Survey Protection” 2017. Qualtrics Support. Accessed April 10, 2017. https://www.qualtrics.com/support/survey-platform/survey-module/survey-options/survey-protection/

U.S. Department of Education. 2014. Institute of Education Sciences, National Center for Education Statistics. IPEDS: Integrated Postsecondary Education Data System.

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? : Department of Economics, Williams College.