The Relationship Between Siblings' College Choices:
Evidence from One Million SAT-Taking Families*

Joshua Goodman
Harvard University
joshua_goodman@hks.harvard.edu

Michael Hurwitz
College Board
mhurwitz@collegeboard.org

Jonathan Smith
College Board
jsmith@collegeboard.org

Julia Fox
College Board
jfox@collegeboard.org

February 2015

ABSTRACT

Recent empirical work has demonstrated the importance both of educational peer effects and of various factors that affect college choices. We connect these literatures by highlighting a previously unstudied determinant of college choice, namely the college choice made by one's older sibling. Data on 1.6 million sibling pairs of SAT-takers reveals that younger and older siblings' choices are very closely related. One-fifth of younger siblings enroll in the same college as their older siblings. Compared to their high school classmates of similar academic skill and with observably similar families, younger siblings are about 15-20 percentage points more likely to enroll in four-year colleges or highly competitive colleges if their older siblings do so first. These findings vary little by family characteristics. Younger siblings are more likely to follow the college choices of their older siblings the more they resemble each other in terms of academic skill, age and gender. We discuss channels through which older siblings' college choices might causally influence their younger siblings, noting that the facts documented here should prompt further research on the sharing of information and shaping of educational preferences within families.

The truth is that if Princeton hadn't found my brother as a basketball recruit and if I hadn't seen that he could succeed on a campus like that, it never would have occurred to me to apply to that school, never.

--Michelle Obama

Introduction

For decades, researchers from various disciplines have tried to model how students make college enrollment decisions. Such disciplines include economics (Fuller, Manski and Wise, 1982), sociology (Hearn, 1991), and education (Jackson, 1978). The modeling problem has, however, proven difficult, for at least three reasons. First, there are thousands of colleges, each with numerous attributes. Second, students have heterogeneous preferences for college enrollment and for those college attributes. Third, students differ in the extent to which they have accurate information about potential colleges. Many of these factors are unobservable to the econometrician modelling college choice.

A few unsurprising characteristics of college have emerged from this literature as important to the college decision. First, the cost of college and the availability of financial aid are important factors in students' decisions, particularly for low-income students (Avery and Hoxby, 2004; Dynarski, 2003; Hurwitz, 2012). Second, proximity to colleges increases the likelihood that students enroll as students, and particularly low-income students, prefer colleges closer to home (Hossler et al., 1989; Leppel, 1993; DesJardins et al. 1999). Third, college quality has become an increasingly important determinant of students' enrollment choices (Long, 2004), with small changes in college rankings affecting the number of applicants to a given college (Luca and Smith, 2013). Fourth, the quality of college amenities, such as dormitories and student activities, also matters to many students, with only high-achieving ones exhibiting demand for academic quality (Jacob et al., 2013).

Other determinants of college choice are harder to explain from a model of fully rational behavior on the part of students. High-achieving low-income students do not apply to or enroll in the same quality colleges as their higher income peers, despite the fact that the students would likely pay very little at these selective institutions (Hoxby and Avery, 2012). Many students apply only to the number of colleges for which it is free to send their test scores, such that even an elimination as small as $6 in cost can substantially change students' college choices (Pallais, 2013). Colleges receive substantially fewer applications when they increase their application fees by a few dollars or add an admission essay (Smith et al., 2014), but more applications when their sports teams succeed (Pope and

Pope, 2009). Relatively small amounts of merit aid can induce students into colleges of dramatically lower quality, harming their own graduation rates (Cohodes and Goodman, 2014). The fact that relatively small interventions, such as information mailings with application fee waivers (Hoxby and Turner, 2013), help with the completion of financial aid forms (Bettinger et al. 2012), or mandatory college entrance exams (Goodman, 2013; Hurwitz et al., 2015; Hyman, 2014; Klasik, 2013) can increase enrollment suggests that such suboptimal behavior is likely driven by a combination of information gaps and behavioral biases (Dillon and Smith, 2013).

Given the volume of research into determinants of college choice, it is therefore remarkable how little the economics of education literature has focused on the influence of family members themselves. A rich descriptive literature in education does consider the association between parental education, parental involvement and college choice of children (Choy, 2001; Perna and Titus, 2005), though siblings are rarely the focus of such literature. Consideration of families are, of course, implicit in much of the aforementioned economic research, in that most analyses control for or even estimate the impact of family factors such as parental income and education. Some papers exploit family structure in their analyses, using twin or other sibling fixed effects to account for selection bias when estimating returns to college quality (Ashenfelter and Krueger, 1994; Behrman, Rosensweig, and Taubman, 1996; Rouse, 1999; Lindahl and Regner, 2005; Smith, 2013). Yet others study birth order effects on educational attainment, though these often focus on differential sources of parental investment (Behrman and Taubman, 1986; Black, Devereaux, and Salvanes, 2005; Kantarevic and Mechoulan, 2006; Booth and Kee, 2009; Hotz and Patano, 2013).

It is perhaps even more remarkable that little has been written on the particular influence of siblings on each other's educational decisions. Though a fairly extensive literature documents sibling influences on risky behaviors such as smoking and drinking (Altonji, Cattan and Ware, 2010), we are aware of only three papers that attempt to measure the influence of siblings on each other's educational decisions. Using the NLSY79, Oettinger (2000) argues that older siblings' high school graduation status influences the high school graduation status of younger siblings, addressing endogeneity of the former by instrumenting with gender, family structure and unemployment rates. Loury (2004) estimates that, controlling for a host of other variables, African-Americans' college enrollment rates are substantially higher when they have older siblings who have enrolled in college. Using Danish data, Joensen and Nielsen (2013) show that quasi-experimental variation in older siblings' access to advanced math and science coursework alters the coursework choices of younger siblings.

Effects of other sorts of peers have, of course, been extensively documented (Sacerdote, 2011). The now vast literature on peer effects rarely considers siblings as peers, instead studying interactions between classmates, schoolmates or roommates. That literature most frequently estimates impacts of peers on student achievement or behavior, rarely if ever using college choice as an outcome. We therefore connect the literature on college choice to the literature on peer effects by carefully investigating the relationship between siblings' college enrollment decisions. To do so, we use data on the SAT scores and college choices of the universe of SAT-takers from the 2004-11 high school graduation cohorts. Among the approximately 10 million students in those cohorts, we identify 1.6 million pairs of siblings by matching students on last names and home addresses. We then analyze simple college choice models in which the younger siblings' enrollment choices are regressed on a rich set of demographic and academic skill controls, as well as on variables measuring the college enrollment choices of their older siblings. We also explore the extent to which the relationship between siblings' college choices varies by siblings' similarities in terms of academic skill, age and gender.

We show that younger and older siblings' choices are very closely related. One-fifth of younger siblings enroll in the same college as their older siblings. Compared to their high school classmates of similar academic skill, younger siblings are about 16 percentage points more likely to enroll in four-year colleges and 19 percentage points more likely to enroll in highly competitive colleges if their older siblings do so first. The quality of college selected by an older sibling is strongly predictive of the quality chosen by a younger sibling. These findings vary little by family income, race, parental education, or proximity to four-year colleges. Younger siblings are more likely to follow the college choices of their older siblings the more they resemble each other in academic skill, age and gender. Our hope is that these results may improve the targeting of college choice interventions and, more importantly, prompt further research on the sharing of information and shaping of educational preferences within families.

We turn now to a description of the data. After that, we explain in detail how we estimate the relationship between siblings' college choices and discuss the magnitude of these estimates. We then explore whether such estimates vary by the similarity of the siblings. Finally, we discuss a number of theoretical reasons why siblings' college choices might affect each other. We argue that the relationships estimated here are at least partly causal. We conclude with implications for future work.

Data and Summary Statistics

Our data set comes from the College Board's (CB) universe of SAT-takers in the 2004-11 high school graduation cohorts. Every year, approximately 1.5 million high school students take the SAT, a standardized test often required for college admissions.[1] The test has both a math and critical reading section, each graded on a scale of 200 to 800 for a maximum possible total score of 1600.[2] We observe those test scores, as well as a rich set of variables self-reported by students, including high school GPA, gender, race/ethnicity, parental education and parental income.[3] We also observe up to 30 Score Sends, the process by which a student requests that CB officially provide his or her SAT score to a given college. Score Sends have been shown to be good proxies for actual college applications (Card and Krueger, 2005; Pallais 2013). The CB data also include each student's full name and home address, which we use to identify siblings, as well as the high school attended.

The data on SAT takers has been merged with National Student Clearinghouse (NSC) data that tracks postsecondary enrollment information for more than 94 percent of students enrolled in U.S. postsecondary institutions.[4] Data from the NSC allow us to observe which college, if any, a student enrolls in after high school graduation. Using data from the Integrated Postsecondary Education Data System (IPEDS), we further characterize each college as a two-year or four-year institution and, for four-year colleges, by the average SAT score of incoming students. We also consider as an outcome attendance at one of the roughly 200 colleges classified as "highly competitive" based on the top two categories in the Barron's Admissions Competitive Index ("most competitive" and "highly competitive").

Students do not list siblings on their SAT registration forms. We identify each student's siblings as those who share the student's last name and home address.[5] Because the data contain eight cohorts and students generally take the SAT around age 17, sibling pairs in our data can be at most about eight years apart in age. Our method of identifying siblings is unlikely to falsely identify two students as siblings given how unusual it would be for two individuals close in age to share a last name and home address without being siblings. The method does, however, fail to match siblings if their families have changed home address between the times the two siblings took the SAT. Furthermore, we miss any siblings who graduated from high school before 2004 or after 2011. We focus on the younger sibling as

---

[1] The ACT is the other dominant college admissions test. A list of states where a majority of students take the SAT over the ACT can be found in Clark, Rothstein and Schanzenbach (2008).

[2] In 2005, CB added a writing section. For continuity across the sample, and because admissions put most weight on the math and critical reading sections, we only consider those two sections.

[3] The self-reported variables are sometimes missing. Rather than drop those observations, we generate indicators for those instances.

[4] A large fraction of non-participating colleges are for-profit institutions.

[5] For home address, we use city, state, and the first five characters of the street, including street number.

the unit of analysis because we are interested in the potential influence of older siblings. Our final analytic sample consists of younger siblings from the high school classes of 2005-11 for whom we could find older siblings in the classes of 2004-10.

We present summary statistics in Table 1, where the first column contains all 10 million SAT-takers and the second column contains the 1.6 million SAT-takers we identify as younger siblings. Panels A and B list some of the demographic variables and academic skill measures we observe. About 5% of students fail to report race, nearly half fail to report parental income, and 10% fail to report parental education. [6] Compared to the full population of SAT-takers, younger siblings are less likely to be black or Hispanic, to come from families earning less than $50,000 a year, and to have parents with no college education. Younger siblings have slightly higher GPAs and higher SAT scores than the overall population of SAT-takers.[7] That younger siblings are more advantaged demographically and academically is likely driven by the fact that we can only identify siblings if they come from families with multiple children taking the SAT and whose addresses are stable over time. Disadvantaged families are less likely to satisfy both of these conditions.

Panel C of Table 1 shows the college enrollment outcomes we analyze. We define all such outcomes using the college each student enrolls in the fall after their high school graduation. 73% of younger siblings enroll in four-year colleges, 21% enroll in two-year colleges, and 7% do not enroll in any college or are missing college enrollment data, perhaps by attending a college not included in the NSC. One-fifth enroll in a highly competitive college, as defined by Barron's. Finally, conditional on enrolling in a four-year college, the average SAT scores of incoming students at younger siblings' enrolled colleges is 1130.

Some of the younger siblings in our data have multiple older siblings we can identify. Among the sets of siblings identified in the data, 90% are pairs, 9.4% are triads and just under 1% are sets of four or more. To define a unique older sibling for estimation purposes, we can associate each younger sibling either with the oldest identifiable sibling or with the older sibling closest in age (i.e. the youngest older sibling). We choose to use the oldest sibling as the relevant older sibling. Though not reported here, none of our results are substantially affected by this choice, in part because so many of our younger siblings have only a single older sibling in the data.

Panel D of Table 1 shows various ways in which we can compare younger and older siblings, both in terms of demographic and academic background and in terms of college choices. Comparing

---

[6] Thought not listed here, about 3% report "Other race" in both the full and younger siblings sample.

[7] The nearly 40 point difference in SAT scores represents roughly 0.2 standard deviations during this time period.

siblings' characteristics allows us to explore whether the relationship between siblings' colleges choices is stronger when the siblings are more similar in terms of gender, age or academic skill. We observe, for example, that 51% of the younger siblings have the same gender as their older siblings. The average age difference between siblings in our data is 2.8 years. More than half of sibling pairs differ in their composite (M+CR) SAT scores by 100 points, or roughly half of a standard deviation in the overall population. We will estimate below the extent to which sibling differences in gender, age and academic skill relate to differences in siblings' college choices.

To further motivate this exploration, we note the striking results in panel D concerning the absolute proportion of younger siblings who make similar college choices as their older siblings. 31% of younger siblings apply to the college their older sibling attended, as measured by SAT Score Sends. A remarkable 19% of younger siblings enroll in the same college as their older sibling. 69% of younger siblings enroll in the same the same type of college as their older sibling, where type is defined as 4-year, 2-year or no college. Finally, 36% of younger siblings enroll in a college in the same Barron's category as their older siblings. All of this suggests that siblings' college choices are strongly related. The analysis below attempts to unpack some of the determinants of that relationship.

The Predictive Power of Older Siblings' College Choices

We focus on simple estimates of the relationship between younger and older siblings' college choices. To do so, we consider the following OLS regression for each college choice outcome *Y* of younger sibling *i* in high school *s* and graduation year *t*:

$$Y_{ist} = \beta SibCollegeChoice + \alpha X_i + \gamma Z_i + \delta_{st} + u_{ist} \qquad (1)$$

Equation 1 includes characteristics of the younger sibling (*X)*, such as a full set of indicators for gender, race/ethnicity, parental income and parental education, and continuous measures for the younger sibling's high school GPA and SAT score. Also included are controls (*Z)* for the academic skill of the older sibling, in the form of his or her high school GPA and SAT score.[8] These potentially control for within family-specific factors not accounted for by the previously mentioned controls, as well as potentially differing parental investment by birth order. δ_{st} represents a set of high school by year fixed effects, so that the subsequent analyses always compare students to their high school classmates. This eliminates

---

[8] When these variables are missing, a dummy variable is created to indicate as such.

any bias arising from factors that vary across high schools, such as geography or the quality of guidance counselors.

The central innovation of this paper is to include in each regression a measure of the older sibling's college choice, such as an indicator for whether the older sibling enrolled in a four-year college. The coefficient of primary interest β, estimates the extent to which older siblings' college choices are related to the college choices of their younger siblings, conditional on all of the other factors previously mentioned. Finally, we cluster standard errors by family to account for the fact that some families contain more than one younger sibling in our data.

Table 2 shows the regression estimates from equation 1. The first column's outcome is an indicator for the younger sibling's enrollment in any four-year college. Unsurprisingly, both measures of a student's own academic skill, high school GPA and SAT score, strongly predict four-year college enrollment.[9] Interestingly, when an older sibling has a higher GPA and SAT score, younger siblings are slightly less likely to enroll in a four-year college, perhaps a mechanical result of simultaneously controlling for the older sibling's college choice as well. The variable of primary interest is whether the older sibling enrolled in a four-year college, in panel B. Conditional on their own characteristics, their own academic skill, and the academic skill of their older siblings, and compared to their own high school classmates, younger siblings are a substantial 16.2 percentage points more likely to enroll in a four-year college if their older sibling did so. This coefficient is much larger in magnitude than the coefficients on any other indicator variable, including demographic controls not shown, suggesting that older siblings' four-year college enrollment is a comparatively strong predictor of younger siblings' four-year college enrollment. Remarkably, these fairly rich controls for student background and academic skill, as well as high school-by-year fixed effects, explain only 31 percent of the variation in four-year college enrollment.

We obtain similar results if we instead use measures of college quality as outcomes. The estimates in column 2 suggest that, conditional on all else, a younger sibling is 18.7 percentage points more likely to enroll in a highly competitive college if his or her older sibling did so. Although not displayed, a similar result appears when considering the younger sibling enrolling in the most competitive category (Barron's highest category) after the older sibling does so. In column 3, we estimate the relationship between the quality of siblings' chosen colleges as measured by institution-

---

[9] Using SAT percentiles instead of raw scores yields qualitatively similar results.

level SAT scores.[10] The relevant coefficient implies that an older sibling enrolling in a college with an average SAT score 100 points higher is associated with an increase of 14.4 SAT points in the average score of the younger sibling's chosen college. This strong relationship between the quality of siblings' college choices can be seen in the unconditional data shown in Figure 1, which simply plots the relationship between the institution-level SAT scores of siblings' chosen colleges.[11]

In the final two columns of Table 2, we use as outcomes measures of the match quality between younger siblings and their chosen colleges. We define two measures of match quality, undermatch and overmatch.[12] Students whose math SAT score is greater than the 75th percentile SAT score of their college peers are considered "undermatched". Similarly, students whose math SAT score is less than the 25th percentile SAT score of their college peers are considered "overmatched". Column 4 suggests that, conditional on all else, a younger sibling is a substantial 6.4 percentage points more likely to undermatch if his or her older sibling also undermatched. The relationship between siblings' overmatch probabilities is a similar 5.8 percentage points, as shown in column 5.[13] These coefficients are larger in magnitude than any of the other binary demographic variables controlled for in the background, suggesting that the sibling gap in match quality exceeds any observed gender, race or income gaps. In short, these measures suggest that older siblings' college match quality is strongly predictive of the college match quality of their younger siblings.

In Table 3, we explore whether the observed relationships between younger and older siblings' college choices vary by the characteristics of the families from which they come. To do so, we re-run the regressions from Table 2, separating the sample by income, race, parental education, and the proximity of a four-year public college. For each regression, we then list in Table 4 only the coefficient measuring the relationship between siblings' college choices (i.e. the one listed in panel B in Table 2). Columns 1 and 2, which divide the sample into families reporting less than and more than $50,000 in annual income, show no substantial differences in the magnitudes of the estimated sibling relationships. Columns 3 and 4, which divide the sample into black and Hispanic families compared to white families,

---

[10] In some instances, one or both siblings did not have values for the average SAT score of first four-year college attended, either because the sibling lacked enrollment data, attended a two-year college or attended a four-year college without IPEDS-reported average SAT scores. We assigned those missing values an average SAT of college enrolled the lowest average SAT of 675 and separately included dummies for whether this value was assigned to the younger sibling and/or older sibling.

[11] For graphical purposes, Figure 1 excludes the observations where either a younger or older sibling has no reported average SAT of college enrolled.

[12] This undermatch and overmatch terminology first appears in Bowen, Chingos, and McPherson (2009).

[13] Results generated by defining undermatch and overmatch as being more than 100 SAT points away from the college average SAT score look quite similar in magnitude.

similarly show no systematic or large differences by race in these relationships. In columns 5 and 6, we split the sample into students with and without parents who have a bachelor's degree. Once again, there is not a large difference in estimates regardless of the outcome considered. In columns 7 and 8 we divide the sample into students living within 25 miles of a public four-year college and those who are not. There is little evidence of substantial heterogeneity by proximity to public four-year colleges. This implies that the observed correlations may not simply be driven by geographic constraints.

Do Similar Siblings Have Stronger College Choice Relationships?

Having documented the strong relationship between the college choices of younger and older siblings, we now explore whether siblings' tendency to emulate each other depends upon the extent to which the siblings are similar in terms of academic skill, gender and age. We do this by running regressions of the form:

$$Y_{ist} = \beta_1 SkillDiff + \beta_2 GenderDiff + \beta_3 AgeDiff + \alpha X_i + \gamma Z_i + \delta_{st} + u_{ist} \quad (2)$$

Here, the outcome *Y* indicates whether the younger and older sibling have made the same college choice, such as applying to the same college, enrolling in the same college, enrolling in the same type of college (4-year/2-year/none) or enrolling in a college in the same Barron's category. As before, we control for demographic characteristics and academic skills of the two siblings, as well as high school by year fixed effects. To measure large differences in academic skill between siblings, we define *SkillDiff* as an indicator for whether the absolute value of the difference between the siblings' SAT scores is greater than 100.[14] We define *GenderDiff* as an indicator for the two siblings' genders differing and *AgeDiff* as a continuous measure of the siblings' age difference in years.[15]

Table 4 shows the results of these regressions, with panel A showing controls for own and older sibling's academic skills. Those estimates suggest that higher academic skills generally predict a higher likelihood of following an older sibling's college choices. Panel B highlights the differences between siblings. The first row implies that, conditional on all else, younger siblings whose SAT scores differ by

---

[14] Other functional forms of the academic difference were tested. This parameterization highlights the importance of large differences in skills between siblings.
[15] We observe few consistent differences between sister-sister and brother-brother pairs, so that this specification is the simplest that captures the interesting features of the data.

more than 100 points from their older siblings' scores are six percentage points less likely to apply to and enroll in the same colleges as those older siblings. They are 10 percentage points less likely to enroll in a college of the same Barron's category as their older sibling. The second rows implies that, across all of these outcomes, having an older sibling of the opposite gender decreases the probability of following that sibling's college choice by about two percentage points. The last row of panel B implies that each year of age difference between siblings reduces by 0.5-0.8 percentage points the likelihood of a younger sibling following an older sibling's college choice. This effect is relatively linear across the distribution of age differences observed in our data. Taken as a whole, these results suggest that the more similar siblings are to each other in terms of academic skill, gender and age, the more likely a younger sibling is to make the same college decisions as his or her older sibling. This is consistent with Joensen and Nielsen (2013), who also find that siblings close in age and of the same gender have stronger influences over each other.

Discussion and Conclusion

This paper documents a number of previously unknown facts about intra-family patterns in college enrollment. First, many younger siblings apply to and enroll in the same colleges as their older siblings. Second, even controlling for a rich set of covariates, older siblings' college enrollment decisions are strongly predictive of their younger siblings' decisions about whether to enroll and which quality of college to enroll in. Third, these strong relationships between siblings' college choices vary little by income, race or proximity to public four-year colleges. Fourth, younger siblings are more likely to follow the college choices of their older siblings the more they resemble each other in terms of academic skill, age and gender.

These facts, taken as a whole, are consistent with the possibility that the college decisions of older siblings influence the college decisions of younger siblings. There is, however, a potential non-causal explanation for these patterns, namely that the covariates available to us for this analysis are insufficient to control for fundamental differences between families that determine college enrollment choices. These could include differences in educational preferences, information about college or the labor market, or access to credit. It may be that siblings simply have the same preferences for factors such as college quality and distance from home that result from a shared environment. If the available covariates do not completely absorb such inter-family differences, then the strong relationship between siblings' choices may be partly picking up those unobserved differences. If so, older siblings' college

choices reveal something about a family's type, in which case it is unsurprising that such choices then predict those of younger siblings.

There are, however, at least five channels through which older siblings' college choices might causally influence their younger siblings. First, an older sibling's application and enrollment experiences may provide information to the younger sibling that would otherwise have been costly or even impossible to obtain. This could include information about the application process, the probability of admission, the net price of enrollment, or the quality of the actual enrollment experience. Recent evidence that informational interventions (Hoxby and Turner, 2013) and exposure to teachers or schoolmates who attended selective colleges (Bulman et al., 2014; Hoxby and Avery, 2012) may affect college application and enrollment choices strongly suggests that students often operate without full information about available college options. Older siblings may thus lower the cost of obtaining such information in a classical choice model or, in a more behavioral model of choice, may increase the salience of a given college or set of colleges. The relevant information here may be about the college the older sibling enrolls in or may be about colleges more generally. Second, a younger sibling may derive specific benefits from being enrolled at the same college as an older sibling. Siblings may enjoy each other's company and thus derive utility from being on the same campus. They may study together or help each other with course selection, improving their educational experience. They may live or carpool together, reducing the overall costs of enrollment. The third channel involves college pricing. Federal financial aid is, for example, calculated in part based on the number of children in the family currently enrolled in college. Having an older sibling enrolled in any college can therefore increase the amount aid for which a younger sibling is eligible, reducing the net price of college. Younger siblings may also benefit from the fact that some colleges offer tuition discounts if multiple siblings attend the same campus.[16] A fourth channel worth mentioning comes from the fact that some colleges are more likely to admit legacy students than observationally similar non-legacy students (Hurwitz, 2011). At such colleges, an older sibling's enrollment often qualifies a younger sibling as such a legacy student.

The first two channels mentioned above, namely lowered information costs and the impacts of having an older sibling on the same campus, could in theory either raise or lower the probability of a younger sibling following an older sibling's decision. New information could reveal positive or negative

---

[16] Though we cannot find any definitive estimate of the number of colleges that offer sibling discounts, multiple recent news articles suggest the practice is fairly common. For example, see U.S. News and World Report's "Some Colleges Discount Tuition for Siblings", by Katy Hopkins, December 11, 2012, available at http://www.usnews.com/education/best-colleges/paying-for-college/articles/2012/12/11/some-colleges-discount-tuition-for-siblings.

facets of the application or enrollment process. Younger siblings might benefit from having older siblings on campus or might find that proximity unpleasant. Conversely, the price and legacy channels have a theoretically unambiguous sign, as they can only serve to increase the probability of a younger sibling following an older sibling's choice. Finally, a fifth channel, namely the wealth effect of having an older sibling enrolled in college, unambiguously implies a negative impact on the probability of the younger sibling's enrollment. Given that some families are likely credit constrained with respect to the total budget they can spend on their children's college education, an older sibling's enrollment may exhaust funds that would otherwise have been able to finance a younger sibling's enrollment.

In short, there are a number of theoretical reasons why we might expect to find a causal impact of an older sibling's college choice on a younger sibling. Though we cannot rule out the possibility that the relationships we have documented are simply correlational, two sets of facts documented here suggest that the non-causal story cannot fully explain the strength of the observed correlations. First, the lack of heterogeneity in these sibling relationships by proximity to public four-year colleges is inconsistent with the possibility that inter-family differences in geographic access to colleges is driving our results. This rules out the explanation that students simply choose the nearest college and that, for siblings, that nearest college is always the same. The lack of heterogeneity by income and race also make it less likely that these relationships are purely about unobserved inter-family differences, as the magnitude of those unobserved differences would need to be quite similar across income and race categories to generate results of such similar magnitudes.

Second, the fact that younger siblings more strongly follow the choices of more similar older siblings seems difficult to attribute to unobserved inter-family differences. In particular, the fact that being of the same gender increases the probability of following an older sibling's college choice is inconsistent with that story, given the low probability that the gender composition of siblings is highly correlated with inter-family differences in educational preferences (i.e. conditional on having two children, the gender of each is random). That siblings closer in age are more likely to make similar college choices is also difficult to explain away with such unobserved family factors. A much more likely explanation is that younger siblings of the same gender as or close in age to their older siblings are more likely to value the older siblings' information about the college application process, information about the experience of attending a particular college, or physical proximity on campus. All of these possibilities suggest a causal influence of the older sibling's college choices.

Ultimately, the estimates presented in this paper cannot rule out the possibility that older siblings' college choices are simply proxies for unobserved differences between families. If so, then at

the very least this suggests that such differences are not well-captured by existing variables available to researchers in most administrative data sets, itself an interesting fact. It also suggests that older siblings' college choices can be used to help better predict the choices of younger siblings. Numerous recent interventions, such as Hoxby and Turner (2013), have been targeted at "at-risk" students, but identifying exactly which students are most at risk poses a challenge if model explanatory power is weak and it is clear that older sibling choices have explanatory power. We therefore have some evidence that high school counselors and schools might find it worthwhile to flag for extra attention students whose older siblings have made apparently poor college decisions.

More work is, however, needed to establish the extent to which the college choices of older siblings influence the choices of their younger siblings. The primary empirical challenge is finding an exogenous source of variation in the college choices of older siblings. Researchers have begun to find sources of exogenous variation driving students' enrollment choices generally (Goodman et al., 2014). Modern administrative data sets, which are large and rich with detail, make it increasingly likely that such variation can be combined with information on family structure to identify the causal impact of older siblings' choices. Identifying which channels, if any, are responsible for such impacts will be crucial to making such research findings useful to policymakers.

References

Altonji, J. G., Cattan, S., and I. Ware. (2010). "Identifying sibling influence on teenage substance use," National Bureau of Economic Research, Working Paper 16508.

Ashenfelter, O. and A. Krueger (1994). "Estimates on the Return to Schooling From a New Sample of Twins," *American Economic Review*. 84: 1157-1173.

Avery, C. and C. Hoxby (2004). "Do and Should Financial Aid Decisions Affect Students' College Choices?" in Caroline Hoxby, ed. *College Choices: The New Economics of Choosing, Attending, and Completing College*. University of Chicago Press.

Behrman, J., M. Rosensweig, and P. Taubman (1996). "College Choice and Wages: Estimates Using Data on Female Twins," *The Review of Economics and Statistics*. 78(4): 672-685.

Behrman, J., and P. Taubman (1986). "Birth Order, Schooling, and Earnings," *Journal of Labor Economics*. 4: S121-45.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). "The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment," *Quarterly Journal of Economics*, 127, 1205–1242.

Black, S, P. Devereaux, and K.Salvanes (2005). "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *Quarterly Journal of Economics*. 120(2): 669-700.

Booth, A. and H.J. Kee (2009). "Birth order matters: the effect of family size on educational attainment". *Journal of Population Economics*. 22, 367-397.

Bowen, W.G., Chingos, M., and McPherson, M. (2009). *Crossing the finish line*. Princeton, NJ: Princeton University Press.

Bulman, G., C. Hoxby, and J. Meer (2014). "Teach For America and Students' College Application Choices," Working paper.

Card, D. and A. Krueger (2005). "Would the Elimination of Affirmative Action Affect Highly Qualified Minority Applicants? Evidence From California and Texas," *Industrial and Labor Relations Review*, 56: 416-434.

Choy, S. (2001). "Students Whose Parents Did Not Go to College: Postsecondary Access, Persistence, and Attainment." Findings from the Condition of Education.

Clark, M., J. Rothstein and D.W. Schanzenbach. (2008). "Selection Bias in College Admissions Test Scores." Working Paper 14265, National Bureau of Economic Research.

Cohodes, S. and J. Goodman (2014). "Merit Aid, College Quality and College Completion: Massachusetts' Adams Scholarship as an In-Kind Subsidy," *American Economic Journal: Applied Economics*.

DesJardins, S., H. Dundar, and D. Hendel (1999). Modeling the college application decision process in a land-grant university, " *Economics of Education Review*, " 18: 117-132.

Dillon, E. and J. Smith (2013). "The determinants of mismatch between students and colleges," NBER Working Paper 19286.

Dynarski, S. (2003). "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion," *American Economic Review*, 93(1): 279-288

Fuller, W. C., Manski, C. F., and Wise, D. A. (1982). New evidence on the economic determinants of postsecondary schooling choices. *Journal of Human Resources*, 17(4): 477–498.

Goodman, J., Hurwitz, M. and Smith, J. (2014). College Access, Initial College Choice and Degree Completion. Harvard Kennedy School Faculty Research Working Paper Series, RWP14-030.

Goodman, S. (2013). "Learning from the Test: Raising Selective College Enrollment by Providing Information ." Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2013-69.

Hearn, J. (1991). "Academic and nonacademic influences on the college destinations of 1980 high school graduates," *Sociology of Education*. 64(3): 158-171.

Hossler, D., Braxton, J. and Coppersmith, G. (1989) "Under-standing student college choice," In *Higher Education: Handbook of Theory and Research* (Edited by J. C. Smart), Vol. 5, pp. 231–288. New York: Agathon Press.

Hotz, J. and J. Patano (2013). "Strategic Parenting, Birth Order and School Performance," NBER Working Paper #19542.

Hoxby, C., and C. Avery (2012). "The Missing 'One-Offs': The Hidden Supply of High-Achieving, Low Income Students," NBER Working Paper No. w18586.

Hoxby, C., and S. Turner (2013) "Expanding College Opportunities for High-Achieving, Low Income Students," SIEPR Discussion Paper No. 12-014.

Hurwitz, M. (2011). "The impact of legacy status on undergraduate admissions at elite colleges and universities," *Economics of Education Review*. 30(3): 480-492.

Hurwitz, M. (2012). "The impact of institutional grant aid on college choice," *Educational Evaluation and Policy Analysis*. 34(3): 344-363.

Hurwitz, M., J. Smith, S. Niu, and J. Howell (2105). " The Maine Question: How is Four-Year College Enrollment Affected by Mandatory College Entrance Exams?" *Educational Evaluation and Policy Analysis*. 37(1): 138-159.

Hyman, J. (2014). "ACT for All: The Effect of Mandatory College Entrance Exams on Postsecondary Attainment and Choice." University of Michigan Working Paper.

Jackson, G. (1978). "Financial aid and student enrollment," *Journal of Higher Education*. 49(6): 548-574.

Jacob, B., McCall, B., and Stange, K. M. (2013). "College as Country Club: Do Colleges Cater to Students' Preferences for Consumption?" National Bureau of Economic Research Working Paper 18745.

Joensen, J. and H. Nielsen (2013). "Peer Effects in Math and Science." Working Paper.

Kantarevic, J., Mechoulan, S. (2006). "Birth order, educational attainment, and earnings: an investigation using the PSID," *Journal of Human Resources*. 41, 755–777.

Klasik, D. (2013). "The ACT of enrollment: The college enrollment effects of required college entrance exam taking." *Educational Researcher,* 42, 151–160.

Leppel, K. (1993). "Logit estimation of a gravity model of the college enrollment decision," *Research in Higher Education*, 34(3): 387–398.

Lindahl, L., H. Regner. (2005). "College Choice and Subsequent Earnings: Results Using Swedish Sibling Data," *Scandinavian Journal of Economics*, 107(3): 437-457.

Long, B. (2004). "How have college applications changed over time? An application of the conditional logistic choice model," *Journal of Econometrics*. 121: 271-296.

Loury, L. (2004). "Siblings and gender differences in African-American college attendance," *Economics of Education Review*, 23(3), 213-219.

Luca, M. and J. Smith (2013). "Salience in Quality Disclosure: Evidence from the U.S. News College Rankings," *Journal of Economics and Management Strategy*, 22(1): 58-77.

Oettinger, G. (2000). "Sibling similarity in high school graduation outcomes: Causal interdependency or unobserved heterogeneity?," *Southern Economic Journal*: 631-648.

Pallais, A. (2013). "Small Differences that Matter: Mistakes in Applying to College," NBER Working Paper #19480.

Perna, L. W., & M.A. Titus. (2005). "The Relationship between Parental Involvement as Social Capital and College Enrollment: An Examination of Racial/Ethnic Group Differences." *Journal of Higher Education*. 76(5): 485-518.

Pope, D. and J. Pope (2009). "The Impact of College Sports Success on the Quantity and Quality of Student Applications," *Southern Economic Journal*. 75(3): 750-780.

Rouse, C. (1999). "Further Estimates of the Economic Return to Schooling From a New Sample of Twins," *Economics of Education Review*. 18(2): 149-157.

Sacerdote, B. (2011). "Peer effects in education: How might they work, how big are they and how much do we know thus far?" *Handbook of the Economics of Education*, 3, 249-277.

Smith, J. (2013). "Ova and Out: Using Twins to Estimate the Educational Returns to Attending a Selective College," *Economics of Education Review*, 36: 166-180.

Smith, J., M. Hurwitz, and J. Howell (2014). "Screening Mechanisms and Student Responses in the College Market," Working paper.

Figures and Tables

Figure 1 - Average SAT of Siblings' Enrolled Colleges
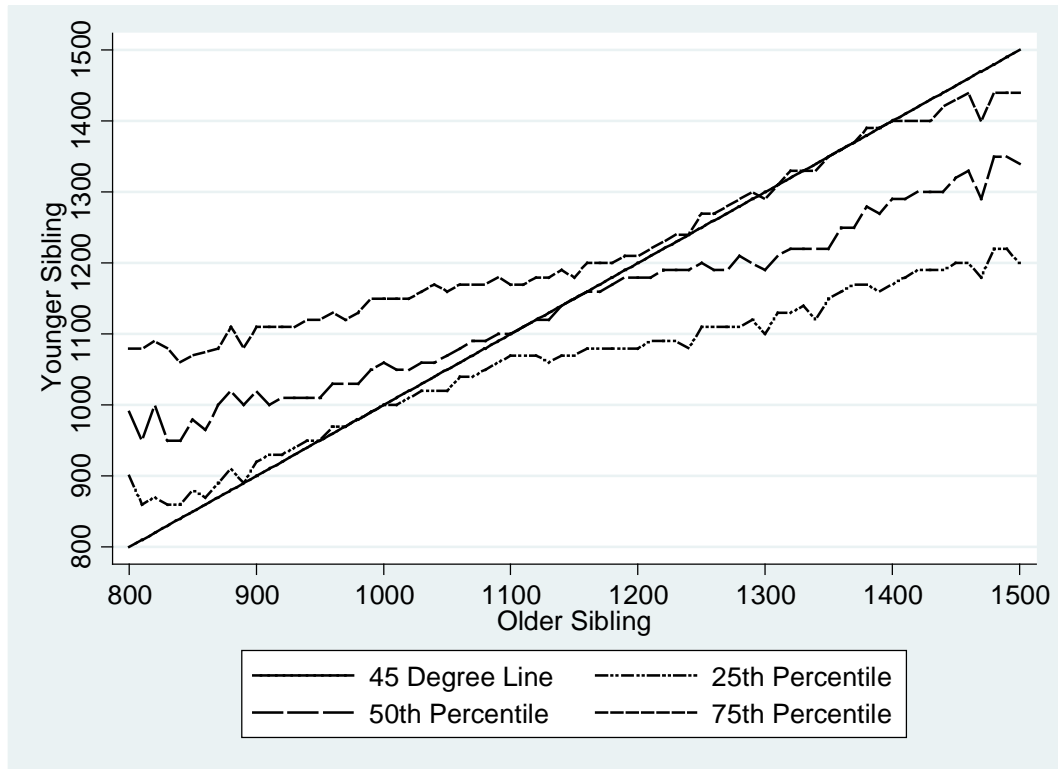
Notes: Includes all identified siblings who took the SAT between 2004 and 2011. Only includes younger and older sibling pairs who both attend a four-year college with reported average SAT scores. The older siblings college's average SAT is rounded to the nearest ten SAT points and are plotted on the x-axis. For all older siblings with a given average college SAT score, the 25th, 50th, and 75th percentiles of younger sibling's average SAT of enrolled colleges are calculated.

Figure 2 - Difference in Siblings' SATs and Average SATs of Enrolled Colleges
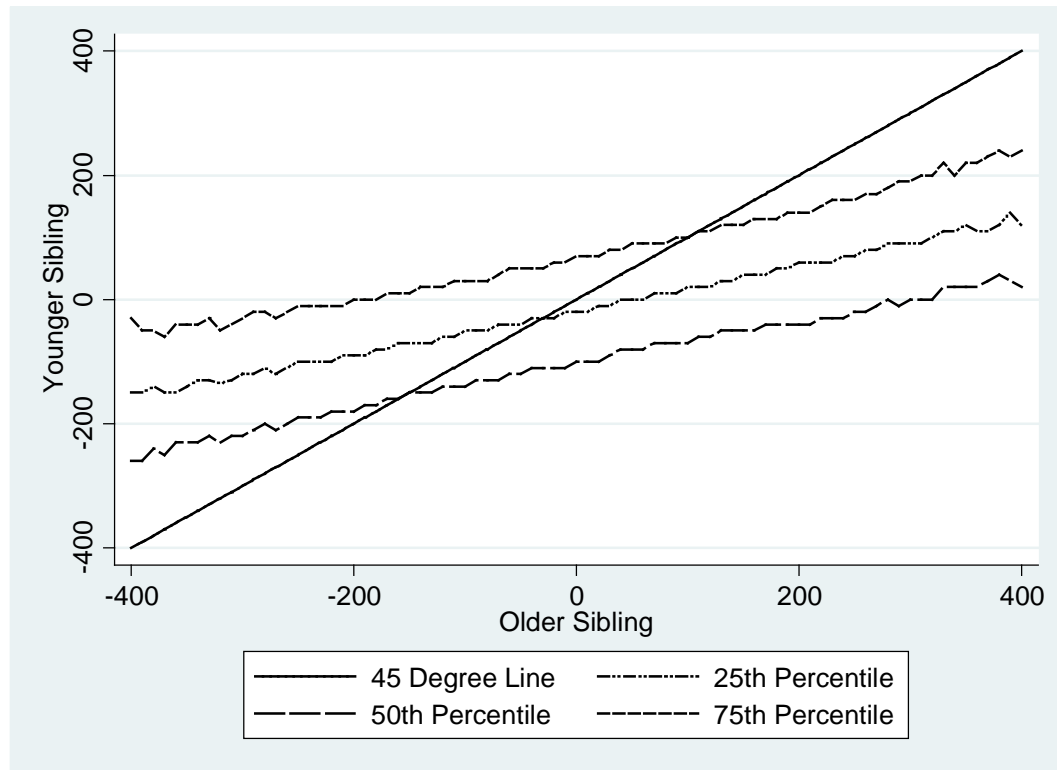

Notes: Includes all identified siblings who took the SAT between 2004 and 2011. Only includes younger and older sibling pairs who both attend a four-year college with reported average SAT scores. The older siblings college's average SAT is rounded to the nearest ten SAT points, subtracted from their own SAT score, and are plotted on the x-axis. For all older siblings with a given average difference, the 25th, 50th, and 75th percentiles of younger sibling's differences are calculated.

	All students (1)	Younger siblings (2)
Table 1 - Summary Statistics		
(A) Demographic characteristics		
Female	0.538	0.522
White	0.573	0.682
Black	0.119	0.065
Hispanic	0.128	0.093
Asian	0.087	0.087
Parental income: $0-$50,000	0.193	0.109
Parental income: $50,000-$100,000	0.208	0.198
Parental income: $100,000+	0.152	0.204
Parental income: Missing	0.447	0.490
Parental education: High school or less	0.164	0.108
Parental education: Some college or A.A.	0.245	0.201
Parental education: B.A. or higher	0.490	0.603
Parental education: Missing	0.102	0.087
(B) Academic skill		
High school GPA[1]	3.323	3.390
SAT score (Math + Critical Reading)	1016	1055
(C) College enrollment		
Four-year college	0.648	0.725
Two-year college	0.255	0.206
No college	0.097	0.069
Highly competitive college (Barrons' Most or Highly Compeitive	0.159	0.195
SAT score of enrolled college[2]	1118	1130
Undermatches (own SAT - college SAT > 100)	0.465	0.473
Overmatches (college SAT - own SAT >100)	0.160	0.161
Undermatches (own SAT > college 75th percentile)	0.445	0.457
Overmatches (own SAT < college 25th percentile)	0.153	0.153
(D) Relation to older sibling		
Same gender		0.510
Age difference		2.791
|Own SAT - older sibling SAT| > 100 SAT Points		0.563
Applies to older sibling's college[3]		0.309
Enrolls in same college[4]		0.187
Enrolls in same type (4-year/2-year/none)		0.687
Enrolls in same Barron's category[4]		0.358
N	10,044,488	1,614,007

Notes: The sample in column 1 includes all SAT test-takers from the high school graduation cohorts of 2005-2011. Column 2 includes only students who could be identified as younger siblings. Demographic characteristics and high school GPA are self-reported by students to the College Board. College enrollment outcomes come from the National Student Clearinghouse.

1. Among students with reported GPA (94%).
2. Among students who enroll in college with reported SAT.
3. Equals zero if older sibling did not enroll in college.
4. Equals zero if younger or older sibling did not enroll in college.

Table 2 - Younger Siblings' College Choices

	Enrolls in four-year college (1)	Highly competitive college (2)	Average SAT of college (3)	Undermatches (4)	Overmatches (5)
(A) Academic skill measures					
High school GPA	0.151***	0.103***	22.703***	-0.056***	0.060***
	(0.001)	(0.001)	(0.140)	(0.001)	(0.001)
SAT score	0.058***	0.068***	16.395***	0.123***	-0.088***
	(0.000)	(0.000)	(0.054)	(0.000)	(0.000)
Older sibling's high school GPA	-0.012***	-0.011***	-1.168***	-0.000	0.001
	(0.001)	(0.001)	(0.141)	(0.001)	(0.001)
Older sibling's SAT score	-0.010***	-0.008***	-1.143***	-0.012***	0.007***
	(0.000)	(0.000)	(0.050)	(0.000)	(0.000)
(B) Older sibling's college choice					
Older sibling enrolls in four-year college	0.162***	--	--	--	--
	(0.001)	--	--	--	--
Older sibling enrolls in highly competitive college	--	0.187***	--	--	--
	--	(0.001)	--	--	--
Average SAT of older sibling's enrolled college	--	--	0.144***	--	--
	--	--	(0.001)	--	--
Older sibling undermatches	--	--	--	0.062***	--
	--	--	--	(0.001)	--
Older sibling overmatches	--	--	--	--	0.057***
	--	--	--	--	(0.001)
R-squared	0.311	0.377	0.904	0.640	0.341

Notes: Standard errors are clustered at the family level and in parentheses (* p<.10 ** p<.05 *** p<.01). The sample includes all SAT test-takers from the high school graduation cohorts of 2005-2011 who could be identified as younger siblings (N = 1,614,007). Older siblings are defined as the oldest sibling observed in the data. In column 1, the outcome is an indicator for the younger sibling enrolling in any four-year college. In column 2, the outcome is an indicator for enrollment in a college ranked as most or highly competitive by Barron's. In column 3, the outcome is the average SAT score of the four-year college enrolled in, where colleges missing SAT scores and non-enrollee are assigned the lowest reported college average SAT of 675. In columns 4 and 5, the outcomes are indicators for undermatch, defined as a student's SAT score exceeding the 75th percentile SAT score of his enrolled college, and overmatch, defined as the 25th percentile SAT score of a student's enrolled college exceeding his own SAT score. The 25th and 75th percentile SATs of enrolled colleges are assigned as 440 and 790, respectively, when missing. All regressions include high school by year fixed effects. Also included but not shown are controls for gender, race, parental income and parental education.

Table 3 - Heterogeneity by Family Characteristics

Y = Younger sibling's college choice	Income Less Than $50k (1)	Income More Than $50k (2)	Black or Hispanic (3)	White (4)	At least one parent has BA degree (5)	No parents have BA degree (6)	4-year public within 25 miles (7)	No 4-year public within 25 miles (8)
Older sibling's college choice								
Enrolls in four-year college	0.167***	0.156***	0.160***	0.156***	0.155***	0.163***	0.159***	0.175***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
Enrolls in highly competitive college	0.165***	0.183***	0.157***	0.186***	0.191***	0.163***	0.184***	0.205***
	(0.004)	(0.001)	(0.003)	(0.001)	(0.001)	(0.002)	(0.001)	(0.003)
Average SAT score of enrolled college	0.123***	0.137***	0.126***	0.143***	0.147***	0.132***	0.143***	0.144***
	(0.003)	(0.001)	(0.003)	(0.001)	(0.001)	(0.002)	(0.001)	(0.003)
Undermatches	0.049***	0.055***	0.062***	0.057***	0.063***	0.053***	0.063***	0.054***
	(0.003)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Overmatches	0.052***	0.053***	0.052***	0.054***	0.060***	0.050***	0.057***	0.055***
	(0.003)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)

Notes: Each coefficient is from a separate regression. Standard errors are clustered at the family level and in parentheses (* $p<.10$ ** $p<.05$ *** $p<.01$). The sample includes all SAT test-takers from the high school graduation cohorts of 2005-2011 who could be identified as younger siblings and self-reported income, race/ethnicity, or zip code. Older siblings are defined as the oldest sibling observed in the data. All regressions include high school by year fixed effects. Also included but not shown are controls for gender, race, parental income, parental education, own SAT score and high school GPA, and older sibling's SAT score and high school GPA.

Table 4 - Sibling Skill, Gender and Age Differences

	Applies to same college as older sibling (1)	Enrolls in same college as older sibling (2)	Enrolls in same type (4-yr/2-yr/none) (3)	Enrolls in same Barron's category (4)
(A) Academic skill measures				
High school GPA	0.039***	0.022***	0.084***	0.036***
	(0.001)	(0.001)	(0.001)	(0.001)
Own SAT score	0.012***	-0.005***	0.020***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Older sibling's high school GPA	0.072***	0.012***	0.068***	0.012***
	(0.001)	(0.001)	(0.001)	(0.001)
Older sibling's SAT score	0.020***	-0.006***	0.020***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)
(B) Differences between siblings				
|Own SAT - older sibling SAT| > 100	-0.063***	-0.057***	-0.053***	-0.095***
	(0.001)	(0.001)	(0.001)	(0.001)
Siblings' gender differs	-0.024***	-0.024***	-0.011***	-0.023***
	(0.001)	(0.001)	(0.001)	(0.001)
Age difference	-0.006***	-0.008***	-0.005***	-0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
R-squared	0.165	0.123	0.174	0.095

Notes: Standard errors are clustered at the family level and in parentheses (* $p<.10$ ** $p<.05$ *** $p<.01$). The sample includes all SAT test-takers from the high school graduation cohorts of 2005-2011 who could be identified as younger siblings (N = 1,614,007). Older siblings are defined as the oldest sibling observed in the data. Each column represents a single regression, where the outcome is an indicator for the younger sibling making the same college choice as the older sibling. All regressions include high school by year fixed effects. Also included but not shown are controls for gender, race, parental income and parental education.