Information Inequality in Major Choices*

Xinyao Qiu†

October 31, 2023

[Click Here for Latest Version]

Abstract

I study disparities in college major choices across students from different socioeconomic backgrounds and analyze their implications for intergenerational income mobility. One potential explanation for these disparities is differential access to information about majors' academic content and personal fit. To explore the role of information frictions on major choices, I use administrative data from the centralized college application system in China. Consistent with the information inequality hypothesis, I document that students of low socioeconomic status (SES) are 21.6% (3.16 percentage points) more likely than their high-SES peers to choose majors that are familiar to them from their high school curricula. Further support for the information inequality hypothesis comes from a survey experiment in which high school students report their expectations about college majors and from information spillovers among high school classmates. To discuss the economic consequences, I calibrate a model of major choice and find that, because of information inequality, low-SES students face higher mismatch rates and lower future incomes than their high-SES peers. Counterfactual analyses indicate that information interventions and affirmative action policies can effectively narrow the income gap across socioeconomic backgrounds.

JEL classification: I23, I24, J3

*I am grateful to Luigi Pistaferri for invaluable guidance and support. I thank Ran Abramitzky, Caroline Hoxby, Petra Persson, Maya Rossin-Slater, Alessandra Voena, and seminar participants at Stanford for helpful discussions, comments, and questions. I also thank Hongbin Li and Xiaoyang Ye for kindly providing data. This research is supported by the George P. Shultz Dissertation Support Fund and the E.S. Shaw and B.F. Haley Fellowship for Economics through grants to the Stanford Institute for Economic Policy Research and by the Institute for Research in the Social Sciences at Stanford University and the Stanford Center on China's Economy and Institutions.

†Department of Economics, Stanford University. Email: xyqiu@stanford.edu.

1 Introduction

The choice of a college major is a decision with important economic consequences. The literature has documented substantial variation in labor market returns across different majors, even after accounting for college selectivity (Altonji et al., 2012, 2016; Hastings et al., 2013; Kirkeboen et al., 2016; Andrews et al., 2017; Bleemer and Mehta, 2022). Despite the importance of this choice, students face informational barriers when choosing majors, including imperfect information about tuition and future earnings (e.g., Hastings et al., 2016), career prospects (Conlon and Patel, 2022), and their own fit with the major. Hence, it is possible that students make suboptimal major choices with long-term economic consequences.

Not all students are impacted equally by information frictions. There is evidence that the incidence of information frictions falls more on students from disadvantaged socioeconomic backgrounds (Hastings et al., 2015, 2016). This raises a series of empirical and policy-related questions: Do students from different socioeconomic backgrounds differ in their choice of major due to unequal access to information? Is higher education fully effective in promoting intergenerational mobility when students from disadvantaged backgrounds are more likely to make underinformed major choices? Can policies be designed to effectively mitigate the consequences of unequal access to information?

This study answers these questions by analyzing socioeconomic disparities in students' tendencies to choose majors that resemble (in content if not name) the subjects studied in high school in the context of China's centralized college application system. Unlike students in the U.S., but similarly to those in most other countries (Bordon and Fu, 2015), Chinese students decide on their majors right after high school, at the time of college application. This early timing makes applicants more vulnerable to information frictions and more likely to rely on high school experience when making major choices.[1] This study uses large-scale and high-quality administrative data from China, but the prevalence of similar application processes in other countries means that the results are likely to have implications for understanding inequities in college major choices in a wide range of settings.

I classify college majors into two groups: ones that resemble specific high school subjects in content, and ones that are unrelated to the high school curriculum. Short of a better nomenclature, I refer to them as "familiar" and "unfamiliar" majors, respectively. In China, the high school curriculum is standardized, with all students studying the same set of subjects: Chinese Language, English,

[1] This contrasts with the experience of most U.S. students, who have the opportunity to sample foundational courses across various fields before committing to a specific major. However, some universities and departments in the U.S. impose admissions and grade requirements to restrict entry into certain competitive majors. Students aiming to enroll in these majors often need to make their decisions early as well.

Mathematics, Physics, Chemistry, Biology, History, Geography, and Politics. Within this context, examples of familiar majors include Chinese Language, Mathematics, Physics, and History, whereas examples of unfamiliar majors include Finance, Psychology, Computer Science, and Media Studies. Intuitively, students with limited information would rely more heavily on their high school experience when deciding which major to pursue and hence would be more likely to choose a "familiar" major.

This study comprises three parts. First, I document that students of low socioeconomic status (SES) are significantly more likely to apply to and enroll in familiar majors than their high-SES counterparts, despite the fact that familiar majors offer on average lower labor market returns than unfamiliar majors. Second, I provide both experimental and observational evidence suggesting that information inequality between high- and low-SES students is the primary driver of the observed disparity in major choice, whereas differences in application strategies, risk attitudes, or preferences are not important factors. Finally, by calibrating a major choice model, I quantify the economic consequences of the disparity in major choice and evaluate the effectiveness of information interventions and affirmative action policies.

For my main analyses, I use administrative records on college application and admission from Ningxia province in China between 2014 and 2018.[2] The primary strength of the data is that they include students' submitted rank-order lists (ROLs), which provide direct insight into their preferences for majors. I supplement the ROL data with additional administrative data that cover applicants from all of China between 1999 and 2003. Compared with the ROL data, this supplementary data source is appealing for its national coverage, but it does not include students' submitted rank-order lists. I complement the administrative data sets with survey data from the Chinese College Student Survey (conducted between 2010 and 2015), which contains richer information on students' family backgrounds and post-college plans. Across these data sources, I use students' urban–rural status as a proxy for their socioeconomic backgrounds: urban for high-SES students and rural for low-SES students. The findings remain robust when I use alternative measures of SES, such as parental job sector and household income.

Using the administrative data, I document a large disparity in choice of major between students from different socioeconomic backgrounds. Specifically, I categorize a major as familiar if its name overlaps with the name of a high school subject. When comparing students with *similar* demographic characteristics, college entry exam performance, and probability of admission, I find that low-SES students are 21.6% more likely to choose a familiar major than their high-SES counterparts. The

[2]Ningxia is one of China's 34 province-level administrative divisions with a population of approximately 7.2 million. While this represents approximately 0.5% of China's total population, it is larger than the population of Norway.

observed disparity in major choice holds important implications for income inequality, as I show in my data that familiar majors yield lower labor market returns on average than unfamiliar majors. This gap in returns between familiar and unfamiliar majors is of a similar magnitude for both high- and low-SES students, which suggests that the observed socioeconomic disparity in major choice is not driven by differential labor market returns.

There are several possible explanations for the documented disparity in major choice. It is possible that students from high-SES backgrounds have better access to information about unfamiliar majors than low-SES students through channels outside high school, such as family members and career counselors. It is also possible that high- and low-SES students differ in their application strategies, risk attitudes, or preferences for majors. In the second part of my study, I provide both experimental and observational evidence suggesting that information inequality is an important driver of the observed disparity in major choice.

I first conduct a survey experiment to directly measure students' information about different majors. In the survey, students are instructed to assess their fit with specific majors and to report their confidence levels in these assessments. I use the self-reported confidence as a measure of students' information about the majors, under the assumption that when students are more informed about a major, regardless of whether they assess the major as a good fit, they are more confident about the accuracy of the assessment. For a randomly selected subset of students in the sample, I implement an additional information treatment aimed at "familiarizing" them with the unfamiliar majors before the assessment. In contrast, the control group does not receive such information. By focusing on the control group only, I find that students are less confident about the accuracy of their assessment of unfamiliar majors than familiar majors, and that the gap is larger for low-SES students. The results confirm that students, especially low-SES students, are less informed about unfamiliar majors than about familiar majors. Moreover, by comparing the responses of the treated students with those of the control group, I find that the information treatment can successfully narrow the information gap between familiar and unfamiliar majors, with a more significant treatment effect observed among low-SES students.

I also provide observational evidence that connects a lack of information with choosing more familiar majors. Leveraging the random assignment of students into high school classes, I find that a student becomes more likely to apply to an unfamiliar major if her class has a larger share of high-SES classmates. The peer effects are more pronounced when the high-SES classmates share the same gender or ethnic group with the focal student, as they are likely to interact with her

more frequently. The results are consistent with an information spillover story: low-SES students become more likely to choose unfamiliar majors after obtaining information about unfamiliar majors from their high-SES peers.[3] In addition to the experimental and observational evidence supporting information inequality, I conduct supplementary analyses to show that alternative explanations, including differences in application strategies, risk attitudes, and preferences, are unlikely to be the main cause of the observed disparity in major choice.

In the final part of the paper, I analyze the economic consequences of information inequality by building a model in which students make their choice of major to maximize their expected utility. A key set of parameters in the model is the amount of information that students have about each major, which depends on the student's socioeconomic background and the major's familiarity. I estimate these parameters by matching the model's simulations with empirical data on students' major choices and the average earnings of each major. The estimates are consistent with the existence of information inequality: students from low-SES backgrounds are less informed about unfamiliar majors than their high-SES counterparts and are, on average, 11.9% more likely to experience mismatches in the enrolled major.

Furthermore, the estimation results indicate that familiar majors, compared with unfamiliar ones, have lower labor market returns, even after adjusting for selection on students' observed abilities. As a result, information inequality causes income inequality because less informed low-SES students choose familiar majors that have lower returns. Specifically, by expressing income as a function of students' major choice and abilities, the model predicts a 1.4% income gap between the high- and low-SES college graduates (equivalent to a lifetime income gap of 64,000 CNY), of which 70.0% can be explained by the disparity in major choice caused by information inequality. Counterfactual analyses suggest that information interventions and affirmative action admission policies can effectively reduce the income gap and thereby promote intergenerational mobility.

Literature Review

My study contributes to the literature that documents socioeconomic disparities in college major choices. Using data from the U.S., the literature finds that low-income students tend to choose majors that are associated with higher returns and less risky careers (Saks and Shore, 2005; Ma, 2009). However, studies based on systems where students choose majors *before* entering college find that low-income students are matched to less selective majors with lower earnings potential (Delaney

[3]Placebo analyses provide evidence against the possibility that the results are driven by non-random sorting of students into different classes or pure conformity among peers.

and Devereux, 2020; Campbell et al., 2022).[4] These findings are consistent with my research, which shows that students from disadvantaged backgrounds face additional barriers when they have to choose a major immediately after high school.

By introducing the concept of major familiarity, my work is very close in spirit to the research that examines the effects of previous exposure on major choices. Joensen and Nielsen (2016), Fricke et al. (2018), and Patterson et al. (2021) document, in different settings, that past experience with courses or tasks related to a major increases students' probability of choosing that major later in college. My study extends this literature by documenting that the impact of previous exposure on the choice of major is larger for students from disadvantaged backgrounds. These students rely more heavily on their personal educational experiences, as they lack access to other information sources when choosing majors.

This study also contributes to the literature that explores factors behind students' choices of major (e.g., Altonji et al., 2012; Beffy et al., 2012; Wiswall and Zafar, 2015a; Altonji et al., 2016; Patnaik et al., 2020; Abramitzky et al., 2022; Ersoy and Speer, 2022), with a focus on the role of imperfect information (Hastings et al., 2016). I find that students respond to the information treatment in the survey experiment and change their choice of major when they obtain information from peers. This evidence adds to the literature that studies how access to information affects students' major choices, in both experimental (Wiswall and Zafar, 2015b; Hastings et al., 2015; Barone et al., 2017; Baker et al., 2018) and non-experimental settings (Zafar, 2011; Stinebrickner and Stinebrickner, 2014).[5]

This study further adds to the literature on within-classroom peer effects (e.g., Hoxby, 2000; Carman and Zhang, 2012; Lavy et al., 2012; Burke and Sass, 2013; Le and Nguyen, 2019) and spillover effects in educational decisions (Bobonis and Finan, 2009; Joensen and Nielsen, 2018; Gurantz et al., 2020; Abramitzky et al., 2021; Aguirre and Matta, 2021; Dahl et al., 2021; Barrios-Fernández, 2022; Chesney, 2022). Related to the finding of within-family spillovers in major choices by Altmejd et al. (2021), I provide evidence that students' choice of major is also affected by information spillovers

[4] A substantial body of literature studies socioeconomic disparities in college applications and admissions. Low-income students have a lower likelihood of enrolling in colleges (e.g., Boneva and Rauh, 2017), and along the intensive margin, they are admitted to less selective colleges (e.g., Smith et al., 2013; Black et al., 2015; Dillon and Smith, 2017). The disparities are attributed to factors such as a lack of accurate information regarding the costs and benefits of college (Scott-Clayton, 2012; Ehlert et al., 2017; Peter and Zambre, 2017; Bleemer and Zafar, 2018; Lergetporer et al., 2021) and the non-transparent financial aid policies (Bettinger et al., 2012; Hoxby and Avery, 2012; Hoxby and Turner, 2013, 2015; Dynarski et al., 2021; Burland et al., 2022).

[5] Beyond the choice of major, there is a broad literature that studies the impact of information frictions on school choice (Kapor et al., 2020; Artemov, 2021; Bucher and Caplin, 2021; Arteaga et al., 2022; Grenet et al., 2022) and the effect of information interventions on guiding educational decisions (e.g., Jensen, 2010; Oreopoulos and Dunn, 2013; McGuigan et al., 2016; Carrell and Sacerdote, 2017; Bleemer and Zafar, 2018; Bonilla-Mejía et al., 2019; Kerr et al., 2020; Mulhern, 2021; Ainsworth et al., 2023).

from their high school classmates.

Finally, this study extends an emerging literature that studies the impact of higher education on upward social mobility (Zimmerman, 2019; Chetty et al., 2020; Barrios Fernández et al., 2021; Kaufmann et al., 2021; Jia et al., 2022). By shifting the focus from elite college admissions to choice of major, my study discusses how higher education may not successfully serve the purpose of promoting intergenerational mobility if low-SES students have to make major choices with more limited information than their high-SES counterparts. To quantify the inequality consequences, I build a discrete choice model where students make major choices that maximize their expected utility from labor market returns, taking self-selection based on observed exam scores into consideration (Arcidiacono, 2004; Altonji et al., 2012; Bordon and Fu, 2015; Kinsler and Pavan, 2015). Using this framework, I analyze the labor market consequences of information frictions and how information inequality translates into intergenerational income inequality through the choice of major.

The rest of the paper is organized as follows. Section 2 explains the relevant Chinese institutional background, and Section 3 describes the data sources and the key variables. Section 4 presents the SES disparity in major choice. Section 5 argues that information inequality explains the observed disparity, and Section 6 discusses the alternative explanations. Section 7 builds a major choice model for the welfare and counterfactual analysis. Section 8 concludes.

2 Institutional Background

2.1 High School Education

In China, high schools (grades 10 to 12) offer two academic tracks for their students: the science track or the humanities track.[6] Although the width and depth of students' learning for each subject vary depending on their academic tracks, all students, regardless of their academic track, study the same set of nine core subjects in high school: Chinese Language, English, Mathematics, Physics, Chemistry, Biology, History, Geography, and Politics. Students are required to pass examinations for these subjects to graduate from high school.

2.2 The National College Entrance Examination

All students who wish to enter college are required to take the National College Entrance Examination (NCEE) in the last year of high school.[7] During the once-a-year examination, students are tested

[6] In a few provinces, some cohorts of students do not declare academic tracks.

[7] A negligible share of students with exceptional talent are admitted to colleges without taking the NCEE.

on Chinese Language, English, Mathematics, and subjects specific to their academic tracks: Physics, Chemistry, and Biology for the students on the science track; or History, Geography, and Politics for the humanities track.[8] The exam score determines the student's eligibility for admission to colleges and majors. Because the exam, grading, and subsequent college application and admission process are administered at the provincial level, students' NCEE scores are comparable within the same province–year–academic track.

2.3 Colleges and Majors

There are over 1,000 postsecondary institutions in China. They are classified vertically into four tiers of decreasing quality: public elite colleges (Tier 1), public non-elite colleges (Tier 2), private colleges (Tier 3), and other vocational or for-profit institutions (Tier 4). In the baseline analysis of this study, I focus on students' application to the first two tiers, and argue that, among students from low-SES backgrounds, even those with good enough academic performance to qualify for public colleges are heavily impacted by information frictions. Approximately 1,500 different majors are offered by these institutions, among which 465 are offered by public colleges.[9] The Chinese Ministry of Education groups these majors into ten main categories, including History, Science, Art and Humanities, Law and Politics, Education, Engineering, Agriculture, Economics, Healthcare, and Management.

2.4 Application and Admission

College applications and admissions are centralized and are administered at the province level. Before the application season starts, colleges make plans and set an admission quota for each major, which is the maximum number of students that they plan to admit from each province and academic track.[10] The admission quota of each college–major–province–academic track is available to the public.

There are multiple rounds of application and admission: one round for each college tier. Students not admitted in the earlier rounds apply for college–majors in the subsequent rounds. In each round, eligible students fill out their rank-order lists with college–major programs. In Ningxia province of China, where the main analysis of my study is based, a student can list up to four colleges and up to six majors within each college on her ROLs. This allows a total of 24 positions on each

[8] In recent years, some provinces initiated a reform that allows students to self-select the subjects on which they are tested. However, the samples in this study are not impacted.

[9] The number of distinct majors is large because sometimes very similar majors are named and coded differently by different colleges. For example, Biology, Biotechnology, Bioinformatics, and Biological Engineering are listed under different codes.

[10] Selective colleges generally admit students from all provinces across the country, whereas less selective colleges primarily admit students from local and nearby provinces (Yang, 2021).

student's list. Refer to Appendix Figure F1 for a rank-order list template. Within each round, a centralized matching mechanism assigns students to their admitted college–major programs based on the students' priority scores, the submitted rank-order lists, and the pre-determined admission quota. The priority score is the student's NCEE score.[11]

The matching algorithm varies across provinces and years (Chen and Kesten, 2017). Ningxia province adopts a constrained deferred acceptance algorithm. Under this mechanism, the student with the highest score is admitted to her first choice college–major, and the other students are admitted to their most preferred choice on the ROL that has unfilled quota after the admission of students with higher priority scores. Ex post, each college–major program has an admission cutoff that is jointly determined by students' preferences and admission quotas. Students can be admitted to the program on their ROLs only if their scores are at or above its cutoff. In other words, students with the same scores have an equal probability of being admitted to any given college–major program.

Because each student is admitted to no more than one college–major, it is expected that students will enroll in the college–major they have been admitted to. Once enrolled, very few students switch majors during college, and dropout rates are extremely low.[12][13] College tuition and fees are fixed by the government at a level that is affordable to most families (approximately 800 USD per year), with limited variation across different colleges and majors. All these features make China an attractive setting to study college and major choices.

3 Data and Key Variables

3.1 Data Sources

I use data from multiple sources, including administrative records of college application and admission, the Chinese College Student Survey, and a survey experiment that I conducted myself.

3.1.1 ROL Data

The primary data set that I use is individual-level administrative data that cover the universe of students from Ningxia province in China who applied to college during the years 2014–2018. Ningxia

[11] Only a small fraction of students, such as ethnic minority students, children of military casualties, or those with exceptional talents, receive bonus points added to their NCEE scores.

[12] Most colleges have policies in place that restrict the percentage of students who are allowed to change majors during their undergraduate studies to a range of $\leq 5\%$ to $\leq 20\%$. Moreover, students who plan to change majors often encounter additional requirements related to their NCEE scores and college GPAs.

[13] In China, the higher education system upholds rigorous admission standards alongside relatively lenient graduation requirements. Consequently, China has one of the lowest college dropout rates worldwide, with fewer than 1% of enrolled students failing to complete their bachelor's degrees (Marioulas, 2017; Jia and Li, 2021).

is a relatively low-income province situated in the northwest of China. The data set has four components: the applicants' personal and academic profiles, their submitted rank-order lists, admission results, and the basic information for each college–major program. The data are referred to as ROL data because they contain students' submitted rank-order lists, which are unique to this data set.

The applicants' personal characteristics include their gender, ethnicity, county of residence, and variables that reflect their SES, including urban–rural status and parental job sector. The applicants' academic profiles include their high school and high school class, academic track, exam performances in the NCEE, and number of bonus points awarded (if any). Most importantly, the data contain students' submitted rank-order lists with each college–major program to which they applied, allowing me to directly observe each applicant's choice of major. The data also contain students' admission results, including whether they are admitted and, if so, to which college–major program. For each college–major program, I additionally observe its basic information, including the tier and location of the college, the category of the major, and the admission quota and tuition level of the college–major pair.[14] The main sample of analysis in this study consists of the students who apply to public colleges in the first two rounds from the ROL data.

3.1.2 National Data

To make sure any disparity in major choice that I am able to capture is generalizable to the entire country, I supplement the ROL data with the administrative records of the universe of college applicants in China between 1999 and 2003. The data are appealing due to their national coverage, compared with the ROL data, which cover only college applicants from Ningxia province. The national data contain variables similar to those in the ROL data, but unfortunately, they do not include the students' rank-order lists. Therefore, to study students' major choices using the national data, I have to restrict the sample of analysis to the students who are admitted to a college–major program and analyze their admitted majors, as the admission partially reflects their choice of major.[15]

3.1.3 Chinese College Student Survey

For the analyses that require richer information on students' family backgrounds and the analyses that involve students' paths after college, I use data from the Chinese College Student Survey (CCSS).

[14] The information on students' high schools and high school classes is available between 2017 and 2018. The admission quota of each college–major program is available for the period between 2016 and 2018.

[15] We observe similar levels of admission disparity (i.e., the difference in admission likelihood for familiar majors between high- and low-SES students) when we compare the data from Ningxia and the national data (Appendix Table F2). This suggests that the findings obtained from analyzing the rank-order list data in Ningxia are applicable nationwide.

The survey was conducted between 2010 and 2015 and covers a representative sample of senior college students who took the NCEE during the 2006–2011 period.[16] In addition to the variables available in the administrative data, the CCSS contains more self-reported information on students' family backgrounds, such as household income. Moreover, because the survey recruits senior college students who are about to graduate, I am able to observe students' post-college plans, including the wages of the best job offers among those who have secured employment and the graduate schools the students plan to attend among those who intend to pursue advanced degrees.

3.1.4 Survey Experiment

Despite the rich information, the administrative records and the CCSS data do not allow us to directly observe how informed students are about different majors. To establish information inequality as the mechanism behind the observed disparity in major choice, I conducted a survey experiment in February of 2022 in one high school in China. Approximately 700 senior high school students participated in the survey, in which I directly measured their information and knowledge on some selected majors, with and without an information treatment. The survey design is described in detail in Section 5.1. I additionally obtained administrative records from the high school that contain the students' demographic and academic performance information.

3.2 Key Variables

The key variables for characterizing socioeconomic disparities in major choice are the familiarity of the major, which describes the proximity between the major and subjects in high school curricula, and socioeconomic status of the students.

3.2.1 Major Familiarity

I define major familiarity as an indicator that summarizes whether the major is closely related to a subject in the high school curriculum. As mentioned in Section 2, in China, all high schools offer nine core subjects, while the choices for college majors are much more abundant. Nevertheless, as we can tell from their names, many majors (e.g., History, Applied Mathematics) are tied to specific high school subjects and can thus be classified as familiar. Specifically, I consider three definitions of major familiarity. Because all Chinese students study the same set of subjects during high school, the defined major familiarity is free from endogenous selection.

[16] See Jia and Li (2021) for a detailed description of how the survey was conducted.

Narrowly Defined Major Familiarity Under the narrow definition, I classify majors with the exact same names as high school subjects as familiar. Mathematics, Biology, and Chinese Language are examples of familiar majors due to the existence of high school subjects with the same name. In contrast, Statistics, Economics, and Computer Science are examples of unfamiliar majors, as there are no high school subjects with these names. In column (1) of Table 1, I list examples of most popular familiar and unfamiliar majors according to this narrow definition.

Broadly Defined Major Familiarity The narrow definition can sometimes be too strict. Majors such as Applied Mathematics and Biotechnology are categorized as unfamiliar under the narrow definition due to a lack of exact match, despite their high relevance to the high school subjects of Mathematics and Biology. Therefore, I build a broader definition of major familiarity: a major is classified as familiar if its name overlaps with the name of one high school subject. Under this definition, majors such as Applied Mathematics, Applied Physics, and Biotechnology are classified as familiar majors. In column (2) of Table 1, I present more examples of familiar and unfamiliar majors based on this broader definition. The majors labeled as familiar are a superset of the familiar majors under the previous narrower definition. Throughout this study, I use the broad definition of major familiarity as my preferred measure, while demonstrating the robustness of the results across alternative definitions.

Continuously Defined Major Familiarity I also build a continuous measure of major familiarity capturing variations in students' exposure to different college majors beyond the simple comparison to high school subject names. In Appendix A.2, I explain how the continuous measure is constructed and provide examples. Column (3) of Table 1 displays the continuous familiarity index of the most popular majors. In most cases, the three measures of familiarity align closely. The distribution of the continuous familiarity index, in relation to the two discrete familiarity measures, can be seen in Appendix Figure F2.

Discussion As major familiarity is the key concept introduced in this study, I provide further discussion on the interpretation of major familiarity.

All three definitions of major familiarity here rely mostly on the name of the college major. This is because names are very salient during the process of choosing a major while the actual content of a major may not be as obvious. Specifically, in the Chinese context, when applicants are filling out their rank-order lists, there is not even a short description or any additional information available

from the application platform about the majors other than their names. Therefore, it is difficult to have a good understanding of the content of different majors at the time of college application, especially for poorly informed high school students without any college experience.

Furthermore, major familiarity is to be interpreted in a relative sense. When a major is classified as unfamiliar, it does not preclude the possibility that part of its content is covered in high school curricula. The only assumption required in my study is that high school education provides students with more exposure to familiar majors than to unfamiliar majors.[17] Conversely, when a major is classified as familiar, it does not imply that students have perfect information about the major. The information that students gain about familiar majors through high school education might be incomplete.[18] However, the overall assumption remains: students are more exposed to familiar majors than to unfamiliar ones during their high school education. Finally, major familiarity describes students' exposure to the major through high school curricula, and it does not apply to students' information obtained through other channels. It is possible that a student is quite familiar with an unfamiliar major because her parents work in a related field; this possibility does not contradict the narrative of the study.

Income Gap between Familiar and Unfamiliar Majors Does the distinction between familiar and unfamiliar majors matter? In Table 2, I show that familiar majors are associated with lower labor market returns than unfamiliar majors. The analysis utilizes the CCSS data, which include self-reported monthly income from students who have secured employment after graduating from four-year public colleges.[19] Column (1) of Table 2 reveals that the average income of students graduating from familiar majors is 18.7% lower than that of students who graduate from unfamiliar majors. In column (2), after accounting for major category and college fixed effects, there remains a 7.93% gap in average income between familiar and unfamiliar majors.

The estimated gap includes both the disparity in the labor market returns between familiar and unfamiliar majors and the selection effects arising from differences in the characteristics of enrolled students. To mitigate the selection effects, I further include controls for all observable student

[17] For example, despite being classified as an unfamiliar major, Materials Science appears in part of the high school Chemistry curriculum, where students learn some basic properties about materials. However, the only assumption in my study is that students overall have a better understanding of the concepts and principles of Chemistry than of those of Materials Science because of the breadth and depth of learning.

[18] For example, what students learn in their high school Mathematics class can be very different from what they will learn in college if they major in Mathematics. Enjoying and performing well in high school Mathematics do not guarantee continued interests or academic success in college Mathematics.

[19] Appendix Table F1 shows that, after controlling for major category and college fixed effects, there is no significant difference is the likelihood of receiving a job offer among students graduating with familiar and unfamiliar majors.

characteristics in column (3). More specifically, I adjust for the students' province–year–academic track fixed effect, log household income as a proxy for socioeconomic status, demographics (gender and ethnic group), NCEE performances and bonus points, rankings in college GPA, tuition levels, and job location. It is worth noting that by controlling for students' NCEE scores, I am essentially comparing students with equal admission probabilities. The result indicates that, after all observables are accounted for, the return to familiar majors is 5.03% lower than that to unfamiliar majors.

In summary, columns (1)–(3) of Table 2 reveal that when we directly compare the observed average incomes, as a prospective college applicant might do, and when we conduct a kitchen sink regression that better controls for students' self-selection, familiar majors exhibit significantly lower labor market returns than unfamiliar majors.

In column (4), I additionally include the interaction term between major familiarity and the student's socioeconomic status. The result suggests that the gap in labor market returns between familiar and unfamiliar majors does not systematically differ between high- and low-SES students. Consequently, we can rule out the possibility that differential labor market returns across SES causes the observed disparity in major choice. The higher likelihood of low-SES students than of high-SES students to choose familiar majors is not caused by incentives tied to labor market returns.

3.2.2 Socioeconomic Status

Urban–Rural Status In the baseline analysis, I use applicants' urban–rural status as a measure for their socioeconomic backgrounds. Urban–rural status stems from the Chinese *hukou* system, also known as the household registration system. The system, first introduced in 1950s, classifies each citizen into either an agricultural or a non-agricultural *hukou*; these categories are commonly referred to as rural and urban. The urban–rural status is a high-quality measure of SES in the Chinese context for two reasons. First, because the *hukou* system is such a fundamental element of the social structure in China, the urban–rural indicator is available in all data sets used in this study. Therefore, using urban–rural status as the preferred measure of SES makes the baseline results easily comparable across different data sources.

Second, and more importantly, the urban–rural indicator alone can effectively capture a large fraction of the variation in students' family background because of the vast urban–rural differences in household income, wealth, and well-being in China (Park, 2008). Since its first introduction, the *hukou* system has been linked to social policy and benefits, giving urban residents access to better employment opportunities, education, healthcare, and social welfare (Cheng and Selden, 1994;

Young, 2013; Chan, 2015). Recent years have witnessed plans to alleviate the urban–rural disparity, but the inequality between urban and rural *hukou* holders persists. For example, in 2016, the average disposable income per capita of urban residents in Ningxia province is 176% higher than that of rural residents. For college students from the CCSS data, the average (median) household income of urban students is 83% (67%) higher than that of their rural counterparts.

Alternative Measures I employ alternative measures of SES, including local GDP per capita, household income, parental job sector, and parental education levels. Specifically, I use more granular measures of socioeconomic status, including county-level GDP per capita from the ROL data and household income from the CCSS data, to confirm the robustness of the baseline results. Additionally, for the survey experiment in which the subjects are from the same high school and for the analysis of information spillovers within the same high school class, alternative measures of SES are helpful because there is sometimes limited variation in the urban–rural status among students within the same high school. As a result, I use parental education level measured in the survey and parental job sector obtained from the administrative records as proxies for socioeconomic status for the analyses, leveraging variation beyond urban–rural status.

3.3 Summary Statistics

In Appendix A, I discuss how the samples are constructed in detail. Appendix Table F2 reports the summary statistics, including the sample average of the key variables for the full sample and separately for urban and rural students.

4 Descriptive Evidence of Disparity in Major Choice

I document the SES disparity in applications to and enrollment in familiar majors by providing graphical and regression evidence from the ROL and national data. The baseline results show that low-SES students, when compared to their high-SES peers with similar academic and demographic characteristics, are 3.16 percentage points (21.6% relative to the sample mean) more likely to apply to familiar majors even though they offer lower labor market returns.

4.1 Graphical Results

I first graphically present the socioeconomic disparity in applications to and enrollment in familiar majors. To offer a comprehensive perspective, I start with students' admission records from the

national data. Given that each student is admitted to only one college–major, the admission records can be regarded as equivalent to enrollment records.

In Figure 1(a), I plot the share of students enrolled in familiar majors against their NCEE performances with 95% confidence intervals separately for urban and rural students. I measure the exam performances by converting the NCEE scores to rankings within the province–year–academic track cell.[20] According to the admission mechanism, students with the same NCEE performances (from the same province–year–academic track cell) have the same priority scores and, therefore, are qualified for admission to the same set of college–major programs. However, Figure 1(a) shows very clearly that applicants from rural areas (of low SES) are much more likely to enroll in familiar majors than their urban counterparts with the same exam performances and thus the same feasible choice set, suggesting the existence of an SES disparity in major choice. The observed SES disparity is large and consistent for applicants across the entire score distribution, except for those in the top 5% of the distribution.

Table F3 shows large variation in the share of familiar majors across the ten different major categories. There are virtually no familiar majors in categories such as Engineering, Education, Agriculture, Economics, Healthcare, and Management, while the History category contains mostly familiar majors. A natural question is whether the SES disparity observed in Figure 1(a) is driven by urban and rural students enrolling in different major categories or whether the gap exists even within the same major category. To answer this question, Figures 1(b)–(d) plot the SES disparity in enrollment in familiar majors separately for each of the three major categories that contain non-negligible shares of both familiar and unfamiliar majors, i.e., Art & Humanities, Law & Politics, and Science. Although the shapes of the curves slightly differ, the gap between urban and rural students' share of enrollment in familiar majors remains substantial within each major category, which suggests that the SES disparity is not driven by urban and rural students enrolling in different major categories.

However, students' admissions to and enrollment in a major do not reflect only their own choice of major. At the application stage, a student lists multiple majors on the rank-order list, but the specific major to which she is admitted depends on factors such as other students' rank-order lists (demand) and the pre-allocated admission quotas (supply). Hence, to test whether the enrollment gap in Figure 1 comes from differences in major choices between low- and high-SES students, I use submitted rank-order list from the ROL data. As noted above, this data set is for students in Ningxia province who

[20] As discussed in Section 2.2, the NCEE scores are comparable only among students from the same province–year–academic track cell, and it is the relative ranking within the cell that matters for admission.

applied to four-year public colleges between 2014 and 2018. Using the ROL data, I separately plot for high- and low-SES students their applications to and enrollment in familiar majors against their exam performances. Specifically, in Figure 2(a), I plot the average familiarity of the students' top choice majors.[21] This graph confirms the existence of a disparity in major choice between urban and rural applicants. In Figure 2(b), I plot whether the student enrolls in a familiar major, the same variable as in Figure 1. Comparing Figures 2(b) and 1(a), we can observe very similar patterns, although Figure 2(b) is slightly noisier because of the small sample size. As robustness checks, I use alternative measures of SES in Appendix Figure F3 and observe similar disparities.

4.2 Regression Results

To control for additional sources of heterogeneity in choosing familiar majors, I conduct regressions using the ROL data. Specifically, I look at the familiarity of all majors listed at different positions of the ROLs, by running the following regression:

$$F_{pi} = \beta_0 + \beta_1 \text{SES}_i + \delta_p + X_i\eta + Z_{pi}\gamma + \varepsilon_{pi}, \tag{1}$$

for each major choice listed at position p of student i's ROL ($p = 1, 2, \cdots, 24$).[22] Therefore, the unit of observation for the regression is each position on the rank-order list of each student. The outcome F_{pi} is a measure of familiarity for the major listed at position p. For example, F_{1i} equals one if student i's top choice is a familiar major and zero otherwise. The key independent variable SES_i denotes the student's socioeconomic status proxied by her urban–rural status. I include the position fixed effect δ_p so that the choices of major are directly comparable. The covariates X_i include student characteristics and their NCEE performances, while the covariates Z_{pi} include the characteristics of the college–major pair listed in position p of student i's ROL. The standard error ε_{pi} is clustered at the student level.

Table 3 reports the results. Column (1) suggests that, for each position on the rank-order list, rural students are 3.84 percentage points more likely to list a familiar major than their urban counterparts. In column (2), I control for student-level characteristics, including year–academic track fixed effects and the student's gender, ethnic group (Han or minority), and eligible bonus points. I further incorporate the student's NCEE performance in column (3), which is converted to dummy

[21] Note that each student can list up to four colleges in one application round; thus, a student can have at most four first-choice majors in one application round.

[22] There are 24 positions on the ROL for each student to fill out in one application round. An average student fills out 18 out of the 24 positions. Students who are not admitted by Tier 1 colleges and apply to Tier 2 colleges go through two application rounds with 48 positions to fill out in total.

variables for each score percentile. Column (4) additionally controls for college–major characteristics, including the tuition level and a selectivity index measured by the admission cutoff of the college–major from the previous year. The estimate in column (4) suggests that, after adjusting for students' demographic characteristics and exam performances, tuition, and selectivity, rural students are still 3.16 percentage points more likely to choose familiar majors than urban students.[23] This translates to a 21.6% disparity in major choice between urban and rural students, relative to the sample mean of 14.6 percentage points. This is noteworthy considering that familiar majors on average yield lower labor market returns. In column (5), I restrict the sample to only each student's top choice. The estimate suggests that the disparity in major choice becomes even more pronounced when I focus specifically on the most preferred choice.

Robustness The results are robust to using alternative samples, alternative measures of major familiarity and socioeconomic status—Appendix B provides details.

5 Information Inequality Channel

In this section, I show that information inequality by socioeconomic status is an important channel for the observed disparity in major choice. I first present results from a survey experiment that I designed to measure students' information about different majors. The results confirm the presence of information inequality across socioeconomic backgrounds. Then, I present evidence on within-classroom information spillovers, which indicates that obtaining information from social networks increases students' likelihood of choosing unfamiliar majors.

5.1 Survey Evidence

To explore the reasons behind the observed disparity in major choice, I design a survey experiment for a sample of senior students from a Chinese high school. The survey experiment directly measures students' information on familiar and unfamiliar majors and provides a simplified information treatment.[24] While the experiment is run on a small sample, it reveals two key findings: 1) students have more information about familiar majors than about unfamiliar majors, and 2) the information gap between familiar and unfamiliar majors is larger for low-SES students.

[23] To benchmark the magnitude, I compare it with a more widely recognized disparity in major choice: the gender gap in STEM enrollment. Using data from Canada (Card and Payne, 2021) and Ireland (Delaney and Devereux, 2019), the literature finds a 4–5 percentage point gender gap in choosing STEM majors among STEM-ready students.

[24] Since the students were instructed to complete the survey online at home using either their phones or computers, spillovers of the information treatment are not a major concern.

5.1.1 Survey Design

The survey experiment was conducted in February 2022 in a sample of senior high school students from a Chinese high school located in Chongqing, China. The response rate was 30% (677 students completed the survey out of 2,256 enrolled senior students). Appendix C contains details about the experiment design. Importantly, I collaborated with the high school administration to link students' survey responses to their administrative records, which contain students' basic information such as gender, academic track, and their academic performance in the high school entrance examination.

In the first part of the survey, students answer baseline questions about their demographic characteristics, including their urban–rural status and their parents' highest education level, as well as questions that assess their attitudes and strategies toward college applications and major choices in general. As the entire sample consists of students from a single high school and the high school is located in a downtown area of a major city, a majority of the students are urban residents. Therefore, in the following analysis of the survey results, I use parental education level as a proxy for SES. Specifically, a student is classified as being of high socioeconomic status if at least one of her parents has a bachelor's degree (or higher).

The second part of the survey measures students' information on college majors. Each student is assigned a pair of majors (X,Y) and is expected to answer questions about the two majors. I choose major pairs (X,Y) such that X is a familiar major (according to the broad definition) and Y is an unfamiliar major. I ensure that both majors X and Y are closely related to high school subject S. Specifically, I consider the following three major pairs:

	College Major X	College Major Y	High School Subject S
1)	Mathematics	Statistics	Mathematics
2)	Applied Physics	Automation	Physics
3)	Chinese Language	Editing and Publishing	Chinese Language

To make the survey questions more relevant, I assign major pairs to students based on their academic tracks. Science-track students are randomly assigned to one of the first two pairs of majors, while all humanities-track students are assigned the third pair. Given the assigned major pair, students first answer a set of questions about one of the majors. The questions are about the students' interest in and aptitude for that major, and are subsequently repeated for the other major in the pair. The order in which the two majors are presented (X versus Y) is randomized.

To better identify the role of information inequality, I implement a simple experiment with my

survey sample. In particular, I assign students to either a treatment or a control group, differing in how the two majors are introduced. For students randomly assigned to the control group, the prompt that they see on the screen is: "The following questions are about college major X (Y)." For the students assigned to the treatment group, however, the message is: "The following questions are about college major X (Y). College major X (Y) is closely related to high school subject S." The additional sentence provided to the treated students seems redundant for familiar majors X. For example, it might say, "The following questions are about the college major Mathematics. The college major Mathematics is closely related to the high school subject Mathematics." For unfamiliar majors Y, instead, it provides students with a simple information treatment that points out the relationship between the unfamiliar major and a related high school subject and is intended to familiarize the students with unfamiliar majors. For example, the prompt could be: "The college major Statistics is closely related to the high school subject Mathematics."

While the information intervention used in this survey experiment is simplified, which may complicate generalization to other unfamiliar majors that lack a direct connection to high school subjects, it is important to highlight that the treatment directly captures the concept of major familiarity. By establishing connections between unfamiliar majors and high school subjects, the intervention is particularly suitable for this study because it is able to offer valuable insight into the mechanism underlying the observed disparity in major choice.

5.1.2 Measure of Information

To measure students' information on the assigned major pair, I focus on the following four questions.

1. Do you think you will have a strong interest in major X (Y)? [Yes or No]
2. How confident are you about your answer to the previous question? [on a scale of 1 to 5].
3. Do you think you will do well if you study major X (Y)? [Yes or No]
4. How confident are you about your answer to the previous question? [on a scale of 1 to 5].

Questions 1 and 3 ask about the student's self-perceived interests and abilities for majors X and Y. These are the typical questions that students would ask themselves when evaluating their match with the major. How well informed the students are about the majors is reflected in both the bias of their responses and the degree of uncertainty associated with those responses. As the researcher, I am unable to directly observe their true fit with the majors, which complicates the quantification of bias in their answers. Therefore, instead of analyzing students' answers to the two questions per se, I use how certain they are about their answers to measure their information. A student with

richer information about a major will be more confident in accurately assessing whether the major fits her interests and abilities. Hence, after questions 1 and 3, I ask about students' confidence in their answers (questions 2 and 4), as a measure of students' information on majors X and Y. Due to variations in students' personalities and their individual interpretations of the survey question scales, comparing self-reported confidence levels across individuals is challenging. So my analysis focuses on comparing the self-reported confidence *within* individual between the familiar and unfamiliar majors.

5.1.3 Evidence of Information Inequality

I start by comparing students' information on familiar and unfamiliar majors in the absence of the information treatment. Specially, I focus on students in the control group and explore the within-individual difference in their reported confidence levels between the familiar and unfamiliar majors.

In the top panel of Figure 3, separately for low- and high-SES students in the control group, I plot the shares of students who report being more confident in accurately assessing their interests about the familiar major in the assigned pair, being equally confident about the assessments of both majors, and being more confident about the assessment of the unfamiliar major. Similarly, in the second panel of Figure 3, I plot students' confidence levels for their assessments of ability.

First, Figure 3 shows that students are more informed about familiar majors than unfamiliar majors, as there are more students who are confident about their assessment of familiar majors than unfamiliar majors. For example, when we look at students' answers to question 2, 42% of low-SES students respond that they are more confident about their assessment of the familiar major, and 37% of students report equal confidence levels, while only 21% of students report higher confidence level for assessing the unfamiliar major. I use the discrepancy between 42% and 21% to measure the extent of the information gap, and this difference is significantly larger than zero. It shows that students have more information about familiar majors than about unfamiliar majors, consistent with the definition of major familiarity. In Appendix Figure F4, I validate that the information discrepancy between familiar and unfamiliar majors originates from unequal exposure during high school, which again is consistent with how major familiarity is defined.[25]

Second, when comparing the responses of high- and low-SES students, Figure 3 illustrates that the information gap between familiar and unfamiliar majors is more pronounced among low-SES stu-

[25] In particular, I ask students whether they have heard about majors X and Y through four different channels: family members, media and the Internet, high school peers, and high school teachers. The findings indicate that students are equally likely to learn about familiar and unfamiliar majors through family and media. However, they are substantially more likely to gain knowledge about familiar majors than about unfamiliar majors from high school teachers and classmates.

dents. For high-SES students, the gap is smaller and no longer statistically significant. Specifically, the shares of high-SES students who are more confident about unfamiliar majors are 25% and 26% in their assessment of interests and abilities, respectively, which are both larger than the corresponding shares for low-SES students (21% and 15%). This shows that high-SES students are more informed than their low-SES peers about unfamiliar majors. Intuitively, it is because learning about unfamiliar majors requires information sources beyond high school classes, such as social networks, career counselors, or online platforms. These resources are more accessible for high-SES students.

5.1.4 Effects of Information Treatment

In Figure 4, I plot the effect of the information treatment on students' reported confidence levels. Instead of splitting the control group into low- and high-SES students as in Figure 3, I include the full sample and split it by control and treatment groups. Figure 4 shows that a higher share of the treated students is more confident about their assessment of unfamiliar majors than the share of those in the control group. The treatment leads to a reduced information gap between familiar and unfamiliar majors, which suggests that mentioning the most relevant high school subjects helps familiarize students with unfamiliar majors.

To quantify the treatment effect, I adopt a difference-in-differences design and run the following regression:

$$Y_{ij} = \beta_0 + \beta_1 \mathbb{1}\{F_j = 0\} + \beta_2 \mathbb{1}\{\text{treat}_i = 1\} + \delta \mathbb{1}\{F_j = 0\} \times \mathbb{1}\{\text{treat}_i = 1\} + X_i\eta + \varepsilon_{ij}, \qquad (2)$$

where the outcome variable Y_{ij} is whether the student i is confident in accurately assessing major j. $\mathbb{1}\{F_j = 0\}$ and $\mathbb{1}\{\text{treat}_i = 1\}$ are indicators for whether major j is an unfamiliar major and whether student i is assigned to the treatment group. δ is the coefficient of interest, which captures the treatment effect of mentioning the most relevant high school subject on students' confidence levels when they evaluate unfamiliar relative to familiar majors. Other control variables X_i include gender, urban–rural status, parental education, academic track, and high school entrance exam performances of student i. In addition, I include a dummy variable for whether the student answers questions about the familiar major first within the major pair (X,Y) and the major pair (X,Y) fixed effects. The standard error ε_{ij} is clustered at the student level.

Table 4 reports the estimated information treatment effects. In columns (1) and (4), I present the regression results using the full sample, separately for the assessments of interests and abilities. The negative coefficient on $\mathbb{1}\{F_j = 0\}$ confirms that students are less informed about unfamiliar

than about familiar majors. The coefficient on the interaction term indicates strong and significant treatment effects. Treated students are more confident than students in the control group in accurately assessing whether they will have an interest in or aptitude for the unfamiliar major. I also explore heterogeneous treatment effects by student SES. In columns (2), (3), (5), and (6), I split the sample by whether the student is of high or low socioeconomic status as measured by their parental education levels. The coefficients reveal that the treatment effect is concentrated among students from low-SES backgrounds, indicating that the information treatment is particularly beneficial for them. This finding further supports the assumption that low-SES students are more affected by information frictions.

5.2 Evidence on Information Spillovers

Next, using observational data, I examine how obtaining information affects students' choice of major. To do this, I analyze students' major choices in response to information spillovers from their high-SES peers, using the ROL data from 2017 and 2018. These data enable me to identify each student's high school classmates—those peers whom they interact with daily.[26] In principle, most high schools randomly assign students to classes within the cohort–by–academic track.[27] However, it is possible that some high schools sort the students into different classes based on their characteristics. To address this concern, I first exclude cohorts for which I observe large variation in class average exam performances because these cohorts are highly likely to be the ones that sort students into classes based on characteristics such as academic abilities. In later discussion, I will provide additional evidence in favor of random sorting of students into high school classes.

Empirically, I leverage the (possible) random assignment of students to different high school classes within the cohort–by–academic track and examine how a student's major choice is affected by her peers' socioeconomic backgrounds. In the presence of information spillovers, I expect that low-SES students obtain information from their high-SES peers and thus become more likely to apply to unfamiliar majors. Because I am comparing students in the same class but from different socioeconomic backgrounds, I use parental job sector as a proxy for SES instead of the student's urban–rural status. A student is of high (low) SES if her parents work in the formal (informal) sector. The formal sector comprises institutions such as government agencies, educational institutions,

[26] Unlike many high schools in the U.S., in the Chinese context of this study, high school classmates study in the same classroom and attend all classes together throughout their high school years, enabling them to forge a stronger bond with one another. The information on students' high school classes is available only for 2017 and 2018.

[27] A cohort is defined as the group of students entering the same high school in the same year, and a cohort–by–academic track is the group of students who choose the same academic track in the cohort.

healthcare facilities, and corporations.

I estimate the following reduced-form model for spillover effects (Manski, 1995):

$$Y_{ich} = \alpha + \beta_1 z_{ich} + \beta_2 \bar{z}_{-ich} + X_{ich}\Gamma_1 + \bar{X}_{-ich}\Gamma_2 + \psi_h + \varepsilon_{ich}, \tag{3}$$

where the outcome variable Y_{ich} is the average familiarity of the first-choice majors on the ROL of student i from class c in cohort–by–academic track h. The independent variable z_{ich} denotes whether the parents of student i work in the formal sector, and $\bar{z}_{-ich}$ denotes the share of student i's classmates whose parents work in the formal sector. Individual-level characteristics X_{ich} include gender, ethnicity, urban–rural status, whether the student is awarded bonus points, and exam performances (dummies for each score percentile). $\bar{X}_{-ich}$ represents the leave-one-out class-level average of these characteristics, excluding student i. ψ_h is cohort–by–academic track fixed effects. The standard error ε_{ich} is clustered at the cohort–by–academic track level.

Column (1) of Table 5 presents the results. The estimated coefficient β_1 confirms the main finding that high-SES students are less likely to apply to familiar majors. The estimate of coefficient β_2 is negative, suggesting that the probability of choosing familiar (unfamiliar) majors decreases (increases) when the student happens to have a larger share of high-SES classmates, consistent with my prediction of information spillovers. The comparison between the magnitudes of β_1 and β_2 indicates that an increase in a student's share of high-SES classmates by two standard deviations has an impact on major choices similar to the effect of her parents working in the formal rather than the informal sector.

In column (2), I split the student's classmates into those of the same gender as the focal student and those of the opposite gender. In column (3), I split the classmates into those of the same ethnic group as the focal student and those of a different ethnic group. The results in both columns indicate that the student's choice of major are mostly affected by the share of high-SES classmates within the same demographic group and less affected by the socioeconomic backgrounds of classmates from different demographic groups. This finding provides further support for information spillovers because we expect students belonging to the same gender or ethnic group to interact with each other more frequently, leading to increased sharing of information.

However, the coefficient β_2 in the reduced-form regression (3) captures the overall spillover effect, which includes both the contextual effect and the endogenous effect (Manski, 1995). To argue that low-SES students with more high-SES classmates choose unfamiliar majors because they obtain information from their peers (the contextual effect), we must rule out the endogenous effect. That

is, we need to ensure that the increased likelihood of choosing an unfamiliar major is *not* driven by conformity or the utility gains from joint consumption. To verify these alternative explanations, I consider students' choices of college locations as placebo outcomes. If students choose unfamiliar majors to align with their friends, we should expect to see similar behaviors in their choices of college locations. In columns (4)–(6) and columns (7)–(9) of Table 5, I explore spillover effects in students' applying to in-province colleges and colleges located along the east coast. Because Ningxia province is one of the low-income areas in China while the east coast is the most developed region, low-SES students are more likely to apply to in-province colleges and less likely to consider moving to the east coast, as is confirmed by the estimated β_1s. However, we do not see any spillover effects from columns (4)–(9), suggesting that students are not conforming to the location choices of their high-SES classmates. Consequently, it is reasonable to deduce that the students are similarly not following the major choices of their high-SES classmates as an outcome of conformity, which serves as informal evidence against the endogenous effects and in favor of the contextual effects.

At the same time, students' choice of college location can be used as a placebo outcome to further address the concern of non-random sorting into high school classes. If non-random sorting is indeed behind the observed spillover effects in major choices, it should have similar effects on college choices. The fact that we do not see spillovers in college location choices reassures us that the documented spillovers are not caused by systematic sorting into high school classes on the basis of academic abilities or other characteristics.

Another alternative explanation for the documented spillovers is that there are class-specific unobservable factors related to major choices. For example, classes with more parents working in the formal sector might be more likely to be assigned better teachers, who could provide students with more information about college majors. However, this "common shock" story cannot account for the observation that the spillover effects are stronger among students within the same demographic group than among students in the same class but belonging to different demographic groups. Therefore, it is unlikely to be the primary cause of the observed information spillovers.

In summary, I find that students with more high-SES classmates are more likely to apply to unfamiliar majors and that the spillover effects are stronger if the high-SES classmates are in the same demographic group as the focal student. This finding suggests that obtaining information from social networks can effectively increase students' likelihood of choosing unfamiliar majors and further supports the hypothesis that low-SES students' avoidance of unfamiliar majors is caused by imperfect information.

6 Alternative Channels

I argue that other factors, including differences in application strategies, risk attitudes, and preferences, are unlikely to be the main drivers of the documented SES disparity in major choice. In this section, I briefly outline the empirical evidence. Appendix D provides a more detailed discussion.

Different Application Strategies High-SES students often have more sophisticated application strategies. If unfamiliar majors have less competition and lower admission thresholds than familiar ones, they are more likely to attract these strategically sophisticated high-SES students. To empirically evaluate this alternative channel, I compare the competitiveness of familiar and unfamiliar majors using the ROL data. The results reported in Appendix Table F10 suggest no significant difference in competitiveness between familiar and unfamiliar majors. Thus, a difference in application strategies is not the main factor behind the observed disparity in major choice.

Different Risk Attitudes Risk attitudes can be an important factor for major choices. To investigate whether students from different socioeconomic backgrounds exhibit different risk attitudes toward the choice of major, I asked three risk-related questions in the survey experiment. The results, presented in Appendix Table F11, indicate that risk attitudes do not significantly differ between high- and low-SES students.

Different Preferences High- and low-SES students may differ in their preferences for familiar majors. To test this hypothesis, I use the CCSS data, which provide information on students' post-college plans. I find, in Appendix Table F12, that low-SES students studying familiar majors in college are 13.5 percentage points more likely to switch to a different major in graduate school, a pattern not observed among high-SES students. This suggests that low-SES students might regret their choice of familiar majors, rather than having an inherent preference for them.

7 Economic Consequences of Disparities in Major Choice

I build a major choice model to analyze the economic consequences of information inequality, including the disparity in match quality and future incomes between high- and low-SES students.

7.1 Model Setup

7.1.1 Simplifying Assumptions

As discussed in Section 2, the real-world college application system requires students to choose colleges and majors jointly. To focus on major choices, I assume, in the model, that students' choice of major is independent of their college choices. This simplification is justified by the empirical observation that most students list very similar majors across different colleges on their rank-order lists. Specifically, Appendix Figure F5 shows that more than 60% of students in the estimation sample apply to similar first-choice majors for at least half of the colleges listed on their rank-order lists. By imposing this assumption, I can focus exclusively on modeling students' major choices without needing to consider their choice of college.

In addition, I simplify the choice set. Instead of modeling how students fill out the entire ROL, the model focuses on each student's first-choice major at her first-choice college. Moreover, the choice set is restricted to three major categories that contain non-negligible shares of both familiar and unfamiliar majors—Art & Humanities, Law & Politics, and Science. Furthermore, I treat all familiar (unfamiliar) majors within the same category as a single major for simplification. As a result, there are a total of six different majors (two from each of the three categories) for each student to choose from. The simplification keeps the model tractable while still capturing the essence of students' major choice decisions.

7.1.2 Estimation Sample

The estimation sample comprises $5,801$ students who applied to public colleges in Ningxia province in 2018. Following the simplifying assumptions, the major choice that I model is students' decision on the first-choice major of their first-choice college, and the sample is restricted to the students who choose majors from one of the three major categories. For each student in the sample, the model incorporates demographic variables, including gender and urban–rural status as a proxy for SES, as well as academic information including high school academic track and NCEE scores.

7.2 Students' Major Choices

Within the model, students choose their major with the objective of maximizing their expected utility, which depends primarily on future income.

7.2.1 Income

At the time of major choices, students predict the future incomes associated with graduating with each respective major. I use y_{im} to denote the monthly income of student i if she studies major m in college, and I model log income $\log(y_{im})$ as the sum of a major-specific component r_m and an idiosyncratic component v_{im}. The latter is further decomposed into the return to idiosyncratic abilities $g_m(s_i)$, where s_i denotes the ability vector of student i, and the idiosyncratic error e_{im}, which is unknown to the students when they apply:

$$\log(y_{im}) = r_m + v_{im} = r_m + g_m(s_i) + e_{im}. \tag{4}$$

Return to Major The major-specific component r_m captures the return to the major. It differs from the empirically observed average income, as it adjusts for selection on abilities. Following the literature (e.g., Arcidiacono, 2004; Bordon and Fu, 2015), the model assumes that r_m is known to the students but is unknown to the researcher and needs to be estimated.[28]

Return to Abilities Each student i has her idiosyncratic ability vector s_i measured by the NCEE performances. The ability vector $s_i = [s_i^a, s_i^v]$ consists of two dimensions (Kinsler and Pavan, 2015): analytical ability s_i^a, which is approximated by students' performance in the Mathematics subject, and verbal ability s_i^v, which is approximated by students' performance in Chinese Language and English. The model assumes linear returns to the abilities, where the returns ω_m^a and ω_m^v are major-specific: $g_m(s_i) = \omega_m^a s_i^a + \omega_m^v s_i^v$. This allows students to enjoy higher returns if they choose a major that fits their idiosyncratic abilities.

Idiosyncratic Error The idiosyncratic error e_{im} is the discrepancy between student i's predicted income for major m and the actual income that she will achieve. The error term can be attributed to factors such as unobserved abilities and interests, inaccurate information, or pure luck. I assume that the idiosyncratic error e_{im} follows a normal distribution with a mean of zero, and is independently distributed among students and within each student across majors.[29] The variance of e_{im} captures the level of uncertainty that students face when predicting their income $\log(y_{im})$ given the ability vector s_i. I allow the variance to differ between familiar and unfamiliar majors and between high-

[28] According to the empirical findings in column (4) of Table 2, there is no systematic difference in labor market returns between familiar and unfamiliar majors across different socioeconomic backgrounds. Therefore, I do not assume differential returns between high- and low-SES students.

[29] In the model, similar majors are grouped into one major category and treated as one choice, which makes the assumption that the idiosyncratic error e_{im} is independent across majors weaker.

and low-SES students by imposing the following structure:

$$var[e_{im}] = \begin{cases} \sigma_F^2 & \text{major } m \text{ is a familiar major;} \\ \sigma_{Uh}^2 & \text{major } m \text{ is an unfamiliar major and student } i \text{ is from a high-SES background;} \\ \sigma_{Ul}^2 & \text{major } m \text{ is an unfamiliar major and student } i \text{ is from a low-SES background.} \end{cases}$$

Under the hypothesis that students have more information about familiar majors than about unfamiliar majors due to their high school experience and that low-SES students are more uninformed than their high-SES counterparts, we expect $\sigma_F^2 < \sigma_{Uh}^2 < \sigma_{Ul}^2$.[30]

7.2.2 Utility

Utility from Lifetime Income I assume that student i's lifetime income Y_{im} is a function of her monthly income y_{im}, $\log(Y_{im}) = \beta_1 \log(y_{im}) + \beta_0$, and that the student's utility from lifetime income is CRRA

$$U_Y(Y_{im}) = \frac{Y_{im}^{1-\gamma} - 1}{1-\gamma},$$

where I set the relative risk coefficient $\gamma = 1.5$ based on Blundell et al. (1994) and Attanasio and Weber (1995). Since the idiosyncratic error e_{im} is unknown, students choose their major to maximize the ex ante expected utility. Given that $\log(Y_{im}) \sim N(\mu_{im}, \sigma_{im}^2)$, where $\mu_{im} = \beta_1(r_m + g_m(s_i)) + \beta_0$ and $\sigma_{im}^2 = \beta_1^2 var[e_{im}]$, expected utility can be calculated analytically:

$$E[U_Y(Y_{im})] = \frac{e^{\mu_{im}(1-\gamma) + \frac{1}{2}\sigma_{im}^2(1-\gamma)^2}}{1-\gamma}.$$

Overall Expected Utility The model assumes that the overall expected utility of student i choosing major m takes a quasi-linear form:

$$E[U_{im}] = E[U_Y(Y_{im})] + \alpha_1 E[\mathbb{1}\{Admitted_{im}\}] + \alpha_2 \log(Quota_{im}) + \eta_{c(m)}[1, gender_i] + \varepsilon_{im}. \quad (5)$$

In addition to future incomes Y_{im}, the overall utility depends on whether the student is admitted, represented by $\mathbb{1}\{Admitted_{im}\}$. Since students are unaware of their admission results at the time of application, they base their utility on the perceived admission probabilities $E[\mathbb{1}\{Admitted_{im}\}]$. The expected utility also incorporates the admission quota of the major and major category $c(m)$–specific preferences $\eta_{c(m)}$, which interact with student i's characteristics $[1, gender_i]$. The former captures the

[30] Note that $var[e_{im}]$ being larger for low-SES students and for unfamiliar majors does not necessarily mean $var[\log(y_{im})]$ is larger. $var[e_{im}]$ represents the students' uncertainty about the residualized return conditional on major return r_m, ability s_i, and return to skills $g_m(\cdot)$, while $var[\log(y_{im})]$ is the unconditional variance of log income.

prevalence of the major, and the latter allows male and female students to have different preferences for major categories. ε_{im} denotes student i's idiosyncratic taste for major m and is assumed to follow a type 1 extreme value distribution.

Maximization Problem Student i chooses one major $m^*(i)$ from the choice set to maximize the expected utility $E[U_{im}]$. I assume that students are rational agents, holding rational expectations about their future incomes, admission probabilities, and utilities given the information that they have.

$$m^*(i) = \arg\max_{m \in \mathrm{M}} E[U_{im}].$$

7.3 Admission Process

The model primarily focuses on students' major choices, so I simplify the admission process. I model a student's probability of being admitted as a function of the major to which she applies and her academic track, as the admission quota varies by major and by academic track. This implies that, conditional on the chosen major and academic track, the student's admission probabilities are independent of other characteristics, including exam performances. Imposing this assumption significantly simplifies both the parameter structure and the simulation of the admission process while still representing a reasonable approximation of reality. Empirically, there is no strong correlation between students' exam performances and their admission probabilities. This is primarily due to the wide range of college choices available, which allows students to choose colleges with admission probabilities that closely align with their preferred risk levels, provided that the student falls within a range that is not at the extreme ends of the score distribution. In other words, having higher scores does not lead to a higher likelihood of being admitted because students with higher scores tend to aim higher and choose more selective options. Using the ROL data, I indeed observe limited variation in the probability of admission to the first-choice major of the first-choice colleges across the distribution of students' scores, except for the students in the top and bottom ventiles, as is shown in Appendix Figure F6.

7.4 Estimation

Using the method of simulated moments, I estimate the return to majors $\{r_m\}$, parameters $\boldsymbol{\theta} = \{\beta_0, \beta_1, \alpha_0, \alpha_1, \omega_m^a, \omega_m^v, \eta_{c(m)}, \sigma_F^2, \sigma_{Uh}^2, \sigma_{Ul}^2\}$, and the admission probability $E[\mathbb{1}\{Admitted_{im}\}]$ for each major m and academic track of student i.

The estimation proceeds in two nested steps. In the first step, I assume that the return to majors

$\{r_m\}$ is known and estimate the remaining parameters by fitting a multinomial logistic model, such that the log likelihood of students choosing their observed major choice is maximized. In the second step, I use the method of simulated moments to estimate the return to majors $\{r_m\}$. This involves matching the simulated average income of students admitted to each major with the empirically observed average income of each major. Given $\{r_m\}$, the simulated average income is calculated based on the parameters estimated in the first step. The observed average income is calculated from major-specific average income data from the Chinese Four-Year College Graduates Employment Annual Report (2020). Appendix E provides a more detailed description of the estimation process.

7.5 Estimation Results

Information Inequality The information inequality between high- and low-SES students is captured by the level of uncertainty that they experience when predicting their future incomes using observed abilities, specifically, parameters σ_F^2, σ_{Uh}^2, and σ_{Ul}^2. To ease interpretation, I convert the estimated parameters to an unfamiliarity penalty $p_x\% = -\frac{1-\gamma}{2}(\sigma_{Ux}^2 - \sigma_F^2)$ separately for high- and low-SES students, $x \in \{h, l\}$.[31] The unfamiliarity penalty captures the idea that, for a student to be indifferent between an unfamiliar and a familiar major, the expected income Y_{im} of the unfamiliar major must be $p\%$ higher to offset the uncertainty caused by the limited information, assuming that all other factors are equal. A higher unfamiliarity penalty means the student is less informed about unfamiliar majors. The estimated parameters suggest that the unfamiliarity penalty for the low-SES students is 11.05% higher (SE = 4.24%) than that for their high-SES peers, indicating that their information about unfamiliar majors is significantly more limited. The result confirms the existence of information inequality, and is consistent with the empirical finding that low-SES students are less likely to choose unfamiliar majors than their high-SES counterparts.

Return to Major The return to majors $\{r_m\}$ is an important set of parameters for discussing economic consequences. In the left panel of Figure 5, I plot the empirically observed average log income $\overline{\log(y_m)}$ for each major m from the three major categories. The graph shows that students with Law & Politics or Science degrees on average have higher incomes than students with Art &

[31] To calculate the uncertainty penalty, I equalize the expected utility from lifetime income for an unfamiliar major ($E[U_Y(Y_{Ux})]$) with that for a familiar major $E[U_Y(Y_F)]$, assuming that the other components of the overall utility function are the same between the two majors, separately for high- and low-SES students $x \in \{h, l\}$:

$$\frac{e^{\mu_{Ux}(1-\gamma)+\frac{1}{2}\sigma_{Ux}^2(1-\gamma)^2}}{1-\gamma} = \frac{e^{\mu_F(1-\gamma)+\frac{1}{2}\sigma_F^2(1-\gamma)^2}}{1-\gamma} \iff E[\log(Y_{Ux}) - \log(Y_F)] = \mu_{Ux} - \mu_F = -\frac{1-\gamma}{2}(\sigma_{Ux}^2 - \sigma_F^2)$$

Humanities degrees and, more relevantly to this study, that students who study unfamiliar majors have higher incomes than those who study familiar majors within the same category. However, the extent to which the differences in average incomes can be attributed to different returns or selection effects is not obvious.

Therefore, in the right panel of Figure 5, I plot the model-estimated return to majors, i.e., r_m, that adjusts for students' selection based on the observed analytical and verbal abilities. Although the magnitude of the differences diminishes when I account for selection effects, the estimates still support the finding that familiar majors yield lower labor market returns than unfamiliar majors. It is also reassuring that the magnitude is very similar to that of the coefficients in Table 2, where I use observational data from CCSS to estimate the difference in labor market returns between familiar and unfamiliar majors.

Inequality Implications The documented information inequality and differential returns by major familiarity lead to inequality across socioeconomic backgrounds along the dimensions of match quality and future incomes.

First, because low-SES students are less informed, they are more likely to be mismatched. A student is considered mismatched when the major she enrolls in is different from the major she would choose, assuming that all majors are considered "familiar" to her, i.e., $\sigma^2_{Ul} = \sigma^2_{Uh} = \sigma^2_F$. In other words, the benchmark for assessing mismatch is a hypothetical scenario where all students, regardless of their socioeconomic backgrounds, have access to sufficient information about every major during their high school education, such that there is no need to distinguish between familiar and unfamiliar majors. According to the model estimates, 37.9% of the high-SES students are classified as mismatched due to their lack of information about unfamiliar majors ($\hat{\sigma}^2_{Uh} > \hat{\sigma}^2_F$). The corresponding figure for low-SES students is significantly higher at 49.6%, because their access to information about unfamiliar majors is even more limited than that of the high-SES students ($\hat{\sigma}^2_{Ul} > \hat{\sigma}^2_{Uh} > \hat{\sigma}^2_F$). In other words, information inequality leads to the gap in match quality, as evidenced by a higher share of low-SES students choosing a familiar major than in the full information scenario.

Second, there will be a disparity in future incomes between high- and low-SES students because familiar majors, which are more likely to be chosen by low-SES students, yield lower labor market returns. Based on the model estimates and after adjusting for gender and academic track, the average future income of low-SES students is 1.4% lower than that of high-SES students. To provide an interpretation of the magnitude, we can consider that approximately 20% of low-SES students experience labor market returns that are around 7% lower than those of high-SES students as a

result of choosing familiar majors suboptimally. This result indicates that higher education does not promote intergenerational mobility to the extent anticipated in an ideal scenario where all students, regardless of their socioeconomic backgrounds, have equal access to information.

7.6 Counterfactual Policy Experiments

I investigate counterfactual policy experiments, including information interventions targeting low-SES students and affirmative action admission policies, as potential methods for mitigating inequality. The counterfactual analyses model changes in students' application decisions and admission probabilities in response to the policy experiments, while abstracting from general equilibrium effects by assuming that the labor market returns for the majors remain constant. The simplification is reasonable as long as the binding admission quotas are not relaxed—the number of students graduating from each major remains unchanged despite the counterfactual policies.

7.6.1 Information Interventions

I consider a hypothetical information intervention that targets the low-SES students and can reduce the information inequality between low- and high-SES students about unfamiliar majors. Utilizing the model, I analyze the impact of this intervention on students' choice of major and their future incomes. The results are presented in Figure 6, with the x-axis representing the percentage of information inequality closed by the intervention. The scale ranges from 0, indicating the status quo, to 100, indicating a complete closure of the information gap, corresponding to the case where σ^2_{Ul} is set equal to σ^2_{Uh}.

In Figure 6(a), I plot the impact of information interventions on the average log income separately for the treated low-SES students and the untreated high-SES students. Under the information intervention, more low-SES students choose unfamiliar majors, which have higher labor market returns. Meanwhile, due to the competitive nature of college applications, the intervention has negative crowding-out effects on the high-SES students. The increased competition reduces their likelihood of being admitted to high-return unfamiliar majors, thereby decreasing their future incomes. When the information inequality is fully closed, the income gap between high- and low-SES students can be reduced by 70.0%. This highlights the economic consequences of information inequality. Of the two factors (choice of major and abilities) modeled in the income equation (4), differences in major choice caused by information inequality explain 70% of the income gap between high- and low-SES students. Differences in abilities explain only the remaining 30%. In Figure 6(b), I show how infor-

mation interventions affect the match quality. Low-SES students are less likely to be mismatched when they become better informed about unfamiliar majors, while high-SES students experience a moderate increase in mismatch rates because of the crowding-out effects.

7.6.2 Affirmative Action Admission Policy

Despite the effectiveness, information interventions can sometimes be hard to implement. Therefore, I consider affirmative action admission as an alternative policy experiment. This approach has very recently been adopted by the Chinese Ministry of Education with the aim of incentivizing low-SES students to apply for certain majors. Specifically, I analyze a counterfactual scenario where a designated portion of the admission quota for unfamiliar majors is allocated exclusively for the admission of low-SES students. A hypothetical example could be the government announcing that 10% of the admission quota for unfamiliar majors is allocated exclusively to low-SES students. In this scenario, all students would compete for the 90% of the quota seats, while the remaining 10% would be reserved specifically for low-SES students who apply but are not accepted.

The affirmative action policy increases the perceived (and actual) probability of admission to unfamiliar majors for the low-SES students, consequently providing them with a stronger incentive to apply for these majors. Further, the policy ensures that a greater number of low-SES students who apply to unfamiliar majors can secure admission.[32] This is consistent with empirical findings from the U.S. indicating that affirmative action encourages qualified underrepresented minority students to apply to selective colleges (e.g., Bleemer, 2022) and increases their enrollment (e.g., Backes, 2012; Hinrichs, 2012).

As more low-SES students study unfamiliar majors in college, it increases their labor market returns while generating negative crowding-out effects on their high-SES counterparts. This is demonstrated in Figure 7, where I plot separately for low- and high-SES students how their future incomes and match quality change in response to the designated percentage of the admission quota. Figure 7(a) suggests that the income disparity across socioeconomic backgrounds can be fully closed when approximately 18.9% of the admission quota for unfamiliar majors is reserved exclusively for low-SES students.[33] Alternatively, if the objective is to compensate for the income gap created by information

[32] For example, the model predicts that when 10% of the admission quota for unfamiliar majors is reserved for low-SES students, the share of low-SES students applying for unfamiliar majors increases by 3.2% (from 18.4 pp to 19.0 pp), and the probability of admission, conditional on a student's applying to unfamiliar majors, increases by 11.4% (from 62.0 pp to 69.9 pp).

[33] This analysis compares the average income, conditional on admission to four-year public colleges, across SES. The required percentage is smaller if we consider the average income unconditional on admission, because the affirmative action policy increases the admission likelihood for low-SES students while decreasing it for high-SES students.

inequality, specifically to close the income disparity by 70.0%, we can designate 14.8% of the admission quota to low-SES students, as is shown by the dashed line in sub-figure (a). However, Figure 7(b) indicates that the affirmative action plan has limited impact on improving low-SES students' match quality. While it increases the likelihood of low-SES students studying *an* unfamiliar major, it does not necessarily ensure that they are matched with *the* unfamiliar major that they would have chosen if provided sufficient information. Meanwhile, high-SES students face large crowding-out effects, as a larger share of them are compelled to choose familiar majors due to the reduced probability of admission to unfamiliar majors.

8 Conclusion

Higher education plays a significant role in fostering human capital development and is expected to promote intergenerational mobility by offering equal opportunities to students from diverse socioeconomic backgrounds. However, in this study, I document that—within the context of an exam-based centralized college application system—the application decisions of students from disadvantaged backgrounds are disproportionately impacted by information frictions. Specifically, using administrative data from Chinese college applicants, I find that low-SES students are 21.6% more likely to choose familiar majors, i.e., majors that resemble high school subjects, than their high-SES counterparts with similar demographic characteristics, exam performances, and admission probabilities. Since familiar majors on average have lower labor market returns than unfamiliar majors, the observed disparity in major choice creates gaps in future incomes across socioeconomic backgrounds. This diminishes the potential for college education to serve as a vehicle for intergenerational mobility.

Information inequality is the leading factor driving the observed disparity in major choice. In particular, I present evidence from a survey experiment and from within-classroom information spillovers. In the survey experiment, I confirm the existence of an information gap between familiar and unfamiliar majors, and find that low-SES students exhibit a greater lack of information regarding unfamiliar majors than their high-SES counterparts. The evidence on information spillovers connects the lack of information with the choice of major. Leveraging random assignment of students into high school classes, I find evidence indicating that students obtain information about unfamiliar majors from their high-SES classmates and become more likely to choose such majors.

Finally, I discuss the economic consequences of the observed socioeconomic disparity in major choice that results from information inequality. Specifically, I calibrate a model in which students choose majors to maximize their expected utility based on the information that they possess about

each major, and the information depends on the student's SES and the major's familiarity. The estimation results show that, due to a lack of information about unfamiliar majors, low-SES students face an 11.9% higher likelihood of being mismatched. Moreover, as familiar majors generally offer lower labor market returns, low-SES students expect a 1.4% lower average income than their high-SES peers, which is equivalent to a $64,000$ CNY ($8,800$ USD) gap in lifetime incomes.

By specifically focusing on the choice of familiar majors, this study illustrates how students from different socioeconomic backgrounds are differentially impacted by information frictions. Undoubtedly, there exist various other dimensions of "familiarity" beyond the resemblance between a major and high school curricula, including whether any family members have pursued the major or whether students possess knowledge about the occupations associated with it. Low-SES students may also encounter information disadvantages along these dimensions, further exacerbating the challenges that they face in making an informed major choice. Consequently, the presence of information inequality within the higher education system undermines its ability to effectively promote intergenerational mobility and calls for information interventions targeting low-SES students or affirmative action admission policies.

References

Abramitzky, Ran, Victor Lavy, and Maayan Segev, "The effect of changes in the skill premium on college degree attainment and the choice of major," *Journal of Labor Economics*, 2022.

_ , _ , and Santiago Pérez, "The long-term spillover effects of changes in the return to schooling," *Journal of Public Economics*, 2021, *196*, 104369.

Aguirre, Josefa and Juan Matta, "Walking in your footsteps: Sibling spillovers in higher education choices," *Economics of Education Review*, 2021, *80*, 102062.

Ainsworth, Robert, Rajeev Dehejia, Cristian Pop-Eleches, and Miguel Urquiola, "Why do households leave school value added on the table? The roles of information and preferences," *American Economic Review*, 2023, *113* (4), 1049–1082.

Altmejd, Adam, Andrés Barrios-Fernández, Marin Drlje, Joshua Goodman, Michael Hurwitz, Dejan Kovac, Christine Mulhern, Christopher Neilson, and Jonathan Smith, "O brother, where start thou? Sibling spillovers on college and major choice in four countries," *The Quarterly Journal of Economics*, 2021, *136* (3), 1831–1886.

Altonji, Joseph G, Erica Blom, and Costas Meghir, "Heterogeneity in human capital investments: High school curriculum, college major, and careers," *Annual Review of Economics*, 2012, *4* (1), 185–223.

_ , Peter Arcidiacono, and Arnaud Maurel, "The analysis of field choice in college and graduate school: Determinants and wage effects," in "Handbook of the Economics of Education," Vol. 5, Elsevier, 2016, pp. 305–396.

Andrews, Rodney J, Scott A Imberman, and Michael F Lovenheim, "Risky business? The effect of majoring in business on earnings and educational attainment," 2017. *NBER Working Paper 23575*.

Arcidiacono, Peter, "Ability sorting and the returns to college major," *Journal of Econometrics*, 2004, *121* (1-2), 343–375.

Arteaga, Felipe, Adam J Kapor, Christopher A Neilson, and Seth D Zimmerman, "Smart matching platforms and heterogeneous beliefs in centralized school choice," *The Quarterly Journal of Economics*, 2022, *137* (3), 1791–1848.

Artemov, Georgy, "Assignment mechanisms: common preferences and information acquisition," *Journal of Economic Theory*, 2021, *198*, 105370.

Attanasio, Orazio P and Guglielmo Weber, "Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey," *Journal of Political Economy*, 1995, *103* (6), 1121–1157.

Backes, Ben, "Do affirmative action bans lower minority college enrollment and attainment?: Evidence from statewide bans," *Journal of Human Resources*, 2012, *47* (2), 435–455.

Baker, Rachel, Eric Bettinger, Brian Jacob, and Ioana Marinescu, "The effect of labor market information on community college students' major choice," *Economics of Education Review*, 2018, *65*, 18–30.

Barone, Carlo, Antonio Schizzerotto, Giovanni Abbiati, and Gianluca Argentin, "Information barriers, social inequality, and plans for higher education: Evidence from a field experiment," *European Sociological Review*, 2017, *33* (1), 84–96.

Barrios-Fernández, Andrés, "Neighbors' effects on university enrollment," *American Economic Journal: Applied Economics*, 2022, *14* (3), 30–60.

Beffy, Magali, Denis Fougere, and Arnaud Maurel, "Choosing the field of study in postsecondary education: Do expected earnings matter?," *Review of Economics and Statistics*, 2012, *94* (1), 334–347.

Bettinger, Eric P, Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu, "The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment," *The Quarterly Journal of Economics*, 2012, *127* (3), 1205–1242.

Black, Sandra E, Kalena E Cortes, and Jane Arnold Lincove, "Academic undermatching of high-achieving minority students: Evidence from race-neutral and holistic admissions policies," *American Economic Review*, 2015, *105* (5), 604–610.

Bleemer, Zachary, "Affirmative action, mismatch, and economic mobility after California's Proposition 209," *The Quarterly Journal of Economics*, 2022, *137* (1), 115–160.

_ **and Aashish Mehta**, "Will studying economics make you rich? A regression discontinuity analysis of the returns to college major," *American Economic Journal: Applied Economics*, 2022, *14* (2), 1–22.

_ **and Basit Zafar**, "Intended college attendance: Evidence from an experiment on college returns and costs," *Journal of Public Economics*, 2018, *157*, 184–211.

Blundell, Richard, Martin Browning, and Costas Meghir, "Consumer demand and the lifecycle allocation of household expenditures," *The Review of Economic Studies*, 1994, *61* (1), 57–80.

Bobonis, Gustavo J and Frederico Finan, "Neighborhood peer effects in secondary school enrollment decisions," *The Review of Economics and Statistics*, 2009, *91* (4), 695–716.

Boneva, Teodora and Christopher Rauh, "Socio-economic gaps in university enrollment: The role of perceived pecuniary and non-pecuniary returns," *Available at SSRN 3106691*, 2017.

Bonilla-Mejía, Leonardo, Nicolas L Bottan, and Andrés Ham, "Information policies and higher education choices experimental evidence from Colombia," *Journal of Behavioral and Experimental Economics*, 2019, *83*, 101468.

Bordon, Paola and Chao Fu, "College-major choice to college-then-major choice," *The Review of economic studies*, 2015, *82* (4), 1247–1288.

Bucher, Stefan F and Andrew Caplin, "Inattention and inequity in school matching," 2021. *NBER Working Paper 29586*.

Burke, Mary A and Tim R Sass, "Classroom peer effects and student achievement," *Journal of Labor Economics*, 2013, *31* (1), 51–82.

Burland, Elizabeth, Susan Dynarski, Katherine Michelmore, Stephanie Owen, and Shwetha Raghuraman, "The power of certainty: Experimental evidence on the effective design of free tuition programs," 2022. *NBER Working Paper 29864*.

Campbell, Stuart, Lindsey Macmillan, Richard Murphy, and Gill Wyness, "Matching in the dark? Inequalities in student to degree match," *Journal of Labor Economics*, 2022, *40* (4), 807–850.

Card, David and A Abigail Payne, "High school choices and the gender gap in STEM," *Economic Inquiry*, 2021, *59* (1), 9–28.

Carman, Katherine Grace and Lei Zhang, "Classroom peer effects and academic achievement: Evidence from a Chinese middle school," *China Economic Review*, 2012, *23* (2), 223–237.

Carrell, Scott and Bruce Sacerdote, "Why do college-going interventions work?," *American Economic Journal: Applied Economics*, 2017, *9* (3), 124–151.

Chan, Kam Wing, "Five decades of the Chinese hukou system," in "Handbook of Chinese migration," Edward Elgar Publishing, 2015, pp. 23–47.

Chen, Li and Juan Sebastián Pereyra, "Self-selection in school choice," *Games and Economic Behavior*, 2019, *117*, 59–81.

Chen, Yan and Onur Kesten, "Chinese college admissions and school choice reforms: A theoretical analysis," *Journal of Political Economy*, 2017, *125* (1), 99–139.

Cheng, Tiejun and Mark Selden, "The origins and social consequences of China's hukou system," *The China Quarterly*, 1994, *139*, 644–668.

Chesney, Alexander J, "Should I get a master's degree?: Evaluating peer effects on education investment decisions in the workplace," *Economics of Education Review*, 2022, *91*, 102329.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan, "Income segregation and intergenerational mobility across colleges in the United States," *The Quarterly Journal of Economics*, 2020, *135* (3), 1567–1633.

Conlon, John J and Dev Patel, "What jobs come to mind? stereotypes about fields of study," 2022. *Working Paper*.

Dahl, G, Dan-Olof Rooth, and Anders Stenberg, "Intergenerational and sibling peer effects by gender in high school majors," 2021. *NBER Working Paper 27618*.

Delaney, Judith M and Paul J Devereux, "Understanding gender differences in STEM: Evidence from college applications," *Economics of Education Review*, 2019, *72*, 219–238.

_ **and** _ , "Choosing differently? College application behavior and the persistence of educational advantage," *Economics of Education Review*, 2020, *77*, 101998.

Dillon, Eleanor Wiske and Jeffrey Andrew Smith, "Determinants of the match between student ability and college quality," *Journal of Labor Economics*, 2017, *35* (1), 45–66.

Dynarski, Susan, CJ Libassi, Katherine Michelmore, and Stephanie Owen, "Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students," *American Economic Review*, 2021, *111* (6), 1721–56.

Ehlert, Martin, Claudia Finger, Alessandra Rusconi, and Heike Solga, "Applying to college: Do information deficits lower the likelihood of college-eligible students from less-privileged families to pursue their college intentions?: Evidence from a field experiment," *Social science research*, 2017, *67*, 193–212.

Ersoy, Fulya and Jamin D Speer, "Opening the black box of college major choice: Evidence from an information intervention," in "*2022 APPAM Fall Research Conference*" APPAM 2022.

Fernández, Andrés Barrios, Christopher Neilson, and Seth D Zimmerman, "Elite universities and the intergenerational transmission of human and social capital," *Available at SSRN 4071712*, 2021.

Fricke, Hans, Jeffrey Grogger, and Andreas Steinmayr, "Exposure to academic fields and college major choice," *Economics of Education Review*, 2018, *64*, 199–213.

Grenet, Julien, YingHua He, and Dorothea Kübler, "Preference discovery in university admissions: The case for dynamic multioffer mechanisms," *Journal of Political Economy*, 2022, *130* (6), 1427–1476.

Gurantz, Oded, Michael Hurwitz, and Jonathan Smith, "Sibling effects on high school exam taking and performance," *Journal of Economic Behavior & Organization*, 2020, *178*, 534–549.

Hastings, Justine, Christopher A Neilson, and Seth D Zimmerman, "The effects of earnings disclosure on college enrollment decisions," 2015. *NBER Working Paper 21300*.

Hastings, Justine S, Christopher A Neilson, and Seth D Zimmerman, "Are some degrees worth more than others? Evidence from college admission cutoffs in Chile," 2013. *NBER Working Paper 19241*.

_ , _ , Anely Ramirez, and Seth D Zimmerman, "(Un)informed college and major choice: Evidence from linked survey and administrative data," *Economics of Education Review*, 2016, *51*, 136–151.

Hinrichs, Peter, "The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities," *Review of Economics and Statistics*, 2012, *94* (3), 712–722.

Hoxby, Caroline and Sarah Turner, "Expanding college opportunities for high-achieving, low income students," *Stanford Institute for Economic Policy Research Discussion Paper*, 2013, *12* (014), 7.

Hoxby, Caroline M, "Peer effects in the classroom: Learning from gender and race variation," 2000. *NBER Working Paper 7867*.

_ and Christopher Avery, "The missing "one-offs": The hidden supply of high-achieving, low income students," 2012. *NBER Working Paper 18586*.

_ and Sarah Turner, "What high-achieving low-income students know about college," *American Economic Review*, 2015, *105* (5), 514–17.

Jensen, Robert, "The (perceived) returns to education and the demand for schooling," *The Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

Jia, Ruixue and Hongbin Li, "Just above the exam cutoff score: Elite college admission and wages in China," *Journal of Public Economics*, 2021, *196*, 104371.

_ , _ , and Lingsheng Meng, "Can elite college education change one's fate in China?," *Available at SSRN 4113768*, 2022.

Joensen, Juanna Schrøter and Helena Skyt Nielsen, "Mathematics and gender: Heterogeneity in causes and consequences," *The Economic Journal*, 2016, *126* (593), 1129–1163.

_ and _ , "Spillovers in education choice," *Journal of Public Economics*, 2018, *157*, 158–183.

Kapor, Adam J, Christopher A Neilson, and Seth D Zimmerman, "Heterogeneous beliefs and school choice mechanisms," *American Economic Review*, 2020, *110* (5), 1274–1315.

Kaufmann, Katja Maria, Matthias Messner, and Alex Solis, "Elite higher education, the marriage market and the intergenerational transmission of human capital," *Working Paper*, 2021.

Kerr, Sari Pekkala, Tuomas Pekkarinen, Matti Sarvimäki, and Roope Uusitalo, "Post-secondary education and information on labor market prospects: A randomized field experiment," *Labour Economics*, 2020, *66*, 101888.

Kinsler, Josh and Ronni Pavan, "The specificity of general human capital: Evidence from college major choice," *Journal of Labor Economics*, 2015, *33* (4), 933–972.

Kirkeboen, Lars J, Edwin Leuven, and Magne Mogstad, "Field of study, earnings, and self-selection," *The Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

Lai, Fang, Elisabeth Sadoulet, and Alain de Janvry, "The adverse effects of parents' school selection errors on academic achievement: Evidence from the Beijing open enrollment program," *Economics of Education Review*, 2009, *28* (4), 485–496.

Lavy, Victor, M Daniele Paserman, and Analia Schlosser, "Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom," *The Economic Journal*, 2012, *122* (559), 208–237.

Le, Kien and My Nguyen, "'Bad Apple' peer effects in elementary classrooms: the case of corporal punishment in the home," *Education Economics*, 2019, *27* (6), 557–572.

Lergetporer, Philipp, Katharina Werner, and Ludger Woessmann, "Does ignorance of economic returns and costs explain the educational aspiration gap? Representative evidence from adults and adolescents," *Economica*, 2021, *88* (351), 624–670.

Ma, Yingyi, "Family socioeconomic status, parental involvement, and college major choices–gender, race/ethnic, and nativity patterns," *Sociological Perspectives*, 2009, *52* (2), 211–234.

Manski, Charles F, *Identification problems in the social sciences*, Harvard University Press, 1995.

Marioulas, Julian, "China: A world leader in graduation rates," *International Higher Education*, 2017, (90), 28–29.

McGuigan, Martin, Sandra McNally, and Gill Wyness, "Student awareness of costs and benefits of educational decisions: Effects of an information campaign," *Journal of Human Capital*, 2016, *10* (4), 482–519.

Mulhern, Christine, "Changing college choices with personalized admissions information at scale: Evidence on Naviance," *Journal of Labor Economics*, 2021, *39* (1), 219–262.

Oreopoulos, Philip and Ryan Dunn, "Information and college access: Evidence from a randomized field experiment," *The Scandinavian Journal of Economics*, 2013, *115* (1), 3–26.

Park, Albert, "Rural-urban inequality in China," *China urbanizes: Consequences, strategies, and policies*, 2008, pp. 41–63.

Patnaik, Arpita, Matthew J Wiswall, and Basit Zafar, "College majors," 2020. *NBER Working Paper 27618*.

Patterson, Richard W, Nolan G Pope, and Aaron Feudo, "Timing matters: Evidence from college major decisions," *Journal of Human Resources*, 2021, pp. 0820–11127R1.

Peter, Frauke H and Vaishali Zambre, "Intended college enrollment and educational inequality: Do students lack information?," *Economics of Education Review*, 2017, *60*, 125–141.

Saks, Raven E and Stephen H Shore, "Risk and career choice," *The BE Journal of Economic Analysis & Policy*, 2005, *5* (1).

Scott-Clayton, Judith, "Information constraints and financial aid policy," 2012. *NBER Working Paper 17811*.

Smith, Jonathan, Matea Pender, and Jessica Howell, "The full extent of student-college academic undermatch," *Economics of Education Review*, 2013, *32*, 247–261.

Stinebrickner, Ralph and Todd R Stinebrickner, "A major in science? Initial beliefs and final outcomes for college major and dropout," *Review of Economic Studies*, 2014, *81* (1), 426–472.

Wiswall, Matthew and Basit Zafar, "Determinants of college major choice: Identification using an information experiment," *The Review of Economic Studies*, 2015, *82* (2), 791–824.

_ and _, "How do college students respond to public information about earnings?," *Journal of Human Capital*, 2015, *9* (2), 117–169.

Yang, Yu Alan, "Place-Based college admission, migration and the spatial distribution of human capital: Evidence from China," *Working Paper*, 2021.

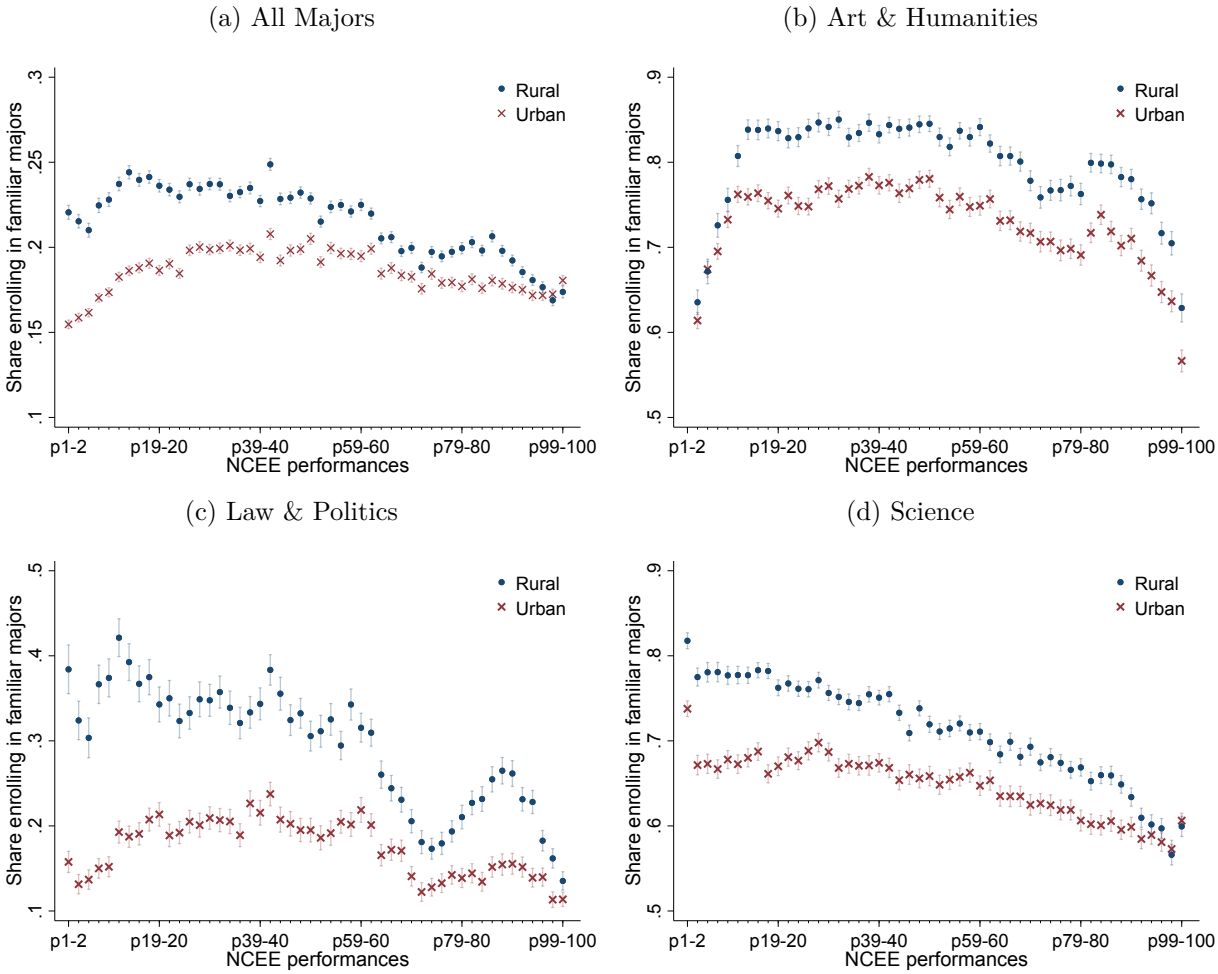
Young, Jason, "China's hukou system," *Basingstoke: Palgrave Macmillan*, 2013, *10*.

Zafar, Basit, "How do college students form expectations?," *Journal of Labor Economics*, 2011, *29* (2), 301–348.

Zimmerman, Seth D, "Elite colleges and upward mobility to top jobs and top incomes," *American Economic Review*, 2019, *109* (1), 1–47.

Figures and Tables

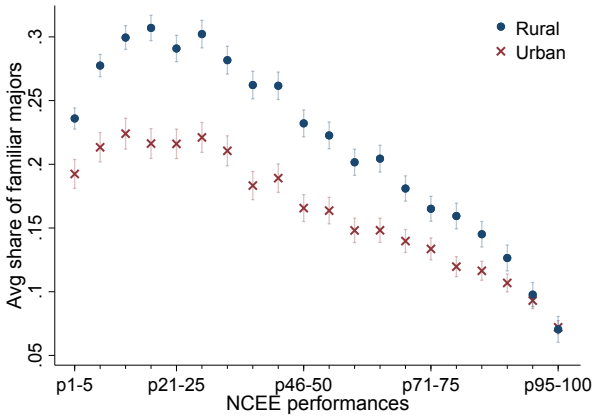
Figure 1: Enrollment in Familiar Majors by Urban–Rural Status, National Data



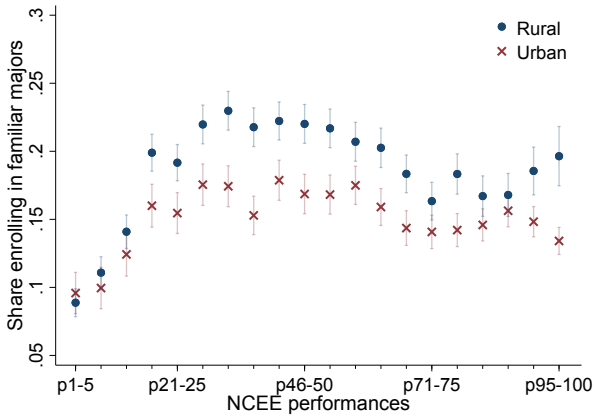
Notes: This figure plots the share of students who enroll in familiar majors by their NCEE performances with 95% confidence intervals separately for rural and urban students. Students' NCEE performances are converted into 50 groups, with group 1 containing those who score in the 1st and 2nd percentiles and group 50 containing those who score in the top 2% of the distribution. The sample of sub-figure (a) includes all college applicants who enrolled in four-year public colleges between years 1999 and 2003 from the national data. Sub-figures (b), (c), and (d) specifically focus on subsets of college applicants who enroll in majors within the Art & Humanities, Law & Politics, and Science categories, respectively. These three major categories are chosen because they contain a significant share of both familiar and unfamiliar majors.

Figure 2: Applications to and Enrollment in Familiar Majors by Urban–Rural Status, ROL Data

(a) First-Choice Majors, Urban–Rural Status

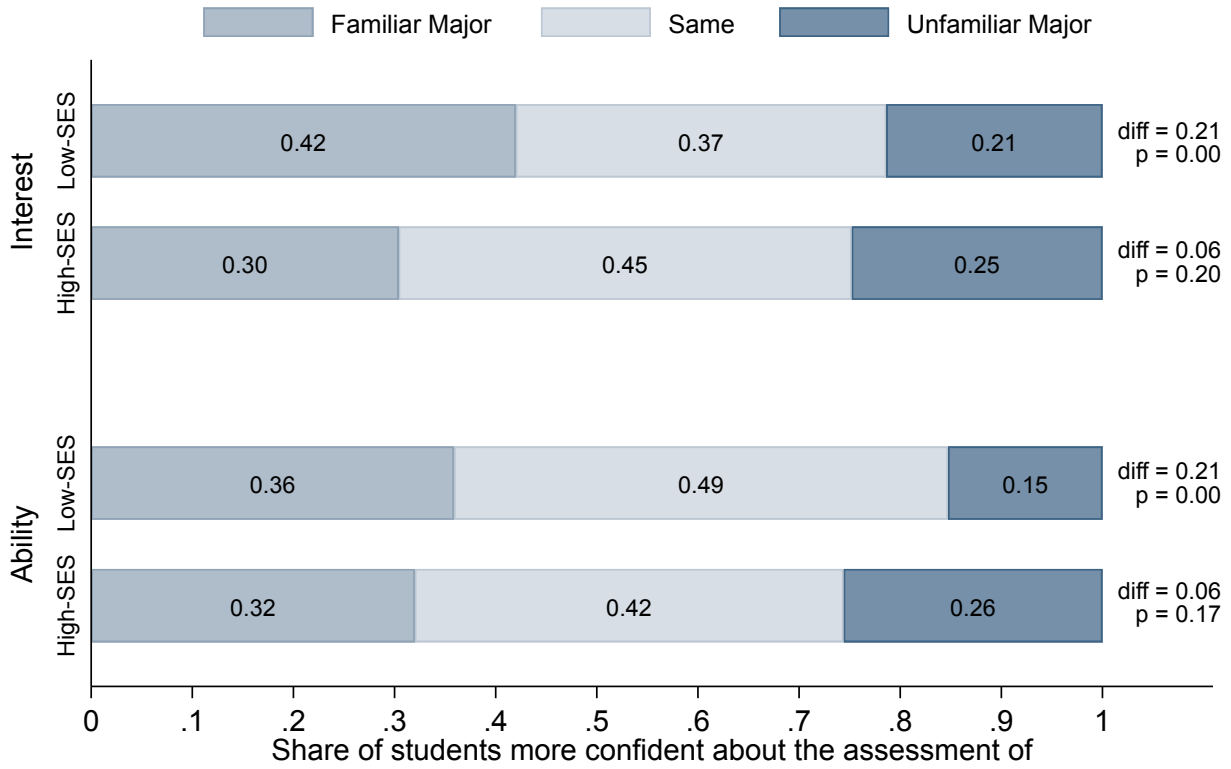


(b) Enrolled Major, Urban–Rural Status



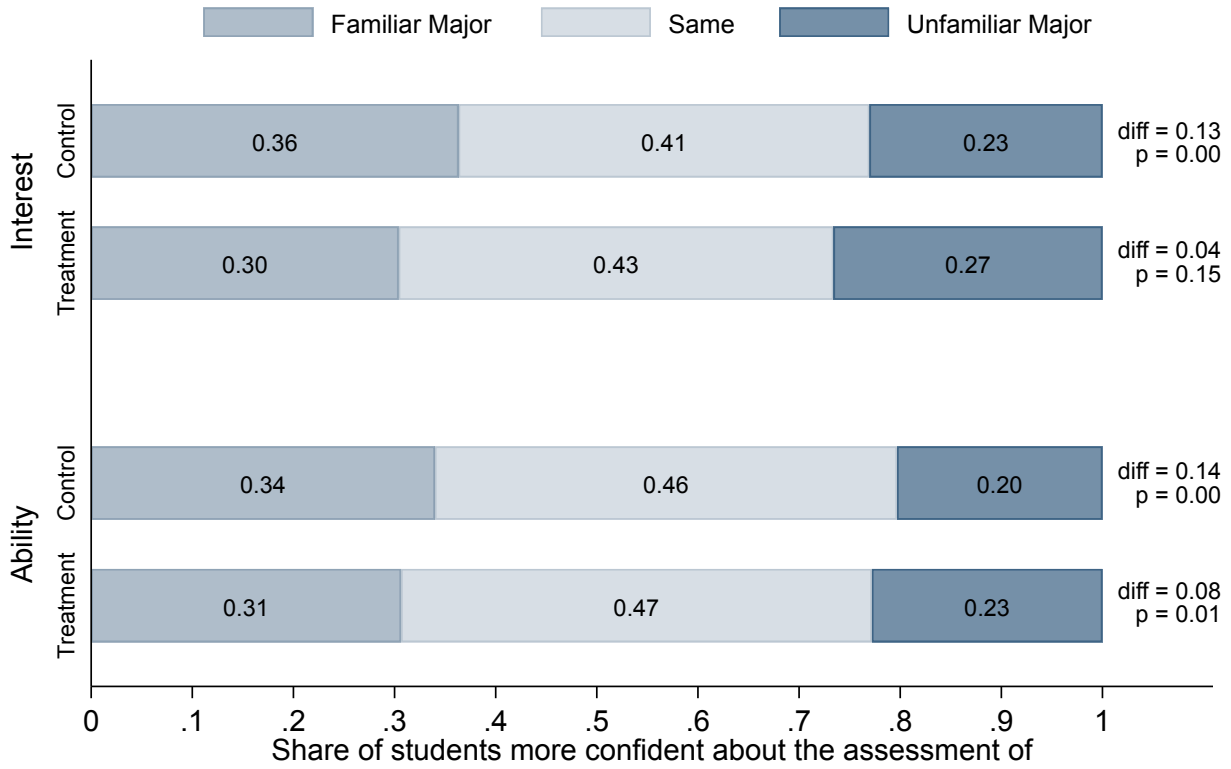
Notes: This figure plots students' applications to and enrollment in familiar majors by their NCEE performances with 95% confidence intervals separately for rural and urban students. Students' NCEE performances are converted into 20 ventiles, with group 1 containing those who score between the 1st and 5th percentiles and group 20 containing those who score in the top 5% of the distribution. The sample includes the students who applied to four-year public colleges in Ningxia province between 2014 and 2018 from the ROL data. Sub-figure (a) plots the average share of familiar majors among each student's first-choice majors, and sub-figure (b) plots the share of students who enroll in familiar majors.

Figure 3: SES Disparity in Confidence Levels, without Information Treatment



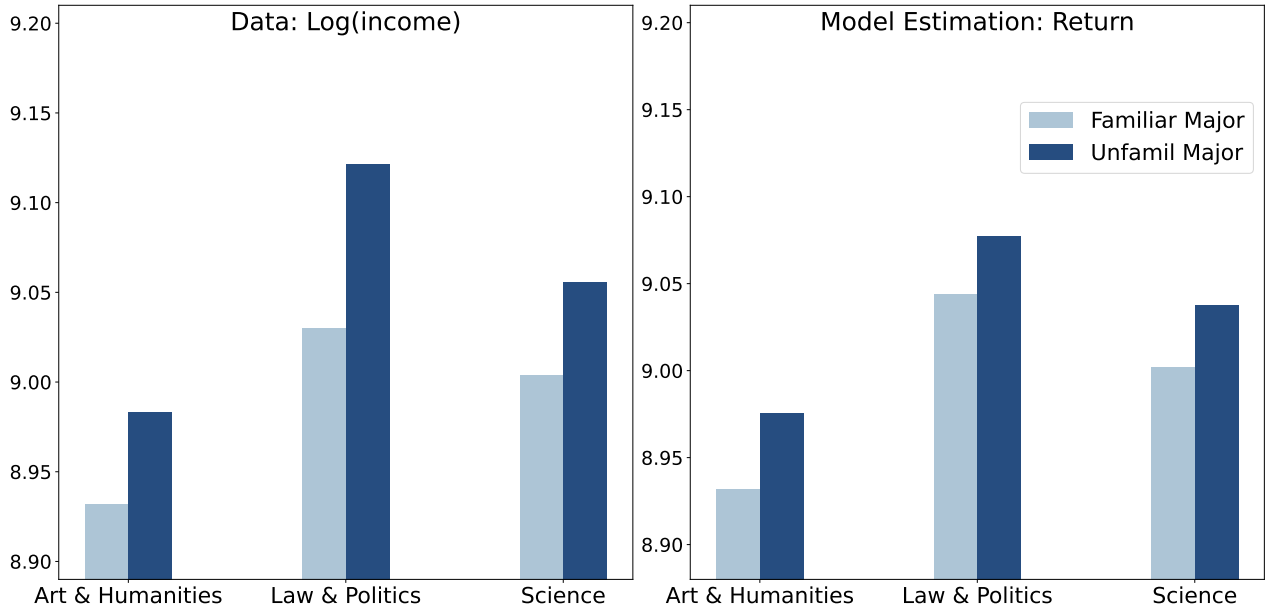
Notes: This figure plots the shares of students who reported being more confident in accurately assessing the familiar major in the assigned pair, being equally confident in assessing both majors, and being more confident in accurately assessing the unfamiliar major, separately by SES and by assessment of interests and abilities. The sample includes the students randomly assigned to the control group in the survey experiment. The numbers on the right side show the difference between the share of students who expressed more confidence about the assessment of the familiar major versus the share who expressed more confidence about the assessment of the unfamiliar major, together with the corresponding p-values from a t-test where the null hypothesis is that the share of students who are more confident about the assessment of the familiar major is no larger than the share of students who are more confident about the assessment of the unfamiliar major.

Figure 4: Effect of Information Treatment on Confidence Levels



Notes: This figure plots the shares of students who reported being more confident in accurately assessing the familiar major in the assigned pair, being equally confident in assessing both majors, and being more confident in accurately assessing the unfamiliar major, separately for the students randomly assigned to the control and treatment groups in the survey experiment and by assessment of interests and abilities. The numbers on the right side show the difference between the share of students who expressed more confidence about the assessment of the familiar major versus the share who expressed more confidence about the assessment of the unfamiliar major, together with the corresponding p-values from a t-test where the null hypothesis is that the share of students who are more confident about the assessment of the familiar major is no larger than the share of students who are more confident about the assessment of the unfamiliar major.

Figure 5: Empirical Average Log Income and Estimated Return to Majors

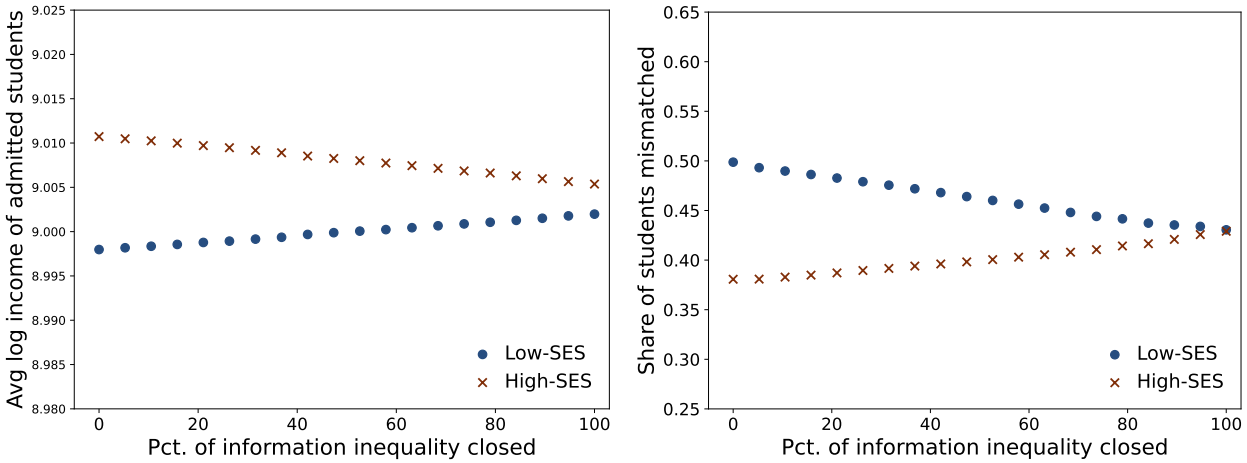


Notes: The figures plots the empirical average log income and estimated return to major separately for the familiar and unfamiliar majors from the three major categories: Art & Humanities, Law & Politics, and Science. The left panel plots the empirical average log income calculated from the Chinese Four-Year College Graduates Employment Annual Report (2020), and the right panel plots the return to major r_m estimated in the model.

Figure 6: Effects of Information Interventions on Future Incomes and Match Qualities

(a) Average Log Incomes

(b) Mismatch Rates

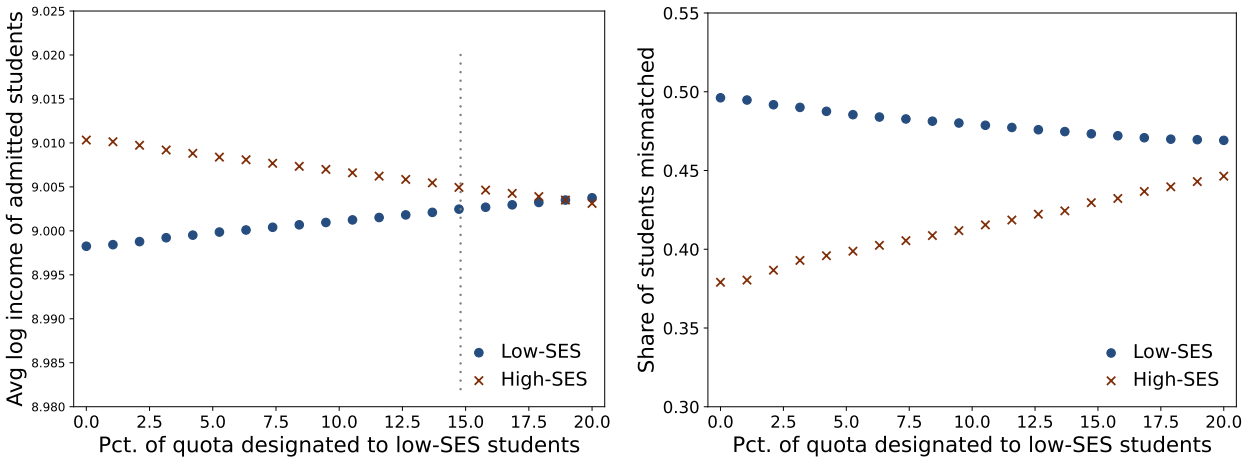


Notes: The figures plot separately for low- and high-SES students the effects of an information intervention that targets low-SES students on the simulated average log incomes (sub-figure (a)) and mismatch rates (sub-figure (b)) for the admitted students. The x-axis denotes the percentage of the information inequality between low- and high-SES students that is closed by the information intervention.

Figure 7: Effects of Affirmative Action Policies on Future Incomes and Match Qualities

(a) Average Log Incomes

(b) Mismatch Rates



Notes: The figures plot separately for low- and high-SES students the effects of an affirmative action policy on the simulated average log incomes (sub-figure (a)) and mismatch rates (sub-figure (b)) for the admitted students. The affirmative action policy allocates a designated portion of the admission quota for unfamiliar majors exclusively to low-SES students. The x-axis denotes the percentage of the admission quota for unfamiliar majors designated to low-SES students. The dashed line in sub-figure (a) denotes the designated percentage required to compensate for the income gap created by information inequality between low- and high-SES students.

Table 1: Examples of Familiar and Unfamiliar Majors, under Three Definitions of Major Familiarity

Major	(1) Narrow	(2) Broad (Preferred)	(3) Continuous
Chinese Language	1	1	0.25
English	1	1	0.22
Mathematics	1	1	0.80
Applied Mathematics	0	1	0.24
Physics	1	1	0.37
Chemistry	1	1	1.00
Biology	1	1	0.60
Biotechnology	0	1	0.02
Geography	1	1	0.30
History	1	1	0.80
Politics	1	1	0.17
Computer Science	0	0	0.05
Mechanical Engineering	0	0	0.08
Statistics	0	0	0.18
Psychology	0	0	0.01
Media Studies	0	0	0.03
Economics	0	0	0.39
Accounting	0	0	0.02
Marketing	0	0	0.00
Archeology	0	0	0.01
Clinical Medicine	0	0	0.00
Nursing	0	0	0.00

Notes: The table provides examples of major familiarity under three definitions of major familiarity: the narrow definition, the broad definition, and the continuous index.

Table 2: Income Gap between Familiar and Unfamiliar Majors

	(1)	(2)	(3)	(4)
Familiar Major	-0.1866***	-0.0793***	-0.0503**	-0.0509*
	(0.0172)	(0.0219)	(0.0217)	(0.0260)
Familiar × High-SES				0.0017
				(0.0316)
Student Char			x	x
Category FE		x	x	x
College FE		x	x	x
Mean(Y)	7.735	7.735	7.729	7.729
N	13,720	13,709	11,024	11,024

Notes: This table presents the gap in log monthly income between familiar and unfamiliar majors. The sample includes the students graduating from four-year public colleges who have secured employment in the CCSS data. The outcome variable is the self-reported log income of their best job offer, and the key independent variable is whether the student studies a familiar major in college. I additionally include major category and college fixed effects in column (2) and control for student characteristics in column (3). The student characteristics include year–academic track fixed effects, gender, ethnic group, eligible bonus points, and NCEE performances, as well as the student's log household income, GPA ranking in college, college tuition, and the location of the best job offer. In column (4), I add the interaction term between major familiarity and the student's socioeconomic status. Robust standard errors are reported in parentheses. Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table 3: Applications to Familiar Majors by Urban–Rural Status

	(1)	(2)	(3)	(4)	(5)
Rural	0.0384***	0.0398***	0.0322***	0.0316***	0.0557***
	(0.0009)	(0.0008)	(0.0009)	(0.0008)	(0.0021)
Position FE	x	x	x	x	First Only
Student Char		x	x	x	x
NCEE Score			x	x	x
Coll-Maj Char				x	x
Mean(Y)	0.146	0.146	0.146	0.146	0.188
N	2,457,800	2,457,800	2,457,800	2,457,800	159,545

Notes: This table presents the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student, except in column (5), where the sample is restricted to the first-choice major of the first-choice college for each student. The outcome variable is whether the listed major is classified as a familiar major according to the preferred definition, and the key independent variable is whether the student is a rural resident, as a proxy for the socioeconomic status. I additionally control for position fixed effects, student characteristics, NCEE score, and the characteristics of the listed college–major. The student characteristics include year–academic track fixed effects, gender, ethnic group, and eligible bonus points. The NCEE score is converted to dummy variables for each score percentile. The characteristics of the college–major include its tuition level and selectivity, measured by the admission cutoff in the previous year. Standard errors (in parentheses) are clustered at the individual level.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table 4: Effect of Information Treatment on Confidence Levels

	Interest			Ability		
	(1) Full Sample	(2) High SES	(3) Low SES	(4) Full Sample	(5) High SES	(6) Low SES
Unfamil	-0.1480***	-0.0645	-0.2259***	-0.1210***	-0.0968*	-0.1427***
	(0.0393)	(0.0542)	(0.0562)	(0.0379)	(0.0560)	(0.0516)
Treat × Unfamil	0.1360***	0.0106	0.2621***	0.0986**	0.0545	0.1418**
	(0.0490)	(0.0662)	(0.0725)	(0.0475)	(0.0686)	(0.0666)
Mean(Y)	0.471	0.459	0.484	0.455	0.455	0.456
N	1,346	712	634	1,346	712	634

Notes: This table presents the estimates of equation (2) using the sample of students in the survey experiment. The outcome variable is whether the student is confident in accurately assessing their fit with the major, with the results reported separately for the assessment of interests in columns (1)–(3) and for the assessment of abilities in columns (4)–(6). The independent variables include whether the major is an unfamiliar major, whether the student is randomly assigned to the treatment group, and their interaction term. I additionally control for the student's gender, urban–rural status, parental education, academic track, and high school entrance exam performances, as well as the ordering within the major pair and major pair fixed effects. In columns (2)/(3) and (5)/(6), I split the sample by the student's socioeconomic status, measured by parental education level. Standard errors (in parentheses) are clustered at the individual level.

Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table 5: Evidence on Information Spillovers among High School Classmates

	Share of Familiar Majors			Location: In Province			Location: East Coast		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Parents in Formal Sector	-0.0175***	-0.0174***	-0.0175***	-0.0270***	-0.0270***	-0.0268***	0.0329***	0.0329***	0.0329***
	(0.0040)	(0.0040)	(0.0040)	(0.0042)	(0.0042)	(0.0042)	(0.0045)	(0.0045)	(0.0045)
Share of Classmates with Parents in Formal Sector	-0.0594*			-0.0009			0.0544		
	(0.0301)			(0.0360)			(0.0336)		
Share of Formal Sector, Same Gender Group		-0.0420**			0.0015			0.0339	
		(0.0190)			(0.0198)			(0.0205)	
Share of Formal Sector, Diff Gender Group		-0.0183			-0.0033			0.0189	
		(0.0181)			(0.0204)			(0.0202)	
Share of Formal Sector, Same Ethnic Group			-0.0452**			0.0096			0.0087
			(0.0172)			(0.0190)			(0.0219)
Share of Formal Sector, Diff Ethnic Group			-0.0153			-0.0055			0.0240
			(0.0199)			(0.0226)			(0.0193)
Mean(Y)	0.162	0.162	0.162	0.234	0.234	0.234	0.361	0.361	0.361
N	18,348	18,348	18,328	18,348	18,348	18,328	18,348	18,348	18,328

Notes: This table presents the estimates of equation (3) using the ROL data from 2017 and 2018, because the information on students' high school classes is available only for 2017 and 2018. I exclude the cohorts in which the variation in average NCEE performances across classes is larger than 50%. The outcome variable in columns (1)–(3) is the share of familiar majors among the first-choice majors for each student. The outcome variables are the share of colleges in the student's rank-order list that are within Ningxia province for columns (4)–(6) and the share of colleges in the student's rank-order list that are located along the east coast for columns (7)–(9). The key independent variables are the formality of the student's parental job sector as a proxy for socioeconomic status and the share of the student's classmates whose parents work in the formal sector. I control for student characteristics including gender, ethnic group, urban–rural status, whether the student is awarded bonus points, and NCEE score converted to dummy variables for each score percentile, as well as the leave-one-out class-level average of these characteristics. In columns (2), (5), and (8), I split the class by whether students are of the same gender group as the focal student. In columns (3), (6), and (9), I split the class by whether students are of the same ethnic group as the focal student. I additionally include cohort–by–academic track fixed effects, and standard errors (in parentheses) are clustered at the cohort–by–academic track level.

Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

ONLINE APPENDIX

A Sample Construction and Key Variables

A.1 Sample Construction

ROL Data Using the ROL data, I construct a sample of college applicants from Ningxia province during the period of 2014 to 2018. I restrict the sample to high school students who apply for general admission to four-year public colleges while excluding applications to private colleges, vocational colleges, or other programs designated for students with certain characteristics such as special arts talents. For a small subsample of high school students who apply to both public colleges and private or vocational colleges, I focus solely on their applications to public colleges. Similarly, for students who apply to both the general admission and special programs, I examine only their applications for general admission.

Panel A of Appendix Table F2 reports the summary statistics for the sample constructed with the ROL data. Out of the total of 117, 714 high school students, 56, 335 (47.9%) are urban residents and thus are classified as high-SES students. A share of 55.2% of the sample is female, and 33.4% of the students are ethnic minorities. The share of ethnic minorities is much higher than the national average (8.49%) because Ningxia has historically been home to various minority ethnic groups, particularly the Hui ethnic group. Moreover, the statistics suggest that a substantial majority of ethnic minorities in Ningxia reside in rural areas. On average, rural students receive more bonus points as ethnic minorities. However, they tend to have slightly lower average exam performances. The statistics also indicate that in comparison to urban students, rural students are more inclined to apply to familiar majors and are more likely to be admitted to those majors.

National Data I construct an alternative sample using the national data from 1999 to 2003. Since the students' submitted rank-order lists are not available in the national data, I am able to analyze only the students' major of enrollment instead of their initial choice of major. Consequently, the sample is limited to students admitted by four-year public colleges. The summary statistics are reported in panel B of Appendix Table F2. The sample consists of approximately 6 million high school students, with 53.0% being urban students.

The proportion of female students in panel B is notably smaller than that in panel A, particularly among rural students. This difference actually signifies the progress made in achieving gender equality in China over time, as more female students, especially those from disadvantaged backgrounds, have

gained access to higher education. Meanwhile, in comparison to that in panel A, the proportion of ethnic minorities in the sample moves closer to the national average. Another notable difference with respect to the ROL sample is that the average exam performance of rural students is now higher than that of urban students. This arises from the fact that the national sample includes only students admitted to four-year public colleges, whereas the ROL sample encompasses all applicants. This suggests an urban–rural disparity in application strategies, as rural students need higher NCEE scores to secure admission than their urban counterparts. What remains consistent across both samples is the observation that rural students have a higher tendency to enroll in familiar majors than urban students.

CCSS Data I constructed a third sample using data from the Chinese College Student Survey (CCSS), which covers a representative group of college students who applied to college between 2006 and 2011. Similarly to the national data, the sample data are restricted to students enrolled in four-year public colleges, and only information on the students' enrolled majors instead of their initial choice of major is available. The final sample consists of $36,672$ students, of which 44.6% are urban students. The summary statistics, presented in panel C of Appendix Table F2, exhibit patterns similar to those observed in panel B, which includes the presence of the urban–rural disparity in enrollment in familiar majors.

A.2 Key Variables

Continuously Defined Major Familiarity Both the narrowly and broadly defined measures of major familiarity are based on the names of high school subjects. They do not account for the fact that some topics covered in high school are not fully represented by their subject names. For example, although students acquire some basic statistics concepts through their high school Mathematics curriculum, Statistics is classified as an unfamiliar major under the previous two definitions. To address this concern, I build a continuous measure of major familiarity that better captures the variation in students' exposure to different college majors during high school, going beyond simple comparisons to high school subject names.

Specifically, I utilize the National High School Curriculum Guidelines published by the Chinese Ministry of Education in 2017, which lists the topics required to be covered in each high school subject. For each college major, I count how frequently its name appears in the guidelines, and use the frequency as an approximation for students' exposure to this major during high school. I convert

the frequency into a continuous familiarity measure by rescaling it to an index between 0 and 1. The college major that appears most often (i.e., Chemistry) has a familiarity index of 1, and the majors that do not appear in the guidelines at all (e.g., Marketing and Nursing) have a familiarity index of 0. Other majors such as Statistics, Materials Science, and Economics, have indices between 0 and 1. Although there is no subject named Statistics in the high school curriculum, the key word "Statistics" appears quite often in the guideline for high school Mathematics and, thus, is not completely unfamiliar. Similarly, Materials Science appears multiple times in the guideline for high school Chemistry, and Economics appears in the guidelines for high school Politics and History. For most majors, the three measures of familiarity are highly consistent. Appendix Figure F2 shows the distribution of the continuous familiarity index by the other two discrete measures of major familiarity.

B Robustness of Disparity in Major Choice

The results reported in Table 3 are robust to the use of alternative definitions of major familiarity and alternative measures of socioeconomic status. In Appendix Table F4, I report the regression results of equation (1) estimated with different definitions of major familiarity (the narrow definition and the continuous index). The estimated coefficients have magnitudes and significance levels similar to those of the estimates in Table 3, suggesting that the results are robust under alternative definitions of major familiarity. In Appendix Table F5, I report the regression results of equation (1) using alternative proxies for socioeconomic status (county-level log GDP per capita, whether the parents work in the informal sector, and log household income). All in all, the different definitions of SES deliver results similar to the baseline: students from low-SES backgrounds are more likely to choose familiar majors than high-SES students.

In Appendix Table F6, instead of including all listed majors in the regression analysis, I present regression estimates that focus on the first-choice majors of all the listed colleges (panel A) and on the first-choice major of the first-choice college (panel B). The estimated results indicate that the disparity in choice of major across socioeconomic statuses is even more pronounced when we specifically examine the top, most preferred choices.

In Appendix Table F7, I report the regression results of equation (1) while additionally including major category fixed effects, college fixed effects, and both. The inclusion of major category and college fixed effects reduces the estimated SES disparity. This is partly mechanical given the presence of some major categories and colleges that predominantly offer familiar or unfamiliar majors. Despite

this, there remains a noticeable 6.6% disparity in major choice by socioeconomic status among students who choose the same college and major category. In Appendix Table F8, I report the regression results of equation (1), while additionally including county fixed effects and high school fixed effects. There remains a substantial disparity in major choice even with the county or high school fixed effects. The results suggest that living in the same county or attending the same high school does not fully close the gap in information and choice of major.

Furthermore, Appendix Table F9 presents evidence of a comparable urban–rural disparity in students' enrolled majors based on the national data spanning 1999 to 2003. Together with panel C of Appendix Table F5 that uses survey data from a national representative sample of college applicants between 2006 and 2011, the results demonstrate the prevalence of the observed SES disparity. The finding is not unique to a particular time period, region, or sample.

C Survey Implementation

The survey was conducted in February 2022 at Bashu High School in Chongqing, China. This survey was part of a comprehensive initiative launched by the high school administration to gather information from senior students. The aim was to gain insights into their daily routines, study schedules, and attitudes toward college applications and major choices, and to gather any feedback that they had for the school. As a consultant, I was responsible for designing the module on college applications and major choices. In the module, I measured students' information on a selected set of majors and conducted the randomized experiment, as described in Section 5.1.

I collected students' responses to the questions in the college application and major choice module, while also obtaining students' demographic information and academic records, including their performance on the high school entrance exam, from the high school administration. All 2,256 senior students enrolled in the school were instructed by their teachers to complete the survey online using either their phones or computers. Out of these students, 1,187 initiated the survey, and 677 successfully completed it while correctly responding to the comprehension questions. The comprehension questions were presented to the students following a prompt that introduced the specific major to which the subsequent questions pertained. The purpose was to assess the students' attentiveness and recall of the major discussed, ensuring that they were actively engaged and had memorized the information. It is worth mentioning that the results presented in Section 5.1 remain robust even when I include students who answered the comprehension questions incorrectly.

D Alternative Channels

In this section, I provide more details about why differences in application strategies, risk attitudes, and preferences are unlikely to be the main drivers of the documented SES disparity in major choice.

D.1 Different Application Strategies

The first alternative explanation for the observed disparity in major choice is differences in application strategies across socioeconomic backgrounds. There is evidence in the literature that low-SES students have lower levels of sophistication in their application strategies and tend to make more strategic errors (Lai et al., 2009; Chen and Pereyra, 2019). It is possible that unfamiliar majors, compared with the familiar ones, may have less competition and lower admission thresholds, which could attract high-SES students who are more strategically sophisticated.

To empirically verify this alternative channel, I compare the admission competitiveness of familiar and unfamiliar majors using the ROL data. Specifically, I regress the competitiveness on major familiarity, controlling for the year–academic track, application round, college, and major category fixed effects. I build three measures of competitiveness, including the number of applicants relative to admission quota, the percentage of the admission quota filled, and the admission threshold. A major is considered more competitive if it has a higher applicants-to-quota ratio, a greater proportion of the admission quota filled, and a higher admission threshold. The regression results reported in Appendix Table F10 suggest no significant difference in competitiveness between familiar and unfamiliar majors. If anything, familiar majors are marginally less competitive than unfamiliar majors, as they on average receive a slightly smaller number of applicants relative to their admission quotas.

D.2 Different Risk Attitudes

Differences in risk attitudes between high- and low-SES students could be another explanation for the observed disparity in major choice. Compared with familiar majors, unfamiliar majors are riskier choices. As is suggested by the findings from the survey experiment, it is more difficult for students to accurately evaluate whether an unfamiliar major fits their interests and abilities, relative to a familiar major. It is possible that high-SES students are more inclined to choose unfamiliar majors not only because they are better informed and face less uncertainty but also because of their higher risk tolerance levels. For example, high-SES students might have better financial support and safety

net from their families and, therefore, be more able to manage the possibility that the major that they enroll in turns out to be a poor fit.

To investigate whether students from different socioeconomic backgrounds exhibit different risk attitudes when facing majors with high levels of uncertainty, I incorporated three risk-related questions into the survey experiment.

1. Will you consider applying to a major when you are not sure whether it is a good fit for your interests or abilities?
2. Are you concerned about the possibility that the major you choose turns out to be a poor fit for you?
3. Do you agree that you can still have a successful career even if the major you study in college does not fit your interests or abilities?

To determine whether there exists an SES gap in risk attitudes, I conduct a student-level regression analysis in which the primary independent variable is the student's SES, measured by parental education, while accounting for other covariates such as gender, urban–rural status, academic track, and academic performance. The results of the regression analysis, presented in Appendix Table F11, suggest that there are no significant differences in the risk attitudes of high- and low-SES students, including their willingness to apply to high-risk majors, their level of concern regarding risks associated with major choices, and their perception of the long-term consequences of studying high-risk majors. Hence, it is improbable that the observed SES disparity in the choices of familiar majors is primarily driven by divergent risk attitudes across socioeconomic backgrounds.

D.3 Different Preferences

I also consider the possibility that the observed SES disparity in choosing familiar majors is mainly due to differing preferences between low- and high-SES students. In other words, I consider the possibility that low-SES students are more inclined to choose familiar majors because they derive greater utility from studying them.

However, measuring preferences empirically is inherently challenging. Therefore, I present indirect evidence against this alternative explanation by examining major changes from college to graduate school.[34] There are various reasons why students might change their majors when pursuing graduate studies, with one of the most common factors being the realization that their undergraduate major is not the best fit for them. If low-SES students' tendency to choose familiar majors is *not* solely driven by their personal preferences, then we would anticipate a higher share of low-SES students

[34]Note that it is very uncommon for students to change majors during college in China.

who choose familiar majors to later experience regret and ultimately switch majors, in contrast to the scenario in which the disparity in major choice can be fully explained by differences in preferences.

For the analysis, I use the CCSS data, as they contain information on students' post-college plans, and I restrict the sample to those who pursue graduate school after completing their undergraduate degree. In particular, I regress whether the student switches to a different major in graduate school on whether her major in college is a familiar major, her SES measured by household income, the interaction term between the two, and control variables including gender, ethnicity, the NCEE performances, any awarded bonus points, and province–year–academic track fixed effects. In addition, I include fixed effects for the student's college and her undergraduate major category to mitigate the concern that sometimes students choose one area of study in college because it prepares them for a different area of study in graduate school.[35]

Appendix Table F12 presents the regression results. According to column (1), students who study familiar majors in college are more likely to switch to a different major during graduate school.[36] However, this gap is notably smaller for students of high SES. In columns (2) and (3), I analyze high- and low-SES students separately. The results show that only low-SES students who study familiar majors during college are significantly more likely to switch to a different major in graduate school than those who study unfamiliar majors. Under the mild assumption that low-SES students are not more inclined to change their preferences, this result is consistent with the information friction hypothesis—low-SES students choose familiar majors because of a lack of information at the time of college application, but they obtain more information during college and switch to a better-fitting major for graduate school. It also contradicts the alternative explanation that the observed SES disparity in choosing familiar majors stems entirely from differences in preferences.

E Model Estimation

I estimate the unknown parameters by matching the students' simulated choice of major with their actual choice and the simulated average income of each major with the empirically observed average income. The unknown parameters include return to majors $\{r_m\}$, other parameters in students' expected utility $\boldsymbol{\theta} = \{\beta_0, \beta_1, \alpha_0, \alpha_1, \omega^a_m, \omega^v_m, \eta_{c(m)}, \sigma^2_F, \sigma^2_{Uh}, \sigma^2_{Ul}\}$, and the perceived admission probability $\{AdmProb_{t(i)m}\} = E[\mathbb{1}\{Admitted_{im}\}]$.

[35] For example, some students study Mathematics in college and plan to study Financial Engineering in graduate school.

[36] The CCSS data do not provide information on the specific majors that students pursue in graduate school, only on their major categories. Based on the available data, a conservative estimate is that at least half of the students who study familiar majors in college and switch majors end up studying unfamiliar majors during graduate school.

For each student i in the model estimation sample, I observe her major choice m_i^o and characteristics $\boldsymbol{X_i}$. The latter include student i's academic track $t(i)$, exam performances s_i, gender, and urban–rural status as a proxy for SES. Given a set of the unknown parameters $\{r_m, \boldsymbol{\theta}, AdmProb_{t(i)m}\}$, we are able to calculate the probability of student i choosing the major m_i^o: $p(\boldsymbol{X_i}, m_i^o | r_m, \boldsymbol{\theta}, AdmProb_{t(i)m})$. When summing across the sample of students, the log likelihood function can be expressed as:

$$LL = \sum_i \log p(\boldsymbol{X_i}, m_i^o | r_m, \boldsymbol{\theta}, AdmProb_{t(i)m}). \tag{6}$$

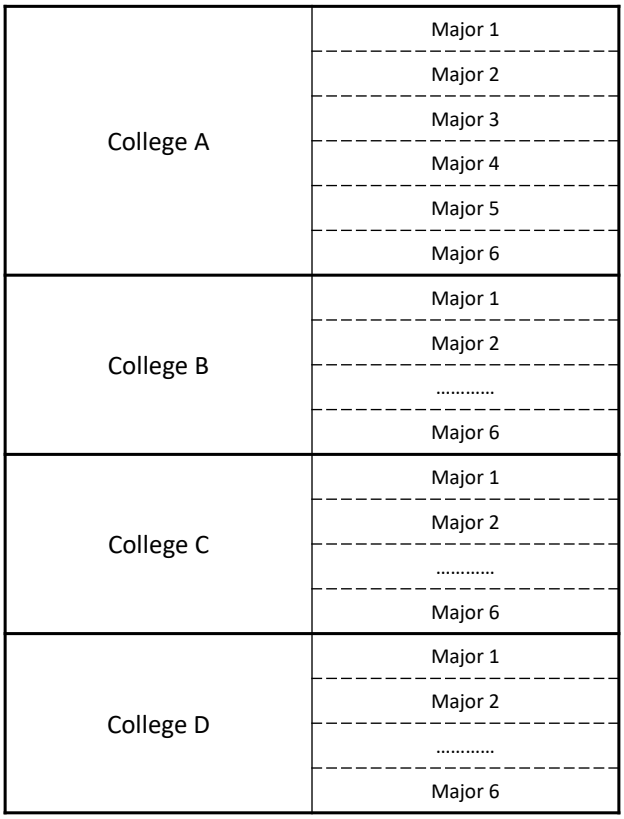
The estimation proceeds in two nested steps. In the first step, I take the return to majors $\{r_m\}$ as given and estimate parameters $\boldsymbol{\theta}$ and admission probabilities $\{AdmProb_{t(i)m}\}$ as functions of $\{r_m\}$. Specifically, I look for $\boldsymbol{\theta}$ and $\{AdmProb_{t(i)m}\}$ such that (1) $\boldsymbol{\theta}$ maximizes the log likelihood function (6) given $\{r_m\}$ and $\{AdmProb_{t(i)m}\}$ and (2) $\{AdmProb_{t(i)m}\}$ equals the admission probabilities calculated based on the simulated choice of major $\tilde{m}_i(\boldsymbol{X_i}, \varepsilon_{im} | r_m, \boldsymbol{\theta}, AdmProb_{t(i)m})$ and admission quotas, where ε_{im} is a random sequence of students' idiosyncratic preferences.

In the second step, I estimate the return to majors $\{r_m\}$ using the method of simulated moments. The first step allows me to express the parameters $\boldsymbol{\theta}$ and admission probabilities $\{AdmProb_{t(i)m}\}$ as functions of $\{r_m\}$. Therefore, for each set of $\{r_m\}$ and a random sequence of idiosyncratic preferences ε_{im}, I am able to simulate students' major choices $\tilde{m}_i(\boldsymbol{X_i}, \varepsilon_{im} | r_m)$ and the admission process that follows. With the simulated choices of major and admission outcomes, I calculate the average income of students graduating from each major. Using the method of simulated moments, I estimate the return to majors $\{r_m\}$ such that the simulated average income of each major matches the empirically observed average income.

Combining the two steps, I obtain estimates of the return to majors r_m, parameters $\boldsymbol{\theta}$, and admission probabilities $\{AdmProb_{t(i)m}\}$ so that the students' simulated choices of major match their actual choices and the simulated average income of each major matches the empirically observed average income.

F Additional Results

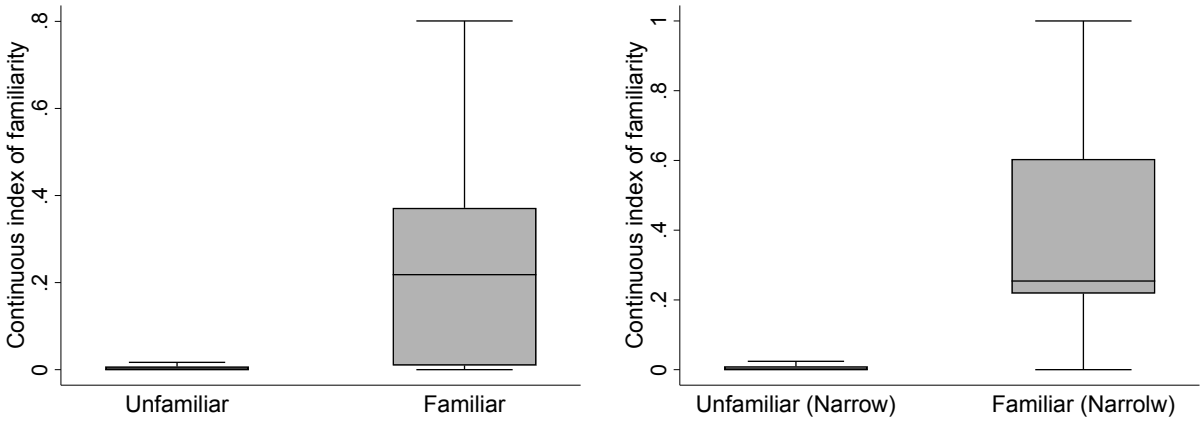
Figure F1: Template of Rank-Order List



Notes: This figure illustrates a template of the rank-order list to be completed by college applicants from Ningxia province between 2014 and 2018. In each application round, an applicant can list up to four colleges, and within each college, up to six majors. The rank-order list is structured with the college first and the major second. Applicants have the option to leave some positions blank.

Figure F2: Distribution of Continuous Index of Familiarity by Discrete Major Familiarity

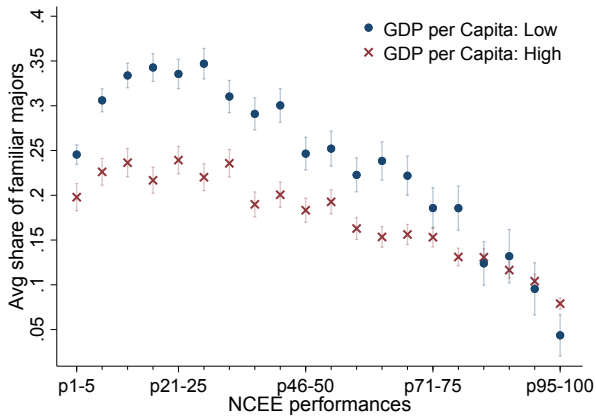
(a) By Major Familiarity, Broadly Defined (b) By Major Familiarity, Narrowly Defined



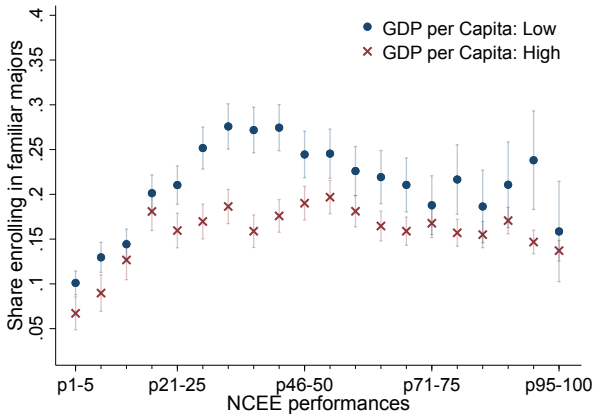
Notes: This figure cross-validates the continuous index of familiarity with the discretely defined major familiarity. Specifically, sub-figure (a) plots the distribution of the continuous index of familiarity separately for the broadly defined familiar and unfamiliar majors. Sub-figure (b) plots the distribution of the continuous index of familiarity separately for the narrowly defined familiar and unfamiliar majors. The box graph shows the 25th percentile, median, and 75th percentile, all weighted by the size of the major, and includes the lower and upper bounds.

Figure F3: Applications to and Enrollment in Familiar Majors by Alternative Proxies for SES, ROL Data

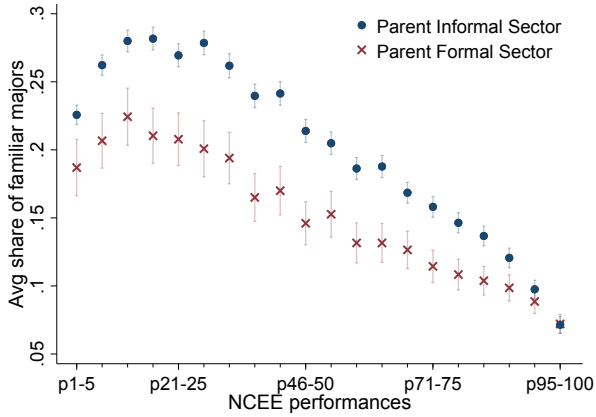
(a) First-Choice Majors, County-Level GDP per Capita



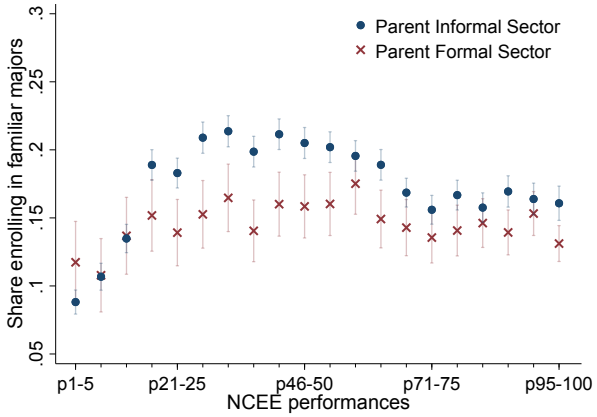
(b) Enrolled Major, County-Level GDP per Capita



(c) First-Choice Majors, Parental Job Sector

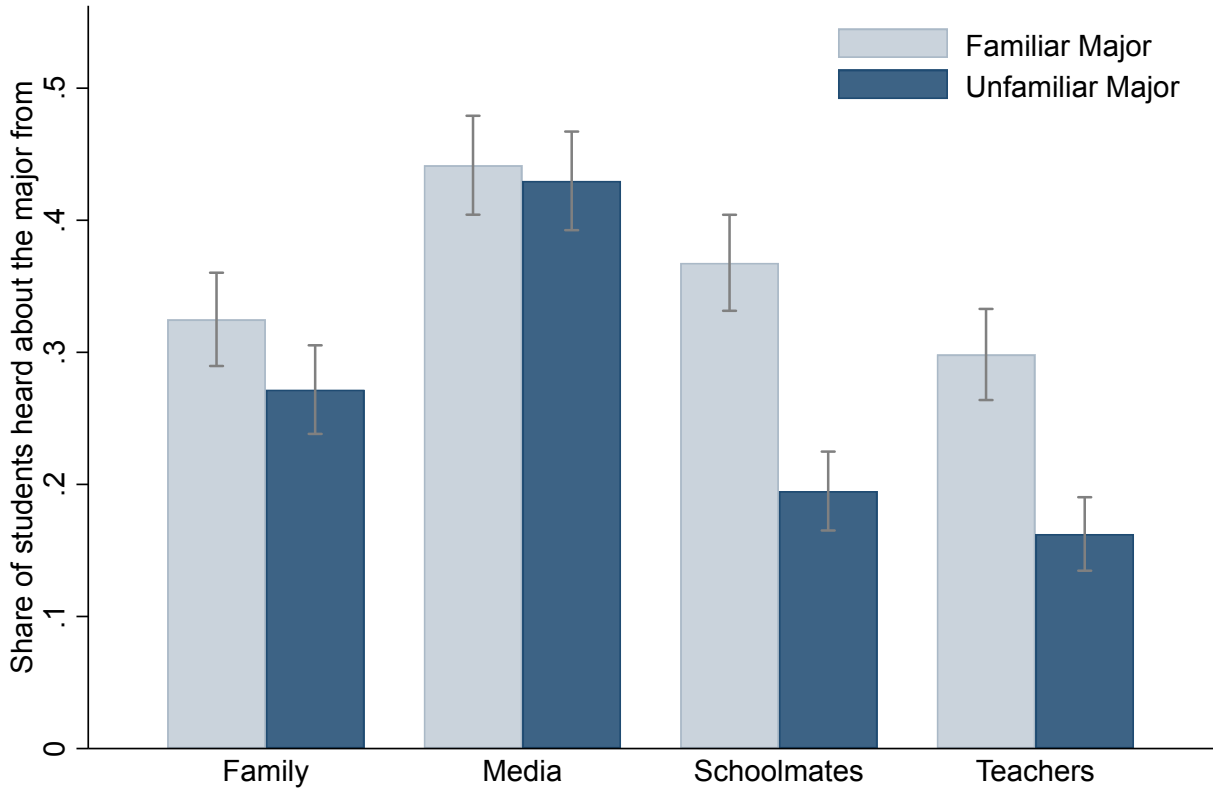


(d) Enrolled Major, Parental Job Sector



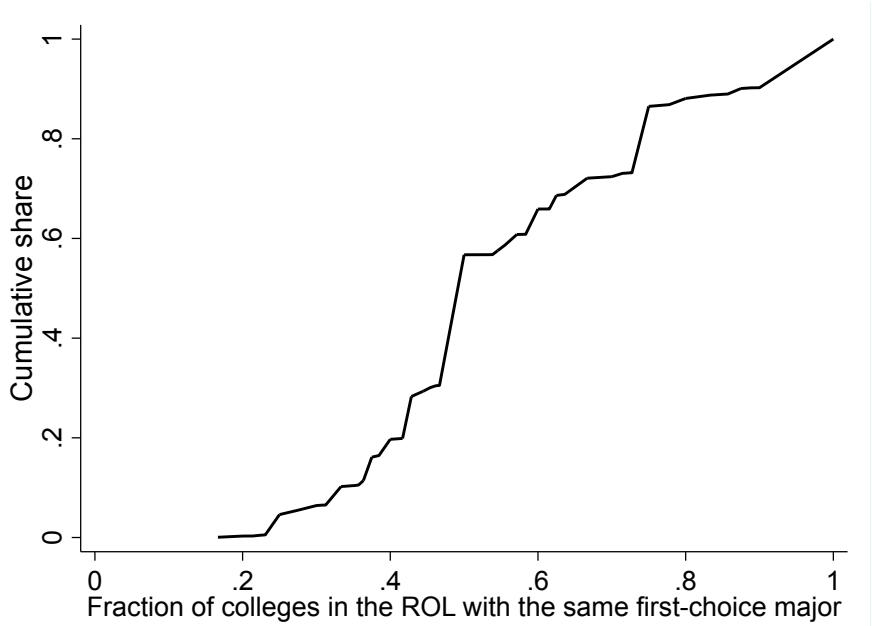
Notes: This figure plots separately for low- and high-SES students their applications to and enrollment in familiar majors by NCEE performances with 95% confidence intervals. Students' socioeconomic status is proxied by local GDP per capita in sub-figures (a) and (b) and by the formality of parental job sector in sub-figures (c) and (d). Students' NCEE performances are converted into 20 ventiles, with group 1 containing those who score between the 1st and 5th percentiles and group 20 containing those who score in the top 5% of the distribution. The sample includes the students who applied to four-year public colleges in Ningxia province between 2014 and 2018 from the ROL data. Sub-figures (a) and (c) plot the average share of familiar majors among each student's first-choice majors, and sub-figures (b) and (d) plot the share of students who enroll in familiar majors.

Figure F4: Sources of Knowledge by Major Familiarity



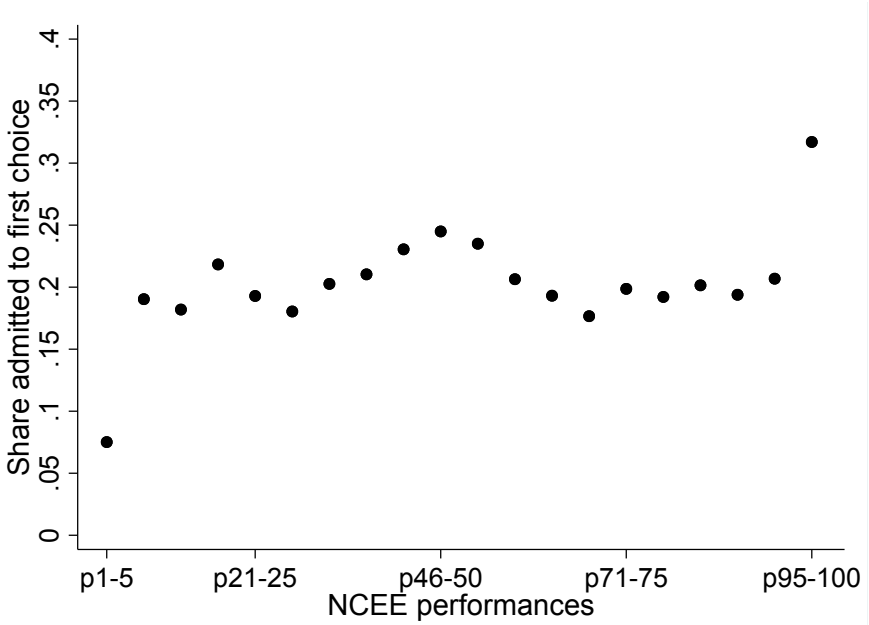
Notes: This figure plots the share of students who reported having heard about familiar and unfamiliar majors through four channels—family members, media and the Internet, high school peers, and high school teachers—in the survey experiment and presents the 95% confidence intervals.

Figure F5: Cumulative Distribution of the Fraction of Colleges on the ROL with the Same First-Choice Major



Notes: This figure plots the cumulative distribution of the fraction of colleges on the ROL with the same first-choice major, where a major is re-defined as major category by familiarity. The sample is the model estimation sample that includes the students who applied to public colleges in Ningxia province in 2018 and chose majors from one of the three major categories (Art & Humanities, Law & Politics, and Science) as the first-choice major of the first-choice college.

Figure F6: Share of Students Admitted to the First-Choice Major of the First-Choice College by Exam Performances



Notes: This figure plots the share of students admitted to their first-choice major of the first-choice college against their NCEE performances. The sample includes the students who applied to four-year public colleges in Ningxia province between 2014 and 2018 from the ROL data. The NCEE performances are converted to 20 ventiles by ranking students within the same year–academic track. Group 1 contains those who score between the 1st and 5th percentiles and group 20 contains those who score in the top 5% of the distribution.

Table F1: Employment Gap between Familiar and Unfamiliar Majors

	(1)	(2)	(3)	(4)
Familiar Major	-0.0703***	-0.0160	-0.0102	-0.0207
	(0.0156)	(0.0218)	(0.0217)	(0.0252)
Familiar × High-SES				0.0253
				(0.0314)
Student Char			x	x
Category FE		x	x	x
College FE		x	x	x
Mean(Y)	0.753	0.753	0.764	0.764
N	16,490	16,476	13,040	13,040

Notes: This table presents the gap in the likelihood of receiving a job offer among students graduating with familiar and unfamiliar majors. The sample includes the students graduating from four-year public colleges who plan to enter the labor market straight after college. The outcome variable is whether the student reports having received a job offer at the time of the survey, and the key independent variable is whether the student studies a familiar major in college. I additionally include major category and college fixed effects in column (2) and control for student characteristics in column (3). The student characteristics include year–academic track fixed effects, gender, ethnic group, eligible bonus points, and the NCEE performances, as well as the student's log household income, GPA ranking in college, and college tuition. In column (4), I add the interaction term between major familiarity and the student's socioeconomic status. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F2: Summary Statistics of the Student Samples: ROL Data, National Data, and CCSS Data

Panel A: ROL Data			
	Full Sample	Urban Students	Rural Students
Share female	0.552	0.551	0.552
Share ethnic minority	0.334	0.245	0.416
Share science-track	0.732	0.728	0.735
Avg bonus points	6.576	4.693	8.304
Avg NCEE score (in percentiles)	79.264	82.218	76.553
Share choosing familiar majors	0.194	0.154	0.231
Share admitted to familiar majors	0.170	0.152	0.187
Number of students	117,714	56,335	61,379
Panel B: National Data			
	Full Sample	Urban Students	Rural Students
Share female	0.400	0.454	0.340
Share ethnic minority	0.069	0.072	0.067
Share science-track	0.650	0.643	0.659
Avg bonus points	1.136	1.124	1.151
Avg NCEE score (in percentiles)	80.560	79.727	81.502
Share admitted to familiar majors	0.199	0.185	0.215
Number of students	5,901,726	3,129,900	2,771,826
Panel C: CCSS Data			
	Full Sample	Urban Students	Rural Students
Share female	0.446	0.483	0.415
Share ethnic minority	0.051	0.060	0.044
Share science-track	0.664	0.622	0.702
Avg bonus points	0.676	0.804	0.595
Avg NCEE score (in percentiles)	58.207	57.418	59.199
Share admitted to familiar majors	0.157	0.143	0.167
Number of students	36,672	16,345	18,443

Notes: The table presents the summary statistics of the student samples constructed using the ROL data, national data, and CCSS data. It reports the sample average of the key variables separately for the full sample, urban students, and rural students.

Table F3: Size and Share of Familiar Majors by Major Categories

	National Data		ROL Data	
Major Category	# of Enrollees	Share Familiar	# of Applicants	Share Familiar
History	47,281	0.938	6,744	0.957
Art & Humanities	575,213	0.744	62,257	0.708
Science	757,811	0.682	79,851	0.752
Law & Politics	309,920	0.219	33,023	0.287
Engineering	2,267,880	0.053	471,768	0.034
Education	89,644	0.006	20,033	0.000
Agriculture	162,257	0.002	19,097	0.002
Economics	351,088	0.000	88,191	0.017
Healthcare	442,167	0.000	89,724	0.001
Management	864,840	0.000	185,347	0.000

Notes: The table presents the size and the share of familiar majors for each major category. The left panel reports the number of students who enroll in majors within each major category and share of students who enroll in familiar majors within each major category for the sample of students who attend four-year public colleges in the national data. The right panel reports the number of students who apply to majors within each major category and share of students who apply to familiar majors within each major category for the sample of students who apply to four-year public colleges in the ROL data.

Table F4: Applications to Familiar Majors by Urban–Rural Status, Alternative Measures of Major Familiarity

	(1)	(2)	(3)	(4)	(5)
Panel A: Narrowly Defined Major Familiarity					
Rural	0.0199***	0.0219***	0.0185***	0.0183***	0.0328***
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0017)
Position FE	x	x	x	x	First Only
Student Char		x	x	x	x
NCEE Score			x	x	x
Coll-Maj Char				x	x
Mean(Y)	0.086	0.086	0.086	0.086	0.126
N	2,457,800	2,457,800	2,457,800	2,457,800	159,545
Panel B: Continuously Defined Major Familiarity					
Rural	0.0186***	0.0180***	0.0156***	0.0150***	0.0296***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0012)
Position FE	x	x	x	x	First Only
Student Char		x	x	x	x
NCEE Score			x	x	x
Coll-Maj Char				x	x
Mean(Y)	0.083	0.083	0.083	0.083	0.101
N	2,457,800	2,457,800	2,457,800	2,457,800	159,545

Notes: This table presents the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student, except in column (5), where the sample is restricted to the first-choice major of the first-choice college for each student. The outcome variable in panel A is whether the listed major is classified as a familiar major according to the narrower definition, and the outcome variable in panel B is the continuous index of major familiarity. The key independent variable is whether the student is a rural resident, as a proxy for socioeconomic status. I additionally control for position fixed effects, student characteristics, NCEE score, and the characteristics of the listed college–major. The student characteristics include year–academic track fixed effects, gender, ethnic group, and eligible bonus points. The NCEE score is converted to dummy variables for each score percentile. The characteristics of the college–major include the tuition level and selectivity, measured by the admission cutoff in the previous year. Standard errors (in parentheses) are clustered at the individual level.

Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F5: Applications to and Enrollment in Familiar Majors by Alternative Measures of SES

	(1)	(2)	(3)	(4)	(5)
Panel A: County Level Log GDP per Capita, ROL Data					
Log GDP Per Capita	-0.0234***	-0.0238***	-0.0177***	-0.0173***	-0.0330***
	(0.0008)	(0.0007)	(0.0007)	(0.0007)	(0.0017)
Position FE	x	x	x	x	First Only
Student Char		x	x	x	x
NCEE Score			x	x	x
Coll-Maj Char				x	x
Mean(Y)	0.146	0.146	0.146	0.146	0.188
N	2,457,800	2,457,800	2,457,800	2,457,800	159,545
Panel B: Parental Job Sector, ROL Data					
Informal Sector	0.0351***	0.0343***	0.0258***	0.0252***	0.0412***
	(0.0011)	(0.0009)	(0.0009)	(0.0009)	(0.0023)
Position FE	x	x	x	x	First Only
Student Char		x	x	x	x
NCEE Score			x	x	x
Coll-Maj Char				x	x
Mean(Y)	0.146	0.146	0.146	0.146	0.188
N	2,457,800	2,457,800	2,457,800	2,457,800	159,545
Panel C: Log Household Income, CCSS Data					
Log HH Income	-0.0294***	-0.0160***	-0.0150***	-0.0142***	
	(0.0033)	(0.0033)	(0.0032)	(0.0032)	
Student Char		x	x	x	
NCEE Score			x	x	
Coll-Maj Char				x	
Mean(Y)	0.161	0.161	0.161	0.161	
N	27,799	27,799	27,799	27,799	

Notes: Panels A and B of this table present the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student, except in column (5), where the sample is restricted to the first-choice major of the first-choice college for each student. The outcome variable is whether the listed major is classified as a familiar major according to the preferred definition. The key independent variable is county-level log GDP per capita in panel A and the informality of parental job sector in panel B, both as proxy for socioeconomic status. I additionally control for position fixed effects, student characteristics, NCEE score, and the characteristics of the listed college–major. The student characteristics include year–academic track fixed effects, gender, ethnic group, and eligible bonus points. The NCEE score is converted to dummy variables for each score percentile. The characteristics of the college–major include the tuition level and selectivity, measured by the admission cutoff in the previous year. Standard errors (in parentheses) are clustered at the individual level. The sample of panel C is the students who attend four-year public colleges from the CCSS data. The outcome variable is whether the student enrolls in a familiar major in college, and the key independent variable is the self-reported log household income. I control for the same set of covariates as in panels A and B, except for the selectivity of the college–major. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F6: Applications to Familiar Majors by Urban–Rural Status, First-Choice Majors

	(1)	(2)	(3)	(4)
Panel A: First-choice Majors Only				
Rural	0.0593***	0.0605***	0.0497***	0.0484***
	(0.0016)	(0.0015)	(0.0015)	(0.0015)
Position FE	x	x	x	x
Student Char		x	x	x
NCEE Score			x	x
Coll-Maj Char				x
Mean(Y)	0.181	0.181	0.181	0.181
N	582,716	582,716	582,716	582,716
Panel B: First-choice Majors of First-choice Colleges Only				
Rural	0.0728***	0.0725***	0.0571***	0.0557***
	(0.0021)	(0.0020)	(0.0021)	(0.0021)
Student Char		x	x	x
NCEE Score			x	x
Coll-Maj Char				x
Mean(Y)	0.188	0.188	0.188	0.188
N	159,545	159,545	159,545	159,545

Notes: This table presents the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student. The sample in panel A is restricted to the first-choice majors on the rank-order lists, and the sample in panel B is restricted to the first-choice majors of the first-choice colleges on the rank-order lists. The outcome variable is whether the listed major is classified as a familiar major according to the preferred definition, and the key independent variable is whether the student is a rural resident, as a proxy for socioeconomic status. I additionally control for position fixed effects (in panel A), student characteristics, NCEE score, and the characteristics of the listed college–major. The student characteristics include year–academic track fixed effects, gender, ethnic group, and eligible bonus points. The NCEE score is converted to dummy variables for each score percentile. The characteristics of the college–major include the tuition level and selectivity, measured by the admission cutoff in the previous year. Standard errors (in parentheses) are clustered at the individual level.

Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F7: Applications to Familiar Majors by Urban–Rural Status, Controlling for Major Category and College Fixed Effects

	(1)	(2)	(3)	(4)
Rural	0.0316***	0.0182***	0.0103***	0.0096***
	(0.0008)	(0.0004)	(0.0006)	(0.0004)
Position FE	x	x	x	x
Student Char	x	x	x	x
NCEE Score	x	x	x	x
Coll-Maj Char	x	x	x	x
Category FE		x		x
College FE			x	x
Mean(Y)	0.146	0.146	0.146	0.146
N	2,457,800	2,457,800	2,457,799	2,457,799

Notes: This table presents the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student. The outcome variable is whether the listed major is classified as a familiar major according to the preferred definition, and the key independent variable is whether the student is a rural resident, as a proxy for socioeconomic status. In addition to position fixed effects, student characteristics, NCEE score, and the characteristics of the listed college–major, I include major category fixed effects in column (2), college fixed effects in column (3), and both in column (4). Standard errors (in parentheses) are clustered at the individual level.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F8: Applications to Familiar Majors by Urban–Rural Status, Controlling for County and High School Fixed Effects

	(1)	(2)	(3)
Rural	0.0316***	0.0269***	0.0237***
	(0.0008)	(0.0009)	(0.0016)
Position FE	x	x	x
Student Char	x	x	x
NCEE Score	x	x	x
Coll-Maj Char	x	x	x
County FE		x	
HS FE			x
Mean(Y)	0.146	0.146	0.163
N	2,457,800	2,457,800	1,043,164

Notes: This table presents the estimates of equation (1) using the rank-order lists of the students who apply to four-year public colleges in the ROL data. The unit of observation is each position on the rank-order list of each student. The outcome variable is whether the listed major is classified as a familiar major according to the preferred definition, and the key independent variable is whether the student is a rural resident, as a proxy for socioeconomic status. In addition to position fixed effects, student characteristics, NCEE score, and the characteristics of the listed college–major, I include county fixed effects in column (2) and high school fixed effects in column (3). Because students' high school information is available only for years between 2017 and 2018, the sample size is smaller in column (3). Standard errors (in parentheses) are clustered at the individual level.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F9: Enrollment in Familiar Majors by Urban–Rural Status, National Data

	(1)	(2)	(3)	(4)
Rural	0.0301***	0.0426***	0.0386***	0.0386***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Student Char		x	x	x
NCEE Score			x	x
Coll-Maj Char				x
Mean(Y)	0.199	0.199	0.199	0.199
N	5,901,726	5,901,726	5,901,726	5,901,726

Notes: This table presents the estimates of equation (1) using the sample of students who attend four-year public colleges in the national data. The outcome variable is whether the student enrolls in a familiar major according to the preferred definition, and the key independent variable is whether the student is a rural resident, as a proxy for socioeconomic status. I additionally control for student characteristics, NCEE score, and the selectivity of the enrolled college–major. The student characteristics include province–year–academic track fixed effects, gender, ethnic group, and eligible bonus points. The NCEE score is converted to dummy variables for each score percentile. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F10: Difference in Major Competitiveness by Familiarity

	(1) # of App to Quota	(2) Adm Quota Filled	(3) Admission Cutoff
Familiar Major	-1.5310*	-0.0093	0.6227
	(0.8569)	(0.0090)	(0.4132)
Category FE	x	x	x
College FE	x	x	x
Mean(Y)	24.053	0.940	491.178
N	17,808	17,808	17,524

Notes: This table presents the difference in competitiveness between familiar and unfamiliar majors. The sample includes the college–major pairs in four-year public colleges between 2016 and 2018 from the ROL data, as the admission quota is available for the period between 2016 and 2018. The outcome variables in columns (1)–(3) are the number of applicants relative to admission quota, the percentage of the admission quota filled, and the admission threshold, and the key independent variable is whether the major is classified as a familiar major according to the preferred definition. I additionally include year–academic track, application round, major category, and college fixed effects. The regression is weighted by the admission quota of the major. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F11: SES Disparity in Risk Attitudes, Survey Data

	(1) Will Not Apply	(2) Feel Concerned	(3) Hard to Succeed
High SES	0.0329	-0.0601	-0.0120
	(0.0384)	(0.0381)	(0.0310)
Mean(Y)	0.357	0.354	0.187
N	673	673	673

Notes: This table presents the difference in risk attitudes between high- and low-SES students using the sample of students from the survey experiment. The outcome variables in columns (1)–(3) are their willingness to apply to high-risk majors, the level of concern regarding risks associated with choice of major, and their perception of the long-term consequences of choosing high-risk majors. The key independent variable is the student's socioeconomic status measured by parental education level. I additionally control for the student's gender, urban–rural status, academic track, and high school entrance exam performances. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.

Table F12: Changes in Major for Graduate School, CCSS Data

	(1) Full Sample	(2) High SES	(3) Low SES
Familiar Major	0.1102**	-0.0074	0.1350**
	(0.0464)	(0.0398)	(0.0532)
Familiar × High SES	-0.0971*		
	(0.0548)		
Mean(Y)	0.335	0.357	0.305
N	6,143	3,188	2,862

Notes: This table presents the difference in the likelihood of switching to a different major in graduate school between the students who study familiar and unfamiliar majors in college. The sample includes the students who pursue graduate school after completing their undergraduate degree from the CCSS data. The outcome variable is whether the student switches to a different major in graduate school. The key independent variable is whether the undergraduate major is classified as a familiar major according to the preferred definition. In column (1), I also control for the student's socioeconomic status and its interaction with major familiarity. In columns (2) and (3), I conduct the regression separately for the high- and low-SES students, where the student's socioeconomic status is measured by whether the household income is above median. I additionally control for the student's gender, ethnic group, eligible bonus points, NCEE performances, and province–year–academic track fixed effects, as well as major category and college fixed effects. Robust standard errors are reported in parentheses.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, and ***: $p < 0.01$.