Place-Based College Admission, Migration and the Spatial Distribution of Human Capital: Evidence from China

Yu (Alan) Yang*

University of Wisconsin-Madison

April 8th, 2021

Abstract

How to allocate college seats across regions is an important but often controversial question. It may imply a policy tradeoff between efficiency in human capital production and equality of opportunities for people growing up in different places. Regional disparity in college access also impacts the spatial distribution of college-educated workers and thus affects regional inequality in development. This paper studies the province-based college admission quotas in China and evaluates potential reforms. Despite longstanding concerns about the unequal quota allocation across provinces, little is known because China’s college entrance exam is not comparable across provinces. Combining individual-level administrative data and college student surveys, I estimate a structural model of college and migration choice under quota constraints, together with a measurement model that can recover the nationally comparable distribution of pre-college human capital in each province. I find substantial gaps of pre-college human capital among college applicants across provinces, but these gaps are not well reflected by the admission quotas. A purely merit-based nationwide admission without provincial quotas increases efficiency at the cost of larger regional inequality: it increases college-student sorting and total human capital output during college, but decreases the college opportunity in less developed provinces and the share of college graduates working in those provinces.

*Email: alan.yang@wisc.edu. I am deeply grateful to Jeff Smith, Chris Taber and Matt Wiswall for their guidance and encouragement. I thank Naoki Aizawa, Peter Arcidiacono, Chao Fu, Long Hong, John Kennan, Jingnan Liu, Corina Mommaerts, Xiaoyang Ye, Anson Zhou, and seminar/conference participants at UW-Madison, Peking GSM, Jinan IESR, CUHK Shenzhen, AASLE 2020 Job Market Conference for helpful comments and suggestions. I am grateful to Xiaoyang Ye, Jinghuan Shi, Fei Guo and Lan Duo for their help during data application. All errors are mine. [Please click here for the latest version.]
1 Introduction

Where people grow up is one of the most salient factors affecting opportunities and success later in life (Chetty et al., 2014), and the difference in college access across areas is found to play an important role (Hillman, 2016; Klasik et al., 2018; Ishimaru, 2020). College admission in many countries, at least to some extent, is based on residence. For example, public universities in the US charge much higher tuition to out-of-state students and, in some states, impose out-of-state enrollment caps or guarantee admission to high-achieving in-state students.1 In China, colleges directly set admission quotas for each province. However, how best to allocate college seats across regions is often a difficult policy question. Should college seat allocation be purely merit-based or equalized across places? When there exists regional inequality in economic development and pre-college education quality, a purely merit-based college admission may worsen the college opportunities faced by students in poor regions and decrease the intergenerational mobility.

This paper studies the spatial variation in college access in China. This issue is particularly important in China because China’s spectacular economic growth has been accompanied by a substantial increase in regional inequality (Kanbur and Zhang, 2005; Xu, 2011; Zhang and Zou, 2012). At the same time, China has one of the largest college markets in the world. In each year, about ten million Chinese students take the National College Entrance Examination (NCEE, or Gaokao in Chinese) to compete for three million four-year college seats (Chen and Kesten, 2017). China’s college admission is place-based. The current pre-college education system is decentralized at the provincial level and the education authority in each province administers their own version of the NCEE. As a result, unlike the college entrance exams in many countries (e.g. the SAT or ACT in the US), China’s NCEE is not nationally comparable. To allocate college seats without the ability to directly compare students across provinces, China uses a provincial quota system, in which each college is allowed to set an admission quota (i.e. how many students to admit) for every province in each year. Each province then runs an independent centralized admission system based on its NCEE scores.

One natural question is how well the current provincial quota system performs. A longstanding

---

1For example, the University of California System has an 18% cap, and the University of Virginia has a one-third cap on out-of-state students. Public college systems in Texas and California offer guaranteed admission to in-state high school students if they are ranked above certain thresholds in the graduating classes.
concern about this admission system is the observed large variation and potential unfairness of quota allocations across provinces (e.g. Fu, 2013; Guo et al., 2018). The average number of total four-year college seats received by each province ranged from 0.21 to 0.49 per NCEE exam taker during 2006-2011, and this large variation has persisted to the present. If admission quotas are allocated based purely on merit, the marginal student admitted from each province should have the same achievement level. However, the non-comparability of the NCEE scores makes it difficult to compare students across provinces, thus it is not clear whether the observed quota variation is driven by the difference in student quality.

Importantly, the "quality" of college seat allocation depends on policy objective functions. There may exist a policy tradeoff between maximizing human capital production and reducing regional inequality in college opportunities. Moreover, in China’s college admission system, a quota is not simply a college seat, but an assignment of students from their home province to a college location, potentially in a different province. Students value both college quality and location (Du and Zhong, 2019). College location is particularly relevant for Chinese students given the migration restrictions under China’s hukou system. Many studies suggest that college location influences subsequent work location choices (Groen, 2004; Sjoquist and Winters, 2014; Kennan, 2018; Huang et al., 2019). Therefore, the college quota allocation also affects the future spatial distribution of high-skilled workers, and human capital distribution plays an important role in economic growth and regional inequality (Fleisher et al., 2010; Ha et al., 2016).

This paper is the first to examine how the college admission quota allocation shapes both the college choice and migration decision of millions of college students in China every year, and what the implications are if the current province-based admission using quotas is replaced by a purely merit-based nationwide admission. Despite its importance, these questions remain largely unexplored due to a lack of data to track college students both before and after college, and a clear method to compare student quality across provinces. There is a small group of recent studies on

---

2China’s college capacity keeps expanding over years, but the variation in admission quota allocation is not reduced. In 2020, the total number of four-year college seats received by each province ranged from 0.35 to 0.76 per NCEE exam taker, according to the China Higher-Education Student Information platform (http://www.chsi.com.cn), an official website run by the Ministry of Education of China.

3The hukou system in China is an official household residency registration system. It grants access to major local public services, such as education, medical insurance and other social security programs. Location change of the hukou, especially across provinces, is restricted under China’s longstanding internal migration control. For more details of the hukou system, see Song (2014) and Colas and Ge (2019).

4Most of the literature on China’s college market focus on 1) the matching mechanism and its impact on mismatch
China’s college quota allocation. Guo et al. (2018) overcome the non-comparability of the NCEE scores by giving a standardized test to a sample of college students in six provinces. They find a solely merit-based admission would significantly change the current regional student body composition in their sample. Chan et al. (2019) estimate an equilibrium model of the quota-setting game among the top-100 colleges in China. By analyzing the quota-setting behaviors, they find that elite colleges systematically prefer students from richer provinces with better pre-college education. Pu (2020) analyzes the college admission in three Chinese provinces and shows that pooling provincial quotas increases average student welfare if the student quality distribution is the same across provinces.

This paper makes three main contributions compared to the existing studies. First, I compile a novel dataset by combining individual-level administrative data of all Chinese college students admitted during 2006-2011 with college student survey data for the same six cohorts. This dataset tracks student choices in college admission, during college studies and after graduation, and contains detailed student and college characteristics. It permits an analysis of the college admissions of all four-year colleges in all provinces. This has not been done in previous studies but is crucial for analyzing the distributional effect in college and migration choices at the national level.

Second, I build and estimate a measurement model to directly compare college applicants across provinces. The measurement model utilizes the information from both the NCEE scores and college GPAs in a dynamic latent factor framework (Cunha and Heckman, 2008; Cunha et al., 2010). It enables me to recover the distribution of pre-college human capital of students in each province that is comparable nationwide. I find substantial differences in students’ pre-college human capital distribution across provinces; the difference of the mean between the top and bottom provinces is nearly 1.3 national standard deviation. Moreover, these differences are not well aligned with the observed admission quota allocation. To the best of my knowledge, this paper is the first study to provide a comparable skill measure for all college applicants in China. Compared to the approach in Guo et al. (2018) that administers a separate test to a sample of colleges and provinces, the measurement model developed in this paper is easier and less costly to implement.

(Chen and Kesten, 2017; Wu and Zhong, 2014; Bo et al., 2019; Ha et al., 2019; Cao, 2020; Wu and Zhong, 2020), 2) the rural-urban inequality in college access (Li et al., 2015; Ye, 2018), and 3) the effect of college expansion since 1990s on various outcomes such as employment (Li et al., 2014; Ou and Zhao, 2016), labor market return (Huang et al., 2020), technology adoption and firm productivity (Che and Zhang, 2018; Feng and Xia, 2018), and economic growth (Li et al., 2017).
Third, given the comparable measure of student quality, I estimate a dynamic structural model of college choice under admission quota constraints and post-college migration choice. I contribute to the existing studies on college quotas (Guo et al., 2018; Chan et al., 2019; Pu, 2020) by allowing for heterogeneous individual preferences and highlighting the distributional effect of the place-based quotas on both college access and residential choice. I then use the model to evaluate a counterfactual merit-based nationwide admission without provincial quotas. The merit-based admission increases student-college sorting and, as a result, increases total human capital output by 1.7% as measured by initial earnings, and increases the average student welfare by 1.2%. However, the merit-based admission decreases college opportunities in less developed provinces by 10% and decreases the share of college graduates working in those provinces by 14% when the quota constraint on where for college is removed. This finding highlights a policy tradeoff between college education efficiency and regional inequality. In the counterfactual analysis, the existing quota allocation is exogenously changed from a social planner’s perspective. In this sense, it complements the general equilibrium approach in Chan et al. (2019) which adopts a decentralized perspective and focuses on colleges’ behavior in a game-theoretical framework.

This paper also contributes to several strands of literature in addition to those directly about China’s college admission. First, it is related to the broad literature on place-based policies and their efficiency implication (Glaeser and Gottlieb, 2008; Busso et al., 2013; Kline and Moretti, 2014; Lu et al., 2019; Fajgelbaum and Gaubert, 2020). Recent work has started to explicitly examine the equity-efficiency tradeoff in these place-based policies (Gaubert et al., 2020). I contribute to this new direction in the literature by studying the effect of a place-based education policy on human capital production efficiency and regional inequality. Second, this paper contributes to the literature on affirmative action policies in college admission, which is primarily race-based, and the debate over the quality-fit tradeoff in these policies. See Arcidiacono and Lovenheim (2016) for a recent review. I provide place-based evidence and show that the consideration of inequality in college access may have implications for not only individual students but also regions. Third, this paper connects the college choice literature with the literature on migration. Existing studies that link the two have worked on the effect of local government subsidies for college (Kennan, 2018), out-of-state tuition distortion (Knight and Schiff, 2019), and state-based merit aids (Sjoquist and

---

5Barrow et al. (2020) study a place-based affirmative action policy in Chicago’s high school admission.
Winters, 2014; Fitzpatrick and Jones, 2016). I complement these studies by examining the effect of a quantity-based policy (i.e. admission quotas) instead of prices on the interplay between college choice and migration.

The remainder of the paper proceeds as follows: Section 2 introduces the institutional background and provides some basic data patterns of college choice and migration. Section 3 presents the structural model of college and residential choices. Section 4 presents the measurement model that recovers the comparable distribution of pre-college human capital. Section 5 describes data, estimation strategy and procedure. Estimation results are presented in Section 6. Section 7 examines counterfactual college seat allocations and their impacts. Section 8 concludes.

2 Institutional background and basic data patterns

2.1 College admission and provincial quotas in China

There are 2,359 postsecondary education institutions in China as of 2011 (Ministry of Education, 2011). 1,131 of them are four-year colleges, most of which are public schools. The rest are three-year vocational and technical schools. Four-year colleges are classified into three tiers by the Ministry of Education based on school quality and schools in higher tiers have priority in admitting students (Bo et al., 2019; Jia and Li, 2020). The top tier is generally considered as elite colleges. Most of them belong to the Project-211, which was initiated in 1995 by the Ministry of Education with generous funding and policy support to raise the education and research standards of high-level universities. The second tier includes non-elite public colleges that are administrated by provincial governments but are still widely recognized and are able to recruit students nationally. Schools in the third tier mostly admit students from their own province or surrounding provinces, and many of them are private schools. College tuition in China is fixed by the government at an affordable level to most families and does not vary by student’s residence location.

All four-year colleges and three-year schools admit students through the National College En-
trance Examination (NCEE) based solely on the NCEE scores.\textsuperscript{8} The NCEE in most provinces has two tracks based on high school curriculum: sciences and humanities. Students in both tracks take the NCEE tests on Chinese, English, and Mathematics (the difficulty level is lower for the humanities track). In addition, students in the sciences take tests on Physics, Chemistry, and Biology, while students in the humanities take History, Geography, and Political Science.\textsuperscript{9} Students specialized in arts or sports follow a different admission procedure and are not studied in this paper. Since 2001, the provincial education authority is allowed to design their own high school curriculum. As a result, the NCEE tests for each track and subject are also designed and administrated at the provincial level, and are not comparable across province, track and year. Some provinces do not have sufficient resources to develop their own NCEE tests for some or all subjects, in which case the Ministry of Education provides a set of general versions for those provinces to choose from. Even in these cases, the grading is still organized by each province independently.

After the NCEE is held in June every year, each province runs an independent centralized college admission. Most provinces adopt a hybrid of the Boston and Deferred Acceptance mechanism with a restricted application list, usually allowing for 6-8 schools (Chen and Kesten, 2017; Bo et al., 2019). Students also provide a preference list of 3-5 college majors within each school. The admission algorithm operates at the school level. According to the listed ranking, the algorithm sends students to a school as long as there are still available seats in any major. After receiving them, the school assigns students to a major given the NCEE score and the listed major preference. If a student’s proposed majors cannot be fulfilled, the school assigns her to other available majors following some school-specific rules.\textsuperscript{10}

China’s college admission features a province-based admission quota system (Kang, 2005). In each year, the Ministry of Education coordinates with the National Development and Reform Commission to determine how many students each college can admit during the next NCEE. Each school then allocates their new college seats (i.e. admission quotas) to various provinces and tracks before the NCEE. Almost all colleges allocate quotas to more than one province. Top-tier

\textsuperscript{8}Some top colleges are also allowed by the Ministry of Education to admit students through self-organized tests and interviews, but the scale is small. Less than 2% of college students were admitted through this channel during the period studied in this paper.

\textsuperscript{9}A few provinces do not separate tracks in high school curriculum and their NCEE has only one track.

\textsuperscript{10}Students can refuse the college major assignment outside their listed majors when submitting the application. In this case, they will be returned by the school and will not be considered by other schools in the same round of admission. Given this large risk, students are generally advised not to refuse the major adjustment.
colleges generally recruit students from all provinces, while bottom-tier colleges mainly recruit students from the local and surrounding provinces. All admission quotas are defined at both the school and major levels. For example, Tsinghua University may announce to admit 10 students into the Economics major from the humanities track in Guangdong Province. The quotas received by a province are a set of available college seats (located in various provinces) that only opens to students in that province. It determines not only the number of students that can go to college, but also the distribution of student migration flows into each province for college studies. In general, each college has autonomy in allocating quotas, but both the central and provincial governments are able to influence such decisions (Pan and Cui, 2017).

2.2 Data patterns of quotas, college choice and migration

Figure 1 shows the overall allocation of admission quotas across provinces. Each bar represents the total number of the four-year college admission quotas received by each province divided by the number of NCEE exam takers in that province, averaged between 2006 and 2011. In provinces ranked among the top, more than 45% of the NCEE exam takers have a chance to go to a four-year college. This number substantially decreases to lower than 25% among provinces at the bottom.

The spatial disparity of college access is positively correlated with the economic development, as shown in the top panel of Figure 2. The bottom panel of Figure 2 shows that the geographical distribution of colleges is also skewed towards the more economically developed provinces. Almost all second- and third-tier colleges in China are governed and funded by the provincial government. Even the top-tier schools, which are governed by the central government, often receive funds and favorable policies (e.g. free land) from the local government of the place where they are located (Pan and Cui, 2017). As a result, the local government can potentially have a large influence in determining how to allocate admission quotas, and on average a large share of the capacity is kept for local students. Between 2006 and 2011, first-tier colleges on average allocated 45% of their quotas to the local province, and the second- and third-tier allocated 75% and 81% to the local province, respectively. In this paper, I do not consider who makes the decision of the quota allocation, but rather take the allocation as given when modeling student choices. In the counterfactual analysis, I change the quota system exogenously from a social planner’s perspective.

The quota allocation determines the migration flow of students from each province at the col-
lege stage. As shown by previous studies (Huang et al., 2019; Sun et al., 2020), where students go for college is correlated with the subsequent work location choice after graduation. Table 1 shows that, between 2006 and 2011, among the 69.1% of students who attended a college in their home province, 74.1% stayed after graduation and 25.9% migrated to another province. Among the rest who attended a college outside their home province, 29.0% moved back home after college, 33.0% stayed at the college location, and the rest migrated to a third province. This suggests that students not only have home attachment but also seem to prefer the place where they live and study for four years during college. The structural model introduced in the next section will carefully model these features and try to understand how students make college and residential choices under the admission quota constraints.

3 The structural model of college and residential choices

This section develops a dynamic discrete choice model of college choice and subsequent residential location choice after graduating from college. The quality distribution of college applicants in each province is taken as given in the structural model. The next section will discuss how to recover these distributions using a measurement model and how the structural model developed in this section is integrated with the measurement model in estimation. The structural model has three stages.

Stage 1: High school graduates apply to four-year colleges in the centralized admission in their residence province, subject to provincial admission quota constraints, and attend the one that admits them.

Stage 2: At college graduation, students choose to either go to graduate school or enter the labor market. If choosing graduate school, subsequent choices are not tracked, and the model ends immediately with a terminal value. If choosing to enter the labor market right after college, students move to stage 3.

Stage 3: College graduates choose a province in which to live and work.
3.1 Primitives

Denote the set of all provinces as $\mathcal{N}$. Each province has a cohort of high school graduates with a different size and pre-college human capital distribution, and a set of four-year colleges with different quality and capacity. Denote the set of all four-year colleges in the country as $\mathcal{S}$. To reduce the notational burden, individual subscripts on all variables are suppressed.

3.1.1 Initial condition

Consider one cohort of high school graduates in year $t$. Each individual enters the model at the end of high school with the following initial conditions:

$$\Omega = \{ h, z, \theta^0, NCEE_{hzt}, \tau \}.$$

Each individual takes the NCEE of track $z \in \{\text{sciences, humanities}\}$\footnote{Track $z_i$ was chosen two years before taking the NCEE, and is taken as exogenous in the model.} in her home province $h \in \mathcal{N}$ and year $t$. The test score $NCEE_{hzt}$ is a function of her pre-college human capital $\theta^0$ plus a measurement error. Denote the distribution of $\theta^0$ in each province as $F_h(\theta^0)$. $F_h(\theta^0)$ and its relationship with $NCEE_{hzt}$ will be specified in detail in the measurement model in Section 4. $\theta^0$ is known to individuals but not to the econometrician.

The other heterogeneity that is known to individuals but not to the econometrician is the idiosyncratic location preference $\tau$. Each province has a common amenity value, but an individual’s preference may deviate from these common values. $\tau$ captures this idiosyncratic deviation. Allowing for this unobserved location preference to be persistent across model stages is potentially important for isolating the causal effect of college location on subsequent residential choices. $\tau$ is independent from $\theta^0$ and follows a finite mixture distribution of discrete types: $\tau \in \{\emptyset\} \cup \mathcal{N}$. To keep the structural model tractable, I impose that the first type, $\tau = \emptyset$, does not have preference for any province, while each of the remaining types has a preference for the corresponding province in $\mathcal{N}$ but no preference for the other provinces. Students may have such a location preference if they have family connections in a certain place. They may also establish an attachment to a location either because they have been there previously or due to other reasons, but they may not necessar-
ily have any particular knowledge about elsewhere. Denote the type probability for \( \tau = \emptyset \) as \( \pi_\emptyset \). Each of the remaining types has an equal type probability \((1 - \pi_\emptyset)/N\), where \( N \) is the number of provinces in the set \( \mathcal{N} \). The magnitude of this idiosyncratic preference is \( \chi \) for college location and \( \rho \chi \) for work location after college, where \( \rho \) is a scalar. \( \chi \) and \( \rho \) are constant across provinces.

### 3.1.2 Flow payoff of college

The flow payoff of attending college \( s \in S \) located in province \( j_s \in \mathcal{N} \) is:

\[
  u^c_s = Z_s' \gamma^c_1 + D(j_s, h)' \gamma^c_2 + \gamma^c_{3,j_s} + \chi 1(j_s = \tau) + \xi^c_s,
\]

where \( Z_s \) includes a college quality measure \( V_s \), college eliteness (Project-211 or not), college type (comprehensive, STEM-, humanities-, or economics-focused college) and an indicator of being located in the capital city of a province. \( D(j_s, h)' \gamma^c_2 = \gamma^c_{21} 1(j_s = h) + \gamma^c_{22} d(j_s, h) + \gamma^c_{23} d(j_s, h)^2 \), where \( d(\cdot, \cdot) \) is the distance (in 1,000 km) between two provinces. The first term is an indicator for attending college in the home province, so \( \gamma^c_{21} \) captures home preference. The quadratic distance terms capture both migration cost (monetary and psychological) and preference for adjacent places that share similar climate and culture. \( \gamma^c_{3,j_s} \) is the provincial fixed effect and represents the common amenity value in college location \( j_s \). The idiosyncratic preference has two parts. The first part is the individual’s location preference for province \( j_s = \tau \).\(^{12}\) The second part, \( \xi^c_s \), is an idiosyncratic preference for college \( s \) and follows the Type-I extreme value distribution with standard deviation \( \sigma^c \).

I assume all students prefer four-year college over any outside options, so the outside option is not modeled. College major is also abstracted away. As discussed in Section 2, it is possible to be admitted by a school even when the quota of the proposed major is already fulfilled, if students allow the school to assign them to other available majors.

---

\(^{12}\)An alternative approach to model unobserved location preference in the literature is to have an idiosyncratic scalar in front of the amenity level \( \gamma^c_{3,j_s} \) (Diamond, 2016; Huang et al., 2019). It imposes that the idiosyncratic location preference is proportional to the ranking of the common amenity values in each province, which is less flexible than the mixture distribution specification.
3.1.3 Life-time payoff of graduate school

After graduating from college \( s \), students can enter the labor market \((g = 0)\) or go to graduate school \((g = 1)\). If choosing to go to graduate school directly after college \((g = 1)\), the life-time payoff takes the following parsimonious form:

\[
\begin{align*}
    u^g_s &= \gamma^g_0 + \gamma^g_1 \theta^0 S + \gamma^g_2 V_s + \gamma^g_3 \theta^0 V_s + Z^g_s \gamma^g_4 + \xi^g_s
    \end{align*}
\]  

(3.2)

\( \xi^g \) follows the Type-I extreme value distribution with a standard deviation \( \sigma^g \). It will realize and be observed at graduation in stage 2 before students choose between working and graduate school. The capacity and admission of graduate schools are not modeled but are implicitly captured in Equation (3.2). To fit the data, the estimated Equation (3.2) should yield a lower payoff from graduate studies for students with lower human capital and from lower-quality colleges thus make them less likely to attend graduate school.\(^1\)

3.1.4 Life-time payoff of working after college graduation

Instead of graduate school, if the individual enters the labor market \((g = 0)\) and chooses to work in province \( k \in \mathcal{N} \), the life-time payoff is:

\[
\begin{align*}
    u^w_{sk} &= \sum_{t=1}^{T} \beta^{t-1} \ln C_t + D(k, j_s) \gamma^w_1 + D(k, h) \gamma^w_2 + \gamma^w_{3,k} + \rho \chi(k=\tau) + \xi^w_k
    \end{align*}
\]

(3.3)

The first term is the life-time utility from consumption, where the utility in each period equals the log period consumption. Assume college graduates have access to a perfect credit market so that they can borrow or save without restriction at a given interest rate.\(^2\) The consumption utility

---

\(^1\)Graduate school admission in China has two main channels. The first is the autonomous admission run by each school or department. It relies on recommendations and self-organized interviews and is mainly available to top students in the first-tier colleges. The second channel is the test-based national admission open to all students.

\(^2\)The financial relationship between young adults and their parents is typically very close in China. For example, many Chinese parents will help their children to buy their first house (Wei et al., 2017). Free borrowing and saving is also a common assumption used in the related literature. See, for example, Arcidiacono (2005) and Kennan and Walker (2011).
maximization implies the same consumption level in all post-college periods:

\[
\sum_{t=1}^{T} \beta^{t-1} C_t = E \left[ \sum_{t=1}^{T} \beta^{t-1} W_{skt} \right],
\]

(3.4)

\[
C_t = E \left[ \sum_{t=1}^{T} \beta^{t-1} W_{skt} \right] / \sum_{t=1}^{T} \beta^{t-1}, \ \forall t.
\]

(3.5)

\(W_{skt}\) is the earnings in province \(k\) in year \(t\) after graduation, and the expectation is taken with respect to the distribution of the future earning shocks.

The distance function \(D(k, j_s)\) in Equation (3.3) is intended to capture the causal effect of college location on residential choices. College location may matter for two main reasons. First, searching for jobs near the college location may have some labor market advantage (Huang et al., 2019). Students may gain more information on the local than the non-local labor market after spending four years studying and living there. It is also less costly to take internships and job interviews within a certain area near the college location. The second is the migration cost, which is broadly defined. People may get used to and like the place where they have spent four years living, and leaving the college location incurs both monetary and psychological migration costs, as well as a one-time cost of throwing away the location-specific capital. Given the effect of college location on residential choices, the relative attractiveness of the graduate school option may be higher for students attending a college in a less preferred location.

China’s current hukou policy in general gives fresh four-year college graduates a lot of freedom to choose their first residence location and obtain the local hukou,\(^{15}\) except in Beijing and Shanghai, which have a much stricter requirement when offering hukou to migrant workers (Zhang et al., 2019; Ge and Wu, 2020). Inside \(D(k, h)\), there is an extra interaction between an indicator function \(1(k=h)\) and an indicator for \(h\) being Beijing or Shanghai, to capture the special hukou restriction in these two places for migrants and thus the advantage of being a local student or resident. After college graduation, individuals can freely choose the working province \(k \in \mathcal{N}\). However, once

---

\(^{15}\)A common method to obtain a local hukou is through the local worker service center run by the government, especially for fresh college graduates who are employed by private firms and/or rent apartments. When they purchase local housing at a later time, they can change their hukou address from the worker service center to their own residence location.
residence is established after college, further inter-provincial migrations are abstracted away. This is consistent with China’s hukou and migration restriction on workers other than fresh college graduates (Zhang et al., 2019). An alternative interpretation is that starting residence and career in a province leads to a specific expected distribution of subsequent migration choices. In this distribution, most people are permanent stayers, as observed in the data, but some people, especially those in certain places like Beijing and Shanghai, may be more likely to leave later if they cannot obtain a local hukou. Such distributions associated with each initial residence choice are not explicitly modeled, but the expected life-time payoff based on these expected distributions are absorbed by the provincial fixed effect terms, $\gamma_{3,k}^w$.

Similar to the college stage, the idiosyncratic preference in Equation (3.3) also contains two parts. The first part captures the dynamic correlation in the location preference through the unobserved type $\tau$. Students may like a place and, therefore, choose it for both college and working. Alternatively, students may choose to study in one place just because they prefer to live there after college and attending college in that place can help them find a job there. Therefore, it is important to control for such selection in order to identify the causal effect of college location on work location choices. The second part of the idiosyncratic preference, $\xi_k^w$, follows the Type-I extreme value distribution with a standard deviation $\sigma_w$. $\xi_k^w$ is realized and observed in stage 3 before students choose their residential province $k$. Students do not observe $\xi_k^w$ in stage 2. $\xi_k^w$ and $\xi^g$ are assumed to be independent from each other for analytical tractability.

### 3.1.5 Earnings

Lastly, log earnings in province $k$ in year $t$ after graduation are specified as follows:

$$
\ln W_{skt} = \delta_0 + \delta_1 \theta^0_{\text{pre-college human capital}} + \delta_2 V_s + \delta_3 \theta^0_{\text{college quality}} V_s + g_{kt} + \varepsilon_{kt} \sim N(0,\sigma^2_\varepsilon) .
$$

(3.6)

College graduates are paid based on their human capital in the labor market. Earnings are a function of the student’s pre-college human capital, college quality and their interaction. Earnings shock $\varepsilon_{kt}$ is assumed to be i.i.d. for tractability. The specification in Equation (3.6) implicitly incorporates the human capital production during college. Theoretically, the earnings shock $\varepsilon_{kt}$ also contains
the human capital production shock during college, but they cannot be separately identified. From the student’s perspective, the human capital production shock during college cannot be observed when they choose which college to attend and remains zero in expectation. Thus, from an ex-post point of view, \( \varepsilon_{kt} \) may follow a distribution of \( N(\zeta_s, \sigma^2_{\varepsilon}) \), in which \( \zeta_s \) is the realized human capital production shock in college \( s \). However, ex-ante, \( \zeta_s \) is zero in expectation.

Earnings after college also depend on local productivity; \( g_{kt} \) captures the average life-cycle earnings profile of college-educated workers in province \( k \). The earnings should be considered as the average across industries, given that the industry structure in local labor market is not included in the model.

### 3.2 Individual’s problem

The individual’s problem is solved backwards. After college graduation, if a student chooses to enter the labor market in stage 2 \((g = 0)\), she will subsequently choose a province in which to live and work in stage 3. The individual’s problem in stage 3 is:

\[
\max_{k \in \mathcal{N}} \pi^{w}_{sk} + \xi^{w}_k, \tag{3.7}
\]

where the \( \pi \) term represents the deterministic part of the corresponding payoff. \( \xi^{w}_k \) is realized and observed at the beginning of stage 3 before choosing \( k \). Given that \( \xi^{w}_k \) follows the Type-I extreme value distribution, the probability of choosing province \( k \) in stage 3 conditional on graduating from college \( s \), \( p^{w}_{sk} \), is given by

\[
p^{w}_{sk} = \frac{\exp\left(\frac{\pi^{w}_{sk}}{\sigma^{w}}\right)}{\sum_{k' \in \mathcal{N}} \exp\left(\frac{\pi^{w}_{sk'}}{\sigma^{w}}\right)}, \tag{3.8}
\]

Let \( EV^{w}_{s} \) denote the expected payoff to enter the labor market directly after college \( s \), in which the expectation is taken with respect to \( \xi^{w} \):

\[
EV^{w}_{s} = \bar{\gamma} + \sigma^{w} \ln \left[ \sum_{k' \in \mathcal{N}} \exp\left(\frac{\pi^{w}_{sk'}}{\sigma^{w}}\right) \right], \tag{3.9}
\]

where \( \bar{\gamma} \) is the Euler constant.

In stage 2, college graduates choose between graduate school \((g = 1)\) and entering the labor
market \((g = 0)\):

\[
\max_{g \in \{0,1\}} \bar{\pi}_g^g (g = 1) + EV_w^w (g = 0) + \xi^g
\]

\(\xi^g\) is realized and observed before choosing \(g\), while \(\xi^w\) is not observed in this stage. Given that \(\xi^g\) follows the Type-I extreme value distribution and is independent from \(\xi^w\), the probability of choosing graduate school over working after college \(s\), \(p_s^{g=1}\), is given by

\[
p_s^{g=1} = \frac{\exp\left(\bar{\pi}_g^g / \sigma^g\right)}{\exp\left(\bar{\pi}_g^g / \sigma^g\right) + \exp\left( EV_w^w / \sigma^g\right)}. \quad (3.10)
\]

Let \(EV_s\) denote the expected life-time payoff of graduating from college \(s\), in which the expectation is taken with respect to \(\xi^g\):

\[
EV_s = \overline{\gamma} + \sigma^g \ln \left[ \exp\left(\bar{\pi}_g^g / \sigma^g\right) + \exp\left( EV_w^w / \sigma^g\right) \right]. \quad (3.11)
\]

In stage 1, each college \(s \in S\) allocates an admission quota \(Q_s^{hzt}\) to province \(h\) track \(z\) in year \(t\).\(^{16}\) Given the allocated set of quotas \(\{Q_s^{hzt}\}_{s \in S}\), each province runs an independent centralized admission based on \(NCEE_{hzt}\). Each individual submits an application list \(L\) to maximize the expected life-time utility:

\[
\max_L \sum_{s \in S} \left( \bar{\pi}_s^c + \xi_s^c + \beta^4 EV_n^s \right) P \left( \text{admitted by } s \mid L, NCEE_{hzt} \right). \quad (3.12)
\]

When making college application decisions, students know the distribution of current and previous years’ NCEE scores and previous years’ admission cutoffs of each college in their province. Following the literature, I assume individuals take expected admission cutoffs as given and so imagine that their choices will not affect the equilibrium in large markets (Calsamiglia et al., 2020). Given that the main interest of this paper is student choices of college and residential locations rather than admission mechanism and its associated algorithm, details about how students form the probability of being admitted by each college given the application list are not needed for identification and thus not modeled.

\(^{16}\)\(Q_s^{hzt}\) is a set of college seats containing various college majors, but the major choice is not modeled in this paper.
3.3 Identification of college choices

The identification builds on the asymptotic stability assumption of the admission outcome (Fack et al., 2019; Agarwal and Somaini, 2020). (Strict) Stability means every student is admitted by her favorite ex-post feasible school given realized admission cutoffs. It allows for identification and estimation of individual preferences without the need of modeling the choice of application list and solving for the equilibrium of college application (Fack et al., 2019).

Under the stable admission outcome, the ex-post feasible college set, denoted as $S^f$, can be recovered by comparing an individual’s NCEE score with ex-post admission cutoffs. Note that all admission quota constraints are already embedded in these equilibrium admission cutoffs. The individual’s problem in Equation (3.12) is then reduced to a standard discrete choice problem: to choose college $s$ from her feasible set $S^f$ to maximize her expected life-time utility:  

$$\max_{s \in S^f} u_c^s + \xi_c^s + \beta^4 EV_s.$$  

(3.13)

Since $\xi_c^s$ follows the Type-I extreme value distribution, this gives the standard logit-type choice probability:  

$$p_c^s = \frac{\exp\left(\frac{(u_c^s + \beta^4 EV_s)}{\sigma_c}\right)}{\sum_{s' \in S^f} \exp\left(\frac{(u_{c s'} + \beta^4 EV_{s'})}{\sigma_c}\right)}.$$  

(3.14)

Denote this as the stability-based estimator.  

The stability-based estimator, utilizing the ex-post feasible college set, is essentially a fuzzy regression discontinuity (RD) design. Take the identification of the college location’s impact on residential location choice as an example. When unobserved heterogeneity in location preferences are controlled for through individual types, the term $D(k, j_s)$ in the payoff of working in province $k$ in Equation (3.3) is identified by comparing students whose NCEE scores are equal to or slightly

---

17 Infeasible colleges effectively have a flow payoff of negative infinity: $\pi^c(s|\Omega_i) = -\infty$ if $s \notin S^f_i$.

18 An alternative way to analyze the application stage is to assume students can apply to all colleges. After receiving the admission outcome (acceptance or rejection) from all colleges, students choose one college to attend from those that accept them. These colleges equivalently form the feasible set $S^f$.

19 Conceptually, the shock $\xi_c^s$ captures the idiosyncratic preference for each college $s$. In reality, it also contains application mistakes individuals make. There is a small but growing literature showing that students have different levels of sophistication during college application (Luflade, 2017; Calsamiglia et al., 2020), especially among students from different family backgrounds or between urban and rural areas (Ye, 2018). Allowing for correlation between the pre-college human capital $\theta^0$ and the shock $\xi_c^s$ would substantially complicate the analysis, if not make it impossible, given that the distribution of $\theta^0$ is not directly observed and needs to be recovered during estimation.
above the admission cutoff of college $s_1$ to students whose NCEE scores are slightly below. The former students are eligible for all colleges that the latter students are eligible for, but the former are also eligible for this extra college $s_1$. Assuming $s_1$ is located in province $j_1$, all else equal, the former students are (weakly) more likely to attend colleges in $j_1$ compared to the latter students, because one more college is available to choose in $j_1$. Of course, the RD here is necessarily fuzzy. Importantly, the example above is just one place where the discontinuity occurs. The discontinuity will happen at every admission cutoff level, and all of them together provide identification for the causal effect of the college location on residential location choices, captured by $D(k, j_s)$.

Admittedly, the assumption of strictly stable college admission outcomes is not realistic. However, as shown in Fack et al. (2019), college admission using a single priority index (the NCEE score in this paper) in a large market (as in China) is likely to be asymptotically stable. As the number of students and the capacity (thus quota) of each college go to infinity proportionally, the fraction of students not matched with their favorite feasible school (due to an imperfect ability to predict admission cutoffs or other randomness) converges to zero. Fack et al. (2019) proves that the stability-based estimator specified in Equation (3.14) is consistent under asymptotic stability.\footnote{Fack et al. (2019) provide numerical evidence suggesting that typical real-life college markets are sufficiently large for a good asymptotic approximation.}

3.4 Identification of other parameters

Utility parameters for graduate school and residential choices are identified following the standard discrete choice literature (McFadden, 1974; Rust, 1994). The parameter in front of the lifetime utility from consumption in $u^w$ in Equation (3.3) is normalized to 1. Given this normalization, utility parameters determine the relative difference in payoffs across choice alternatives and, thus, the probability of choosing each alternative. Because the idiosyncratic preference shocks, $\xi$’s, follow the Type-I extreme value distribution and are independent from each other, we have the logit-type choice probabilities specified in the individual’s problem. Matching these choice probabilities with observed choice distribution identifies these utility parameters.

Unobserved heterogeneity in location preference is identified by the dynamic correlation of location choices, once the effect of college location on work location choices is pinned down by the fuzzy RD discussed above. The amenity terms in $u^c$ and $u^w$ capture the common value
of each location and the distance term between college location and residential location capture the average locational correlation. If we observe a subgroup of people systematically choose a particular college location \( \tau \) with a probability higher than that implied by the common values, and at the same time they disproportionally choose to work there after college, it suggests these people have an unobserved preference for this location. The share of these subgroups identify the type probability \( \pi_\tau \). The choice probability of college location that deviates from the level implied by common location values identify the magnitude of the unobserved preference \( \chi \). Lastly, the dynamic locational correlation exhibited in those subgroups helps identify the scaler parameter \( \rho \).

4 Measuring the distribution of student quality across provinces

This section builds a measurement model to recover the distribution of students’ pre-college human capital \( \theta^0 \) in each province that is comparable nationwide, overcoming the non-comparability of the NCEE scores. Recovering a comparable distribution of college applicants in each province permits a direct evaluation of the admission quota allocation from a merit perspective. Moreover, it is the key prerequisite for evaluating any counterfactual admission policies.

4.1 Measurement model

There are two measures of the human capital level of each student: the NCEE score received in province \( h \) track \( z \) in year \( t \), and the cumulative college GPA at graduation in college \( s \). To restore the comparability of the NCEE scores, the main intuition is to find a measure that is comparable across provinces. College GPAs can help compare students coming from different provinces but enrolled in the same college. Consider two students, A and B. Student A has an NCEE score of 600 in \( hzt \), and B has an NCEE score of 550 in \( h'z't \). Both of them attend college \( s \). Both of them have a college GPA of 3.5. Assume that both the NCEE and GPA are functions of human capital, and, for simplicity, ignore all measurement errors and human capital production shocks in this example. First, A and B should have same level of human capital at the end of college, implied by the same college GPA. Furthermore, since they receive the same college education in \( s \), regardless of the human capital production technology during college, their human capital at college entry should also be the same, as measured by their NCEE scores. Thus, the NCEE score of 600 in
$hzt$ should be equivalent to 550 in $h'z't$. Admittedly, both the NCEE scores and college GPAs are noisy measures of human capital and the human capital production during college is stochastic as well, so matching a NCEE score in $hzt$ to another score in $h'z't$ is not as straightforward as discussed above. The latent factor model specified below considers these factors and formalizes the idea of combining the NCEE scores and college GPAs to recover the comparable pre-college human capital.

I view an individual’s latent pre-college human capital $\theta^0$ as an one-dimensional measure that aggregates both the innate ability and learned knowledge and skills upon college entry. Consider $\theta^0$ in a linear factor model with two measures: the NCEE score received in province $h$ track $z$ in year $t$, $NCEE_{hzt}$, and the cumulative college GPA at graduation in college $s$, $GPA_s$.

\begin{align*}
NCEE_{hzt} &= \kappa_{1,hzt} + \lambda_{1,hzt} \theta^0 + \sigma_{1,hzt} \nu_1; \quad (4.1) \\
GPA_s &= \kappa_{2,s} + \lambda_{2,s} \theta^0 + \sigma_{2,s} \nu_2. \quad (4.2)
\end{align*}

In Equation (4.1), the raw NCEE score, $NCEE_{hzt}$, is a function of $\theta^0$ plus a measurement error $\nu_1$. The parameters $\kappa_{1,hzt}$, $\lambda_{1,hzt}$ and $\sigma_{1,hzt}$ are subscripted by $hzt$ to allow for the province-track-year-specific testing technology of the NCEE. Unlike Equation (4.1), which is a pure measurement equation, the GPA equation in Equation (4.2) contains information on both human capital production and measurement during college.\(^{21}\) The parameters $\kappa_{2,s}$, $\lambda_{2,s}$ and $\sigma_{2,s}$ are subscripted by $s$ to allow for school-specific college inputs in both human capital production and GPA measurement technology. The error term $\nu_2$ aggregates the production shock and measurement error.

Equations (4.1) and (4.2) apply to the population of all NCEE takers, including both admitted

\(^{21}\)To see this, consider human capital production during college in school $s$ in the following form:

$$
\theta^c = \tau_{0,s} + \tau_{1,s} \theta^0 + \varepsilon^0_s,
$$

where $\theta^c$ is the latent human capital at the end of college, and the parameters $\tau$’s capture the college-specific production technology. Then, consider college GPA as a noisy measure of human capital at the end of college:

\begin{align*}
GPA_s &= \varphi_{0,s} + \varphi_{1,s} \theta^c + \varepsilon^c_s \\
&= (\varphi_{0,s} + \varphi_{1,s} \tau_{0,s}) + (\varphi_{1,s} \tau_{1,s}) \theta^0 + (\varepsilon^c_s + \varphi_{1,s} \varepsilon^0_s) \\
&= \kappa_{2,s} + \lambda_{2,s} \theta^0 + \sigma_{2,s} \nu_2.
\end{align*}

where the parameters $\varphi$’s allow for college-major-specific measurement technologies of the college GPA. The $\tau$’s and $\varphi$’s would not be separately identified.
(for which college GPA are observed) and not admitted students (if they are admitted). In addition, the parameters in the GPA equation are assumed to be stable across years, so there is no $t$-subscript in Equation (4.2). Ideally, the college GPA should be defined at the level of each college major in each school to account for potentially different measurement across college majors within the same school. The student survey data I use indeed provide information on college majors, but the sample size in each school-major cell is usually too small. Due to this data limitation, college GPA is defined at the school level and I leave the finer calculation of college GPAs for future work when larger-scale data become available.

Following the standard approach in factor model analysis, I normalize $\kappa_{1,(hzt)^*} = 0$ and $\lambda_{1,(hzt)^*} = 1$ for a specific $(hzt)^*$ to pin down the position and scale of the latent human capital $\theta^0$. Both $\nu_1$ and $\nu_2$ have zero mean and unit variance, are i.i.d. across individuals, and are independent from $\theta^0$ and from each other in the population.

4.2 Identification of measurement parameters

Identification of this factor model takes advantage of its structure: the NCEE equation does not depend on $s$, so it can measure students from the same province attending different colleges, while the GPA equation does not depend on $hzt$, so it can measure students enrolled in the same college but originally from different provinces. Under the current college admission regime, students from the same province go to multiple colleges in different locations. At the same time, each college admits students from multiple provinces. Intuitively, students from the same province are effectively measured by multiple colleges using GPAs, with each attended college measuring one part of the students in that province. These various college GPA measures are then linked by the same NCEE measurement in that province. At the same time, each college provides comparability of the NCEE scores of students from different provinces using their own GPA measurement, and

---

22 Students’ pre-college human capital is treated as one-dimensional in this paper and $\theta^0$ is an aggregate measure of this human capital. Thus, $\nu_1$ and $\nu_2$ are taken as pure measurement errors (plus a human capital production shock during college in $\nu_2$). This categorization implies that all other observables, like family income or parental education, only affect $\theta^0$, but not $\nu_1$ or $\nu_2$.

23 One potential concern is that the out-of-province students may systematically perform worse in college than the in-province students, at least initially, when they start the college life far away from home and face a different environment, which will break the independence between the GPA measurement and the origin of students. As a robustness check, I use only the out-of-province students in the measurement model estimation part. The recovered distribution of $\theta^0$ in each province is very similar in magnitudes and preserves the entire ranking of provinces in terms of the provincial mean of $\theta^0$. 

---

21
this NCEE score comparison is done multiple times as well by different colleges and for different parts of the NCEE score distribution based on college selectivity. This identification strategy utilizing the “crossing” structure is similar to that in Abowd et al. (1999) in the worker and firm setting. The key variation in Abowd et al. (1999) for identifying the worker and firm heterogeneity in wages is the between-firm mobility of the individual workers. Analogically, students in this paper are workers that are initially employed by a firm called high school $hzt$ and then moved to the second firm called college $s$.

The measurement model will be estimated using a selected sample. It only contains students who are admitted by a college $s$ so that I observe their college GPAs, and the college $s$ is also an endogenous choice made by students. I assume that students know $\theta^0$ and select colleges based on $NCEE_{hzt}$ (i.e. selection by $\nu_1$). However, students do not observe $\nu_2$ nor anything correlated with $\nu_2$ in any college before attending it, so there is no selection by $\nu_2$. Selection by $\nu_1$ implies that, in general, the conditional distribution of $\nu_1$ does not equal the unconditional distribution:

$$E(\nu_1|hzt, s, \theta^0) \neq E(\nu_1) = 0;$$
$$\text{var}(\nu_1|hzt, s, \theta^0) \neq \text{var}(\nu_1) = 1.$$

While no selection by $\nu_2$ implies

$$E(\nu_2|hzt, s, \theta^0) = E(\nu_2) = 0, \forall hzt, s, \theta^0;$$
$$\text{var}(\nu_2|hzt, s, \theta^0) = \text{var}(\nu_2) = 1, \forall hzt, s, \theta^0.$$

Under these assumptions, the measurement model parameters $\{\kappa, \lambda, \sigma\}$ are identified using the conditional mean, variance and covariance of the NCEE scores and college GPAs for each
combination of the province-track-year \( hzt \) and the attended college \( s \):

\[
E(NCEE_{hzt}|s) = \kappa_{1,hzt} + \lambda_{1,hzt}E(\theta^0|hzt, s) + \sigma_{1,hzt}E(\nu_1|hzt, s);
\]

\[
E(GPA_s|hzt) = \kappa_{2,s} + \lambda_{2,s}E(\theta^0|hzt, s);
\]

\[
\text{var}(NCEE_{hzt}|s) = \lambda_{1,hzt}^2 \text{var}(\theta^0|hzt, s) + \sigma_{1,hzt}^2 \text{var}(\nu_1|hzt, s)
+ 2\lambda_{1,hzt}\sigma_{1,hzt} \text{cov}(\theta^0, \nu_1|hzt, s);
\]

\[
\text{var}(GPA_s|hzt) = \lambda_{2,s}^2 \text{var}(\theta^0|hzt, s) + \sigma_{2,s}^2;
\]

\[
\text{cov}(NCEE_{hzt}, GPA_s) = \lambda_{1,hzt}\lambda_{2,s} \text{var}(\theta^0|hzt, s) + \lambda_{2,s}\sigma_{1,hzt} \text{cov}(\theta^0, \nu_1|hzt, s).
\]

Cancelling out \( E(\theta^0|hzt, s) \) and \( \text{var}(\theta^0|hzt, s) \) gives three moment conditions for each combination of \( hzt \) and \( s \), denoted as (M1) to (M3):

\[
\text{(M1)} \quad E(NCEE_{hzt}|s) = \kappa_{1,hzt} + \frac{\lambda_{1,hzt}}{\lambda_{2,s}}(E(GPA_s|hzt) - \kappa_{2,s}) + \sigma_{1,hzt}E(\nu_1|hzt, s);
\]

\[
\text{(M2)} \quad \text{var}(NCEE_{hzt}|s) = \frac{\lambda_{1,hzt}}{\lambda_{2,s}} \left( \text{cov}(NCEE_{hzt}, GPA_s) - \lambda_{2,s}\sigma_{1,hzt} \text{cov}(\theta^0, \nu_1|hzt, s) \right)
+ \sigma_{1,hzt}^2 \text{var}(\nu_1|hzt, s) + 2\lambda_{1,hzt}\sigma_{1,hzt} \text{cov}(\theta^0, \nu_1|hzt, s);
\]

\[
\text{(M3)} \quad \text{var}(GPA_s|hzt) = \frac{\lambda_{2,s}}{\lambda_{1,hzt}} \left( \text{cov}(NCEE_{hzt}, GPA_s) - \lambda_{2,s}\sigma_{1,hzt} \text{cov}(\theta^0, \nu_1|hzt, s) \right) + \sigma_{2,s}^2.
\]

To illustrate the identification idea, first assume \( E(\nu_1|hzt, s), \text{var}(\nu_1|hzt, s) \) and \( \text{cov}(\theta^0, \nu_1|hzt, s) \) are known. The identification will require: 1) data from at least two different \( hzt \)’s and two different colleges, 2) students from each \( hzt \) attend both colleges, and 3) each college admits students from both \( hzt \)’s. These data will generate 4 combinations of \( hzt-s \) and in total 4*3=12 moments. Under the normalization in which \( \kappa_{1,(hzt)^*} = 0 \) and \( \lambda_{1,(hzt)^*} = 1 \) for a specific \( (hzt)^* \), there are in total 10 measurement parameters, so they are identified. Appendix A gives the complete proof of identification and how the identification started from the initial four \( hzt-s \) cells can be extended to cover all provinces, tracks, years and colleges.

Moment conditions (M1) to (M3) for each combination of \( hzt \) and \( s \) will be used in the generalized method of moments (GMM) estimation of measurement parameters \( \{\kappa, \lambda, \sigma\} \). Given the selection by \( \nu_1 \) in college choices, \( E(\nu_1|hzt, s), \text{var}(\nu_1|hzt, s) \) and \( \text{cov}(\theta^0, \nu_1|hzt, s) \) will be endogenously determined and need to be derived from the structural model. Thus, the measurement
model (using the GMM) will be estimated jointly with the structural model, which is discussed in
detail in the Estimation section.

4.3 Recovering the provincial distribution of $\theta^0, F_h(\theta^0)$

The main goal is to recover the distribution of $\theta^0$ in each province, denoted as $F_h(\theta^0)$. They
will be used in the maximum likelihood estimation of the structural model. I assume the pre-
college human capital $\theta^0$ of the NCEE exam takers in each province follows a normal distribution,
$F_h(\theta^0) \sim N(\mu_h^\theta, (\sigma_h^\theta)^2)$, and this distribution is stable across the six cohorts studied in this pa-
paper.\textsuperscript{24}\textsuperscript{25} The choice of the high school track $z$ is assumed to be independent from $F_h(\theta^0)$, thus
$F_h(\theta^0|z) = F_h(\theta^0)$ for both tracks.\textsuperscript{26} I also assume both error terms in the measurement model, $\nu_1$ and $\nu_2$, follow the standard normal distribution, $N(0, 1)$. Note that each of them will have a scal-
ing parameter, $\sigma$, for each province-track-year or college-major, to account for potentially different variances.

Given these normal distribution assumptions, the linear measurement equation implies $NCEE_{hzt}$
should also be normally distributed. As with most raw test scores, it is difficult to map the scale
and distribution of raw NCEE scores to the underlying latent skills (the $\theta^0$ in this paper). Most
standardized test scores, like the SAT and ACT in the US, use some form of normalization. I nor-
malize and standardize the raw $NCEE_{hzt}$ as followed. First, I convert the raw NCEE scores to
percentile rankings within each $hzt$. Second, these percentile rankings are mapped to the standard
normal distribution and are assigned the values from the inverse cumulative density function of
the standard normal distribution. After this transformation, the normalized NCEE scores in each

\textsuperscript{24}The normal distribution assumption is for identification purpose. In principle, once the measurement parameters
are identified, the distribution of $\theta^0$ can be nonparametrically identified for admitted students in each province by com-
bining the subsamples from all colleges. However, this procedure is tedious in practice and is restricted by the small
sample size of some subsamples. In addition, even if the distribution of the admitted students is nonparametrically
identified, I still need some form of interpolation to get the distribution of all NCEE exam takers.

\textsuperscript{25}China’s compulsory education ends after middle school (the ninth grade). Among the entire cohort born in the
same year, about half attend the regular (academic-oriented) high school. The rest of them either attend vocational high
school or directly enter the labor market. High school education is a prerequisite for the NCEE. Although the four-year
college capacity is only about 30% of the size of total NCEE exam takers in the analyzed period, vocational colleges,
which have a slightly larger capacity than four-year colleges, also admit students only through the NCEE. As a result,
almost all high school graduates take the NCEE. With the same education level, assuming a normal distribution for the
NCEE exam takers is a reasonable choice. However, considering the human capital distribution of the entire cohort at
college-going age is beyond the scope of this paper, given the different level and form of education they receive.

\textsuperscript{26}Although there is clear gender difference in choosing tracks (more female students in humanities), there is no
clear evidence for or against the similar achievement distribution between two tracks.
follow the standard normal distribution. Note that the raw values of the NCEE scores are not important. The purpose of the measurement model is to recover the underlying distribution of \( \theta^0 \), which is anchored to an arbitrarily chosen position and scale because the unobserved \( \theta^0 \) does not have a natural position and scale. Normalizing the values of the NCEE scores will only change the value and interpretation of the measurement parameters \( \{\kappa, \lambda, \sigma\} \).

Now I can use the measurement model to recover the provincial distribution \( F_h(\theta^0) \), i.e. \( \mu_{\theta^0}^h \) and \( \sigma_{\theta^0}^h \) for each province \( h \), because \( F_h(\theta^0) \) is fully determined by \( \{\kappa, \lambda, \sigma\} \) and the distribution of NCEE scores. Given that \( NCEE_{htz} = \kappa_{1,htz} + \lambda_{1,htz} \theta^0 + \sigma_{1,htz} \nu_1 \) and the assumption that \( \nu_1 \sim N(0,1) \) and is independent from \( \theta^0 \),

\[
NCEE_{htz} \sim N(\kappa_{1,htz} + \lambda_{1,htz} \mu_{\theta^0}^h, (\lambda_{1,htz} \sigma_{\theta^0}^h)^2 + \sigma_{1,htz}^2).
\]

Since the normalized NCEE scores in each \( hzt \) follow the standard normal distribution, we have two extra moment conditions (M4) and (M5) for each \( hzt \):

\[
\begin{align*}
(M4) \quad & \kappa_{1,htz} + \lambda_{1,htz} \mu_{\theta}^h = 0; \\
(M5) \quad & (\lambda_{1,htz} \sigma_{\theta}^h)^2 + \sigma_{1,htz}^2 = 1.
\end{align*}
\]

(M4) and (M5) will be used in the GMM estimation of the measurement model as additional moment conditions. Since \( F_h(\theta^0) \) is assumed to be stable across \( z \) and \( t \), these moments impose extra over-identifying restrictions.

5 Data and estimation strategy

5.1 Data source and sample selection

5.1.1 Administrative data of college admission

The first dataset is the administrative records of college admission outcome from the Ministry of Education of China. It covers the universe of students admitted to four-year colleges through the NCEE each year in China. I obtain a 10% random sample of the micro-level student data for six cohorts that were admitted between 2006 and 2011. The sample consists of 1,645,728
students and records each student’s home province, NCEE track, NCEE score and its provincial ranking, as well as the admitted college and major. I drop students admitted by military, police, arts and sports colleges, and restrict the sample and analysis to 30 out of 31 provinces in China, excluding Tibet (Xizang). College admission in Tibet is both special and complicated due to various Tibetan-specific policies. Many Tibetan students study in high schools located in other provinces under affirmative action policies and follow special college admission rules, procedures and cutoffs. Given the small size of the student population in Tibet (on average 0.15% of all NCEE exam takers each year in China) and these complications, Tibet is excluded from the analysis throughout the paper.

5.1.2 The Chinese College Student Survey (CCSS)

The second dataset is the Chinese College Student Survey (CCSS), which is administrated by Tsinghua University and covers a nationally representative sample of the same six cohorts of college students in the administrative data. The CCSS first randomly selects schools from all colleges in China, stratified by college tier and location. Within each college, it further randomly selects and surveys students at the time of college graduation. Similar to the administrative data, I restrict the sample to students who were admitted only through the sciences or humanities track in the NCEE and studied in non-military/police schools. The final sample used in the analysis covers 79 four-year colleges and 30,629 students. Each college was surveyed at least once by the CCSS, and many were surveyed multiple times. On average, 205 students were surveyed each year in each school, which is about 9% of the average cohort size in one college.

One disadvantage of the administrative data of college admission is that it does not track students after entering college, so I do not observe their performance during college or their choices after college. The CCSS supplements the administrative data well. It contains all variables in the administrative records and further collects detailed information on individual characteristics, family background, college performance (including the four-year cumulative GPA), and post-graduation choices. If students enter the labor market after college graduation, the survey collects the first job’s information including location, annual earnings, job and employer characteristics. For students who go to graduate school, the name of the institution is recorded. Unfortunately, the CCSS does not track college graduates beyond the first job after college or survey their choices after grad-
uate school. The latter is not an issue since the structural model ends immediately with a terminal value if students choose graduate studies. For the former, I supplement the CCSS with the China Family Panel Studies (CFPS) to obtain the earnings growth pattern, which is discussed below.

5.1.3 Life-cycle earnings profile by location

I use the China Family Panel Studies (CFPS) to obtain the life-cycle profile of earnings after college. The CFPS is a nationally representative longitudinal survey, similar to the Panel Study of Income Dynamics (PSID) in the US. The baseline survey was launched in 2010 and it started with a sample of 14,608 households and 33,600 individuals. These households and individuals are subsequently surveyed every two years. I use five rounds of the survey from 2010 to 2018.

The CFPS is relatively new and lasts for only nine years, so I adopt a synthetic cohort approach to calculate the life-cycle earnings profile. I restrict the sample to individuals aged between 22 and 62 in each round of the survey, living in urban area, and who are college educated. Constrained by the multi-stage sampling design of the CFPS, the sample size in some provinces is not large enough to get accurate estimates, so I aggregate provinces into the following regions: Beijing and Shanghai, other East provinces, Central, West, and Northeast. I run an individual fixed-effect regression of the log annual earnings on age dummy variables for each region separately to get the estimated coefficients for age dummies. These coefficients are used as the earnings growth term for each province and working period.

5.1.4 College quality measures

I extract the quality measures of each college from the Higher Education Undergraduate Teaching Quality Report from 2014 to 2016, published by the Higher Education Evaluation Center of the Ministry of Education. Each college was required to report a list of statistics on students, faculty, infrastructure, funding and expenditure related to undergraduate teaching. It would be ideal to have these reports covering the time period in the student-level data discussed above, but 2014 to 2016 are the only years that are currently available to researchers. Each raw measure is averaged across

---

27East provinces include Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong and Guangdong. The Central includes Shanxi, Anhui, Jiangxi, Henan, Hubei, Hunan and Hainan. The West includes Inner Mongolia, Guangxi, Chongqing, Sichuan, Guizhou, Yunnan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia and Xinjiang. The Northeast includes Heilongjiang, Jilin and Liaoning.
the three years.

I use the principle component analysis (PCA) to aggregate the college quality measures into an one-dimensional variable. To account for other unobserved aspects of college quality affected by college tier and ownership, I also include indicators for Project-985 schools, Project-211 schools, and private schools in the PCA.\textsuperscript{28} Table A1 in Appendix reports the raw measures and the estimated loading factors of the first principle component associated with each measure. The first principle component can explain 50.15\% of the total variance.\textsuperscript{29} To address occasional missing values in those raw measures, I follow Dillon and Smith (2017) and use the loading factors of the first principle component to construct a weighted average of raw measures. This weighted average is further standardized to have a mean of zero and a standard deviation of one to be used as the one-dimensional college quality measure $V_s$. The histogram of $V_s$ is plotted in Figure A2. Table A2 reports the top-20 schools ranked by the estimated $V_s$ and they correspond well to the general consensus and various college rankings.

Additional details of data sources and sample construction are provided in Appendix B.

5.2 Joint estimation of the structural model and the measurement model

The structural model is estimated via the maximum likelihood estimation (MLE) combining both the administrative data and the CCSS survey data. For individuals in the administrative data, only college choices are observed. The likelihood of choosing college $s$, incorporating the unobserved heterogeneity $\tau$, is given by

$$L^{Admin}(s|\Omega) = \sum_{\tau} \pi_\tau p^c_\tau(\Omega). \quad (5.1)$$

For individuals in the CCSS survey data, I observe their college choice $s$, choice of graduate studies $g$, post-college residential choice $k$ and initial earnings $W_{sk1}$ (when $g = 0$). The likelihood

\textsuperscript{28}Project-985 and Project-211 colleges receive comprehensive support from the central government in teaching and research. These supports may not be fully captured by the measures in teaching quality reports. Private colleges in general receive much less support from the government compared to public schools.

\textsuperscript{29}See Figure A1 in Appendix for the screeplot of the PCA.
of observing these choices is given by

\[
\mathcal{L}^{CCSS}(s, g, k, W_{sk1}|\Omega) = \begin{cases} 
\sum_{\tau} \pi_{\tau} p_{s}^c(\Omega) p_{g}^\tau(\Omega), & \text{if } g = 1; \\
\sum_{\tau} \pi_{\tau} p_{s}^c(\Omega)[1 - p_{g}^\tau(\Omega)] p_{w}^\tau(\Omega) l(W_{sk1}|\Omega), & \text{if } g = 0.
\end{cases}
\] (5.2)

where \(l(W_{sk1}|\Omega)\) is the likelihood of observing \(W_{sk1}\) conditional on \(\Omega\). The discount factor \(\beta\) is set at 0.95 (Kennan and Walker, 2011), and the number of working periods (years) \(T\) is set at 40.

Equations (5.1) and (5.2) are not readily estimated yet, since the pre-college human capital \(\theta^0\) in \(\Omega\) is unobservable to the econometrician. The distribution of \(\theta^0\) in each province, \(F_{h}(\theta^0)\), will be recovered from the measurement model. Denote \(\mathcal{A}\) as the set of individual choices and outcomes observed in the data, so \(\mathcal{A} = \{s\}\) for students in the administrative data and \(\mathcal{A} = \{s, g, k, W_{sk1}\}\) for students in the CCSS survey. Rewrite Equation (5.1) and (5.2) in a general form using \(\mathcal{A}\) and integrate the likelihood over the distribution of \(\theta^0\):

\[
\mathcal{L}(\mathcal{A}|\Omega) = \int l(\mathcal{A}|\theta^0) l(NCEE_{hzt}, GPA_{s}|\theta^0) dF_{h}(\theta^0)
\] (5.3)

The first term inside the integration, \(l(\mathcal{A}|\theta^0)\), is the likelihood of observing the choices and earnings from the structural model conditional on \(\theta^0\). The second term, \(l(NCEE_{hzt}, GPA_{s}|\theta^0)\), is the likelihood of observing the NCEE score and college GPA in the data, constructed from the measurement model. Once the measurement model parameters and \(F_{h}(\theta^0)\) are identified, the structural model can be estimated. However, this is not straightforward because the GMM estimation of the measurement model requires calculating several selection terms from the structural model first, as discussed in Section 4. Given the nature of the two interdependent estimation parts, the measurement model and the structural model are estimated jointly by the following steps:

1. With some initial guess of \(F_{h}^0(\theta^0)\) and \(\{\kappa, \lambda, \sigma\}^0\), estimate the structural model using the MLE to get parameters \(B^0\).

2. Simulate individual choices using \(B^0\), calculate the selection terms in the measurement model: \(E(\nu_1|hzt, s), \text{var}(\nu_1|hzt, s)\) and \(\text{cov}(\theta^0, \nu_1|hzt, s)\).

3. Estimate the measurement parameters again in the GMM using the moment conditions (M1)
to (M5) specified Section 4, this time with the selection terms simulated from the step above. Denote the new estimated parameters as \( \{\kappa, \lambda, \sigma\}^1 \) and \( F^1_h(\theta^0) \).

4. Use \( \{\kappa, \lambda, \sigma\}^1 \) and \( F^1_h(\theta^0) \) to re-estimate the structural model in MLE and get new structural model parameters \( B^1 \).

5. Iterate steps 2-4 until convergence.

I obtain standard errors via bootstrap, given that the whole model is estimated in iteration (McLachlan and Peel, 2004). The bootstrap is blocked at the province-track-year-college level to preserve the fundamental structure of the data and the identification requirement of the measurement model (as discussed in detail in Appendix A). Specifically, to construct the bootstrap sample, I randomly sample students with replacement from the administrative data and the CCSS data and obtain the same sample size as that of the estimated sample in each block. The block bootstrap is done with 100 replications, which has been shown to be sufficient in this case (Efron and Tibshirani, 1993; McLachlan and Peel, 2004).

6 Estimation results

6.1 Recovered pre-college human capital

6.1.1 Provincial distributions

Figure 3 plots the estimated mean \( (\mu^0_h) \) and standard deviation \( (\sigma^0_h) \) of NCEE exam takers’ pre-college human capital in each province. The distribution of \( \theta^0 \) in each province is standardized to set the national mean to zero and standard deviation to one. There is a large variation in average pre-college human capital levels across provinces. The gap of the mean of \( \theta^0 \) between provinces that have the best and worst average college applicants is larger than 1.2 national standard deviations. However, provinces that are ranked lower tend to have higher standard deviations, which means the difference between top students across provinces is smaller than the average.\(^{30}\)

\(^{30}\)One potential concern is that college GPAs cannot fully distinguish top students across provinces if most of them get full grades. However, from the CCSS data, I find very few students hold a four-year GPA of 4.0 and there is no evidence of a ceiling effect around the right tail of the GPA distribution.
Figure 4 plots the provincial mean of $\theta^0$ against the log GDP per capita in 2008 in the top panel and against the log of per student total expenditure averaged over primary, middle and high school in each province in the bottom panel. Perhaps unsurprisingly, the estimated average pre-college human capital level is positively correlated with local economic development and per student expenditure before college, but the correlation is far from perfect. Fixing the GDP or pre-college education expenditure, the variation of the average pre-college human capital level across provinces is still large.

6.1.2 Student quality vs. college quotas

With the recovered $\theta^0$ distribution, I can formally evaluate the observed college quota allocation across provinces from a merit-based perspective. Figure 5 plots the average number of four-year college quotas per NCEE exam taker between 2006 and 2011 against the mean of $\theta^0$ in each province. The quota allocation is positively correlated with the average student quality among the bottom half of provinces according to the ranking of $\theta^0$, but the relationship is very noisy among the top half. The larger size of admission quotas enjoyed by Shanghai (SH) and Beijing (BJ) is well justified by their students’ higher pre-college human capital. The similarly large size of quotas in Liaoning (LN), Tianjin (TJ) and Heilongjiang (HLJ), on the contrary, is not aligned with the average quality of their students, which is only around the national average. Moreover, provinces like Jiangsu (JS) and, especially, Shandong (SD) have high-quality students that are among the top in the country, but receive substantially smaller number of college seats. Overall, the observed college quota allocation across provinces is only merit-based to a limited extent.

A more direct evidence for whether the quota allocation is driven by student merit is to compare the marginal students, that is, the last student admitted from each province, as shown in Figure 6. Following the consumer theory that the marginal utility per dollar should be the same for all goods, if the goal is to let the best group of students in China to attend college up to the college capacity, we should allocate the quotas such that the $\theta^0$ of the marginal student is equalized across provinces. However, the gap of $\theta^0$ among the bottom group of admitted students in each province is large, up to 1.1 national standard deviations.

Lastly, to show why it is important to recover the comparable pre-college human capital levels, Figure A4 in Appendix illustrates the bias of using the raw NCEE scores alone to compare students
It plots the mean of the raw NCEE scores against the mean of $\theta^0$ for students admitted into four-year colleges in each of the $hzt$ group. Even when raw NCEE scores are all rescaled to be between 0 and 1 by dividing the corresponding full score, the relationship between the raw NCEE scores and the comparable $\theta^0$ is very weak. Therefore, recovering a comparable measure of students is essential before examining any counterfactual reforms of admission quotas.

6.2 Structural model parameters

The estimated structural model parameters are reported in Table 2. The parameters are grouped into five panels, which correspond to the flow payoff at college stage, the payoff from working after college, the unobserved heterogeneity in location preference, the life-time value from graduate school, and the earnings equation, respectively.

6.2.1 Preference for college

College preference parameters are reported in the top panel of Table 2. Holding the effect of college quality on earnings constant, attending a better college still generates a higher flow payoff, and the prestige from attending an elite school (Project-211 schools or above) yields an additional value. Comprehensive college, STEM college and Econ college are college type indicators. The omitted reference type is non-STEM college. On average, students prefer comprehensive universities that typically offer all major fields of studies and colleges that focus on economics and business over those focusing on STEM or non-STEM only majors. Students also prefer colleges located in the capital city of a province. Provincial capital cities in China are all big cities and typically the provincial center of economy, culture and population.

6.2.2 Preference for location

To better interpret the estimated location preferences, I convert the parameters in front of the location and distance terms to monetary values as follows. Since the parameter in front of the dis-

\footnote{The type of each college is obtained from the school database in the China Higher-Education Student Information platform (http://www.chsi.com.cn), an official website run by the Ministry of Education of China. There are 12 raw types. I group them into the four categories as followed (raw types in parenthesis): comprehensive college (comprehensive), STEM college (science and technology, agriculture, forestry, medical science), non-STEM college (education, language, law, sports, arts, ethnicity), and econ college (economics and business).}
counted life-time payoff from consumption is normalized to 1 in Equation (3.3), and the smoothed period consumption equals the discounted life-time earnings divided by a sum of the discount factor as specified in Equation (3.5), each extra payoff term $\gamma$ in Equation (3.3) is equivalent to a change $\Delta$ in consumption or earnings:

$$
\sum_{t=1}^{T} \beta^{t-1} \ln C_t + \gamma = \sum_{t=1}^{T} \beta^{t-1} \ln(C_t + \Delta);
$$

$$
\Delta = \left[ \exp\left( \frac{\gamma}{\sum_{t=1}^{T} \beta^{t-1}} \right) - 1 \right] C_t
$$

For each payoff parameter $\delta$, I calculate the $\Delta$ for each individual and then take the average. The parameter is then converted to an annual monetary value in Chinese Yuan (CNY) that is equivalently added to the annual consumption or earnings over the life cycle. The average starting annual salary for college graduates is 27,793 CNY in 2012 value, and the annual consumption (smoothed over the life cycle) averaged across individuals is 71,118 CNY.\(^{32}\)

The converted parameters are given in Table 3. Students show a high level of home attachment at both the college and working stage. The value of staying in the home province for college is equivalent to 1,164 CNY per year\(^{33}\), and the value of working in the home province after college is equivalent to 4,563 CNY per year. If leaving the home province, the additional utility loss is a concave function of the distance between the college or work location and the home province. The interaction between working at the home province and the home province being Beijing or Shanghai\(^{34}\) has a positive and large coefficient. This reflects the higher home preference of local students from these two top Chinese cities and also suggests the advantage of being a local resident under strict migration restrictions. I report the estimated amenity values for each province in Table A4 and A5 in Appendix. College graduates systematically value the more developed East

\(^{32}\)In 2012, 1 USD $\approx$ 6 CNY. The GDP per capita in China and the US in 2012 is 6,316 and 51,603 USD, respectively.

\(^{33}\)Converting parameters in the flow payoff at the college stage requires forward-discounting to the first working period. The value of attending an in-province college is converted to an equivalent increase of the future annual earnings.

\(^{34}\)Beijing and Shanghai are provincial-level municipalities. The other two provincial-level municipalities in China are Tianjin and Chongqing.
provinces, which is consistent with a priori of regional attractiveness, but the amenity values for college location do not exhibit a clear pattern with regard to regions. Such a difference in valuations may reflect student preference at different stages of life (study versus working) with different living arrangements (college dormitory\textsuperscript{35} versus rented/owned housing) or a learning process of amenity difference across locations.

The bottom panel of Table 3 shows that the college location plays an important role in subsequent work location choice, which is consistent with recent evidence both in China (Huang et al., 2019) and the US (Kennan, 2018). Controlling for the unobserved location preference that is persistent across stages, working at the college location generates a value equivalent to 3,790 CNY per year, and migrating to other provinces incurs additional costs that increase with distance. This causal effect of college location may have important policy implications: how to allocate college seats across regions and where these college seats come from can potentially change the spatial distribution of college graduates. Lastly, the estimated size of the idiosyncratic location preference is non-negligible, about 21% of the college location effect (796/3,790), so it is necessary to control for the dynamic correlation of the unobserved location preference.

6.2.3 Human capital production

Back to Table 2, the last two panels of parameters describe how human capital production during college affects the value of graduate studies and the return in the labor market. The parameter of the interaction between pre-college human capital $\theta^0$ and college quality $V_s$ is positive and statistically significant in the return from both graduate school and labor market. This implies complementarity between $\theta^0$ and $V_s$ in college education and is consistent with studies of other countries like the US (Dillon and Smith, 2020). Starting from a nationally average college applicant ($\theta^0 = 0$), if she has a one-standard-deviation higher pre-college human capital ($\theta^0 = 1$), the marginal return from attending a better college is 69% higher in the payoff from graduate studies (0.643/0.932) and 32% higher in post-college earnings (0.026/0.082). Direct evidence on the interaction effect between student and college quality in human capital production is scarce in the literature. To my best knowledge, this is the first study to provide an estimate of the complementarity in China. Interestingly, the estimated magnitude of the complementarity in earnings in China

\textsuperscript{35}Almost all college students in China live in on-campus dormitories for the entire time of college.
is similar to that found in Dillon and Smith (2020) for college graduates in the US.\footnote{Dillon and Smith (2020) find that the marginal return from college quality will increase by 18-37\% if student quality moves from the 50th to the 75th percentile. According to the recovered national distribution of $\theta^0$, $\theta^0 = 0$ is at the 49th percentile and $\theta^0 = 1$ is at the 84th percentile. If $\theta^0$ increases from the 50th to the 75th percentile as in Dillon and Smith (2020), I estimate the marginal return from college quality will increase by 22\% in earnings.}

### 6.3 Model fit

This subsection describes how well the model can fit the observed student choices in each stage. Figure 7 plots the average NCEE scores (all rescaled to be between 0 and 1) of students admitted by each tier of colleges from each province, track and year in model simulation against the data. In college admission that is solely based on test scores, those scores can be seen as student’s purchasing power for college education. As a result, the average test score of admitted students reflects the general attractiveness of a college and student’s preference. Figure 7 shows that the model can reasonably fit the admission results by college quality and location.

Table 4 reports the share of students directly going to graduate school after college graduation, averaged across six cohorts. Because the administrative data do not track students after college admission, this share is calculated based on the CCSS-surveyed schools to check the model fit. The observed share of choosing graduate studies from the CCSS data is adjusted by the sampling weights. The simulated share is calculated by picking up all students who graduate from the CCSS-surveyed schools in model simulation. Given the estimated complementarity between student and college quality in the payoff of graduate schools, the share of students from top-tier colleges continuing with graduate studies is nearly twice of that in the middle-tier and four times of that in the bottom-tier colleges. The model can fit the share of graduate studies in all three tiers of colleges.

For students who enter the labor market after college, Figure 8 plots the provincial distribution of college graduates in the data versus simulation, combining all six cohorts. Same with the share of choosing graduate studies, the work location choices are only observed in the CCSS sample, so I compare the observed and simulated distribution of college graduates based on CCSS-surveyed schools. The model can fit both the ranking and the magnitude of the spatial share in residential choices. Figure 9 further reports the model fit on dynamic correlation of location choices. The left panel shows the share of students working in their home province by each home province $h$,.
and the right panel shows the share of students working in their college province by each college province $j$. Overall, the model can fit the observed spatial dispersion in these features.

7 Counterfactual analysis of a merit-based nationwide admission

China’s current admission quota system has long been criticized for being unfair across regions (e.g. Fu, 2013; Guo et al., 2018). The finding that the current college seat allocation does not fully reflect the student quality difference across provinces indeed supports this criticism. A natural and frequently advocated reform is to run a nationwide admission based on a nationally comparable test, but the potential impact of such a reform has not been systematically evaluated due to the limitation of data and methods. In this section, I use the estimated structural model and the recovered distribution of pre-college human capital in each province to simulate a counterfactual nationwide admission that is purely merit-based and explore its impact on efficiency in college education as well as the spatial distribution of college opportunities and college-educated workers.

7.1 Nationwide admission setup

The cohort of 2008 is taken as the baseline in the counterfactual analysis for comparison. I simulate the same number of the NCEE exam takers in 2008 (in total 9,996,601). Each student’s pre-college human capital $\theta^0$ is randomly assigned according to the provincial distribution $F_h(\theta^0)$. Under the nationwide admission, the province-based NCEE testing and admission quotas are eliminated. Instead, a standardized college entrance exam is given to students in all provinces so that all scores are comparable. This comparable test score is a function of $\theta^0$ plus a measurement error. The variance of this measurement error is imposed to be the average of those estimated in each province’s NCEE from the measurement model, so that this hypothetical national exam has an average testing technology. Based on this test, a nationwide admission is run to admit students up to the existing college capacity in 2008. The nationwide admission uses the Serial Dictatorship algorithm, in which students choose college sequentially based on their national ranking of test scores and everyone knows all previous students’ choices.

Throughout the counterfactual analysis, the distribution of student size and quality ($\theta^0$), as well as college quality ($V_s$) and location, are fixed to the level in 2008. The policy-invariant distribution
of $\theta^0$ rules out the potential response of testing effort to college opportunity changes. Note that the recovered $\theta^0$ should be productive in both the NCEE and during college. Since the test-oriented preparation effort is generally not productive after the test, it would not be a part of the $\theta^0$. In addition, the relationship between effort and prize is usually nonlinear in a contest (Clark and Riis, 1998), so it is difficult to know whether and how testing effort will change in the counterfactual.37 The college quality $V_s$ is also held constant in the counterfactual, which means the peer quality change during the resorting between students and colleges is abstracted away (Dillon and Smith, 2020). I discuss the implication of this assumption when presenting the counterfactual results.

To make the comparison between the baseline and the counterfactual easier, provinces are grouped into four regions: East, Central, West and Northeast.38 The coastal East region is the most economically developed in China. The Central is less developed but has a high population density, and the West region is the least developed in China. The Northeast region is China’s “Rust Belt”. Its economy was previously active as the home of China’s heavy industry but is now in protracted economic decline. Table 5 describes the average college resources and student quality in each region. The per capita college capacity is defined as the number of college seats in colleges located in each region, divided by the number of NCEE exam takers in that province. In the baseline of 2008, provinces in the East on average host a large college capacity of 0.34 seats per capita and those colleges are of the highest average quality. College applicants in the East also have the highest average $\theta^0$. In contrast, the Central and West have much lower college capacity and quality as well as student quality. The Northeast has a historical advantage in college capacity, which, combined with a declined population size, gives it the highest per-capita college capacity in 2008.

### 7.2 Re-distribution of college seats and sorting outcome

Table 6 reports the average changes in college access by region. In the baseline, how many students can go to college (Column (1)) is determined by the total number of admission quotas received by students’ residence province. The quota size is highly correlated with the local college

37 Selection into or out of taking the NCEE is also abstracted away but would not be a big concern for the purposes of this counterfactual. Because such selection mostly happens at the lower end of the student distribution, counterfactual admission quotas for four-year colleges are not likely to affect these students, and this paper does not study vocational college admission.

38 The regions are officially defined by the central government. Many national strategies and policies are based on regions.
capacity reported in Table 5, because on average 63% to 78% of the college capacity is allocated to students in the local province, which is reflected in Column (3). As a result, the average college quota in each of the East provinces was 0.31 per exam taker (Column (1)) in 2008, which was much higher than that in the Central and West. Benefiting from the largest college capacity, the Northeast had the largest college quota in the baseline. Such a quota allocation in the baseline results in unequal quality ($\theta^0$) of admitted students across regions (Column (5)). While an average $\theta^0$ of 0.98 is good enough in the Northeast to be admitted into a college, the standard increases by 0.3 national standard deviation to 1.28 in the East.

Under the counterfactual nationwide admission, the large share of admitted students in the East further increases to 0.38 per exam taker (Column (2)), but more students have to go to other provinces for college (Column (4)) compared to the baseline. This is because the high-achieving students from other regions who find colleges in the East attractive can now be admitted to those colleges under the nationwide admission, but were not originally due to the very low allocated quotas in the baseline. As a result, medium- and low-achieving students in the East lose those college seats in their home province under the nationwide admission. Contrary to the East, both Central and West see a slight decrease in college seats from 0.23 and 0.25 per exam taker in the baseline, which are already low, to 0.21 and 0.23, respectively. The Northeast sees the largest decrease in college opportunities given the lower average student quality and the inability to keep local college capacity for their own students when provincial quotas are removed. Given that the nationwide admission is based on a common standard, the average $\theta^0$ of admitted students is now very close across regions (Column (6)) compared to the baseline, so the admission fairness is restored by merit standard.

These changes of college opportunities under the merit-based nationwide admission are driven by the inequality in the student quality distribution across provinces at the end of high school. Importantly, the student quality, or merit, is measured by the stock of human capital before college. This can be viewed as an outcome which combines the student’s learning ability and the quality of education they have received up to high school graduation. Separating the two is beyond the ability of the measurement framework in this paper. In the counterfactual, the pre-college educa-

---

39Note that in the baseline, the shares of students going to in-province and out-of-province colleges are determined by how many quotas are allocated from in-province and out-of-province colleges.
tion quality and students’ pre-college human capital distribution in each province will be taken as
exogenous.

Eliminating the distortion of provincial quotas can potentially increase the sorting between
students and colleges. Table 7 compares the student-school match between the baseline and the
nationwide admission, following the approach in Dillon and Smith (2017). Students are divided
into four quartiles based on their pre-college human capital $\theta^0$. Colleges are also divided into
four quartiles based on the quality measure $V_s$, but are weighted by their capacity so that each
college quartile contains the same number of college seats. Each quartile cell in Table 7 gives
the percentage of students falling into the corresponding combination of student-college quality
quartiles. Within each cell, the 2008 baseline is reported in the first row and the counterfactual
nationwide admission is reported in the second row in brackets. Overall, students concentrate along
the diagonal, indicating assortative matching, but there is substantial mismatch. Locating in the
three upper-right cells or the three lower-left cells means the distance between student and college
quality is at least two quartiles, indicating substantial overmatch (upper-right cells) or undermatch
(lower-left cells). Under the baseline quota allocation, 5.60% and 5.03% of all students end up
in these two areas, respectively. In contrast, the merit-based nationwide admission reduces the
percentage of large mismatch to 1.75% in overmatch and 0.75% in undermatch. Importantly, the
assortative matching is not perfect even when students are free to choose college according to
their national ranking. As found in the structural model estimates, students value not only college
quality but also other characteristics like location. Another possible reason is that peer quality is
not included in the college quality measure $V_s$. When students and colleges are more assortatively
matched, the peer quality in better schools also increases, which should further increase the sorting.
As a result, the positive effect of the nationwide admission on student-college sorting found in this
analysis should be considered as a lower bound.

7.3 A closer look of winners and losers

It is encouraging that the nationwide admission increases college education efficiency through
higher quality cutoff of the admitted students and better student-college sorting. However, because
the total college capacity is held constant, there are both winners and losers in the counterfactual
system. In Table 8, the NCEE exam takers are grouped by their national percentile of $\theta^0$ and home
region. Column (1) and (2) compare the admission rate between the nationwide admission and the 2008 baseline. Column (3) and (4) compares the average quality of the attended college. High-achieving students (ranked between the 90th and 100th percentile nationally) are all winners. They benefit from the merit-based national competition and attend higher-quality colleges, especially those from the East, where the student quality is much higher. For lower ranked students, college opportunities in general decrease in terms of both quantity and quality in the three non-East regions. The released college seats instead become accessible to students in the East, although as shown in Table 6, many of them are located in other provinces and thus require extra migration at the college stage. Lastly, given that the overall four-year college capacity in 2008 is 0.28 per NCEE exam taker, all students ranked between the 60th and 70th percentile nationally lose their college seats under the nationwide admission.

I further examine the winners and losers among colleges. Table 9 compares the average quality ($\theta^0$) of admitted students by college location and tier. When college seats are open to all students in the country under the nationwide admission, the top-tier colleges in the East admit better students from all regions due to the higher college quality and more attractive location. In contrast, the top-tier colleges in West and Northeast lose some of the good incoming students they originally have in the baseline. Interestingly, the winners and losers among bottom-tier colleges are reversed in terms of location. The bottom-tier colleges in Central, West and Northeast benefit from the much higher national cutoff of student quality under the nationwide admission and admit better students. In comparison, the bottom-tier colleges in the East see a slight quality drop in their incoming students because the college access expands to lower ranked students in the East, but the average student quality is still higher than that in other regions.

7.4 Spatial distribution of college graduates

The re-distributed college seats across provinces further affect students’ post-college choices. Table 10 gives the share of college graduates going to graduate school by college tier under the baseline and the nationwide admission. The capacity of graduate school is assumed to be flexible in the counterfactual. Due to increased sorting between students and colleges and the complementarity in return from graduate studies, more students choose to go to graduate school regardless of college tier, but the magnitude of the increase is small. The offsetting force comes from the re-
distributed migration patterns during college. In the baseline, graduate school is a more attractive option for many students from less developed provinces who have to attend a local college due to quota constraints and thus have a disadvantage in migrating to developed regions. It is less the case under the nationwide admission where these students have better access to colleges in more preferred locations. As a result, the overall share of college graduates choosing graduate studies increases by one percentage point from 16.5% to 17.5%.

Table 11 reports the spatial distribution change of college graduates who choose to enter the labor market in the baseline and the counterfactual. It is important to note that local labor demand is treated as exogenous in the counterfactual and does not respond to the changes in supply of college graduates, so that the wage level for college graduates in each location is fixed (Kennan and Walker, 2011). The result should be considered as capturing the short-run effect if one or two cohorts of college students are affected by the counterfactual admission, which is a small change in labor supply compared to the entire existing workforce. The assumption of exogenous skill prices is also consistent with recent evidence that the return to college education does not respond to changes in local skill supply due to endogenous skill-based technology adoption of firms (e.g. Feng and Xia, 2018).

Columns (1)-(3) of Table 11 report the number of college graduates working in each region. Under the nationwide admission, the number of college graduates choosing to live and work in the developed East region increases by 10%. On one hand, there are more students from other regions attending colleges in the East. Many of them stay there after college as the East provinces are more attractive and graduating from a college there could help them find a job in the same region. On the other hand, although there are also more students from the East having to attend colleges in other regions, as shown in Table 6, most of them move back after college. In contrast, fewer college graduates choose Central and West. The Northeast suffers the most from the human capital supply drop; it loses one third of new college graduates compared to the baseline. This re-distribution of college-educated workers is a combined outcome of the re-distribution of college seats and the subsequent change in migration decisions.

Despite the distributional change, worker quality generally increases in all regions, benefiting from the more efficient human capital production during college under the nationwide admission. Combining these effects, Columns (4)-(6) of Table 11 calculate the sum of the first year’s earnings
of all new college graduates working in each region, which is used as a proxy for total output. The total earnings increase in the East by 12%, but decrease by 11%, 13% and 32% in Central, West and Northeast, respectively. Importantly, at the national level, the merit-based nationwide college admission increases total earnings for the entire cohort by 1.7%, which is a substantial gain in efficiency considering that China’s current GDP annual growth rate is around 6%. Lastly, the average student life-time utility increases by 1.2%. This is a combination of the increased average life-time earnings and the higher average amenity value as more college graduates living and working in the more attractive East region.

7.5 Local subsidy to re-distribute college graduates

Given that the spatial inequality in college graduates’ distribution under the nationwide admission is enlarged, as reported in Table 11, I study whether a place-based subsidy can help the less developed regions (Central, West, and Northeast) to attract more college-educated workers. Since 2016, many local governments in China introduced place-based subsidies to attract new college graduates. Some of these policies provide a one-time cash incentive if students choose to live and work in that location, while some others provide subsidies for renting or purchasing a house. In this subsection, I simulate an one-time region-based subsidy of 10,000 CNY, which is the typical amount observed in current policies, under the nationwide admission analyzed above. The subsidy is provided to any college graduates who choose to work in the subsidized region. The average starting annual salary for the cohort in the 2008 baseline is 27,793 CNY when they entered the labor market in 2012, so the subsidy is about 36% of the starting salary.

Table 12 reports the effect of the region-based subsidy. The counterfactual subsidy is implemented in only one region at a time and the rest of the regions will not be subsidized. Each row in Table 12 reports the result when the subsidy is implemented in that specific region. Overall, I see a modest effect of the 10,000 CNY subsidy on attracting college graduates. The effect is relatively larger when working in Central, West or Northeast is subsidized; 2.5-2.8% more college graduates

40The amount of one-time subsidies for four-year college graduates typically ranges from 5,000 to 15,000 CNY across places.

41All values are in 2012 CNY. The average starting salary of 27,793 CNY is equal to 4,632 USD (1 USD ≈ 6 CNY in 2012). The average starting salary for four-year college graduates in the US is 44,259 USD according to the National Association of Colleges and Employers. This gap is roughly consistent with the GDP per capita difference between China and the US in 2012: 6,316 and 51,603 USD, respectively.
choose to work in the subsidized region (Column (1) in Table 12) compared to the number under
the nationwide admission without subsidies as reported in Table 11 Column (2). The subsidy has a
smaller effect if it is implemented in the East compared to Central, West or Northeast. I then look
at who responds to the subsidy and moves to the subsidized region when a different residential
region would be chosen if there is no subsidy. As reported in Table 12 Column (2), more than half
of the extra workers attracted by Central, West or Northeast grew up in the same region. In this
sense, some of the college students they lose at the college stage under the nationwide admission
can be brought back using subsidies.

However, the real cost of attracting one extra college graduate is very high, because the subsidy
has to be provided to all college graduates working in the subsidized region while most of them
would make the same residential choice even without the subsidy. Dividing the total amount of
subsidy provided in one region by the number of extra college graduates attracted to that region
gives the real cost for one extra worker, which is reported in Table 12 Column (3). The real cost
is 36 to 62 times larger than the nominal 10,000 CNY subsidy per worker. Moreover, the re-
distribution will be zero if all provinces or regions implement the same subsidy and the policy is
canceled out by each other. Since it would be actually easier for a richer region to provide such a
costly subsidy given their higher government revenue, it is important to have coordination at the
central government level if the policy goal is to reduce the brain drain in less developed regions.
Perhaps a more cost-effective approach to implement the subsidy is to focus on a set of cities or
industries instead of the whole province or region, but analyzing such subsidies would require a
spatial model at the city or industry level so I leave them for future research.

8 Conclusion

This paper examines how the spatial allocation of college seats affects college education effi-
ciency and regional inequality by studying China’s province-based admission quotas. I first build
a measurement model to recover the nationally comparable distribution of pre-college human cap-
it among the NCEE exam takers in each province. The recovered student quality is not fully
consistent with the observed admission quota allocation across provinces by merit standard. I then
estimate a structural model of college choice and migration decision and simulate a merit-based na-
tionwide admission. Eliminating provincial quota constraints increases total human capital output during college through better student-college sorting, but this efficiency gain comes together with the cost of the lower college opportunity in less developed regions and the smaller share of college graduates working there, which will further increase regional inequality. I also find region-based subsidy is very costly to re-balance the unequal distribution of college-educated workers under the nationwide admission.

The result in this paper highlights a policy tradeoff between efficiency in human capital production during college and equal distribution of college access and human capital in the labor market. The noisy relationship between the recovered student quality and the observed admission quota size across provinces seems to suggest the current policy objective in China is a mix of efficiency and equality. In particular, although I find a merit-based admission should further reduce the college seats allocated to Central and West, the Ministry of Education is working towards a different direction. Starting from 2008, the Ministry launched a region-based affirmative action, the College Admission Cooperation Plan (Ministry of Education, 2008). The provinces with larger college capacities, mainly in East and Northeast, are required to allocate extra admission quotas to Central and West to increase the college admission rate in those provinces. The discrepancy between this paper’s result and the current policy suggests the importance for future research to understand the quota allocation decision made by government and individual colleges, which is exogenous in this paper. Chan et al. (2019) takes an important step in this direction by modeling the quota allocation of elite colleges in China, although the role of government is not explicitly considered in their analysis. Another direction worth future research is to examine the long-run effect of the spatial distributional change of college graduates due to college admission changes. This places the design of college admission system in a larger policy space and is potentially beneficial towards a more coordinated spatial policy design (Fajgelbaum and Gaubert, 2020; Liu et al., 2018).
References


48


Figures and tables

Figure 1: Four-year college admission quotas per NCEE exam taker by province, 2006-2011

Note: Each bar represents the total number of four-year college admission quotas received by each province divided by the total number of NCEE exam takers in that province, averaged across years between 2006 and 2011.
Figure 2: Four-year college admission quotas and capacity vs. GDP per capita

Note: Provincial GDP per capita series in 2008 are from the National Bureau of Statistics of China. College capacity per capita is defined as the total number of new college seats in all four-year colleges located in a province divided by the total number of NCEE exam taker in that province.
Figure 3: Mean and standard deviation of $\theta^0$ of NCEE exam takers by province

Note: The pre-college human capital $\theta^0$ of NCEE exam takers in each province, $F_h(\theta^0)$, is assumed to follow a normal distribution. Mean and standard deviation of $\theta^0$ in each province are estimated using the measurement model and standardized to set the national mean to zero and national standard deviation to 1. In Appendix, Table A3 reports the exact estimates and standard errors of the mean and standard deviation, and Figure A3 plots the distribution of $\theta^0$ in each province.
Figure 4: Pre-college human capital $\theta^0$ vs. GDP and pre-college education expenditure

Note: Provincial GDP per capita series in 2008 are from the National Bureau of Statistics of China. Per student pre-college education expenditure data in 2008 are from the Ministry of Education of China and are defined as the per student total expenditure averaged over primary, middle and high school in each province.
Figure 5: Mean of $\theta^0$ vs. the average number of four-year college quotas per NCEE exam taker, 2006-2011

Figure 6: $\theta^0$ of the marginally admitted students under college admission quotas in each province
Figure 7: Model fit: average NCEE scores of admitted students

Note: Each dot represents the average NCEE scores of students admitted by each tier of colleges from each province \((h)\), track \((z)\) and year \((t)\) in administrative records and simulation. All raw NCEE scores are rescaled to be between 0 and 1.
Figure 8: Model fit: provincial distribution of college graduates

Note: The distribution is calculated based on the CCSS-surveyed schools only, because the administrative data do not track students after college admission. The observed distribution of college graduates from the CCSS data is adjusted by the sampling weights. The simulated distribution is calculated using all students who graduate from the CCSS-surveyed schools in simulation. The distribution is calculated combining all cohorts ($t$).
Figure 9: Model fit: share of college graduates working in home and college province

Note: Same as Figure 8, the share in each sub-figure is calculated based on the CCSS-surveyed schools only. All numbers are averaged across cohorts ($t$).
Table 1: Data pattern of migration at college and work stages

<table>
<thead>
<tr>
<th>Attend college in</th>
<th>Work in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Home province</td>
</tr>
<tr>
<td>Home province</td>
<td>69.1%</td>
</tr>
<tr>
<td>Non-home province</td>
<td>30.9%</td>
</tr>
</tbody>
</table>

*Note:* Data from the administrative data on college admission between 2006 and 2011 and the Chinese College Student Survey (CCSS) covering the same cohorts. See Subsection 5.1 for more data details.
Table 2: Parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E. (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>College</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>College quality $V_s$</td>
<td>0.027</td>
<td>0.019</td>
</tr>
<tr>
<td>Elite college (Project-211 or above)</td>
<td>0.070</td>
<td>0.013</td>
</tr>
<tr>
<td>Comprehensive college (b)</td>
<td>0.123</td>
<td>0.032</td>
</tr>
<tr>
<td>STEM college</td>
<td>0.081</td>
<td>0.046</td>
</tr>
<tr>
<td>Econ college</td>
<td>0.130</td>
<td>0.050</td>
</tr>
<tr>
<td>College in capital city of a province</td>
<td>0.017</td>
<td>0.010</td>
</tr>
<tr>
<td>College at home province</td>
<td>0.283</td>
<td>0.016</td>
</tr>
<tr>
<td>Distance[home, college] (c)</td>
<td>-0.254</td>
<td>0.055</td>
</tr>
<tr>
<td>Distance[home, college] squared</td>
<td>0.045</td>
<td>0.019</td>
</tr>
<tr>
<td><strong>Working</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work at home province</td>
<td>1.084</td>
<td>0.071</td>
</tr>
<tr>
<td>Work at home province*home is Beijing/Shanghai (d)</td>
<td>1.394</td>
<td>0.171</td>
</tr>
<tr>
<td>Distance[work, home]</td>
<td>-0.963</td>
<td>0.101</td>
</tr>
<tr>
<td>Distance[work, home] squared</td>
<td>0.141</td>
<td>0.046</td>
</tr>
<tr>
<td>Work at college location</td>
<td>0.905</td>
<td>0.059</td>
</tr>
<tr>
<td>Distance[work, college]</td>
<td>-1.008</td>
<td>0.077</td>
</tr>
<tr>
<td>Distance[work, college] squared</td>
<td>0.179</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>Unobserved heterogeneity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of indifferent type ($\tau = \emptyset$)</td>
<td>0.216</td>
<td>0.085</td>
</tr>
<tr>
<td>Size of location preference during college $\chi$</td>
<td>0.011</td>
<td>0.004</td>
</tr>
<tr>
<td>Size scalar for working periods $\rho$</td>
<td>17.640</td>
<td>0.927</td>
</tr>
<tr>
<td><strong>Grad school</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>193.540</td>
<td>0.193</td>
</tr>
<tr>
<td>Pre-college human capital $\theta^0$</td>
<td>0.690</td>
<td>0.094</td>
</tr>
<tr>
<td>College quality $V_s$</td>
<td>0.932</td>
<td>0.034</td>
</tr>
<tr>
<td>$\theta^0*V_s$</td>
<td>0.643</td>
<td>0.050</td>
</tr>
<tr>
<td>Comprehensive college</td>
<td>-0.661</td>
<td>0.089</td>
</tr>
<tr>
<td>STEM college</td>
<td>-0.420</td>
<td>0.066</td>
</tr>
<tr>
<td>Econ college</td>
<td>-2.131</td>
<td>0.074</td>
</tr>
<tr>
<td><strong>Earnings</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>7.019</td>
<td>0.145</td>
</tr>
<tr>
<td>Pre-college human capital $\theta^0$</td>
<td>0.062</td>
<td>0.030</td>
</tr>
<tr>
<td>College quality $V_s$</td>
<td>0.082</td>
<td>0.018</td>
</tr>
<tr>
<td>$\theta^0*V_s$</td>
<td>0.026</td>
<td>0.009</td>
</tr>
<tr>
<td>Log provincial average income</td>
<td>0.264</td>
<td>0.014</td>
</tr>
<tr>
<td>Standard deviation of earning shock $\sigma_e$</td>
<td>0.389</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Note: (a) Standard errors are calculated through bootstrap. (b) Comprehensive colleges typically offer all major fields of study. STEM colleges are technology-/engineering-/science-focused schools. Econ colleges mainly offer majors in economics and business. The omitted reference type is non-STEM colleges that mainly offer field of studies in humanities. (c) All distances are measured in 1,000 km. (d) This interaction is to capture the potential advantage of being a local student in Beijing or Shanghai due to the special Hukou and migration restriction in these two top Chinese cities.
Table 3: Location preferences in monetary value

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Annual value (in 2012 CNY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>College at home province</td>
<td>¥1,164</td>
</tr>
<tr>
<td>Distance[home, college]</td>
<td>¥-845 (1,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-1,322 (2,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-1,432 (3,000km)</td>
</tr>
<tr>
<td>Work at home province</td>
<td>¥4,563</td>
</tr>
<tr>
<td>Work at home province*home is Beijing/Shanghai</td>
<td>¥5,922</td>
</tr>
<tr>
<td>Distance[work, home]</td>
<td>¥-3,245 (1,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-5,335 (2,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-6,269 (3,000 km)</td>
</tr>
<tr>
<td>Work at college location</td>
<td>¥3,790</td>
</tr>
<tr>
<td>Distance[work, college]</td>
<td>¥-3,265 (1,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-5,056 (2,000 km)</td>
</tr>
<tr>
<td></td>
<td>¥-5,382 (3,000 km)</td>
</tr>
<tr>
<td>Unobserved preference for work location</td>
<td>¥796</td>
</tr>
</tbody>
</table>

Note: The estimated payoff parameters are converted to annual monetary values following the procedure described in Section 6.2.

Table 4: Model fit: share of college graduates going to graduate school

<table>
<thead>
<tr>
<th>College tier</th>
<th>Data</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>0.317</td>
<td>0.312</td>
</tr>
<tr>
<td>Middle</td>
<td>0.183</td>
<td>0.182</td>
</tr>
<tr>
<td>Bottom</td>
<td>0.077</td>
<td>0.089</td>
</tr>
</tbody>
</table>

Note: The share is calculated based on the CCSS-surveyed schools only, because the administrative data do not track students after college admission. The observed share of choosing graduate studies from the CCSS data is adjusted by the sampling weights. The simulated share is calculated using all students who graduate from the CCSS-surveyed schools in simulation. All numbers are averaged across cohorts (t).
Table 5: College resources and student quality by region

<table>
<thead>
<tr>
<th>Region</th>
<th>Capacity per capita of local colleges (1)</th>
<th>Average quality ($V_s$) of local college (2)</th>
<th>Average quality ($\theta^0$) of local NCEE exam takers (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>0.34</td>
<td>0.34</td>
<td>0.31</td>
</tr>
<tr>
<td>Central</td>
<td>0.21</td>
<td>-0.14</td>
<td>-0.23</td>
</tr>
<tr>
<td>West</td>
<td>0.23</td>
<td>-0.09</td>
<td>-0.14</td>
</tr>
<tr>
<td>Northeast</td>
<td>0.49</td>
<td>-0.12</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Note: Regions are grouped provinces. College capacity per capita is defined as the number of local college seats per NCEE exam taker in 2008.

Table 6: Change in share and quality of admitted students by home region

<table>
<thead>
<tr>
<th>Home region</th>
<th>No. of admitted p.c.(a)</th>
<th>% attend local college(b)</th>
<th>Average $\theta^0$ of admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline in 2008 (1)</td>
<td>Baseline in 2008 (3)</td>
<td>Baseline in 2008 (5)</td>
</tr>
<tr>
<td></td>
<td>Nationwide admission (2)</td>
<td>Nationwide admission (4)</td>
<td>Nationwide admission (6)</td>
</tr>
<tr>
<td>East</td>
<td>0.31</td>
<td>78%</td>
<td>1.28</td>
</tr>
<tr>
<td>Central</td>
<td>0.23</td>
<td>69%</td>
<td>1.12</td>
</tr>
<tr>
<td>West</td>
<td>0.25</td>
<td>63%</td>
<td>1.08</td>
</tr>
<tr>
<td>Northeast</td>
<td>0.43</td>
<td>73%</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Note: (a) Per capita (p.c.) is defined as the number of local college seats per NCEE exam taker. (b) A local college is the one located at a student's residence province.

Table A6 in Appendix reports the changes in the share of admitted students at the provincial level under the nationwide admission.
Table 7: The joint distribution of student and college quality

<table>
<thead>
<tr>
<th>$\theta^0$ quartile</th>
<th>College quality $V_s$ quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
</tr>
<tr>
<td>Q1 (top)</td>
<td>17.31</td>
</tr>
<tr>
<td></td>
<td>[19.53]</td>
</tr>
<tr>
<td>Q2</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td>[5.36]</td>
</tr>
<tr>
<td>Q3</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>[0.29]</td>
</tr>
<tr>
<td>Q4 (bottom)</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>[0]</td>
</tr>
</tbody>
</table>

Note: In each quartile cell, the first row reports the baseline in 2008, and the second row with brackets reports the counterfactual nationwide admission. Each number is the overall percentage and represents the share of the population in each quartile cell. All percentages add up to 100 in the baseline rows and counterfactual rows, respectively.
Table 8: Winners and losers by home region under the nationwide admission

<table>
<thead>
<tr>
<th>National percentile of $\theta^0$</th>
<th>Home region</th>
<th>Share of admitted</th>
<th>Average $V_\theta$ attended</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Baseline in 2008 (1)</td>
<td>Nationwide admission (2)</td>
</tr>
<tr>
<td>[90,100] East</td>
<td>100%</td>
<td>100%</td>
<td>1.14</td>
</tr>
<tr>
<td>Central</td>
<td>100%</td>
<td>100%</td>
<td>1.35</td>
</tr>
<tr>
<td>West</td>
<td>100%</td>
<td>100%</td>
<td>1.37</td>
</tr>
<tr>
<td>Northeast</td>
<td>100%</td>
<td>100%</td>
<td>1.34</td>
</tr>
<tr>
<td>[80,90) East</td>
<td>89%</td>
<td>100%</td>
<td>0.06</td>
</tr>
<tr>
<td>Central</td>
<td>100%</td>
<td>100%</td>
<td>0.14</td>
</tr>
<tr>
<td>West</td>
<td>100%</td>
<td>100%</td>
<td>0.22</td>
</tr>
<tr>
<td>Northeast</td>
<td>100%</td>
<td>100%</td>
<td>0.19</td>
</tr>
<tr>
<td>[70,80) East</td>
<td>42%</td>
<td>78%</td>
<td>-0.46</td>
</tr>
<tr>
<td>Central</td>
<td>92%</td>
<td>77%</td>
<td>-0.44</td>
</tr>
<tr>
<td>West</td>
<td>81%</td>
<td>77%</td>
<td>-0.33</td>
</tr>
<tr>
<td>Northeast</td>
<td>100%</td>
<td>78%</td>
<td>-0.17</td>
</tr>
<tr>
<td>[60,70) East</td>
<td>8%</td>
<td>0%</td>
<td>-0.67</td>
</tr>
<tr>
<td>Central</td>
<td>9%</td>
<td>0%</td>
<td>-0.57</td>
</tr>
<tr>
<td>West</td>
<td>23%</td>
<td>0%</td>
<td>-0.47</td>
</tr>
<tr>
<td>Northeast</td>
<td>81%</td>
<td>0%</td>
<td>-0.55</td>
</tr>
</tbody>
</table>

*Note: The national percentile is constructed using the national distribution of $\theta^0$. 
### Table 9: Change in quality of admitted students by college region and tier

<table>
<thead>
<tr>
<th>College region</th>
<th>Baseline in 2008</th>
<th>Nationwide admission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-tier (1)</td>
<td>Middle-tier (2)</td>
</tr>
<tr>
<td></td>
<td>Top-tier (4)</td>
<td>Middle-tier (5)</td>
</tr>
<tr>
<td>All</td>
<td>1.81 (+0.06)</td>
<td>1.17 (+0.01)</td>
</tr>
<tr>
<td>East</td>
<td>1.97 (+0.12)</td>
<td>1.27 (+0.00)</td>
</tr>
<tr>
<td>Central</td>
<td>1.64 (+0.04)</td>
<td>1.14 (+0.01)</td>
</tr>
<tr>
<td>West</td>
<td>1.65 (-0.06)</td>
<td>1.09 (+0.01)</td>
</tr>
<tr>
<td>Northeast</td>
<td>1.77 (-0.07)</td>
<td>1.04 (+0.04)</td>
</tr>
</tbody>
</table>

*Note:* Numbers in parentheses are the changes of student quality in the same college region and tier under the nationwide admission, compared to the 2008 baseline.

### Table 10: Share of college graduates going to graduate school by college tier

<table>
<thead>
<tr>
<th>College tier</th>
<th>Baseline in 2008</th>
<th>Nationwide admission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>0.305</td>
<td>0.319</td>
</tr>
<tr>
<td>Middle</td>
<td>0.177</td>
<td>0.178</td>
</tr>
<tr>
<td>Bottom</td>
<td>0.073</td>
<td>0.091</td>
</tr>
<tr>
<td>Average</td>
<td>0.165</td>
<td>0.175</td>
</tr>
</tbody>
</table>
Table 11: Distribution of college graduates and labor market outputs

<table>
<thead>
<tr>
<th>Work region</th>
<th>Number of graduates working in</th>
<th>Total initial earnings (billion CNY)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline in 2008</td>
<td>Nationwide admission</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>East</td>
<td>1,293,127</td>
<td>1,420,158</td>
</tr>
<tr>
<td>Central</td>
<td>340,793</td>
<td>300,331</td>
</tr>
<tr>
<td>West</td>
<td>433,649</td>
<td>374,041</td>
</tr>
<tr>
<td>Northeast</td>
<td>146,160</td>
<td>97,042</td>
</tr>
<tr>
<td>Total</td>
<td>2,213,729</td>
<td>2,191,572</td>
</tr>
</tbody>
</table>

Note: Earnings are measured in 2012 CNY value. 2012 is the year when the cohort admitted in 2008 graduate and enter the labor market.

Table 12: Effect of 10,000 CNY region-based subsidy on the distribution of college graduates

<table>
<thead>
<tr>
<th>Subsidized region</th>
<th>Change in no. of workers</th>
<th>% of movers originally from local region</th>
<th>Real cost (CNY) for one extra worker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>East</td>
<td>1.6%</td>
<td>24.1%</td>
<td>¥622,449</td>
</tr>
<tr>
<td>Central</td>
<td>2.6%</td>
<td>65.1%</td>
<td>¥390,528</td>
</tr>
<tr>
<td>West</td>
<td>2.8%</td>
<td>56.4%</td>
<td>¥367,559</td>
</tr>
<tr>
<td>Northeast</td>
<td>2.5%</td>
<td>72.9%</td>
<td>¥403,657</td>
</tr>
</tbody>
</table>

Note: This counterfactual analysis is based on the nationwide admission and add an one-time region-based subsidy of 10,000 CNY that is provided to any college graduates who choose to work in the subsidized region. The subsidy is available in only one region at a time. Each row reports the result when that specific region is subsidized.
Appendix A  Identification proof for the measurement model

This appendix provides identification proof for the measurement model of the latent pre-college human capital in Section 4.

To reduce notational burdens, ignore NCEE track $z$ and year $t$. The identification will be the same when adding these dimensions back. Consider province $h$ and $h'$, and college $s$ and $s'$:

\[
NCEE_h = \kappa_{1,h} + \lambda_{1,h}\theta_0 + \sigma_{1,h}\nu_1 \quad NCEE_{h'} = \kappa_{1,h'} + \lambda_{1,h'}\theta_0 + \sigma_{1,h'}\nu_1 \\
GPA_s = \kappa_{2,s} + \lambda_{2,s}\theta_0 + \sigma_{2,s}\nu_2 \quad GPA_{s'} = \kappa_{2,s'} + \lambda_{2,s'}\theta_0 + \sigma_{2,s'}\nu_2
\]

Normalize $\kappa_{1,h} = 0$ and $\lambda_{1,h} = 1$. There are four groups of students depending on $h \times s$ combination.

To begin, calculate the variances and covariances of NCEE scores and college GPAs for the four groups:

\[
\begin{align*}
\text{var}(NCEE_h|s) &= \text{var}(\theta_0|h, s) + \sigma_{1,h}^2\text{var}(\nu_1|h, s) + 2\sigma_{1,h}\text{cov}(\theta_0, \nu_1|h, s) \\
\text{var}(NCEE_h|s') &= \text{var}(\theta_0|h, s') + \sigma_{1,h}^2\text{var}(\nu_1|h, s') + 2\sigma_{1,h}\text{cov}(\theta_0, \nu_1|h, s') \\
\text{var}(NCEE_{h'}|s) &= \lambda_{1,h'}^2\text{var}(\theta_0|h', s) + \sigma_{1,h'}^2\text{var}(\nu_1|h', s) + 2\lambda_{1,h'}\sigma_{1,h'}\text{cov}(\theta_0, \nu_1|h', s) \\
\text{var}(NCEE_{h'}|s') &= \lambda_{1,h'}^2\text{var}(\theta_0|h', s') + \sigma_{1,h'}^2\text{var}(\nu_1|h', s') + 2\lambda_{1,h'}\sigma_{1,h'}\text{cov}(\theta_0, \nu_1|h', s') \\
\text{var}(GPA_s|h) &= \lambda_{2,s}^2\text{var}(\theta_0|h, s) + \sigma_{2,s}^2 \\
\text{var}(GPA_s|h') &= \lambda_{2,s}^2\text{var}(\theta_0|h', s) + \sigma_{2,s}^2 \\
\text{var}(GPA_{s'}|h) &= \lambda_{2,s'}^2\text{var}(\theta_0|h, s') + \sigma_{2,s'}^2 \\
\text{var}(GPA_{s'}|h') &= \lambda_{2,s'}^2\text{var}(\theta_0|h', s') + \sigma_{2,s'}^2 \\
\text{cov}(NCEE_h, GPA_s) &= \lambda_{2,s}\text{var}(\theta_0|h, s) + \lambda_{2,s}\sigma_{1,h}\text{cov}(\theta_0, \nu_1|h, s) \\
\text{cov}(NCEE_h, GPA_{s'}) &= \lambda_{2,s'}\text{var}(\theta_0|h, s') + \lambda_{2,s'}\sigma_{1,h}\text{cov}(\theta_0, \nu_1|h, s') \\
\text{cov}(NCEE_{h'}, GPA_s) &= \lambda_{1,h'}\lambda_{2,s}\text{var}(\theta_0|h', s) + \lambda_{2,s}\sigma_{1,h'}\text{cov}(\theta_0, \nu_1|h', s) \\
\text{cov}(NCEE_{h'}, GPA_{s'}) &= \lambda_{1,h'}\lambda_{2,s'}\text{var}(\theta_0|h', s') + \lambda_{2,s'}\sigma_{1,h'}\text{cov}(\theta_0, \nu_1|h', s')
\end{align*}
\]

Canceling out each $\text{var}(\theta_0|h, s)$ in the variance equations using the corresponding covariance equa-
tion gives:

\[
\text{var}(NCEE_h | s) = \frac{1}{\lambda_{2,s}} \text{cov}(NCEE_h, GPA_s) + \sigma^2 \text{var}(\nu_1 | h, s) + \sigma_{1,h} \text{cov}(\theta^0, \nu_1 | h, s)
\]

\[
\text{var}(NCEE_h | s') = \frac{1}{\lambda_{2,s'}} \text{cov}(NCEE_h, GPA_{s'}) + \sigma^2 \text{var}(\nu_1 | h, s') + \sigma_{1,h} \text{cov}(\theta^0, \nu_1 | h, s')
\]

\[
\text{var}(NCEE_{h'} | s) = \frac{\lambda_{1,h'}}{\lambda_{2,s}} \text{cov}(NCEE_{h'}, GPA_s) + \sigma^2 \text{var}(\nu_1 | h', s) + \lambda_{1,h'} \sigma_{1,h'} \text{cov}(\theta^0, \nu_1 | h', s)
\]

\[
\text{var}(NCEE_{h'} | s') = \frac{\lambda_{1,h'}}{\lambda_{2,s'}} \text{cov}(NCEE_{h'}, GPA_{s'}) + \sigma^2 \text{var}(\nu_1 | h', s') + \lambda_{1,h'} \sigma_{1,h'} \text{cov}(\theta^0, \nu_1 | h', s')
\]

\[
\text{var}(GPA_s | h) = \lambda_{2,s} \text{cov}(NCEE_h, GPA_s) - \lambda_{2,s} \sigma_{1,h} \text{cov}(\theta^0, \nu_1 | h, s) + \sigma_{2,s}
\]

\[
\text{var}(GPA_{s'} | h) = \lambda_{2,s'} \text{cov}(NCEE_h, GPA_{s'}) - \lambda_{2,s'} \sigma_{1,h} \text{cov}(\theta^0, \nu_1 | h, s') + \sigma_{2,s'}
\]

\[
\text{var}(GPA_s | h') = \frac{\lambda_{2,s}}{\lambda_{1,h'}} \text{cov}(NCEE_{h'}, GPA_s) - \frac{\lambda_{2,s}^2}{\lambda_{1,h'}} \sigma_{1,h'} \text{cov}(\theta^0, \nu_1 | h', s) + \sigma_{2,s}
\]

\[
\text{var}(GPA_{s'} | h') = \frac{\lambda_{2,s'}}{\lambda_{1,h'}} \text{cov}(NCEE_{h'}, GPA_{s'}) - \frac{\lambda_{2,s'}^2}{\lambda_{1,h'}} \sigma_{1,h'} \text{cov}(\theta^0, \nu_1 | h', s') + \sigma_{2,s'}
\]

Due to selection by \( \nu_1 \) in college choices, \( \text{var}(\nu_1 | \cdot) \) and \( \text{cov}(\theta^0, \nu_1 | \cdot) \) will be derived from the structural model. After they are known, the 8 equations above have 7 unknown parameters: \( \lambda_{1,h'} \), \( \lambda_{2,s} \), \( \lambda_{2,s'} \), \( \sigma_{1,h'} \), \( \sigma_{1,h} \), \( \sigma_{2,s} \) and \( \sigma_{2,s'} \), and they are over-identified.

Next, calculate the expected NCEE scores and college GPAs:

\[
E(NCEE_h | s) = E(\theta^0 | h, s) + \sigma_{1,h} E(\nu_1 | h, s)
\]

\[
E(NCEE_h | s') = E(\theta^0 | h, s') + \sigma_{1,h} E(\nu_1 | h, s')
\]

\[
E(NCEE_{h'} | s) = \kappa_{1,h'} + \lambda_{1,h'} E(\theta^0 | h', s) + \sigma_{1,h'} E(\nu_1 | h', s)
\]

\[
E(NCEE_{h'} | s') = \kappa_{1,h'} + \lambda_{1,h'} E(\theta^0 | h', s') + \sigma_{1,h'} E(\nu_1 | h', s')
\]

\[
E(GPA_s | h) = \kappa_{2,s} + \lambda_{2,s} E(\theta^0 | h, s)
\]

\[
E(GPA_{s'} | h) = \kappa_{2,s'} + \lambda_{2,s'} E(\theta^0 | h, s')
\]

\[
E(GPA_s | h') = \kappa_{2,s} + \lambda_{2,s} E(\theta^0 | h', s)
\]

\[
E(GPA_{s'} | h') = \kappa_{2,s'} + \lambda_{2,s'} E(\theta^0 | h', s')
\]
Canceling out each $E(\theta_0^0 | \cdot)$ in above equations gives:

$$E(GPA_s | h) = \kappa_{2,s} + \lambda_{2,s} \left( E(NCEE_h | s) - \sigma_{1,h} E(\nu_1 | h, s) \right)$$

$$E(GPA_s' | h) = \kappa_{2,s'} + \lambda_{2,s'} \left( E(NCEE_h | s') - \sigma_{1,h} E(\nu_1 | h, s') \right)$$

$$E(GPA_s | h^{'}) = \kappa_{2,s} + \frac{\lambda_{2,s}}{\lambda_{1,h'}} \left( E(NCEE_{h^{'}} | s) - \kappa_{1,h'} - \sigma_{1,h'} E(\nu_1 | h^{'}, s) \right)$$

$$E(GPA_s' | h^{'}) = \kappa_{2,s'} + \frac{\lambda_{2,s'}}{\lambda_{1,h'}} \left( E(NCEE_{h^{'}} | s') - \kappa_{1,h'} - \sigma_{1,h'} E(\nu_1 | h^{'}, s') \right)$$

Since $\lambda_{1,h'}, \lambda_{2,s}, \lambda_{2,s'}, \sigma_{1,h}$ and $\sigma_{1,h'}$ are already identified, $\kappa_{1,h'}, \kappa_{2,s}, \kappa_{2,s'}$ can be identified using the four equations above. Now all parameters in the two provinces and two colleges are identified.

To identify parameters for a third province $h''$, we need some of its students attend an identified college. Assume they attend college $s$. Calculate the conditional mean, variance and covariance of the NCEE score and college GPA:

$$E(NCEE_{h''} | s) = \kappa_{1,h''} + \lambda_{1,h''} E(\theta_0^0 | h'', s) + \sigma_{1,h''} E(\nu_1 | h'', s)$$

$$E(GPA_s | h'') = \kappa_{2,s} + \lambda_{2,s} E(\theta_0^0 | h'', s)$$

$$\text{var}(NCEE_{h''} | s) = \lambda_{1,h''}^2 \text{var}(\theta_0^0 | h'', s) + \sigma_{1,h''}^2 \text{var}(\nu_1 | h'', s) + 2 \lambda_{1,h''} \sigma_{1,h''} \text{cov}(\theta_0^0, \nu_1 | h'', s)$$

$$\text{var}(GPA_s | h'') = \lambda_{2,s}^2 \text{var}(\theta_0^0 | h'', s) + \sigma_{2,s}^2$$

$$\text{cov}(NCEE_{h''}, GPA_s) = \lambda_{1,h''} \lambda_{2,s} \text{var}(\theta_0^0 | h'', s) + \lambda_{2,s} \sigma_{1,h''} \text{cov}(\theta_0^0, \nu_1 | h'', s)$$

Cancelling out $E(\theta_0^0 | h'', s)$ and $\text{var}(\theta_0^0 | h'', s)$ gives the following three equations:

$$E(NCEE_{h''} | s) = \kappa_{1,h''} + \frac{\lambda_{1,h''}}{\lambda_{2,s}} \left( E(GPA_s | h'') - \kappa_{2,s} \right) + \sigma_{1,h''} E(\nu_1 | h'', s)$$

$$\text{var}(NCEE_{h''} | s) = \frac{\lambda_{1,h''}}{\lambda_{2,s}} \left( \text{cov}(NCEE_{h''}, GPA_s) - \lambda_{2,s} \sigma_{1,h''} \text{cov}(\theta_0^0, \nu_1 | h'', s) \right)$$

$$+ \sigma_{1,h''}^2 \text{var}(\nu_1 | h'', s) + 2 \lambda_{1,h''} \sigma_{1,h''} \text{cov}(\theta_0^0, \nu_1 | h'', s)$$

$$\text{var}(GPA_s | h'') = \frac{\lambda_{2,s}}{\lambda_{1,h''}} \left( \text{cov}(NCEE_{h''}, GPA_s) - \lambda_{2,s} \sigma_{1,h''} \text{cov}(\theta_0^0, \nu_1 | h'', s) \right) + \sigma_{2,s}^2$$

Since $\kappa_{2,s}, \lambda_{2,s}$ and $\sigma_{2,s}$ are known, $\kappa_{1,h''}, \lambda_{1,h''}$ and $\sigma_{1,h''}$ are identified once $E(\nu_1 | h'', s)$, $\text{var}(\nu_1 | h'', s)$ and $\text{cov}(\theta_0^0, \nu_1 | h'', s)$ are calculated from the structural model. By the same fashion, parameters for
a third college $s''$ can be identified as long as some of its students come from an identified province.

In China, all colleges admit students from more than one province, and many admit from all provinces. Since the GPA measurement in each college is assumed to be stable over time, the college GPA can link different cohorts of students within the same college. Given these facts and assumptions, the identification chain can be extended to all province-track-years and colleges.
Appendix B  Details on data and sample selection

B.1  Administrative data of college admission

B.1.1  Data availability

There are a few exceptions in the NCEE track setting and data availability:

• Guangdong: No track difference in 2006 (coded as sciences in the data and analysis); two tracks since 2007.

• Jiangsu: No track difference in 2006 and 2007 (coded as sciences in the data and analysis); two tracks from 2008. Due to regulatory restrictions, admission data between 2009 and 2011 are not available.

• Zhejiang: Due to regulatory restrictions, admission data in 2011 are not available.

Note that having missing data for a province in some (but not all) years do not affect the identification of the student quality distribution $F_h(\theta^0)$ in that province. Based on the measurement model, $F_h(\theta^0)$ can be identified using only one year’s data, although data from multiple years provide over-identification and improve estimation efficiency.

B.1.2  NCEE score outliers of admitted students

Some non-sports and non-arts colleges still offer sports- or arts-related majors, and most of their students are admitted through specialty tracks of the NCEE other than the standard sciences and humanities tracks. Many of these students have much lower NCEE scores. To prevent using their scores when forming the admission cutoff of each college, which will otherwise produce extremely low cutoffs, I drop all students in sports- and arts-related majors, which consist of 1.76% of the sample.

B.2  The Chinese College Student Survey (CCSS) data

B.2.1  Sampling issues

I use CCSS graduating class surveys from 2010 to 2015. After 2015, the sampling design was changed. In 2010, the CCSS randomly selected 100 colleges stratified by tier and region
(Municipalities, East other than Municipalities, Central, West, and Northeast), weighted by college capacity. The survey was rolled out among selected colleges during 2010-2015. Unfortunately, due to unexpected funding cut in 2014 and 2015, 10 colleges that were scheduled for the survey in those years didn’t participate. As a result, in total 90 colleges are surveyed, among which 8 are vocational colleges and 82 are four-year colleges. Among the 82 four-year colleges, one is a police school and is dropped from my sample because their students enter the police system after graduation and have restricted work location options. Another two colleges are also dropped because each of them has less than 10 students recorded in the sample due to unexpected issues in survey administration. Therefore, in total 79 four-year colleges are kept in the sample. Each college participated at least one year in the CCSS. Many colleges participated in multiple years.

Since some colleges that were originally selected in the sampling design are dropped from the final sample, I re-construct the school and individual weights to restore the national representativeness of the final sample using the school capacity information from the administrative data.

B.2.2 Sample selection

To begin, the 2010-2015 four-year college sample contains 33,848 students. 1,165 (3.44% of all students) are dropped from the sample because they are admitted through self-admission by each individual college, not through the regular NCEE. Another 2,054 (6.07% of all students) are dropped because they are in specialty tracks of the NCEE, mostly arts or sports. Same with the administrative sample, students from Tibet are also excluded from the CCSS sample. As a result, 30,629 students is left for the final sample and all of them are admitted through the regular NCEE in sciences or humanities track. On average, 205 students are surveyed each year in each school, which is about 9% of the average cohort size in one college.

B.2.3 Post-graduation choice and initial job

The CCSS collected student’s post-graduation choices during the survey, which is about two months before college graduation. For post-graduation choices (all numbers are weighted using adjusted sampling weights), 74.11% choose to enter the labor market after graduation, 19.20% go to graduate school or in preparation for graduate school entrance exam, 2.91% go overseas for study or work, and 3.78% do not have a plan yet. Among who choose to work after graduation,
89.31% have searched for jobs, and 78.41% have at least one offer at the time of the survey. Among who have offers, detailed information of the best job offer are collected. 71.28% have accepted this offer, 21.33% choose not to accept and continue to search, and 7.39% have not decided. Perhaps not surprisingly, the initial earnings in offers that are not accepted are about 13% lower, after controlling for schools, major, NCEE, college GPA, individual and family background. However, the distribution of the job location is very similar comparing offers that are accepted and rejected. Initial earnings are trimmed at the top and bottom 1%.

I use the accepted job offers to estimate the earnings equation and residential location choices in the model. I interpret the wage rate in accepted job offers as an unbiased (but certainly noisy) measure of the post-college human capital. I assume all students will eventually find and accept such a job offer, and treat those who have not at the time of the survey (reject current offers, no offer yet, or not start searching yet) as being in a frictional job search process. In terms of the location choices, given that the distribution of the job location is very similar comparing offers that are accepted and rejected, estimating location preferences using only the sample of accepted job offers is a less concern. Effectively, I assume there is no unobserved heterogeneity affecting the timing of receiving acceptable job offers, and the timing does not systematically varies across regions.
Appendix C  Joint estimation of the structural model and the measurement model

C.1 Integration over $\theta^0$ in the MLE of the structural model

Start with Equation (5.3):

$$\mathcal{L}(A|\Omega) = \int l(A|\theta^0)l(NCEE_{hzt}, GPA_s|\theta^0)dF_h(\theta^0).$$

Because both $\nu_1$ and $\nu_2$ follow standard normal distribution,

$$l(NCEE_{hzt}, GPA_s|\theta^0) = p(NCEE_{hzt}|\theta^0)p(GPA_s|\theta^0),$$

in which

$$p(NCEE_{hzt}|\theta^0) = p(\nu_1 = \frac{NCEE_{hzt} - \kappa_{1,hzt} - \lambda_{1,hzt}\theta^0}{\sigma_{1,hzt}}) = \phi(\frac{NCEE_{hzt} - \kappa_{1,hzt} - \lambda_{1,hzt}\theta^0}{\sigma_{1,hzt}});$$

$$p(GPA_s|\theta^0) = p(\nu_2 = \frac{GPA_s - \kappa_{2,s} - \lambda_{2,s}\theta^0}{\sigma_{2,s}}) = \phi(\frac{GPA_s - \kappa_{2,s} - \lambda_{2,s}\theta^0}{\sigma_{2,s}}).$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution.

The integration over $F_h(\theta^0)$ is approximated using a 7-point Gaussian quadrature over a local region:

$$\left[ \min \left\{(\theta^0|\nu_1 = 1.96), (\theta^0|\nu_2 = 1.96)\right\}, \max \left\{(\theta^0|\nu_1 = -1.96), (\theta^0|\nu_2 = -1.96)\right\} \right].$$

The likelihood of observing both $NCEE_{hzt}$ and $GPA_s$ with a $\theta^0$ outside this region will be lower than 0.05$^2 = 0.0025$, so those values of $\theta^0$ are ignored in integration. Denote the nodes and weights as $\{n_i^\theta, \omega_i^\theta\}_{i=1,\ldots,7}$, we get

$$\int \ldots dF_h(\theta^0) \approx \sum_{i=1}^{7} \omega_i^\theta f_h(n_i^\theta),$$

where $f_h(\cdot)$ is the probability density function of $\theta^0$ in province $h.$
C.2 Combining datasets in the MLE

Likelihood contribution from the administrative data and the CCSS will be combined in the estimation. However, simply adding up $L^{\text{Admin}}$ and $L^{\text{CCSS}}$ will double count the likelihood of college choices. Because the CCSS covers a random sample of individuals in the administrative data, which is the universe of all admitted college students in those cohorts, individuals surveyed by the CCSS will have their $p^s$ in both $L^{\text{Admin}}$ and $L^{\text{CCSS}}$. To avoid such double counting, I “link” the CCSS individuals back to the administrative records by matching their home province, track, NCEE score, admitted college and major.\textsuperscript{A1} When one CCSS student is matched with multiple students with same information in the administrative records, I randomly drop one from the administrative data. Thus, these “linked” individuals are included only in $L^{\text{CCSS}}$ but not $L^{\text{Admin}}$ to take advantage of their complete information from college choice to post-graduation choices while avoiding double counting.

C.3 Iteration in the joint estimation

In the joint estimation:

1. With some guess of $F^0_h(\theta^0)$ and $\{\kappa, \lambda, \sigma\}^0$, estimate the structural model using the MLE to get parameters $B^0$.

   (a) The starting $F^0_h(\theta^0)$ and $\{\kappa, \lambda, \sigma\}^0$ are estimated in GMM using the measurement model without selection terms, i.e. $E(\nu_1|hzt,s) = 0$, $\text{var}(\nu_1|hzt,s) = 1$ and $\text{cov}(\theta^0, \nu_1|hzt,s) = 0$.

   (b) The structural model is estimated in MLE:

   $$L(A|\Omega) = \int l(A|\theta^0)l(\text{NCEE}_{hzt}, GPA_s|\theta^0)dF^0_h(\theta^0),$$

   and get structural model parameters $B^0$.

2. Simulate individual choices using $B^0$, calculate the selection terms in the measurement model: $E(\nu_1|hzt,s)$, $\text{var}(\nu_1|hzt,s)$ and $\text{cov}(\theta^0, \nu_1|hzt,s)$.

\textsuperscript{A1}I do not have administrative identifiers for individuals surveyed in the CCSS and so cannot perform the exact linking.
(a) $E(\nu_1|hzt, s)$, $\text{var}(\nu_1|hzt, s)$ and $\text{cov}(\theta^0, \nu_1|hzt, s)$ are calculated for $hzt$ and $s$ that are covered by the CCSS.

3. Estimate the measurement parameters again in GMM using the moment conditions (M1) to (M5) specified Section 4, this time with the selection terms simulated from the step above. Denote the new estimated parameters as $\{\kappa, \lambda, \sigma\}^1$ and $F^1_{\theta}(\theta^0)$.

4. Use $\{\kappa, \lambda, \sigma\}^1$ and $F^1_{\theta}(\theta^0)$ to re-estimate the structural model in MLE and get new structural model parameters $B^1$.

5. Iterate step 2-4 until convergence.
Appendix D  Additional figures and tables

Figure A1: Screeplot of the principle component analysis for college quality
Figure A2: Histogram of the estimated college quality $V_s$
Figure A3: Probability density function of $\theta^0$ in each province

*Note:* The pre-college human capital $\theta^0$ of NCEE exam takers in each province, $F_h(\theta^0)$, is assumed to follow a normal distribution. Mean and standard deviation of $\theta^0$ in each province are standardized to set the national mean to zero and national standard deviation to 1.
Figure A4: Raw NCEE scores vs. the comparable human capital $\theta^0$ for admitted students

*Note:* Each dot represents the students in $hzt$ admitted by any four-year college $s \in S$. Raw NCEE scores are re-scaled to be between 0 and 1 by dividing the corresponding full score. The expected $\theta^0$ for students in $hzt$, conditional on being admitted by a four-year college, is calculated by integrating over the distribution of the NCEE measurement error $\nu_1$ and incorporating the admission cutoff.
Table A1: Principle component analysis of college quality

<table>
<thead>
<tr>
<th>Raw measure</th>
<th>Loading factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student-faculty ratio</td>
<td>-0.192</td>
</tr>
<tr>
<td>Size of teaching facility, per student</td>
<td>0.268</td>
</tr>
<tr>
<td>Value of teaching/research equipment, per student</td>
<td>0.481</td>
</tr>
<tr>
<td>Number of library books, per student</td>
<td>0.189</td>
</tr>
<tr>
<td>Teaching expenditure - regular, per student</td>
<td>0.268</td>
</tr>
<tr>
<td>Teaching expenditure - experiment, per student</td>
<td>0.300</td>
</tr>
<tr>
<td>Share of courses taught by professor, per student</td>
<td>0.066</td>
</tr>
<tr>
<td>On-time graduation rate</td>
<td>0.042</td>
</tr>
<tr>
<td>Employment rate</td>
<td>0.155</td>
</tr>
<tr>
<td>Project-985 college</td>
<td>0.430</td>
</tr>
<tr>
<td>Project-211 college</td>
<td>0.391</td>
</tr>
<tr>
<td>Private college</td>
<td>-0.308</td>
</tr>
</tbody>
</table>

Table A2: Top-20 colleges ranked by the estimated college quality $V_s$

<table>
<thead>
<tr>
<th>Ranking by $V_s$</th>
<th>College name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tsinghua University</td>
</tr>
<tr>
<td>2</td>
<td>Shanghai Jiaotong University</td>
</tr>
<tr>
<td>3</td>
<td>Peking University</td>
</tr>
<tr>
<td>4</td>
<td>Fudan University</td>
</tr>
<tr>
<td>5</td>
<td>Beihang University</td>
</tr>
<tr>
<td>6</td>
<td>Renmin University of China</td>
</tr>
<tr>
<td>7</td>
<td>Sichuan University</td>
</tr>
<tr>
<td>8</td>
<td>China Agricultural University</td>
</tr>
<tr>
<td>9</td>
<td>Northwestern Polytechnical University</td>
</tr>
<tr>
<td>10</td>
<td>Xi’an Jiaotong University</td>
</tr>
<tr>
<td>11</td>
<td>Zhejiang University</td>
</tr>
<tr>
<td>12</td>
<td>Sun Yat-sen University</td>
</tr>
<tr>
<td>13</td>
<td>Nankai University</td>
</tr>
<tr>
<td>14</td>
<td>University of Science and Technology of China</td>
</tr>
<tr>
<td>15</td>
<td>Harbin Institute of Technology</td>
</tr>
<tr>
<td>16</td>
<td>Tongji University</td>
</tr>
<tr>
<td>17</td>
<td>Tianjin University</td>
</tr>
<tr>
<td>18</td>
<td>Nanjing University</td>
</tr>
<tr>
<td>19</td>
<td>Jilin University</td>
</tr>
<tr>
<td>20</td>
<td>Wuhan University</td>
</tr>
</tbody>
</table>
### Table A3: Distribution of $\theta^0$ of NCEE exam takers by province, $F_h(\theta^0)$

<table>
<thead>
<tr>
<th>Province</th>
<th>Mean ($\mu^0_h$)</th>
<th>Standard deviation ($\sigma^0_h$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai</td>
<td>0.704 (0.052)</td>
<td>0.822 (0.031)</td>
</tr>
<tr>
<td>Shandong</td>
<td>0.642 (0.029)</td>
<td>0.779 (0.029)</td>
</tr>
<tr>
<td>Beijing</td>
<td>0.493 (0.092)</td>
<td>0.935 (0.055)</td>
</tr>
<tr>
<td>Jiangsu</td>
<td>0.474 (0.039)</td>
<td>0.796 (0.018)</td>
</tr>
<tr>
<td>Jilin</td>
<td>0.186 (0.084)</td>
<td>0.880 (0.052)</td>
</tr>
<tr>
<td>Hebei</td>
<td>0.178 (0.097)</td>
<td>0.902 (0.053)</td>
</tr>
<tr>
<td>Tianjin</td>
<td>0.159 (0.043)</td>
<td>1.125 (0.022)</td>
</tr>
<tr>
<td>Shaanxi</td>
<td>0.137 (0.072)</td>
<td>0.867 (0.041)</td>
</tr>
<tr>
<td>Guangdong</td>
<td>0.137 (0.040)</td>
<td>0.813 (0.030)</td>
</tr>
<tr>
<td>Zhejiang</td>
<td>0.134 (0.044)</td>
<td>0.955 (0.023)</td>
</tr>
<tr>
<td>Liaoning</td>
<td>0.093 (0.033)</td>
<td>1.031 (0.019)</td>
</tr>
<tr>
<td>Hainan</td>
<td>0.079 (0.088)</td>
<td>0.959 (0.039)</td>
</tr>
<tr>
<td>Sichuan</td>
<td>0.007 (0.029)</td>
<td>0.923 (0.038)</td>
</tr>
<tr>
<td>Qinghai</td>
<td>-0.014 (0.041)</td>
<td>0.931 (0.040)</td>
</tr>
<tr>
<td>Heilongjiang</td>
<td>-0.048 (0.029)</td>
<td>1.115 (0.022)</td>
</tr>
<tr>
<td>Jiangxi</td>
<td>-0.097 (0.070)</td>
<td>0.946 (0.034)</td>
</tr>
<tr>
<td>Chongqing</td>
<td>-0.099 (0.055)</td>
<td>1.109 (0.029)</td>
</tr>
<tr>
<td>Fujian</td>
<td>-0.109 (0.021)</td>
<td>1.002 (0.020)</td>
</tr>
<tr>
<td>Ningxia</td>
<td>-0.118 (0.034)</td>
<td>0.980 (0.021)</td>
</tr>
<tr>
<td>Xinjiang</td>
<td>-0.136 (0.104)</td>
<td>0.965 (0.051)</td>
</tr>
<tr>
<td>Hubei</td>
<td>-0.141 (0.040)</td>
<td>1.036 (0.023)</td>
</tr>
<tr>
<td>Yunnan</td>
<td>-0.158 (0.071)</td>
<td>0.970 (0.043)</td>
</tr>
<tr>
<td>Guizhou</td>
<td>-0.175 (0.053)</td>
<td>0.898 (0.057)</td>
</tr>
<tr>
<td>Shanxi</td>
<td>-0.175 (0.055)</td>
<td>0.997 (0.034)</td>
</tr>
<tr>
<td>Gansu</td>
<td>-0.185 (0.038)</td>
<td>0.923 (0.019)</td>
</tr>
<tr>
<td>Anhui</td>
<td>-0.232 (0.048)</td>
<td>1.024 (0.022)</td>
</tr>
<tr>
<td>Hunan</td>
<td>-0.260 (0.090)</td>
<td>1.042 (0.050)</td>
</tr>
<tr>
<td>Henan</td>
<td>-0.343 (0.036)</td>
<td>1.021 (0.018)</td>
</tr>
<tr>
<td>Neimenggu</td>
<td>-0.355 (0.042)</td>
<td>1.046 (0.032)</td>
</tr>
<tr>
<td>Guangxi</td>
<td>-0.570 (0.059)</td>
<td>1.073 (0.025)</td>
</tr>
</tbody>
</table>

*Note:* Bootstrapped standard errors are in parentheses. The pre-college human capital $\theta^0$ of NCEE exam takers in each province, $F_h(\theta^0)$, is assumed to follow a normal distribution. Mean and standard deviation of $\theta^0$ in each province are estimated using the measurement model and standardized to set the national mean to zero and national standard deviation to 1.
Table A4: Estimated amenity values for each province

<table>
<thead>
<tr>
<th>Region</th>
<th>Province</th>
<th>College location $\gamma^C_{3,j_s}$</th>
<th>Work location $\gamma^W_{3,k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>Beijing</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Tianjin</td>
<td>-0.010 (0.033)</td>
<td>-0.666 (0.102)</td>
</tr>
<tr>
<td></td>
<td>Hebei</td>
<td>-0.207 (0.029)</td>
<td>-1.036 (0.112)</td>
</tr>
<tr>
<td></td>
<td>Shanghai</td>
<td>-0.092 (0.073)</td>
<td>0.122 (0.088)</td>
</tr>
<tr>
<td></td>
<td>Jiangsu</td>
<td>-0.076 (0.041)</td>
<td>0.145 (0.055)</td>
</tr>
<tr>
<td></td>
<td>Zhejiang</td>
<td>-0.133 (0.041)</td>
<td>0.237 (0.129)</td>
</tr>
<tr>
<td></td>
<td>Fujian</td>
<td>-0.290 (0.058)</td>
<td>-0.103 (0.071)</td>
</tr>
<tr>
<td></td>
<td>Shandong</td>
<td>0.090 (0.012)</td>
<td>-0.196 (0.063)</td>
</tr>
<tr>
<td></td>
<td>Guangdong</td>
<td>-0.177 (0.077)</td>
<td>1.587 (0.126)</td>
</tr>
<tr>
<td>Central</td>
<td>Shanxi</td>
<td>0.035 (0.049)</td>
<td>-1.496 (0.152)</td>
</tr>
<tr>
<td></td>
<td>Anhui</td>
<td>0.127 (0.020)</td>
<td>-1.892 (0.075)</td>
</tr>
<tr>
<td></td>
<td>Jiangxi</td>
<td>-0.027 (0.012)</td>
<td>-2.163 (0.124)</td>
</tr>
<tr>
<td></td>
<td>Henan</td>
<td>-0.052 (0.008)</td>
<td>-1.740 (0.168)</td>
</tr>
<tr>
<td></td>
<td>Hubei</td>
<td>-0.055 (0.021)</td>
<td>-1.724 (0.116)</td>
</tr>
<tr>
<td></td>
<td>Hunan</td>
<td>-0.066 (0.023)</td>
<td>-1.872 (0.077)</td>
</tr>
<tr>
<td></td>
<td>Hainan</td>
<td>0.101 (0.088)</td>
<td>-2.140 (0.090)</td>
</tr>
<tr>
<td>West</td>
<td>Inner Mongolia</td>
<td>-0.250 (0.030)</td>
<td>-1.259 (0.047)</td>
</tr>
<tr>
<td></td>
<td>Guangxi</td>
<td>-0.031 (0.014)</td>
<td>-1.812 (0.029)</td>
</tr>
<tr>
<td></td>
<td>Chongqiong</td>
<td>0.148 (0.036)</td>
<td>-1.621 (0.044)</td>
</tr>
<tr>
<td></td>
<td>Sichuan</td>
<td>0.034 (0.021)</td>
<td>-0.292 (0.026)</td>
</tr>
<tr>
<td></td>
<td>Guizhou</td>
<td>-0.100 (0.039)</td>
<td>-1.225 (0.129)</td>
</tr>
<tr>
<td></td>
<td>Yunnan</td>
<td>-0.066 (0.032)</td>
<td>-1.039 (0.135)</td>
</tr>
<tr>
<td></td>
<td>Shaanxi</td>
<td>0.022 (0.020)</td>
<td>-1.085 (0.299)</td>
</tr>
<tr>
<td></td>
<td>Gansu</td>
<td>0.397 (0.058)</td>
<td>-2.271 (0.094)</td>
</tr>
<tr>
<td></td>
<td>Qinghai</td>
<td>-0.157 (0.043)</td>
<td>-2.745 (0.129)</td>
</tr>
<tr>
<td></td>
<td>Ningxia</td>
<td>0.001 (0.010)</td>
<td>-2.719 (0.108)</td>
</tr>
<tr>
<td></td>
<td>Xinjiang</td>
<td>-0.056 (0.010)</td>
<td>-2.809 (0.369)</td>
</tr>
<tr>
<td>Northeast</td>
<td>Liaoning</td>
<td>0.240 (0.089)</td>
<td>-1.686 (0.099)</td>
</tr>
<tr>
<td></td>
<td>Jilin</td>
<td>0.220 (0.057)</td>
<td>-2.617 (0.106)</td>
</tr>
<tr>
<td></td>
<td>Heilongjiang</td>
<td>-0.037 (0.061)</td>
<td>-2.387 (0.111)</td>
</tr>
</tbody>
</table>

Note: Bootstrapped standard errors are in parentheses. Beijing is the reference group and has a normalized value of 0.
Table A5: Average amenity in monetary values by work region

<table>
<thead>
<tr>
<th>Work region</th>
<th>Mean (in 2012 CNY)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>¥5,020</td>
<td>¥3,004</td>
</tr>
<tr>
<td>Central</td>
<td>¥-2,273</td>
<td>¥870</td>
</tr>
<tr>
<td>West</td>
<td>¥-1,678</td>
<td>¥3,080</td>
</tr>
<tr>
<td>Northeast</td>
<td>¥-3,602</td>
<td>¥1,753</td>
</tr>
</tbody>
</table>

*Note: The amenity value parameter for each province is given in Table A4. These parameters are then converted to annual monetary values following the procedure described in Section 6.2 and averaged across provinces in each region.*
Table A6: Change in the share of admitted students by province

<table>
<thead>
<tr>
<th>Home region</th>
<th>Home province</th>
<th>Baseline in 2008</th>
<th>Nationwide admission</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>East</td>
<td>Beijing</td>
<td>0.43</td>
<td>0.45</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Tianjin</td>
<td>0.49</td>
<td>0.35</td>
<td>-27%</td>
</tr>
<tr>
<td></td>
<td>Hebei</td>
<td>0.26</td>
<td>0.32</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>Shanghai</td>
<td>0.52</td>
<td>0.55</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td>Jiangsu</td>
<td>0.35</td>
<td>0.43</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>Zhejiang</td>
<td>0.37</td>
<td>0.31</td>
<td>-15%</td>
</tr>
<tr>
<td></td>
<td>Fujian</td>
<td>0.30</td>
<td>0.24</td>
<td>-20%</td>
</tr>
<tr>
<td></td>
<td>Shandong</td>
<td>0.25</td>
<td>0.52</td>
<td>105%</td>
</tr>
<tr>
<td></td>
<td>Guangdong</td>
<td>0.28</td>
<td>0.28</td>
<td>2%</td>
</tr>
<tr>
<td>Central</td>
<td>Shanxi</td>
<td>0.21</td>
<td>0.22</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>Anhui</td>
<td>0.23</td>
<td>0.21</td>
<td>-11%</td>
</tr>
<tr>
<td></td>
<td>Jiangxi</td>
<td>0.30</td>
<td>0.24</td>
<td>-20%</td>
</tr>
<tr>
<td></td>
<td>Henan</td>
<td>0.20</td>
<td>0.18</td>
<td>-11%</td>
</tr>
<tr>
<td></td>
<td>Hubei</td>
<td>0.24</td>
<td>0.24</td>
<td>-3%</td>
</tr>
<tr>
<td></td>
<td>Hunan</td>
<td>0.24</td>
<td>0.21</td>
<td>-13%</td>
</tr>
<tr>
<td></td>
<td>Hainan</td>
<td>0.35</td>
<td>0.30</td>
<td>-17%</td>
</tr>
<tr>
<td>West</td>
<td>Inner Mongolia</td>
<td>0.24</td>
<td>0.18</td>
<td>-26%</td>
</tr>
<tr>
<td></td>
<td>Guangxi</td>
<td>0.22</td>
<td>0.14</td>
<td>-38%</td>
</tr>
<tr>
<td></td>
<td>Chongqiong</td>
<td>0.31</td>
<td>0.26</td>
<td>-14%</td>
</tr>
<tr>
<td></td>
<td>Sichuan</td>
<td>0.24</td>
<td>0.26</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Guizhou</td>
<td>0.24</td>
<td>0.20</td>
<td>-18%</td>
</tr>
<tr>
<td></td>
<td>Yunnan</td>
<td>0.29</td>
<td>0.22</td>
<td>-23%</td>
</tr>
<tr>
<td></td>
<td>Shaanxi</td>
<td>0.27</td>
<td>0.30</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>Gansu</td>
<td>0.23</td>
<td>0.20</td>
<td>-11%</td>
</tr>
<tr>
<td></td>
<td>Qinghai</td>
<td>0.30</td>
<td>0.26</td>
<td>-13%</td>
</tr>
<tr>
<td></td>
<td>Ningxia</td>
<td>0.27</td>
<td>0.23</td>
<td>-15%</td>
</tr>
<tr>
<td></td>
<td>Xinjiang</td>
<td>0.26</td>
<td>0.22</td>
<td>-16%</td>
</tr>
<tr>
<td>Northeast</td>
<td>Liaoning</td>
<td>0.42</td>
<td>0.31</td>
<td>-27%</td>
</tr>
<tr>
<td></td>
<td>Jilin</td>
<td>0.41</td>
<td>0.32</td>
<td>-22%</td>
</tr>
<tr>
<td></td>
<td>Heilongjiang</td>
<td>0.46</td>
<td>0.28</td>
<td>-39%</td>
</tr>
</tbody>
</table>