Abstract

Students in China need to choose their academic track (STEM vs Non-STEM) in high school. This choice determines who they will compete with in the college entrance exam (gaokao) and the set of college majors from which they can choose. We study their track choices with a unique dataset. We find that female students are still less likely to choose the STEM track even after accounting for high school entrance exam performance. Rural students are more likely to choose STEM, perhaps due to perceived financial benefits. We then study how the track choice affects gaokao performance and find suggestive evidence of sub-optimal track choice for some male and rural students.

1 Background and Data

Section 2.1 describes track choice in high school in China. Section 2.2 presents our dataset and describes the main variables.

1.1 Background

Held annually, the college entrance exam, commonly known as Gaokao, is almost the only way for Chinese students to gain admittance into college. It takes the form of "3+X" in most provinces in recent years. Under this system, students are required to take tests in three subjects: Chinese, Mathematics and a foreign language (usually English), each of which accounts for 150/750 of the total score. The "X" stands for an additional exam, which is determined by which track, humanities/social sciences or STEM, a student plans to pursue in college. For students who intend to major in the humanities/social sciences, the additional exam tests their knowledge of politics, history, and geography. Students pursuing STEM fields are tested on physics, chemistry, and biology. This "X" accounts for 300/750 of the total score. One more important thing to note here is that even though every student

is required to take math in Gaokao, the math test covers more material and is substantially harder for the STEM track students.

The track choice is crucial for the high school students. It determines who they will compete with in the Gaokao, the set of majors they can choose from when they enter college, and their curriculum during high school.

The college admission in China is conducted through a centralized matching procedure. The preference of the schools over students are fully defined by an identical priority ordering: the college entrance exam score. Students only compete with others in the same track in the same province. Each university has limited capacity, which is predetermined and announced in advance of the college entrance exam. Colleges distribute their quotas to provinces and the quotas are track-specific. For example, even if there are still open slots in humanities/social sciences track at a university, they cannot be filled with STEM-track students.

In the college application process, the students not only need to submit their preferences for the schools but also need to specify a list of preferred majors at each of the schools. Students in the two different tracks choose from different sets of majors. The overlap of the two sets are limited to economics and business in most universities and also include psychology, architecture, and geography in some schools.

To be able to ensure a high rate of college acceptance, most high schools divide a cohort into the two tracks within their first year of high school. In our sample, tracking starts in the second half of the first semester in the first year in one high school in Hebei, starts in the second semester in the other two schools in Hebei and starts in the second year in the high school in Anhui. After choosing the track, the students will only study the three subjects in their track plus the three subjects required for everyone.

1.2 Data

We employ multiple sources of data in our analysis. The primary data source for this study is the administrative record of track choices in four high schools in three counties in two provinces (Hebei and Anhui) from 2014 to 2016. Our data also consists of the students' gaokao test scores, gaokao mock tests scores, high school entrance exam scores (commonly known as zhongkao), and demographic information.

Track Choice and Test Scores

The primary outcome of interest is a student's choice of track in high school, which is a binary variable. We obtain the administrative data in four high schools located in 3 counties in two provinces, Hebei and Anhui. In Hebei, the three schools are in Baoding, a mid-size city. We have the track choice information for those who graduated between 2013 to 2016 in one school and those who graduated in 2015-2016 for the other two schools. In Anhui province, the students in our sample graduated between 2014 and 2016 from a high school in Suzhou.

We mostly obtain the track choice record from a student's Gaokao registration. In other words, what we know is his/her final choice for the Gaokao. As discussed in the earlier section, all students in our sample already made their initial choices after the first year of high school. It is possible to switch the track in the second and third year of high school, but this happens quite rarely. For some years, we have both the initial track choice and the final choice of track and more than % of the students' two choices are consistent.

We construct our dataset by matching a student's track choice with his/her Gaokao score, Gaokao mock tests scores and zhongkao (high school entrance exam) score.

Gaokao score is an outcome measure. In general, colleges are divided into 3 tiers in China: tier 1 ("Ben Yi"), tier 2 ("Ben Er") and tier 3 ("Zhuan Ke"). [1] Each college belongs to one and only one tier. Even though the school quality varies substantially within a tier, the between-tier difference is more salient. When recruiting fresh graduates, it is a common practice for firms to restrict the selection pool to tier 1 and 2 college graduates or even just tier 1. In each year, there are cutoff scores for tier 1 and tier 2 colleges in each province, which enables us to infer whether a student was admitted by a tier 1 or tier 2 college using their Gaokao scores.

Gaokao mock test scores complement the Gaokao scores. Gaokao is a single test which is subject to non-trivial randomness. Students take three mock tests in the final year of high school. Those mock tests mimic the real Gaokao in every single way and is graded by the same pool of graders. The scores in those tests give us a better picture of the students' expected performance. The difference

[1] In some years, there are 4 tiers of college. The extra tier is between "Ben Er" and "Zhuan Ke" and is called "Ben San".

between performance on the Gaokao vs. the mock tests can give us a measure of a student's ability to cope with pressure.

Zhongkao test scores enable us to measure a student's relative ability in STEM versus humanities/social sciences. There is no tracking in zhongkao and everyone takes the same tests in the same city. The students are tested in Chinese, Math, English, Natural Sciences (Physics and Chemistry), Social Sciences (Political Science and History), Geography, and Biology.

Demographic Information

From the students' gaokao registration profile, we gather their gender and birthdate. In addition, we are able to construct categorize them as urban or rural based on their address.

In addition to these three demographic variables, we also have a series of interesting variables for a much smaller sample (N=953). For the students who took Gaokao in 2015 in one of the Baoding schools, we have access to their student profiles (Xue Ji). In the profile, there are three variables of interest: whether the student lives on campus, whether the student is the only child, whether the student got formal education before enrolling in elementary school, and whether the student needs sponsorship.

Surveys

We conducted a survey with X middle school graduates in Baoding in 2017. We link the survey data to the track choices of students who later attended the high schools in our dataset. When they answered the survey, they already took the high school entrance exam, but did not know their scores yet. We asked them to predict their scores to compare those estimates with their realized scores. They were also asked to recall their scores in two zhongkao mock tests and we again compare those recalls with the actual scores.

We also have the students' self-reported first jobs from a survey conducted by one high school in our dataset.

2 Results

We first explore the possible determinants of the students' track choice. Do they mostly focus on just the zhongkao math score as a measure of STEM aptitude or take into account performance in the other subjects as well? Does the propensity to choose the STEM track still differ across genders and background (rural vs. urban) after accounting for zhongkao performance? Are male and female students differentially sensitive to their zhongkao performance when making the track choice?

2.1 Track Choice

2.1.1 Student Profiles and Track Choices

We now turn to the subsample of students for whom we have a more detailed profile. Specifically, we know whether they live on campus or at home, whether they are an only child or not, and whether they self-reported needing a sponsorship for school (which take to be a noisy measure of financial circumstance). The STEM track is typically considered to be the more socially prestigious one and the one that will lead to better financial prospects. Therefore, students who are only children and students who are more financially disadvantaged might experience more pressure to choose that track. Students who live on campus [fill in details].

2.2 Gaokao Performance

standardization:

Gaokao: $G_i^{t,p,s} = \frac{G_i - \bar{G}^{t,p,s}}{\sigma_{G^{t,p,s}}}$

Zhongkao: $Z_i^{t,p} = \frac{Z_i - \bar{Z}^{t,p}}{\sigma_{Z^{t,p}}}$

Table 1: Summary statistics

	Overall	STEM	Non-STEM
female	.524	.473	.625
rural	.755	.772	.722
gaokao score (0-750)	434.742	452.073	400.204
zhongkao score (0-630/800)	565.401	594.825	506.761
Zhongkao math (0-120/150)	100.347	107.761	85.571
Observations	14,031	9,343	4,688

Table 2: Determinants of Track Choice

Dependent variable:	STEM			
	(1)	(2)	(3)	(4)
female	-0.127***	-0.153***	-0.105***	-0.123***
	(0.00690)	(0.014)	(0.013)	(0.016)
rural		0.049***	-0.031***	-0.040***
		(0.011)	(0.010)	(0.013)
rural×female		0.037**	0.053***	0.066***
		(0.017)	(0.015)	(0.018)
math			0.102***	0.100***
			(0.005)	(0.006)
chi			-0.034***	-0.024***
			(0.004)	(0.005)
eng			0.013**	0.019***
			(0.005)	(0.006)
hum			-0.034***	-0.029***
			(0.005)	(0.006)
sci			0.163***	0.150***
			(0.006)	(0.007)
othersubjects			0.025***	0.028***
			(0.004)	(0.005)
Year FE	Y	Y	Y	
School FE	Y	Y	Y	
Class FE				Y
Observations	17,998	16,208	14,031	10,826
R-squared	0.053	0.057	0.291	0.342

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Determinants of Track Choice

Dependent variable:	STEM			
	(1)	(2)	(3)	(4)
female	-0.127***	-0.175***	-0.183***	-0.156***
	(0.007)	(0.020)	(0.025)	(0.033)
rural		-0.030***	-0.031**	-0.024
		(0.011)	(0.013)	(0.032)
female×rural		0.049***	0.053***	0.053***
		(0.017)	(0.019)	(0.020)
math		0.084***	0.075***	0.070***
		(0.011)	(0.013)	(0.013)
chi		-0.040***	-0.041***	-0.041***
		(0.010)	(0.011)	(0.011)
eng		0.039***	0.037***	0.040***
		(0.011)	(0.013)	(0.013)
hum		-0.022**	-0.016	-0.019
		(0.011)	(0.013)	(0.013)
sci		0.139***	0.132***	0.133***
		(0.014)	(0.015)	(0.016)
othersubjects		0.030***	0.034***	0.024***
		(0.007)	(0.008)	(0.008)
female×math		0.028***	0.023**	0.026**
		(0.011)	(0.012)	(0.012)
female×chi		0.002	0.007	0.005
		(0.009)	(0.010)	(0.010)
female×eng		-0.003	0.003	-0.011
		(0.010)	(0.012)	(0.012)
female×hum		-0.025***	-0.022***	-0.024***
		(0.010)	(0.011)	(0.011)
female×sci		0.028**	0.027**	0.018
		(0.012)	(0.013)	(0.014)
female×othersubjects		0.001***	0.001***	0.001***
		(0.000)	(0.000)	(0.000)
Year FE	Y	Y		Y
School FE	Y	Y		Y
Class FE			Y	
Observations	17,998	14,031	10,826	10,826
R-squared	0.053	0.294	0.346	0.297

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Cont. Determinants of Track Choice

Dependent variable:	STEM			
	(1)	(2)	(3)	(4)
rural×math		0.005	0.017	0.006
		(0.012)	(0.013)	(0.013)
rural×chi		0.008	0.019*	0.010
		(0.010)	(0.011)	(0.011)
rural×eng		-0.029**	-0.020	-0.035
		(0.012)	(0.013)	(0.014)
rural×hum		-0.001	-0.003	-0.013
		(0.011)	(0.013)	(0.013)
rural×sci		0.007	-0.002	-0.005
		(0.014)	(0.016)	(0.017)
rural×othersubjects		-0.011	-0.010	-0.010
		(0.008)	(0.009)	(0.009)
classranking				-0.001
				(0.001)
female×classranking				-0.001***
				(0.001)
sexratio				-0.016
				(0.021)
female×sexratio				-0.008
				(0.019)
rural×sexratio				0.032
				(0.020)
rural×classranking				-0.002***
				(0.001)
Year FE	Y	Y		Y
School FE	Y	Y		Y
Class FE			Y	
Observations	17,998	14,031	10,826	10,826
R-squared	0.053	0.294	0.346	0.297

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Determinants of Track Choice

Dependent variable:	STEM		
	(1)	(2)	(3)
female×math	0.028***	0.023**	0.026**
	(0.011)	(0.012)	(0.012)
female×chi	0.002	0.007	0.005
	(0.009)	(0.010)	(0.010)
female×eng	-0.003	0.003	-0.011
	(0.010)	(0.012)	(0.012)
female×hum	-0.025***	-0.022***	-0.024***
	(0.010)	(0.011)	(0.011)
female×sci	0.028**	0.027**	0.018
	(0.012)	(0.013)	(0.014)
female×othersubjects	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
rural×math	0.005	0.017	0.006
	(0.012)	(0.013)	(0.013)
rural×chi	0.008	0.019*	0.010
	(0.010)	(0.011)	(0.011)
rural×eng	0.030**	0.020	0.010
	(0.012)	(0.014)	(0.014)
rural×hum	-0.001	-0.003	-0.013
	(0.011)	(0.013)	(0.013)
rural×sci	0.007	-0.002	-0.005
	(0.014)	(0.016)	(0.017)
rural×othersubjects	-0.011	-0.001	-0.001
	(0.001)	(0.001)	(0.001)
classranking			0.001
			(0.001)
female×classranking			-0.001***
			(0.001)
sexratio			-0.016
			(0.021)
female×sexratio			-0.001
			(0.019)
rural×sexratio			0.032
			(0.020)
rural×classranking			-0.002***
			(0.001)
Year FE	Y		Y
School FE	Y		Y
Class FE		Y	
Observations	14,031	10,826	10,826
R-squared	0.294	0.346	0.297

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Determinants of Track Choice: Student Profile Data

Dependent variable:	STEM	
	(1)	(2)
female	-0.187***	-0.228***
	(0.048)	(0.074)
rural	0.016	0.026
	(0.043)	(0.042)
rural×female	0.132**	0.095*
	(0.056)	(0.057)
oncampus	0.008	-0.111**
	(0.033)	(0.051)
onechild	0.042	0.103***
	(0.029)	(0.039)
needsponsorship	-0.017	0.050
	(0.026)	(0.039)
female×oncampus		0.210***
		(0.066)
female×onechild		-0.119**
		(0.057)
female×needsponsorship		-0.118**
		(0.051)
Individual Controls	Y	Y
Class FE	Y	Y
Observations	953	953
R-squared	0.394	0.406

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Gaokao Score and Zhongkao Score

Dependent variable:	Gaokao Score			
	(1) Overall	(2) Overall	(3) STEM	(4) Non-STEM
math	0.141***	0.145***	0.198***	0.101***
	(0.009)	(0.009)	(0.013)	(0.013)
chi	0.070***	0.055***	0.045***	0.078***
	(0.008)	(0.008)	(0.010)	(0.012)
eng	0.227***	0.205***	0.217***	0.181***
	(0.009)	(0.009)	(0.012)	(0.014)
hum	0.125***	0.133***	0.114***	0.173***
	(0.009)	(0.009)	(0.012)	(0.014)
sci	0.231***	0.256***	0.324***	0.150***
	(0.012)	(0.012)	(0.015)	(0.018)
othersubjects	0.002	0.008	0.003	0.033**
	(0.010)	(0.010)	(0.013)	(0.015)
female		0.124***	0.011	0.097***
		(0.028)	(0.028)	(0.036)
STEM		-0.371***		
		(0.055)		
female×STEM		-0.139***		
		(0.025)		
rural		-0.167***	-0.167***	-0.173***
		(0.021)	(0.025)	(0.036)
rural×female		0.127***	0.110***	0.160***
		(0.026)	(0.032)	(0.043)
Class FE	Y	Y	Y	Y
Observations	17,998	14,030	9,342	4,688
R-squared	0.352	0.574	0.569	0.624

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Confidence and STEM

Dependent variable:	STEM		
	(1)	(2)	(3)
Actual Zhongkao - Estimation	-0.002*	0.000	0.001
	(0.001)	(0.001)	(0.001)
Max Guess- Best Guess			-0.002*
			(0.001)
Best Guess - Min Guess			-0.001
			(0.001)
Actual Mock - Recalled Mock	-0.001	-0.001**	-0.002**
	(0.001)	(0.001)	(0.001)
female	-0.096	-0.086	-0.070
	(0.070)	(0.067)	(0.073)
rural	0.015	0.035	0.074
	(0.091)	(0.082)	(0.090)
rural×female	0.052	-0.007	-0.076
	(0.139)	(0.122)	(0.130)
math		0.009	0.010
		(0.006)	(0.007)
chi		-0.006	-0.003
		(0.007)	(0.007)
eng		-0.002	-0.000
		(0.006)	(0.007)
hum		-0.022***	-0.018**
		(0.007)	(0.007)
sci		0.006	0.008
		(0.006)	(0.006)
othersubjects		0.004	0.003
		(0.004)	(0.005)
risk	0.009	0.012	0.016
	(0.012)	(0.010)	(0.011)
air-purifier	0.004	0.028	0.013
	(0.071)	(0.063)	(0.070)
Observations	151	151	136
R-squared	0.067	0.351	0.352

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Gender Gap in Mock Exams and Gaokao Scores

	Mock Exam 2 (Feb.)			Mock Exam 3 (Apr.)			Gaokao (Jun.)		
	(1) Female	(2) Male	(3) Difference	(4) Female	(5) Male	(6) Difference	(7) Female	(8) Male	(9) Difference
	A. Full sample								
Total	0.097	-0.107	0.204***	0.064	-0.075	0.139***	0.108	-0.118	0.226***
Observations	5,516	4,965		6,808	5,820		7,350	6,681	
	B. Science stream								
Total	0.064	-0.058	0.122***	0.026	-0.024	0.050	0.078	-0.070	0.148***
Chi	0.214	-0.196	0.411***	0.180	-0.170	0.350***	0.200	-0.180	0.380 ***
Math	0.008	-0.007	0.015	-0.033	0.032	-0.065***	0.030	-0.027	0.058
eng	0.224	-0.205	0.429***	0.207	-0.196	0.404***	0.265	-0.238	0.504 ***
sci	-0.027	0.025	-0.052***	-0.070	0.066	-0.137***	-0.029	0.026	-0.056***
	C. Art stream								
Total	0.166	-0.335	0.501***	0.130	-0.254	0.383***	0.152	-0.253	0.405***
chi	0.176	-0.354	0.530***	0.142	-0.278	0.420***	0.142	-0.237	0.380***
math	0.107	-0.215	0.322***	0.076	-0.150	0.226***	0.095	-0.159	0.255***
eng	0.219	-0.440	0.659***	0.174	-0.340	0.514***	0.233	-0.388	0.621***
hum	0.100	-0.201	0.301***	0.083	-0.163	0.246***	0.078	-0.130	0.209***

*** p<0.01, ** p<0.05, * p<0.1.

Table 8: Pressure and Gaokao

Dependent variable:	STEM		
	(1)	(2)	(3)
female	-0.014	0.007	-0.022
	(0.021)	(0.026)	(0.018)
Ave mock	-0.191***	-0.153***	
	(0.008)	(0.008)	
STEM	-0.053	0.349	-0.118
	(0.107)	(0.396)	(0.106)
female×STEM	-0.005	-0.013	-0.005
	(0.019)	(0.023)	(0.016)
rural	-0.034**	-0.044**	-0.042***
	(0.015)	(0.018)	(0.013)
rural×female	0.036**	0.040*	0.046***
	(0.018)	(0.022)	(0.016)
math zhongkao	0.043***	0.034***	0.037***
	(0.007)	(0.008)	(0.006)
lit zhongkao	-0.005	0.008	-0.000
	(0.006)	(0.006)	(0.005)
eng zhongkao	0.040***	0.046***	0.038***
	(0.007)	(0.008)	(0.006)
hum zhongkao	0.023***	0.023***	0.018***
	(0.007)	(0.007)	(0.006)
sci zhongkao	0.095***	0.091***	0.075***
	(0.009)	(0.009)	(0.007)
othersubjects	-0.010	0.006	-0.003
	(0.008)	(0.007)	(0.006)
mock3			-0.145***
			(0.006)
Observations	12,628	10,481	12,628
R-squared	0.158	0.196	0.191

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 9: Confidence in Zhongkao Estimation

	(1) Obs.	(2) Mean	(3) Variance
	A. Full sample		
zhongkao total	429	513.2517	105.7956
estimation survey	287	551.3031	75.91868
estimation official	207	483.2705	110.7621
	B. School 1		
zhongkao total	215	448.0186	110.3759
estimation survey	101	513.7129	95.48846
estimation official	207	483.2705	110.7621
	C. School 2		
zhongkao total	214	578.7897	40.4941
estimation survey	186	571.7151	52.83252

*** p<0.01, ** p<0.05, * p<0.1.

Table 10: Gender Difference in Confidence

Female				
Variable	Obs	Mean	Std. Dev.	P value
zhongkao score - survey est	146	5.342	32.683	0.127
zhongkao score - official est	205	5.117	32.087	0.096
zhongkao math - est math	145	-0.793	7.847	0.219
zhongkao chi - est chi	145	-3.938	7.680	0.000
zhongkao eng - est eng	145	2.924	7.814	0.000
zhongkao sci- est sci	145	-0.069	7.655	0.965
zhongkao hum - est hum	145	4.510	11.106	0.000
Male				
Variable	Obs	Mean	Std. Dev.	P value
zhongkao score - survey est	159	-0.113	28.667	0.723
zhongkao score - official est	208	-0.505	30.999	0.852
zhongkao math - est math	159	-2.132	8.389	0.000
zhongkao chi - est chi	159	-4.774	7.457	0.000
zhongkao eng - est eng	159	1.906	9.702	0.002
zhongkao sci- est sci	159	-1.126	8.514	0.012
zhongkao hum - est hum	159	4.774	9.296	0.000