# Gender Composition and High-Stakes Cognitive Performance: Evidence from a Quasi-Randomized Experiment*

Xuan Li † Xiang Zhou ‡

This version: September 2024

**Abstract:** This paper examines whether gender composition may influence cognitive performance in a real-world high-stakes setting. We use unique administrative data on students taking the college entrance examination in China, who are randomly assigned to test rooms with varying gender compositions. Our findings reveal that an increased presence of male students in the test room leads to a decreased performance of female students, but does not affect males. The presence of males is a widely used cue for triggering stereotype threat in lab experiments, and additional evidence suggests our results are consistent with this concept: female students who are more inclined to endorse the math-gender stereotype are more strongly affected by gender composition. This study identifies a previously unexplored passive gender composition effect, providing new insights into the debate on education policy regarding single-sex versus mixed-sex schools.

**Keywords:** Gender composition, high-stakes test, cognitive performance, stereotype threat

**JEL:** J16, J24, A20

*We are indebted to Raymond Fisman, Kevin Lang and Daniele Paserman for their continuous guidance, advice, and support. We would also like to thank Marcela Gomez-Ruiz, Joshua Goodman, Justin Hong, Hamish Low, Pierre Mouganie, Xavier Ramos, Nina Roussille, Linh Tô, Yuzhao Yang, and all participants at the Gender Reading Group (BU 2024) for helpful comments. We also greatly appreciate the outstanding research assistance provided by Junyu Chen. All errors are our own.

†Department of Economics, Boston University (email: xuanli@bu.edu)

‡Business School of Xiangtan University (corresponding author, email: hoverchow@163.com)

# 1 Introduction

Debates around gender composition and diversity in schools remain contentious (Hoxby, 2000; Halpern et al., 2011; Lavy and Schlosser, 2011; Figlio et al., 2023). Concerns over diversity often point to difficulties posed by intensive daily interactions to underrepresented or discriminated-against groups. In this paper, in contrast to prior work on peer effects, we study whether the mere presence of other groups can impact performance. We do so by examining the effect of gender composition on academic performance in a unique setting characterized by minimal student interaction.

We leverage the Chinese college entrance examination, a high-stakes test in which students are required to complete exams individually. To prevent cheating, education authorities pool all students within a county and then randomly assign them to standardized test rooms, each accommodating a fixed number of students (30 per room). Random assignment leads to exogenously generated variation in gender composition across test rooms. Consequently, students from the same high school class take the same exam, compete with the same set of relevant peers, yet find themselves in different test rooms characterized by varying gender compositions. Throughout the entire testing period, students maintain their assigned seats and complete the tasks independently. Two proctors oversee the entire examination process, minimizing the likelihood of student interaction. Upon the completion of the examination, papers are graded by the provincial education authorities in a double-blind review.

Using administrative data on the entire population of approximately 3,800 students who take college entrance examination in a county in 2019, we first validate the random assignment system. Among 127 test rooms in our sample county, we observe that the gender composition – the proportion of males in the room – varies across test rooms, and the distribution is consistent with the binomial distribution. Moreover, our balance checks indicate that the gender composition of the test room is not correlated with students' demographic characteristics, including ethnicity, age, urban/rural area, and, importantly, their prior academic performance (as indicated by mock exam scores).

After confirming random assignment, we proceed to estimate the impact of test room gender composition on academic performance. We find that a higher proportion of male students in the test room results in lower performance among female students who were admitted

to college (since we do not have test scores for non-admitted students). The performance of male students, by contrast, is unaffected by the gender composition of the test room. For female students, a 10 percentage point increase in the male ratio in the test room corresponds to a reduction of 0.0412 standard deviations in test scores for those in the science track and 0.0441 standard deviations for those in liberal arts. Our findings are robust to the inclusion of high-school-class fixed effects, test-center fixed effects, the inclusion of individual demographic controls, mock exam scores, and additional test-room composition controls.

Our analysis reveals a new channel through which gender composition can affect performance: the mere presence of a large proportion of males in the testing environment can significantly reduce the cognitive performance of females in high-stakes situations. We further explore potential explanations for our findings and offer suggestive evidence that they are consistent with the activation of stereotype threat. We also consider and rule out alternative explanations, including gender differences in competition (Gneezy, Niederle and Rustichini, 2003) and the "acting wife" hypothesis (Bursztyn, Fujiwara and Pallais, 2017).

Stereotype threat is a well-studied concept in social psychology, referring to the fear or anxiety of confirming a negative stereotype about one's social group (Steele and Aronson, 1995; Steele, 1997). The activation of negative stereotypes can consume the cognitive resources required for completing challenging tasks, thereby impairing performance in relevant domains (Schmader, Johns and Forbes, 2008). The activation of negative stereotypes can be subtle, often occurring without conscious awareness. The presence of males is a commonly used cue for activating stereotype threat. In a related lab experiment, Inzlicht and Ben-Zeev (2000) found that the presence of males was sufficient to trigger stereotype threat and impair the performance of female students on a low-stakes math task.

In our setting, stereotype threat predicts that female students will have lower STEM scores, but not lower non-STEM scores, in the presence of more male students, while male students are unaffected by the gender composition. However, the provincial education authority only provides the total scores of students admitted to college to the local education authorities, who are responsible for assisting with the post-admission process. While we successfully demonstrate that male students are unaffected by gender composition, the lack of subject-specific test scores limits our ability to conduct a heterogeneous analysis by subject

for female students.[1] We thus adopt an alternative approach that, in our view, complements the standard method for demonstrating stereotype threat. Our approach is based on the premise that those who believe in the stereotype are more likely to be vulnerable to stereotype threat. Therefore, rather than assuming that all female students are equally affected by stereotype threat, we first use a nationally representative survey to identify which types of female students are more likely to hold gender stereotypes (i.e., agree with the statement that "boys are better at mathematics than girls"). We then investigate whether these girls are also more likely to be affected by the gender composition of the test room during the Chinese college entrance examination.

Using the China Education Panel Survey (2013-2014), we find that 50.5% of female students in their final year of middle school (Grade 9 in China) agree with the statement that "boys are better at mathematics than girls." This is particularly true for girls who had lower STEM scores and higher non-STEM scores in their last midterm exam. Furthermore, we find that female students in classrooms with a higher male ratio are more likely to hold these gender stereotypes. Turning to our exam data, we find that the presence of male students in the test room has a stronger effect on girls with low STEM scores in the mock exam, high non-STEM scores in the mock exam, and those from classes with a higher male ratio. We take these findings as suggesting that female students who are more likely to hold gender stereotypes are also more likely to be affected by the presence of male students in high-stakes exams, consistent with stereotype threat as an explanation for our main results.

This paper contributes to the literature on the impact of the gender composition of peers on academic performance, and to the policy discussion on same-sex schools in general (Jackson, 2012; Eisenkopf et al., 2015; Booth, Cardona-Sosa and Nolen, 2018; Contrerasa et al., 2022; Bostwick and Weinberg, 2022). Using idiosyncratic variation in cohort composition within schools, Hoxby (2000) and Lavy and Schlosser (2011) show that classrooms with more female students yield higher test scores for both boys and girls. However, exploiting similar across-grade variation in peers within schools, Black, Devereux and Salvanes (2013) report opposite effects for men and women: it is beneficial for women to have a greater proportion of female peers, while male students are disadvantaged by their female counterparts. Furthermore, Whitmore (2005), Gong, Lu and Song (2019), and Goulas, Megalokonomou and Zhang

[1] All students take both STEM and non-STEM subjects. In terms of STEM subjects, students in the liberal arts track take only mathematics, while those in the science track take mathematics, physics, chemistry, and biology.

(2024) exploit the random assignment of students to classrooms, but the results are also mixed. Whitmore (2005) demonstrates that female students consistently benefit from a greater proportion of female peers, whereas male students derive such benefits primarily when they are younger. Moreover, both Gong, Lu and Song (2019) and Goulas, Megalokonomou and Zhang (2024) report benefits for both male and female students from a greater presence of female peers. However, Gong, Lu and Song (2019) observes a more substantial benefit for male students, while Goulas, Megalokonomou and Zhang (2024) find that the effect is larger for girls.

Understanding the underlying mechanisms could be crucial for reconciling these findings. Lavy and Schlosser (2011) demonstrates that in a classroom with a higher proportion of female students, a more favorable classroom environment (e.g., reduced disruptions), and enhanced inter-student and student-teacher relationships are observed, without significant individual behavior changes. However, Gong, Lu and Song (2019) indicate that changes in teacher behavior, increased student effort, and a better classroom environment are the main contributing factors in their context. In a similar vein to the literature regarding a teacher's gender or ethnicity (Dee, 2004), the underlying mechanism may operate through both active means (such as behavioral change due to social interaction) and passive ones (such as mere presence). Our work contributes to the literature primarily through its exploration of a unique setting with minimal or even no interaction between the students, thereby allowing us to be the first to identify a passive channel that is currently underexplored. We provide evidence that the mere presence of a significant proportion of males in the test environment has the potential to adversely affect females' cognitive performance.

This paper also joins the recent debate on the replicability of stereotype threat, primarily in social psychology but also in economics. Since the seminal experiments by Steele and Aronson (1995), stereotype threat has been the subject of extensive research. Most early efforts involve small-scale lab experiments and confirm the existence of stereotype threat across varying samples, test difficulties, and stereotype threat cue activations (Nguyen and Ryan, 2008). However, some recent replications using large-scale lab experiments involving more than 1,000 participants report null results for stereotype threat (Paulette C. Flore and Wicherts, 2018; Stoevenbelt et al., 2024), casting serious doubt on the replicability of stereotype threat.

Two key differences distinguish our study from the contributions made by the existing literature. First, although individuals may be especially vulnerable to negative stereotypes in

high-stakes exams, which demand more cognitive resources and can have significant consequences, the discussion of stereotype threat in such contexts is largely absent from the literature. To the best of our knowledge, we provide the first evidence consistent with stereotype threat in a real-world, high-stakes setting, with conditions closely resembling those of a well-designed, large-scale lab experiment. Second, rather than assuming that all girls are equally affected by stereotype threat, we first identify which types of female students are more likely to hold gender stereotypes (i.e., those with low STEM performance, high non-STEM performance, and from classes with a high male ratio), and then demonstrate that these students are also more likely to be affected by stereotype threat. This agnostic approach, linking individual beliefs about the math-gender stereotype to the impact of stereotype threat, offers a complementary method for demonstrating stereotype threat and provides valuable insights into understanding inconsistencies in the literature. For instance, a recent large-scale replication deliberately selected participants from specific groups, such as high-achieving students, who were assumed to be more susceptible to stereotype threat (Paulette C. Flore and Wicherts, 2018). However, as we demonstrate, this assumption may not necessarily hold true. Moreover, our findings suggest that the proportion of students from one's own group within an institution may be key to understanding the seemingly disparity between the significant impact of Black stereotype threat observed in Black students recruited from predominantly white institutions (Steele and Aronson, 1995; Cohen and Garcia, 2005) and the null results found for Black students from a historically Black college (Alston et al., 2022).

# 2 Background and Data

## 2.1 Setting: The National College Entrance Examination

In China, the national college entrance examination holds significant importance as a high-stakes exam, as college admission hinges solely on the scores achieved in this once-a-year assessment. Consequently, the national college entrance examination is widely recognized as a "life-changing event," exerting a lasting impact on both educational trajectories and labor market outcomes (Jia and Li, 2021).

The government places great emphasis on the organization of the national college entrance examination to ensure a smooth and streamlined process. In each county, several high schools are chosen as test centers. All students in the same county will take the national col-

lege entrance examination in these test centers, and each test center has many test rooms. To prevent cheating, some provinces randomly assign students to the test room.[2] This implies that students from the same high school class will sit for the national college entrance examination in different test rooms with varying gender compositions.

The national college entrance examination consists of four sections totaling 750 points: Chinese (150), Mathematics (150), English (150), and a comprehensive test (300). There are generally two tracks: the science track and liberal arts track. The comprehensive test covers physics, chemistry, and biology for science track students, and politics, history, and geography for liberal arts track students. The national college entrance examination extends over two days in early June. During breaks, the majority of students will be taken back home by their parents and typically do not discuss the test with other students in the same test room, as they come from different classes or schools and are unfamiliar with each other (only 13% of students have a classmate in the same testing room). In addition, students do not move across rooms or seats while taking the test. Upon completion, the examination papers undergo computer scanning and are subsequently graded by the provincial education authority.[3]

Students can expect to receive their test scores approximately two weeks following the national college entrance examination. Upon receiving these scores, students will proceed to formulate their application lists. College admission outcomes will be contingent upon both the available quotas and the individual student's ranking among applicants from the same province vying for positions in the same institution. As the competition is within province not county, the national college entrance examination is highly competitive, where a single additional point can significantly alter a student's ranking among their peers within the same province. To illustrate, there are around 210,000 science track students in our sample province in 2019, a student with a score of 500 is ranked 57,838. However, losing just one additional point, his/her ranking would decrease by 782 to 58,620. Therefore, this loss of one additional point places him/her at a disadvantage when choosing a college in comparison with up to 782

[2]The provincial government employs a computer program to randomly assign students to test rooms. This random assignment process is solely conditional on the academic track, ensuring that students in the same test room belong to the same academic track (i.e., science or liberal arts). Specifically, the assignment program employs the following procedure: first, it generates random numbers to rank all students within a county who are in the same academic track. The top thirty students are then assigned to test room 1, the next thirty to test room 2, and so on.

[3]China's political system consists of five levels of state administration: the central government, provinces, municipalities, counties, and townships.

other students.

## 2.2 Data

Our data is collected from the Bureau of Education of a county located in Central China. This administrative dataset encompasses the universe of students who registered for the national college entrance examination in our sample county in 2019. The dataset includes information about gender, ethnicity, age, academic track, high school, high school class, test center, and test room for each student.[4,5]

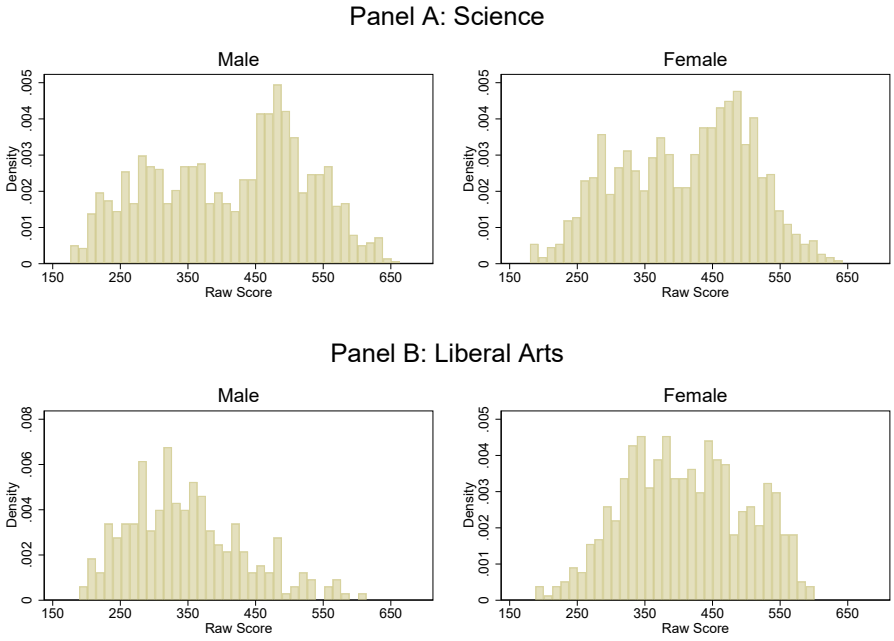
In 2019, our sample county has around 4,000 students from 8 high schools registered for the national college entrance examination. However, due to the administrative process, the provincial education authority only forwards the total test scores of students who have received college offers to the local education authority. Hence, we lack information regarding the test scores of students who haven't received any college offers, mainly due to not meeting the score requirements for their applied colleges. In our sample, the majority of students (over 76%) have received college offers, and we will further discuss this issue and show that the gender composition in the test room is uncorrelated with the college admission in Section 2.3. We also obtained data on the mock exam, which is designed to simulate the national college entrance examination and serves as a useful pre-measured students' academic ability. Since the mock exam is organized by the local government, we have data on all students who took the mock exam, as well as their subject-specific scores. However, the mock exam is less formally organized compared to the national college entrance examination, as students take it in their own schools, and not all students participate.

Figure 1 shows the distribution of raw test scores in our sample county. The test score has substantial variation, with one standard deviation corresponding to 105.8 points in the

[4]Female and male are the only two genders legally recognized in China.

[5]We are by no means the first to utilize the student-level National College Entrance Exam data (Cai et al., 2019). Previous literature utilizes data from a single province or several provinces (Graff Zivin et al., 2020; Kang et al., 2024), with some even employing data from the entire country (between 1999 and 2003) (Li and Qiu, 2023; Li et al., 2024; Guo, Shi and Zhang, 2024). However, our identification necessitates that (1) the sampled area implements random test room assignment during the sampling period, and (2) the data includes detailed administrative information, including the test room. Provinces in China began implementing random test room assignments relatively recently, a practice that extends beyond the sample period covered in existing literature. More importantly, the assignment process is decentralized at the local level, making it challenging to acquire information on each student's test room on a large scale. To the best of our knowledge, the datasets at the provincial or national scale utilized in the existing literature do not include such test room information. Lastly, our resources only allow us to access data in our sample county that meets both of these conditions simultaneously.

Figure 1: Histogram of Raw Score



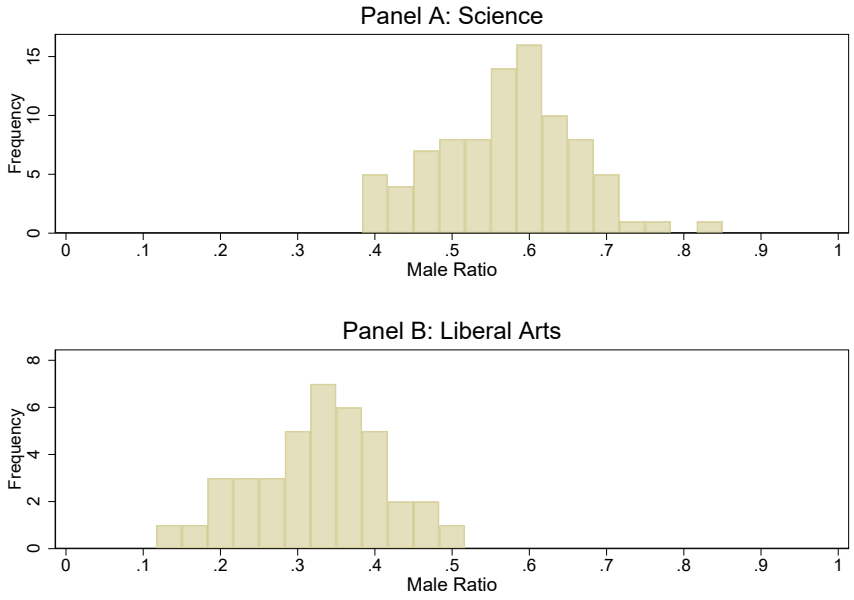

*Notes:* Panel A depicts the histogram of raw score for the science track, and Panel B depicts the histogram of raw score for the liberal arts track. The unit of observation is the student.

science track and 92.3 points in the liberal arts track. Our sample's average test score is largely comparable to the provincial average, but slightly lower. In both the science track and liberal arts track, the provincial average test scores are 437 points (with bonus points), whereas in our county, the average raw test score (without bonus points) for the science track is 411 points, and for the liberal arts track is 392 points.[6] In the science track, male students perform slightly better than female students, with an average (median) score of 412 (431) points for male students and 410 (423) points for female students. In the liberal arts track, female students perform much better than male students, with an average (median) score of 410 (407) points for female students and 348 (334) points for male students.

There are approximately 130 test rooms in our sample county, each accommodating ex-

[6]The bonus point system is a form of affirmative action in China. For example, individuals from minority groups may receive an additional 10 points added to their raw score in the national college entrance exam during the admission process.

Figure 2: Histogram of Gender Composition of the Test Room



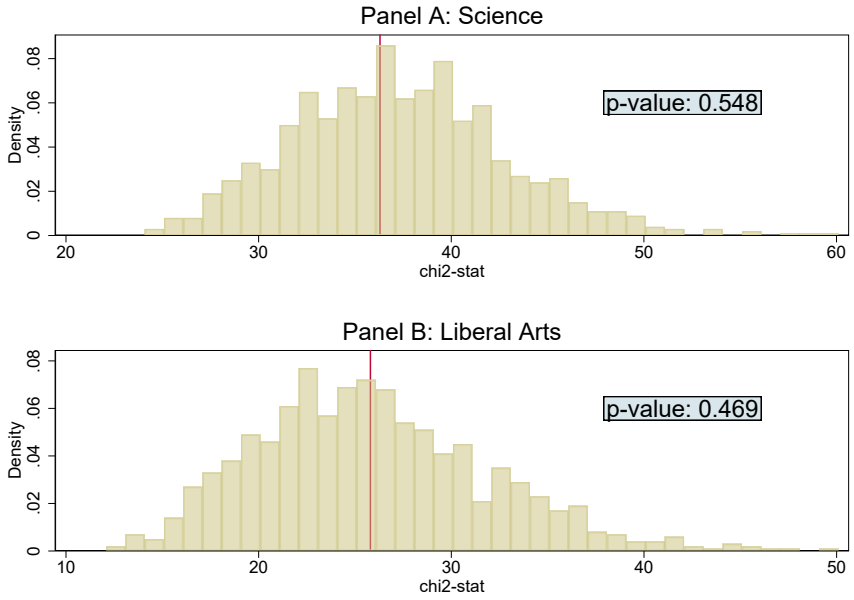

*Notes:* Panel A depicts the histogram of gender composition of the test room for the science track, and Panel B depicts the histogram of gender composition of the test room for the liberal arts track. The unit of observation is the test room.

actly 30 students, located in three test centers.[7] For the science track, there are a total of 88 test rooms, with 1509 male students and 1131 female students. On the other hand, the liberal arts track comprises 39 test rooms, with 382 male students and 788 female students. In Figure 2, we present the histogram of the male ratio in the test rooms. The male ratio is calculated by dividing the number of males in each test room by the room's capacity (i.e., 30), which could vary from zero to one in principal. The significant disparity in male ratios between science and liberal arts tracks implies very distinct distributions of male ratios in test rooms across these tracks. It is likely that girls in liberal arts, where females are predominate, are less likely to be randomly assigned to male-dominated test rooms. Conversely, in the science track, there are likely to be fewer test rooms with a majority of female students. Acknowledging the clear differences in male ratios and score distributions between science and liberal arts tracks, our

[7] As the total number of students is unlikely to be divided evenly by 30, there are three test rooms with capacities lower than 30: two rooms accommodate 13 students each, while one accommodates 10 students. In our main analysis, we do not include these three test rooms to maintain a consistent total number of students in each test room and thereby improve the comparability of our sample. Nevertheless, our findings remain consistent when we include all test rooms in the analysis.

analysis will examine the effect of gender composition on test scores separately for each track.

Figure 3: Histogram of $\chi^2$ Statistic




*Notes:* Panel A depicts the histogram of $\chi^2$ statistic for the number of male students in science track test rooms, derived from 1,000 simulation runs. The vertical red line represents the $\chi^2$ statistic for the number of male students in science track test rooms based on our sample. Panel B depicts the same for the liberal arts track.

Given the random assignment process, it is reasonable to anticipate that the number of male students in the test rooms would largely conform to a binomial distribution (when ignoring correlations across rooms), with parameters $n$ equals 30, and $p$ corresponds to the average male ratio for each academic track. To verify whether the gender distribution across test rooms is consistent with the binomial distribution, we utilize a chi-squared ($\chi^2$) test. Specifically, we simulate the test room assignment 1,000 times using the binomial distribution for 30 individuals, considering the male percentage in each academic track. Figure 3 reports the histogram of $\chi^2$ statistic for the number of male students in test rooms, derived from 1,000 simulation runs. P-values for both the science and liberal arts tracks are larger than 0.45, indicating that the observed test room assignment in our setting align with the binomial distribution.

## 2.3 Balance Checks

To further validate the random assignment of the test room, we investigate whether the gender composition in the test room is associated with individual-level characteristics. Our dataset encompasses basic demographic information that students are required to provide during their registration for the national college entrance examination, including information on ethnicity, age, Communist Youth League membership, class leadership roles, and urban/rural area.[8] Table 1 reports the balance check. Essentially, the gender composition in the test room (i.e., the proportion of males in the room) shows no significant correlation with the basic demographic variables for both male and female students. There is one exception to this observation: female students in test rooms with a higher male ratio are more likely to be Han Chinese on average, though the difference is not significant economically. Nonetheless, we will include individual characteristics as control variables in the regression analysis to account for any observed differences and to improve the precision of the estimation.

Furthermore, we find that students assigned to different test rooms exhibit similar performance on the mock exam. Lastly, we only have access to test scores for those students who have successfully gained admission to college, leading to a smaller sample size for our main analysis.[9] Therefore, it is important for us to examine the relationship between gender composition in the test room and the students' college admission status. In Table 1, we do not identify any noticeable relationship between gender composition and college admission status. The coefficient in Column (7) is very close to zero, and is not statistically significant.

# 3 Identification and Results

## 3.1 Identification

To rigorously quantify the impact of gender composition on cognitive performance, accounting for gender disparities, we estimate the following specifications separately for the science and liberal arts tracks:

[8]The class leaders consist of the class monitor, vice monitor, Youth League secretary, and student union leaders.

[9]According to the administrative procedure, the provincial government only returns those students who have successfully gained admission to college to the local government for admission purposes.

Table 1: Balance Checks

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | | | | Panel A: Male | | | |
| VARIABLES | Han | Age | CYL | Leader | Urban | Mock Exam Score | Admission |
| MaleRatio | -0.00302 | 0.264 | -0.0194 | 0.0153 | -0.144 | -0.147 | 0.0562 |
| | (0.0185) | (0.198) | (0.0657) | (0.0881) | (0.100) | (0.289) | (0.119) |
| Observations | 1,860 | 1,860 | 1,860 | 1,860 | 1,860 | 1,740 | 1,860 |
| Track FEs | X | X | X | X | X | X | X |
| Mean | 0.996 | 17.50 | 0.949 | 0.389 | 0.159 | 0.0692 | 0.741 |
| | | | | Panel B: Female | | | |
| VARIABLES | Han | Age | CYL | Leader | Urban | Mock Exam Score | Admission |
| MaleRatio | 0.00397* | -0.252 | -0.00602 | 0.0439 | 0.00366 | -0.0644 | -0.000647 |
| | (0.00235) | (0.202) | (0.0390) | (0.0812) | (0.0938) | (0.235) | (0.104) |
| Observations | 1,899 | 1,899 | 1,899 | 1,899 | 1,899 | 1,813 | 1,899 |
| Track FEs | X | X | X | X | X | X | X |
| Mean | 0.998 | 17.37 | 0.973 | 0.498 | 0.138 | 0.325 | 0.787 |

*Notes:* This table presents the relationship between the gender composition of the test room and individual-level characteristics. Panel A displays the results for male students, while Panel B shows the results for female students. $MaleRatio$ is the male ratio in the test room, which is calculated by dividing the number of males in each test room by 30 (i.e., the room's capacity). $Han$ is a dummy variable indicating whether the student is of Han ethnicity. $Age$ is the student's age. $CYL$ is a dummy variable indicating whether the student is a member in Communist Youth League. $Leader$ is a dummy variable indicating whether the student is a class leader. $Urban$ is a dummy variable indicating whether the student is from urban area. *Mock Exam Score* refers to the standardized test score from the mock exam. $Admission$ is a dummy variable indicating whether the student is admitted to college. Robust standard errors are reported. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

$$StdScore_{i(hr)} = \beta_1 MaleRatio_r + \beta_2 Female_i + \beta_3 MaleRatio_r * Female_i + \pi_h + \epsilon_i \quad (1)$$

where $StdScore_i$ represents the standardized total score of student $i$ in the national college entrance exam, within the specific academic track (such as science or liberal arts). The variable $MaleRatio_r$ measures the proportion of male students in the test room $r$, which could vary from 0 to 1 in principal. $Female_i$ is a dummy variable indicating whether student $i$ is female. $\beta_3$ is the coefficient of the interaction term between gender composition and the student's gender, capturing the gender disparities in the impact of the test room's gender composition on students' performance in a high-stakes exam. Importantly, we also include high-school-class fixed effects $\pi_h$. In China, high school students attend all their classes with the same group of classmates, who have relatively similar academic background, in an average class size of 50 students. Therefore, high-school-class fixed effects allow for comparisons among more comparable students from the same high school class who are randomly assigned to different test rooms with varying gender compositions. Including high-school-class fixed effects also controls for other class-specific factors, such as the gender ratio in the high school class. Standard errors are clustered at the test-room level.

$$StdScore_{i(hr)} = \beta_1 MaleRatio_r + \beta_2 Female_i + \beta_3 MaleRatio_r * Female_i + \gamma Ind_i + \phi TR_r + \pi_h + \pi_{c(r)} + \epsilon_i \quad (2)$$

Furthermore, to account for individual differences observed in Table 1, we will enhance our baseline specification in our preferred specification by including various individual-level characteristics such as ethnicity, age, Communist Youth League membership, class leadership roles, and urban/rural area. Moreover, as shown in last row of Table 1, male and female students may differ along many dimensions. Therefore, the gender composition of the test room may also be correlated with other characteristics of the test room, which could potentially influence academic performance. To control for differences among other test room characteristics, we introduce a series of test room characteristics, $TR_r$, such as ethnicity composition, average age, Communist Youth League membership ratio, class leader ratio, and urban com-

position.[10] Lastly, we include test-center fixed effects, $\pi_{c(r)}$, in our specification.

## 3.2 Main Results

We begin the analysis by estimating the impact of gender composition on cognitive performance by academic track, as illustrated in Equation 1. Results for students in science track and liberal arts track are reported in Columns (1) and (2) of Table 2, respectively. We do not find the gender composition of the test room affects male student's performance. On average, female students exhibit better academic performance. However, their performance is adversely affected by the presence of a higher proportion of males in the test room, with this effect being statistically significant for those in the science track.

[10] These characteristics of the test room are less apparent to students. While some may be able to infer whether someone is from an urban or rural area based on their appearance or behavior, it is much more difficult compared to identifying gender in this context. Nonetheless, we report the association between gender composition and other test room characteristics in Table A1, and we do not find that a higher male ratio is significantly correlated with other test room characteristics.

Table 2: Gender Composition and Test Score: Main Results

| | Panel A: All | | | | Panel B: Male | | | | Panel C: Female | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| VARIABLES | Std Score | | | | Std Score | | | | Std Score | | | |
| MaleRatio ($\beta_1$) | 0.157 | 0.347 | 0.105 | 0.220 | 0.0918 | 0.0931 | 0.0609 | 0.156 | -0.350* | -0.607** | -0.396** | -0.288*** |
| | (0.189) | (0.403) | (0.185) | (0.444) | (0.198) | (0.574) | (0.178) | (0.129) | (0.185) | (0.281) | (0.172) | (0.105) |
| Female ($\beta_2$) | 0.260* | 0.671*** | 0.241 | 0.672*** | | | | | | | | |
| | (0.151) | (0.160) | (0.148) | (0.145) | | | | | | | | |
| MaleRatio*Female ($\beta_3$) | -0.506** | -0.607 | -0.517** | -0.660 | | | | | | | | |
| | (0.252) | (0.503) | (0.248) | (0.464) | | | | | | | | |
| Mock Exam Std Score | | | | | | | | 0.762*** | | | | 0.837*** |
| | | | | | | | | (0.0220) | | | | (0.0211) |
| $H_0 : \beta_1 + \beta_3 = 0$ | -0.349* | -0.260 | -0.412** | -0.441* | | | | | | | | |
| | (0.203) | (0.378) | (0.191) | (0.260) | | | | | | | | |
| Observations | 1,970 | 877 | 1,966 | 874 | 1,095 | 258 | 1,354 | 1,296 | 871 | 614 | 1,485 | 1,423 |
| R-squared | 0.669 | 0.530 | 0.681 | 0.544 | 0.680 | 0.511 | 0.661 | 0.858 | 0.718 | 0.550 | 0.644 | 0.852 |
| Ind Controls | . | . | X | X | X | X | X | X | X | X | X | X |
| Test-Room Controls | . | . | X | X | X | X | X | X | X | X | X | X |
| Test-Center FEs | . | . | X | X | X | X | X | X | X | X | X | X |
| Class FEs | X | X | X | X | X | X | X | X | X | X | X | X |
| Sample | Science | Liberal Arts | Science | Liberal Arts | Science | Liberal Arts | All | All | Science | Liberal Arts | All | All |

*Notes:* This table presents our main results. $MaleRatio$ is the male ratio in the test room. *Std Score* is the student's standardized test score. $Female$ is a dummy variable indicating whether the student is female. *Mock Exam Score* refers to the standardized test score from the mock exam. Individual controls include ethnicity, age, CYL member, class leader, and urban area. Test-room controls include Han composition, average age, CYL member composition, class leader composition, and the proportion of students from urban area in the test room. Standard errors are clustered at the test-room level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Columns (3) and (4) report the results with additional controls. Our findings reveal that the performance of female students is adversely affected by the presence of a higher share of males in the test room, even in the absence of interaction among students. For female students, a 10 percentage point increase in the male ratio in the test room corresponds to a reduction of 0.0412/0.0441 standard deviations in test scores for those in the science/liberal arts track. Moreover, it indicates that this adverse effect does not extend to male students. Additionally, the results do not exhibit any noteworthy influence of other test-room characteristics on student performance. To give a sense of its practical significance, if we were to relocate a female student from a test room with 15 male students to a test room with 20 male students in the science track (translating to a 16.7 percentage-point rise in the male ratio), her test score would decrease by roughly 7.3 points in the science track (calculated as 0.167 * -0.412 * 105.8). This decrement is non-trivial and could result in one falling behind up to 6,400 other students among 210,000 students from the same province in the college admissions process. Similarly, if we were to relocate a female student in the liberal arts track from a test room with 10 male students to a test room with 15 male students, her test score would decrease by approximately 6.8 points (calculated as 0.167 * -0.441 * 92.3). This decrease could result in a drop of up to 4,100 positions in her provincial ranking out of 150,000 students.

Lastly, in Panel B and Panel C of Table 2, we estimate the impact of gender composition on cognitive performance by gender and by academic track. We again do not find the gender composition of the test room affects male student's performance: the magnitude is small and statistically insignificant. In addition, our evidence suggests that female students' performance is adversely affected by the presence of a higher percentage of males in the test room, despite the small sample size that may have caused the coefficients to differ slightly for the liberal arts track. To increase statistical power, we pool all students by gender. Our main findings remain consistent, even after accounting for students' performance in the mock exam, which represents their pre-measured ability.

# 4 Discussion on Channels

What are the underlying reasons behind the passive influence of gender composition? In this section, we will discuss three possible explanations in detail: the activation of stereotype threat, gender differences in competition, and the "acting wife" hypothesis."

## 4.1 The Stereotype Threat

In a groundbreaking psychology study conducted by Steele and Aronson (1995), they demonstrate that negative stereotypes pertaining to one's group abilities can hinder performance in domains where these stereotypes are relevant, a phenomenon known as "stereotype threat." The activation of negative stereotypes inhibits cognitive performance by consuming the cognitive resources that are required to complete cognitively challenging tasks (Schmader, Johns and Forbes, 2008).[11]

In educational settings, a prevalent gender stereotype is that female students are less talented in science than their male counterparts (Tiedemann, 2000; Leslie et al., 2015). Approximately 70% of more than 500,000 Implicit Association Tests conducted by citizens across 34 countries revealed implicit stereotypes linking science more strongly with males than with females (Nosek et al., 2009). Krendl et al. (2008) use functional magnetic resonance imaging (fMRI) to present neurobehavioral evidence that when women are reminded of math-related gender stereotypes, they do not recruit the neural networks typically associated with mathematical learning. Consequently, their performance in math tasks is adversely affected.[12]

Notably, the activation of negative stereotypes can often be subtle and occur outside of people's conscious awareness. The presence of males is a common trigger for activating stereotype threat in lab experiments (Marx and Roman, 2002). In a related lab experiment, Inzlicht and Ben-Zeev (2000) reveal that the mere presence of males is sufficient to trigger stereotypes and subsequently impair the performance of female students in a math task, even in a low-stakes experimental setting.[13]

In short, stereotype threat theory predicts that female students will perform worse in STEM subjects, but not in non-STEM subjects, when surrounded by a higher proportion of male students, whereas male students' performance remains unaffected by gender composi-

[11] Please refer to Ellemers (2018) for a comprehensive literature review on gender stereotypes in psychology.

[12] Gender stereotypes can shape beliefs and contribute to gender gaps in self-confidence and behavior in cooperative games, particularly when the genders of the partners are revealed to be different (Coffman, 2014; Bordalo et al., 2019).

[13] Inzlicht and Ben-Zeev (2000)'s findings may be open to alternative interpretations due to weaknesses in their experimental design. Since the reward structure is not linked to performance, the level of effort exerted by female participants can also be influenced by the presence of male students. Furthermore, given that the performance results are publicly announced immediately, "acting wife" (Bursztyn, Fujiwara and Pallais, 2017) or the possibility that the psychological cost of public failure may vary depending on the presence of male students could also serve as an explanation.

tion. Our findings confirm that male students are not impacted by the gender balance; however, the absence of subject-specific test scores limits our ability to examine the effect on female students across different subjects. To address this, we introduce an alternative approach based on the intuitive premise that female students who believe in the stereotype should be more susceptible to stereotype threat. Specifically, we begin by using a nationally representative survey to identify which groups of female students are more likely to hold gender stereotypes. We then examine whether these students are disproportionately affected by the gender composition in the test room during the national college entrance examination.

The China Education Panel Survey (CEPS) is a large-scale, nationally representative longitudinal study that began in the 2013-2014 academic year, focusing on middle school students. Using a stratified, multistage sampling design with probability proportional to size (PPS), the CEPS randomly selected a nationally representative sample of approximately 20,000 students from 438 classrooms across 112 schools, spanning 28 county-level units in mainland hina.[14]

Our analysis focuses on female students in their final year of middle school, as this cohort is more comparable to our main analysis. Our sample includes 4,425 female students across 112 schools. We find that 50.5% of female students agree with the statement, "Boys are better at mathematics than girls," lending some support to the notion that girls in our context may be influenced by this gender stereotype. We further investigate how this belief varies among students with different levels of academic ability. One might expect that female students who excel in STEM subjects would be less likely to endorse the math-gender stereotype, whereas those who perform better in non-STEM subjects might be more inclined to agree with it. Using their most recent midterm test scores as a proxy for academic ability, Column 1 of Table 3 shows that girls who perform well in STEM subjects are indeed less likely to agree with the math-gender stereotype. Interestingly, we also observe that girls who excel in non-STEM subjects are more likely to hold this stereotype.[15]

We further investigate how class-level characteristics might be linked to the endorsement of gender stereotypes. Specifically, we focus on two aspects: the proportion of male students in the classroom and whether male students outperform female students in STEM subjects. A

[14]The CEPS is widely used in the literature to study peer and teacher effects (Gong, Lu and Song, 2018, 2019; Yu, 2020; Hill and Zhou, 2023). Website: http://ceps.ruc.edu.cn.

[15]The CEPS was conducted before the end of the semester and only contains information on students' academic performance from their most recent midterm.

Table 3: Individual, Class Characteristics, and Gender Stereotypes

| VARIABLES | (1) | (2) | (3) |
|---|---|---|---|
| | Gender Stereotype | | |
| STEM Exam std score | -0.139*** | -0.160*** | -0.159*** |
| | (0.0133) | (0.0134) | (0.0130) |
| Non-STEM Exam std score | 0.0655*** | 0.0828*** | 0.0832*** |
| | (0.0149) | (0.0145) | (0.0144) |
| Class Male Ratio (Above Median) | | | 0.0449** |
| | | | (0.0200) |
| Class: Males Performing Better in STEM | | | 0.0393 |
| | | | (0.0282) |
| Observations | 4,425 | 4,424 | 4,425 |
| R-squared | 0.039 | 0.144 | 0.098 |
| Class FEs | . | X | . |
| School FEs | . | . | X |

*Notes:* This table presents the relationship between individual characteristics, class characteristics, and gender stereotypes. *Gender Stereotype* is a dummy variable indicating whether a student agrees with the statement that boys are better at mathematics than girls. *STEM Exam Score* refers to the standardized test score for STEM subjects in the midterm exam. *Non-STEM Exam Score* refers to the standardized test score for non-STEM subjects in the midterm exam. *Class Male Ratio (Above Median)* is a dummy variable indicating whether the male ratio in a student's classroom is above the median. *Class: Males Performing Better in STEM* is a dummy variable indicating whether a student is from a classroom where male students, on average, scored higher in STEM subjects than female students in the midterm exam. Standard errors are clustered at the classroom level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

higher proportion of male students may make gender differences more salient, potentially reinforcing the gender stereotype. Additionally, a higher male presence could heighten female students' sensitivity to gender imbalance, making gender composition a particularly potent trigger for stereotype activation during high-stakes exams. We also consider the relative performance in STEM subjects between male and female students within the same classroom. If male students outperform female students, the female students may adjust their beliefs and become more likely to endorse the math-gender stereotype. The results are presented in Column 3 of Table 3. We find that a higher male ratio in the classroom is associated with stronger endorsement of the gender stereotype, whereas the relative performance between boys and girls appears to be less significant. This is likely because girls generally outperform boys in STEM, and in fewer than 40% of classrooms do boys have better STEM scores, with the performance difference being relatively small.

After identifying the types of female students who are more likely to endorse the gender stereotype, we return to our main sample to explore whether these students are disproportionately affected by the gender composition during the high-stakes exam. The results, reported in Table 4, show that girls with low STEM performance, high non-STEM performance in the mock exam, and those from classrooms with a higher male ratio are more affected by the gender composition. Notably, these characteristics align with those of the girls who are more likely to endorse the gender stereotype–a group particularly susceptible to stereotype threat.

Lastly, as demonstrated earlier, relocating a female student, whether in the science or liberal arts track, to a test room with five additional male students reduces her test score by approximately 7 points on average. One concern might be that the effects of gender composition on female students appear similar across both the science and liberal arts tracks, which seems inconsistent with the stereotype threat expectation that female students in the liberal arts track would be more strongly affected. However, it is important to note that the total STEM-related score is 450 points for students in the science track and 150 points for those in the liberal arts track. Therefore, when focusing specifically on stereotype-related domains, the effect is considerably stronger for female students in the liberal arts track, which again aligns with the stereotype threat theory.

Table 4: Gender Composition and Test Score: Heterogeneity

| VARIABLES | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Std Score | | | |
| MaleRatio | -0.418*** | 0.000745 | -0.140 | 0.252 |
| | (0.160) | (0.143) | (0.223) | (0.594) |
| STEM Mock Exam Score | 0.363*** | | | |
| | (0.0796) | | | |
| Non-STEM Mock Exam Score | | 1.044*** | | |
| | | (0.0652) | | |
| MaleRatio*STEM Mock Exam Score | 0.304* | | | |
| | (0.162) | | | |
| MaleRatio*Non-STEM Mock Exam Score | | -0.631*** | | |
| | | (0.132) | | |
| MaleRatio*Class Male Ratio (Above Median) | | | -0.728** | |
| | | | (0.333) | |
| MaleRatio*Class: Males Performing Better in STEM | | | | -1.445 |
| | | | | (1.144) |
| Observations | 1,423 | 1,423 | 1,485 | 1,458 |
| R-squared | 0.764 | 0.812 | 0.645 | 0.634 |
| Ind Controls | X | X | X | X |
| Test-Room Controls | X | X | X | X |
| Test-Center FEs | X | X | X | X |
| Class FEs | X | X | X | X |
| Sample | All | All | All | All |

*Notes:* This table presents results for heterogeneous analysis. $MaleRatio$ is the male ratio in the test room. $Std\ Score$ is the student's standardized test score. *STEM Mock Exam Score* refers to the standardized test score for STEM subjects in the mock exam. *Non-STEM Mock Exam Score* refers to the standardized test score for non-STEM subjects in the mock exam. *Class Male Ratio (Above Median)* is a dummy variable indicating whether the male ratio in a student's classroom is above the median. *Class: Males Performing Better in STEM* is a dummy variable indicating whether a student is from a classroom where male students, on average, scored higher in STEM subjects than female students in the mock exam. Individual controls include ethnicity, age, CYL member, class leader, and urban area. Test-room controls include Han composition, average age, CYL member composition, class leader composition, and the proportion of students from urban area in the test room. Standard errors are clustered at the test-room level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

## 4.2 Alternative Channels

First, one might be inclined to attribute our findings to explanations in the literature on gender differences in competition. Specifically, women may tend to underperform as levels of competitiveness increase or when faced with high time pressure (Gneezy, Niederle and Rustichini, 2003; Ors, Palomino and Peyrache, 2013; Morin, 2015; Cai et al., 2019; Galasso and Profeta, 2024). However, this explanation is unlikely to be the case in our setting. All students are taking the same test and it is well understood by students that they do not compete directly with their fellow students in the same test room, as college admissions are contingent on the performance of all other students in the same province. Therefore, the objective level of competitiveness remains consistent for students across different test rooms.

One could further argue that although the objective set of relevant competitors remains unchanged across test rooms, the gender composition within a room might impact beliefs about the gender composition of the overall pool of competitors, potentially influencing the perceived level of competitiveness and then academic performance (Gomez-Ruiz, Cervini-Plá and Ramos, 2024). While this argument may adequately explain the pattern observed in the science track, it does not align with the pattern noted in the liberal arts track. As shown in section 2.2, female students outperform male students in the liberal arts track, with an average score of 410 points for female students and 348 points for male students. If the presence of more male students in the test room positively correlates with the belief that the exam generally includes more male students, we might expect female students to perform better due to a perceived lower level of competitiveness, which is not consistent with the pattern in Table 2. Moreover, the explanation of gender differences in competition fails to account for our findings that female students, who are more likely to hold gender stereotypes, are more strongly affected by the presence of more male students in the test room. Although the perceived level of competitiveness should remain constant in the same test room, the effect of gender composition depends on their beliefs about gender stereotypes.

Another potential explanation relates to Bursztyn, Fujiwara and Pallais (2017)'s work on "acting wife", which suggests that single women may avoid career-enhancing actions because these actions might signal undesirable traits, such as ambition, to the marriage market. Bursztyn, Fujiwara and Pallais (2017) conduct an experiment among students in an elite US MBA program, asking them to complete a questionnaire about their job preferences for summer

internships. The questionnaire included questions on desired compensation, work hours, and days of travel per month. Answers from students in the control group will only be discussed anonymously, while answers from students in the treatment group will be discussed publicly in the career course. They find that when there is a greater number of (single) male students present, unmarried female students perform less well and present themselves in a less professional manner, signaling desirable traits in the marriage market. An important feature of these signals is their observability: when performance is not expected to be observed by males, unmarried and married female students perform similarly. In our context, the performance – the test score – will only be released in a private manner two weeks after the national college entrance examination. Therefore, it is less likely that females' underperformance in a test room with a higher percentage of male students can be attributed to female students signaling desirable marriage-related characteristics.

A neat division of mechanisms into three categories is, of course, an oversimplification. For example, gender differences in behaviors may reflect conformity to social norms (e.g., the Confucian notion that 'a lack of talent is a virtue in a woman') (Akerlof and Kranton, 2000; Alesina, Giuliano and Nunn, 2013; Bertrand, Kamenica and Pan, 2015). However, the existing literature lacks a conceptual framework linking gender composition to conformity with these norms. It is also important to note that, similar to the "acting wife" hypothesis, conformity to social norms may heavily depend on the observability of behavior—an element not present in our context. Therefore, our findings are most compatible with the stereotype threat explanation based on the evidence at hand.

# 5 Conclusions

In this paper, we use administrative data from a county that randomly assigned students taking the college entrance examination in 2019 to the test room to explore whether the mere presence of a higher portion of males in the working environment has a differential impact on performance based on gender. Our analysis reveals that a higher presence of male students in the test room leads to a reduction in performance among female students. Conversely, the gender composition does not appear to have a significant impact on the performance of male students. Considering the limited interaction among students during the exam, these findings provide empirical support for a previously unexplored channel of gender composition: it can

exert an influence even passively. We further explore the underlying mechanisms of this passive effect and present evidence suggesting that our main findings align with the activation of stereotypes: female students who are more likely to hold gender stereotypes are more strongly affected by the presence of more male students in the test room.

Our findings have important policy implications for the development of an equitable testing system. Most students take standardized tests before finishing high school, and these tests can influence their college admissions and, ultimately, long-term outcomes such as salary (Ebenstein, Lavy and Roth, 2016). These tests are often administered in mixed-gender environments, which may potentially trigger stereotype threat. Furthermore, gender stereotypes are not limited to China but are prevalent globally (Nosek et al., 2009). Therefore, the gender composition of the testing environment could potentially impact a substantial number of female students worldwide, particularly in regions with a male-dominant population or pervasive gender stereotypes. Based on our findings (and with some extrapolation), implementing female-dominated or single-sex test rooms—or even online tests in certain cases—could help mitigate the disadvantages faced by female students in high-stakes academic assessments. In fact, Gomez-Ruiz, Cervini-Plá and Ramos (2024) find that female students performed better when tests were moved online for admissions to Uruguay's Coding Program in 2019, though we cannot dispositively rule out confounding factors, such as changes in the pool of competitors, as males were ineligible for the program that year. More research is needed to better understand potential biases in high-stakes academic assessments.

# References

**Akerlof, George A., and Rachel E. Kranton.** 2000. "Economics and Identity*." *The Quarterly Journal of Economics*, 115(3): 715–753.

**Alesina, Alberto, Paola Giuliano, and Nathan Nunn.** 2013. " On the Origins of Gender Roles: Women and the Plough *." *The Quarterly Journal of Economics*, 128(2): 469–530.

**Alston, Mackenzie, William A. Darity, Catherine C. Eckel, Lawrence McNeil, and Rhonda Sharpe.** 2022. "The effect of stereotypes on black college test scores at a historically black university." *Journal of Economic Behavior  Organization*, 194: 408–424.

**Bertrand, Marianne, Emir Kamenica, and Jessica Pan.** 2015. " Gender Identity and Relative Income within Households *." *The Quarterly Journal of Economics*, 130(2): 571–614.

**Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes.** 2013. "Under Pressure? The Effect of Peers on Outcomes of Young Adults." *Journal of Labor Economics*, 31(1): 119–153.

**Booth, Alison L, Lina Cardona-Sosa, and Patrick Nolen.** 2018. "Do single-sex classes affect academic achievement? An experiment in a coeducational university." *Journal of Public Economics*, 168: 109–126.

**Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer.** 2019. "Beliefs about Gender." *American Economic Review*, 109(3): 739–73.

**Bostwick, Valerie K., and Bruce A. Weinberg.** 2022. "Nevertheless She Persisted? Gender Peer Effects in Doctoral STEM Programs." *Journal of Labor Economics*, 40(2): 397–436.

**Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais.** 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review*, 107(11): 3288–3319.

**Cai, Xiqian, Yi Lu, Jessica Pan, and Songfa Zhong.** 2019. "Gender Gap under Pressure: Evidence from China's National College Entrance Examination." *The Review of Economics and Statistics*, 101(2): 249–263.

**Coffman, Katherine Baldiga.** 2014. " Evidence on Self-Stereotyping and the Contribution of Ideas *." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

**Cohen, Geoffrey L, and Julio Garcia.** 2005. ""I am us": negative stereotypes as collective threats." *Journal of personality and social psychology*, 89(4): 566.

**Contrerasa, Valentina, Chiara Orsinib, Berkay Ozcana, and Johann Koehlera.** 2022. "Effects of team diversity on performance, perceptions, and predictions." *LSE Social Policy Working Paper*, 06-22.

**Dee, Thomas S.** 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *The Review of Economics and Statistics*, 86(1): 195–210.

**Ebenstein, Avraham, Victor Lavy, and Sefi Roth.** 2016. "The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution." *American Economic Journal: Applied Economics*, 8(4): 36–65.

**Eisenkopf, Gerald, Zohal Hessami, Urs Fischbacher, and Heinrich W. Ursprung.** 2015. "Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland." *Journal of Economic Behavior  Organization*, 115: 123–143.

**Ellemers, Naomi.** 2018. "Gender stereotypes." *Annual Review of Psychology*, 69: 275–298.

**Figlio, David, Paola Giuliano, Riccardo Marchingiglio, Umut Ozek, and Paola Sapienza.** 2023. "Diversity in Schools: Immigrants and the Educational Performance of U.S.-Born Students." *The Review of Economic Studies*, rdad047.

**Galasso, Vincenzo, and Paola Profeta.** 2024. "Gender Differences in Math Tests: The Role of Time Pressure." *The Economic Journal*, ueae052.

**Gneezy, Uri, Muriel Niederle, and Aldo Rustichini.** 2003. "Performance in Competitive Environments: Gender Differences*." *The Quarterly Journal of Economics*, 118(3): 1049–1074.

**Gomez-Ruiz, Marcela, María Cervini-Plá, and Xavier Ramos.** 2024. "Do Women Fare Worse When Men Are Around? Quasi-Experimental Evidence."

**Gong, Jie, Yi Lu, and Hong Song.** 2018. "The Effect of Teacher Gender on Students' Academic and Noncognitive Outcomes." *Journal of Labor Economics*, 36(3): 743–778.

**Gong, Jie, Yi Lu, and Hong Song.** 2019. "Gender Peer Effects on Students' Academic and Noncognitive Outcomes: Evidence and Mechanisms." *Journal of Human Resources*.

**Goulas, Sofoklis, Rigissa Megalokonomou, and Yi Zhang.** 2024. "Female Classmates, Disruption, and STEM Outcomes in Disadvantaged Schools: Evidence from a Randomized Natural Experiment." Monash University, Department of Economics.

**Graff Zivin, Joshua, Tong Liu, Yingquan Song, Qu Tang, and Peng Zhang.** 2020. "The unintended impacts of agricultural fires: Human capital in China." *Journal of Development Economics*, 147: 102560.

**Guo, Shiqi, Xinzheng Shi, and Ming-ang Zhang.** 2024. "Sex Imbalance and Female Resilience: Insights from China's College Entrance Examinations." *Available at SSRN 4720789*.

**Halpern, Diane F., Lise Eliot, Rebecca S. Bigler, Richard A. Fabes, Laura D. Hanish, Janet Hyde, Lynn S. Liben, and Carol Lynn Martin.** 2011. "The Pseudoscience of Single-Sex Schooling." *Science*, 333(6050): 1706–1707.

**Hill, Andrew J, and Weina Zhou.** 2023. "Peer discrimination in the classroom and academic achievement." *Journal of Human Resources*, 58(4): 1178–1206.

**Hoxby, Caroline.** 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Papers 7867.

**Inzlicht, Michael, and Talia Ben-Zeev.** 2000. "A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males." *Psychological Science*, 11(5): 365–371.

**Jackson, C. Kirabo.** 2012. "Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in Trinidad and Tobago." *Journal of Public Economics*, 96(1): 173–187.

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

**Kang, Le, Ziteng Lei, Yang Song, and Peng Zhang.** 2024. "Gender Differences in Reactions to Failure in High-Stakes Competition: Evidence from the National College Entrance Exam Retakes." *Journal of Political Economy Microeconomics*, 2(2): 355–397.

**Krendl, Anne C, Jennifer A Richeson, William M Kelley, and Todd F Heatherton.** 2008. "The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math." *Psychological Science*, 19(2): 168–175.

**Lavy, Victor, and Analia Schlosser.** 2011. "Mechanisms and Impacts of Gender Peer Effects at School." *American Economic Journal: Applied Economics*, 3(2): 1–33.

**Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland.** 2015. "Expectations of brilliance underlie gender distributions across academic disciplines." *Science*, 347(6219): 262–265.

**Li, Hongbin, and Xinyao Qiu.** 2023. "Heuristics in Self-Evaluation: Evidence from the Centralized College Admission System in China." *The Review of Economics and Statistics*, 1–36.

**Li, Hongbin, Lingsheng Meng, Kai Mu, and Shaoda Wang.** 2024. "English language requirement and educational inequality: Evidence from 16 million college applicants in China." *Journal of Development Economics*, 168: 103271.

**Marx, David M., and Jasmin S. Roman.** 2002. "Female Role Models: Protecting Women's Math Test Performance." *Personality and Social Psychology Bulletin*, 28(9): 1183–1193.

**Morin, Louis-Philippe.** 2015. "Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform." *Journal of Labor Economics*, 33(2): 443–491.

**Nguyen, Hannah-Hanh D, and Ann Marie Ryan.** 2008. "Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence." *Journal of applied psychology*, 93(6): 1314.

**Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G. Greenwald.** 2009. "National differences in gender–science stereotypes predict national sex differences in science and math achievement." *Proceedings of the National Academy of Sciences*, 106(26): 10593–10597.

**Ors, Evren, Frédéric Palomino, and Eloïc Peyrache.** 2013. "Performance Gender Gap: Does Competition Matter?" *Journal of Labor Economics*, 31(3): 443–499.

**Paulette C. Flore, Joris Mulder, and Jelte M. Wicherts.** 2018. "The influence of gender stereotype threat on mathematics test scores of Dutch high school students: a registered report." *Comprehensive Results in Social Psychology*, 3(2): 140–174.

**Schmader, Toni, Michael Johns, and Chad Forbes.** 2008. "An integrated process model of stereotype threat effects on performance." *Psychological review*, 115(2): 336.

**Steele, Claude M.** 1997. "A threat in the air: How stereotypes shape intellectual identity and performance." *American psychologist*, 52(6): 613.

**Steele, Claude M, and Joshua Aronson.** 1995. "Stereotype threat and the intellectual test performance of African Americans." *Journal of personality and social psychology*, 69(5): 797.

**Stoevenbelt, Andrea H, Paulette C Flore, Jelte M Wicherts, Rachel Bergers, Robert Calin-Jageman, Luis A Gomez, Joachim Hüffmeier, Megan N Imundo, Scott Martin, Steven C Pan, and et al.** 2024. "Registered Replication Report: Johns, Schmader, Martens (2005)."

**Tiedemann, Joachim.** 2000. "Gender-related beliefs of teachers in elementary school mathematics." *Educational studies in Mathematics*, 41(2): 191–207.

**Whitmore, Diane.** 2005. "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment." *American Economic Review*, 95(2): 199–203.

**Yu, Han.** 2020. "Am I the big fish? The effect of ordinal rank on student academic performance in middle school." *Journal of Economic Behavior Organization*, 176: 18–41.

# Appendix

Table A1: Gender Composition and Other Test Room Characteristics

| VARIABLES | (1) HanRatio | (2) Average Age | (3) CYLRatio | (4) ClassLeaderRatio | (5) UrbanRatio |
|---|---|---|---|---|---|
| MaleRatio | -0.00181 | 0.114 | -0.0385 | 0.0184 | -0.0501 |
| | (0.00919) | (0.141) | (0.0381) | (0.0522) | (0.0555) |
| Observations | 127 | 127 | 127 | 127 | 127 |
| R-squared | 0.000 | 0.006 | 0.017 | 0.037 | 0.042 |
| Track FEs | X | X | X | X | X |

*Notes:* The table shows the relationship between gender composition and other test room characteristics. The unit of observation is the test room. $MaleRatio$ is the male ratio, $HanRatio$ is the Han people ratio, $AverageAge$ is the average age of students, $CYLRatio$ is the CYL member ratio, $ClassLeaderRatio$ is the class leader ratio, and $UrbanRatio$ is the proportion of students are from urban area in the test room. Robust standard errors are reported. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.