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

Xuan Li † Xiang Zhou ‡

February 15, 2024

Abstract: This paper examines whether gender composition in the workplace 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, particularly among those pursuing the science track, but it doesn't influence males. This study identifies a previously unexplored passive gender composition effect and offers new perspectives on the debate over single-sex schools and mixed-sex schools.

Keywords: Gender imbalance, high-stakes test, cognitive performance, workplace environment

JEL: J16, J24, A20

*We are indebted to Kevin Lang and Daniele Paserman for their continuous guidance, advice, and support .We would also like to thank Raymond Fisman, Joshua Goodman, Justin Hong, Hamish Low, Linh Tô, Hantao Wu, and Yuzhao Yang for helpful comments. 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

The working environment can potentially influence productivity and cognitive performance in a significant way (Zivin and Neidell, 2012; Adhvaryu, Kala and Nyshadham, 2020; Somanathan et al., 2021). In the field of education, debates around gender composition or the diversity in classrooms or schools remain contentious, largely attributed to the influential role of intensive daily interactions (Halpern et al., 2011; Lavy and Schlosser, 2011; Figlio et al., 2023). Unlike other research on gender peer effects, this paper aims to examine whether gender composition affects academic performance in a unique setting where student interaction is minimal or non-existent.

We leverage the Chinese college entrance examination, a high-stakes exam where students are required to complete exams individually. To prevent cheating, education authorities randomly assign students to standardized test rooms, each accommodating a fixed number of students (i.e., 30 students). Following the random assignment system, the gender composition varies across test rooms. In addition, this system results in students from the same high school class taking the same exam in different test rooms with varying gender compositions. Throughout the entire testing period, students remain in the same seat and complete the task independently. Two proctors oversee the entire examination process, making it unlikely for students to interact with one another. Following the completion of the examination, examination papers are graded by the provincial education authorities in a double-blind review.

Using administrative data on the universe of 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 largely consistent with the binomial distribution. Moreover, our balance checks indicate that the gender composition of the test room is not correlated with students' basic demographic characteristics, including ethnicity, age, and *hukou* status. Having confirmed random assignment, we provide visual evidence as to the gender composition of the test room and academic performance. We find that a higher proportion of male students in the test room correlates with lower performance among female students, especially those pursuing the science track. Meanwhile, we find that the performance of male students remains unaffected by the gender composition of the test room. Finally, we employ a rigorous regression analysis to corroborate the observed pattern. For female students, a 10 percentage point increase in the male ratio in the test room corresponds to a reduction of 0.036 standard deviations in test scores. Our findings remain robust even after incorporating

high-school-class fixed effects, test-center fixed effects, individual demographic controls, and additional test-room composition controls.

Our analysis presents a new channel indicating that the mere presence of a large proportion of males in the working environment can reduce females' cognitive performance in high-stakes situations. We further discuss potential explanations for our findings, including gender differences in competition (Gneezy, Niederle and Rustichini, 2003), the "acting wife" hypothesis (Bursztyn, Fujiwara and Pallais, 2017), and the activation of stereotype threat (Steele and Aronson, 1995; Steele, 1997).

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). 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 (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) both 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 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.

Our paper is also connected to an ongoing body of research focused on developing a fair testing system. The existing literature has demonstrated that performance in high-stakes exams can be easily influenced by temporary factors such as air pollution, temperature, pollen exposure, SNAP timing, local violent crime, and religious participation (Ebenstein, Lavy and Roth, 2016; Bensnes, 2016; Zivin et al., 2020; Park, 2022; Bond et al., 2022; Chang and Padilla-Romo, 2023; Wang, Wang and Ye, 2023). To our knowledge, we are the first to document the importance of the demographic composition of the test room environment: an increased proportion of male students could potentially lead to a decline in cognitive performance among female students.

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 college 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. 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 spans two days in early June, during which students are required to remain in the same test room and occupy the same seat throughout the entire testing period. Upon completion, the examination papers undergo computer scanning and are subsequently graded by the provincial education authority.

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 other students.

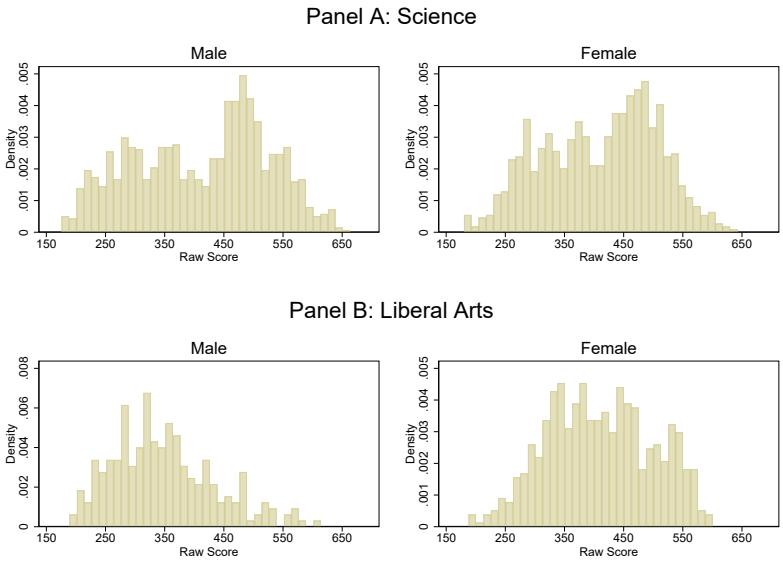
2.2 Data

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. One caveat is that 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.

In 2019, our sample county has around 4,000 students from 8 high schools registered for the national college entrance examination. Figure 1 shows the distribution of raw test scores in our sample. The test score has substantial variation, with one standard deviation corresponding to 105.8 points in the 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.[1] 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.

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.

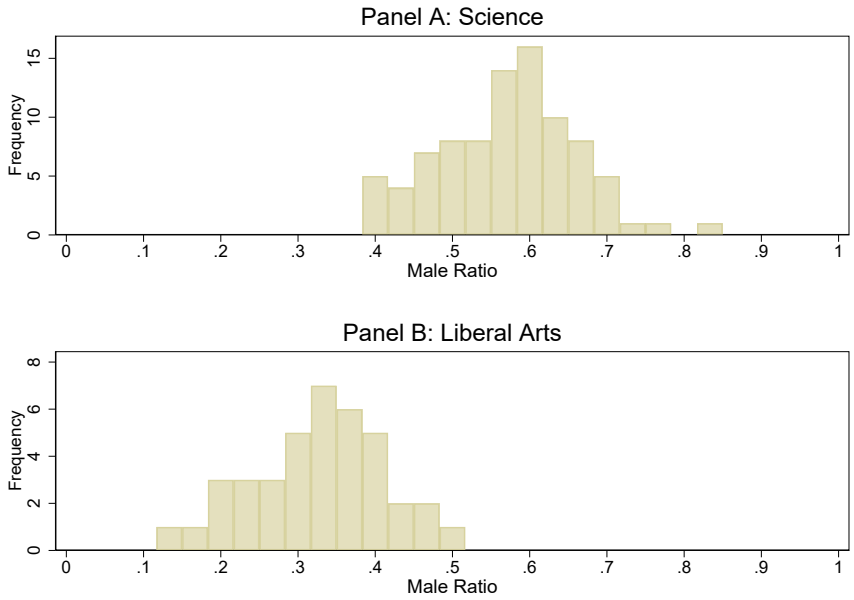
There are approximately 130 test rooms in our sample county, each accommodating exactly 30 students, located in three test centers.[2] The provincial government employs a computer program to randomly assign students to test rooms. This random assignment process is solely conditioanl on the academic track, 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

[1] Bonus points are available to certain students, such as minorities, who may receive a bonus of 10 points.

[2] 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.

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. 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. 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. 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. To verify whether the gender distribution across test rooms is consistent with the binomial distribution, we utilize a chi-squared (χ^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 A1 reports the histogram of χ^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.

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.

Table 1: Balance Checks

	Panel A: Male						Panel B: Female					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES	Han	Age	CYL	Leader	Urban	Admission	Han	Age	CYL	Leader	Urban	Admission
MaleRatio	-0.00302	0.264	-0.0194	-0.0602	-0.144	0.0562	0.00397*	-0.252	-0.00602	-0.0295	0.00366	-0.000647
	(0.0185)	(0.198)	(0.0657)	(0.131)	(0.100)	(0.119)	(0.00235)	(0.202)	(0.0390)	(0.133)	(0.0938)	(0.104)
Observations	1,860	1,860	1,860	1,860	1,860	1,860	1,899	1,899	1,899	1,899	1,899	1,899
Track FEs	X	X	X	X	X	X	X	X	X	X	X	X
Mean	0.996	17.50	0.949	0.389	0.159	0.741	0.998	17.37	0.973	0.498	0.138	0.787

Notes: This table presents the relationship between the gender composition of the test room and individual-level characteristics. $MaleRatio$ 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 has urban *hukou*. $Admission$ is a dummy variable indicating whether the student is admitted to college. Robust standard errors are reported. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

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 *hukou* status. Furthermore, as we only have access to test scores for those students who have successfully gained admission to college, it is important for us to examine the relationship between gender composition in the test room and the students' college admission status.

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. More importantly, we do not identify any noticeable relationship between gender composition and college admission status: the coefficient in Column (12) is very close to zero, and is not statistically significant.

3 Identification and Results

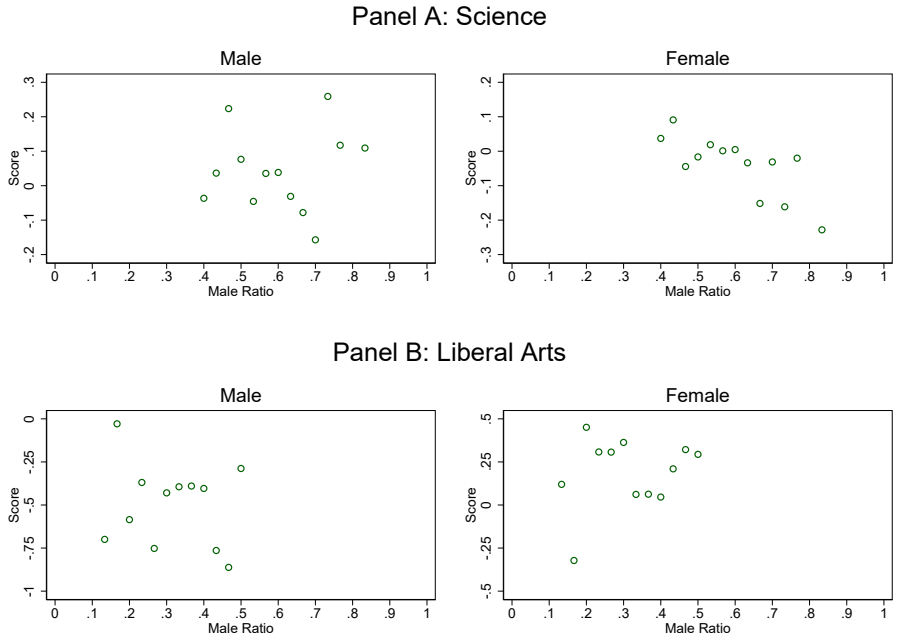
3.1 Preliminary Evidence

Figure 3 offers visual insights into the influence of gender composition on student average performance. The pattern for male students appears noisy and not very clear, while female students, especially those in the science track, seem to experience a decline in performance with an increasing proportion of males in the test room. It's worth noting that the measure of student average performance may be noisy at the extremes of the gender composition distribution in Figure 3, given the limited number of test rooms with either very few male students or very few female students. In the following section, we will employ a rigorous econometric model to quantify the effect of gender composition on student performance.

3.2 Identification

To quantify the effect of gender composition on cognitive performance, we first estimate the following specification separately for males and females, respectively:

Figure 3: Gender Composition and Test Score: Preliminary Evidence



Notes: Panel A shows the relationship between the gender composition of the test room and the average standardized test score for the science track by gender. Panel B shows the same relationship for the liberal arts track by gender.

$$StdScore_i(tr) = \beta_1 MaleRatio_r + \pi_t + \epsilon_i \tag{1}$$

where $StdScore_i(tr)$ represents the standardized total score of student i in the *Gaokao*, within the specific academic track t (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. In the baseline specification, we only include track fixed effects (π_t) to account for the fact that the random assignment process is solely contingent on the academic track. Standard errors are clustered at the test-room level.

In addition to our baseline specification, we will also estimate our preferred specification as follows:

$$StdScore_i(hr) = \beta_1 MaleRatio_r + \gamma Ind_i + \phi TR_r + \pi_h + \pi_{c(r)} + \epsilon_i \tag{2}$$

where Ind_i is a bunch of individual-level characteristics including ethnicity, age, Communist Youth League membership, class leadership roles, and *hukou* status to absorb potential differ-

ences observed in Table 1. In addition, 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 be correlated with other test room characteristics. Table A1 reports the association between gender compositiona and other test room characteristics, and we find that a higher male ratio is associated with a lower class leader ratio. 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 *hukou* composition. We also strengthen the track fixed effects π_t by including high-school-class fixed effects π_h. In China, high school students take all their classes with the same set of classmates. Therefore, high-school-class fixed effects absorb track fixed effects and 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. Lastly, we include test-center fixed effects, $\pi_{c(r)}$, in our specification.

3.3 Main Results

We start the analysis with a simple specification controlling for track fixed effects, as illustrated in Equation 1. Results are reported in Columns (1) and (3) of Table 2. In Columns (2) and (4), we then estimate our preferred specification with individual-level controls, other test room characteristics, high-school-class fixed effects, and test-center fixed effects (i.e., Equation 2). Columns (3) and (4) have virtually identical coefficients, and the precision improves with more controls in Column (4). 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. Moreover, we do not find the gender composition in the test room affects male student's performance: the magnitude is small and statistically insignificant. Additionally, the results do not exhibit any noteworthy influence of other test-room characteristics on student performance.

To offer a clearer perspective on gender disparities, the last two columns of Table 2 introduce an interaction term between gender composition and student's gender, and we investigate how gender composition impacts performance by academic track. Notably, the interaction term is statistically significant in Column (5) and is similar in Column (5) and Column (6). Taken together, the results in Table 2 support the conclusion of an adverse effect on girls that is more strongly supported in science (possibly due to the larger sample size) but is similar in magnitude in both tracks. Additionally, it indicates that this adverse effect does not extend to male students. Taken at face value, if we were to relocate a female student from a test room with 15 male students to a test room with 20 male students (translating to a 16.7 percentage-

Table 2: Gender Composition and Test Score: Main Results

	Panel A: Male		**Panel B: Female**		**Panel C: By Track**	
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Std Score		Std Score		Std Score	
MaleRatio (β_1)	-0.328	0.0981	-0.376	-0.357**	0.157	0.119
	(0.265)	(0.189)	(0.281)	(0.176)	(0.192)	(0.449)
Female (β_2)					0.234	0.600***
					(0.148)	(0.151)
MaleRatio*Female (β_3)					-0.534**	-0.470
					(0.249)	(0.482)
Observations	1,359	1,354	1,488	1,485	1,966	874
R-squared	0.034	0.666	0.013	0.647	0.685	0.547
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
Track FEs	X	X	X	X	X	X
Sample	Male	Male	Female	Female	Science	Liberal Arts
$H_0 : \beta_1 + \beta_3 = 0$	.	.	.	.	-0.376	-0.350
(p-value)	.	.	.	.	(0.057)	(0.250)

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. Individual controls include ethnicity, age, CYL member, class leader, and *hukou*. Test-room controls include Han composition, average age, CYL member composition, class leader composition, and urban *hukou* composition in the test room. Standard errors are clustered at the test-room level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

point rise in the male ratio), her test score would decrease by roughly 6.6 points in the science track (calculated as 0.167 * -0.376 * 105.8). This decrement is non-trivial and could result in one falling behind up to 5,500 other students among 210,000 students from the same province in the college admissions process.

4 Discussion and Conclusion

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, especially those pursuing the science track. Conversely, the gender composition does not appear to have a

significant impact on the performance of male students.

What are the channels that underlie our findings? One might be inclined to attribute our findings to explanations in the literature on gender differences in competition. Specifically, women may experience a tendency to underperform as the competitiveness level increases (Gneezy, Niederle and Rustichini, 2003; Ors, Palomino and Peyrache, 2013; Morin, 2015; Cai et al., 2019). However, this explanation is unlikely to be the case in our setting. The objective level of competitiveness in our sample is fixed since all students are taking the same test. In addition, 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.

Another potential explanation relates to Bursztyn, Fujiwara and Pallais (2017)'s work on "Acting Wife". Bursztyn, Fujiwara and Pallais (2017) 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.

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).[3] 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). 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. Furthermore, the activation of negative stereotypes can often be subtle and occur outside of people's conscious awareness.

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

Relatedly, 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 zero-stakes experimental setting.[4] Our study takes advantage of a real-world high-stakes setting with a continuous measure of gender composition. We find a higher presence of male students in the test room leads to a reduction in performance among female students pursuing the science track, which aligns with the stereotype threat explanation.

Female students in the science track are self-selected to pursue this academic track. Given this self-selection process, the presence of underperformance of this selective group of female students in the presence of a higher proportion of males suggests that susceptibility to male-dominated workplaces is likely to be prevalent and potentially even more pronounced in broader contexts. It is also important to note that, while the impact of gender composition on female students in the liberal arts track is not statistically significant, its magnitude is economically meaningful and comparable to the effect on female science students. In light of the fact that liberal arts students are also required to take the math exam, the stereotype threat explanation may account for part of the impact, but not the full impact since math-related scores make up a smaller portion of liberal arts students' total scores, unless math-related stereotypes take a heavier toll on female liberal arts students.

Due to the limitation of our available data, we cannot provide an informative heterogeneity analysis or conduct a survey to gain insights into students' beliefs. Nevertheless, the fact that we find a group of selective female students underperform in the presence of a higher ratio of males is a novel finding, and this phenomenon would be especially problematic for girls in regions or fields where males substantially outnumber females. Based on our findings (and perhaps with some extrapolation), the implementation of female-dominated or single-sex test rooms could serve as a low-cost strategy to mitigate the disadvantages faced by female students in assessing their academic ability in a high-stakes setting. Lastly, our findings raise some interesting questions about the effect and associated channels of socio-demographic composition in the working environment on performance generally, which deserve further investigation.

[4] Inzlicht and Ben-Zeev (2000)'s findings could potentially be subject to alternative explanations due to 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.

References

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham. 2020. "The light and the heat: Productivity co-benefits of energy-saving technology." *Review of Economics and Statistics*, 102(4): 779–792.

Bensnes, Simon Søbstad. 2016. "You sneeze, you lose:: The impact of pollen exposure on cognitive performance during high-stakes high school exams." *Journal of Health Economics*, 49: 1–13.

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.

Bond, Timothy N., Jillian B. Carr, Analisa Packham, and Jonathan Smith. 2022. "Hungry for Success? SNAP Timing, High-Stakes Exam Performance, and College Attendance." *American Economic Journal: Economic Policy*, 14(4): 51–79.

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.

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.

Chang, Eunsik, and Maria Padilla-Romo. 2023. "When Crime Comes to the Neighborhood: Short-Term Shocks to Student Cognition and Secondary Consequences." *Journal of Labor Economics*.

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.

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

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.

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.

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.

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.

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.

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

Park, R Jisung. 2022. "Hot Temperature and High-Stakes Performance." *Journal of Human Resources*, 57(2): 400–434.

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

Somanathan, Eswaran, Rohini Somanathan, Anant Sudarshan, and Meenu Tewari. 2021. "The impact of temperature on productivity and labor supply: Evidence from Indian manufacturing." *Journal of Political Economy*, 129(6): 1797–1827.

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.

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

Wang, Ao, Shaoda Wang, and Xiaoyang Ye. 2023. "When Information Conflicts with Obligations: the Role of Motivated Cognition." *The Economic Journal*, 133(654): 2533–2552.

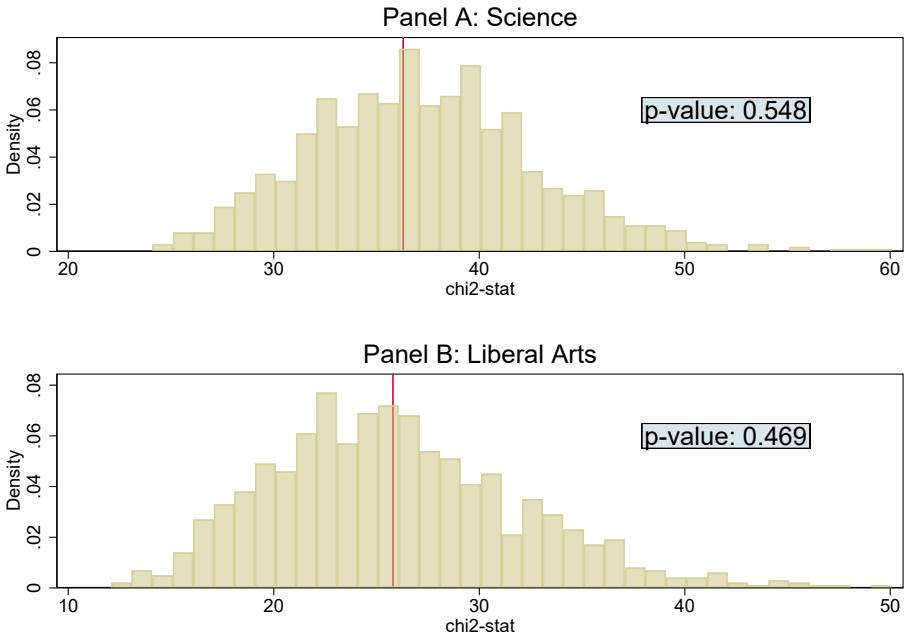
Whitmore, Diane. 2005. "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment." *American Economic Review*, 95(2): 199–203.

Zivin, Joshua Graff, and Matthew Neidell. 2012. "The impact of pollution on worker productivity." *American Economic Review*, 102(7): 3652–3673.

Zivin, Joshua Graff, 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.

Appendix

Figure A1: Histogram of χ^2 Statistic



Notes: Panel A depicts the histogram of χ^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 χ^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.

Table A1: Gender Composition and Other Test Room Characteristics

VARIABLES	(1) MaleRatio
HanRatio	-0.108
	(0.689)
Average Age	0.0415
	(0.0648)
CYLRatio	-0.350
	(0.289)
ClassLeaderRatio	-0.210**
	(0.0983)
UrbanRatio	-0.158
	(0.136)
Observations	127
R-squared	0.648
Track FEs	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 who have urban *hukou* in the test room. Robust standard errors are reported. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.