Gender Gap under Pressure: Evidence from China's National College Entrance Examination *

Xiqian Cai† Yi Lu‡ Jessica Pan§ Songfa Zhong¶

October, 2016

Abstract

This paper examines how female and male examination performance are differentially affected by the degree of competitive pressure faced. Our setting is China's National College Entrance Exam (*Gaokao*) which is widely regarded as the world's most competitive exam. We show that compared to male students, females underperform on the high-stakes *Gaokao*, relative to their performance on the low-stakes mock examination held two months earlier. The gender gap in exam scores is 0.15 standard deviations larger in the *Gaokao* relative to the mock exam. This translates to a 15% decline in the probability that females qualify for a Tier 1 university when moving from a low-stakes setting to a high-stakes setting. To elaborate on the possible mechanisms, we conduct further analyses. First, we observe that the gender gap is similarly observed for mock exams conducted at different time points and self-reported effort does not differ significantly between genders in the months leading up to the *Gaokao*. These suggest that the larger gender gap in *Gaokao* is unlikely due to gender difference in effort provision. Second, for subgroups of students where the stakes matter more, the performance gaps are larger, and we observe a decline in performance among females, coupled with an improvement in performance among males. Third, we find that, compared to males, females perform worse on the afternoon exam in response to negative performance shocks on the morning exam. Overall, our study suggests that gender differences in the response to stress underpins gender differences in performance in high pressure settings.

*We thank seminar participants at the University of Maryland, McMaster University, National University of Singapore, Northwestern, Tsinghua University, Xiamen University, Trans-Pacific Labor Seminar and the SOLE/EALE meetings for helpful comments.

†Xiamen University, Wang Yanan Institute for Studies in Economics. Email: caixiqian@gmail.com

‡National University of Singapore. Email: ecsluyi@nus.edu.sg

§National University of Singapore. Email: jesspan@nus.edu.sg

¶National University of Singapore. Email: ecszs@nus.edu.sg

1 Introduction

A large number of experimental studies suggest that men and women respond differently to competitive pressures. These studies document that women appear to systematically underperform relative to men in competitive settings and that women may simply prefer to opt out of competitions (for examples, see Bertrand, 2010, Gneezy, Niederle and Rustichini, 2003, Gneezy and Rustichini, 2004, Niederle and Vesterlund, 2007, 2011). These studies posit that gender differences in performance and attitudes toward competition may explain an important part of the gender gap in educational choices and labor market outcomes (Buser, Niederle and Oosterbeek, 2014).

A growing line of research has attempted to assess whether such performance differences in response to competition exist in real-world settings. Interestingly, the results are somewhat mixed. Earlier studies by Lavy (2013) and Paserman (2010) examine gender differences in the performance of high school teachers and professional tennis players, respectively, and find little evidence that women perform worse in more competitive settings. More recently, a number of studies focusing on real-world academic settings show that men appear to outperform women when competitive pressures are higher, whereas the reverse holds true in less competitive settings (Azmat, Calsamiglia and Iriberri, 2014, Morin, 2013, Ors, Palomino and Peyrache, 2013, Attali, Neeman and Schlosser, 2011, and Jurajda and Munich, 2011).

In this paper, we utilize a unique dataset from China's National College Entrance Examination (*Gaokao*) and an arguably cleaner empirical setup to examine the extent and mechanisms through which competitive pressure affects the gender gap in academic performance. The *Gaokao* is widely regarded as one of the most competitive examinations in the world - it is practically the only route to admission into universities of higher education and further success in the test-oriented education system of China. Furthermore, the number of exam takers typically exceeds the available places for higher education. The admission rate for candidates sitting for the *Gaokao* is approximately 75% and students' performance on the two-day examination is typically the sole criteria used to determine their placement into one of China's nearly two thousand colleges. Each college has separate cut-offs that determine whether students can qualify for various academic programs and entry into a particular college and major is determined almost exclusively by students' performance on the *Gaokao*. In fact, the examination is so important that a couple of months prior to the actual examination, each province typically runs a mock examination. Importantly, to ensure that the examiners are familiar with the examination protocol and to allow students to gauge their preparedness and relative performance on the exam, the mock examination is administered by the province with the same duration as *Gaokao* and is also marked anonymously by the province.[1]

[1] This is suggested that the observed difference between Gaokao and mock examination is unlikely due to teachers' gender biases in marking exams (Cornwell et al, 2013; Burgess and Greaves, 2013; and Lavy 2008).

Drawing on a dataset that comprises the universe of *Gaokao* takers in Anxi County in 2008, we are able to directly observe the performance of the same individual in the *Gaokao* and the earlier mock examination. By comparing the gender difference in performance in what is essentially the same examination in a high and low-stakes setting, we are able to cleanly estimate of the effect of competitive pressures on the gender gap in performance in an important real world setting.[2] We find strong evidence that females tend to do relatively worse as compared to males on the high-stakes *Gaokao*, relative to their performance on the low-stakes mock examination. While females had a large and statistically 15 point advantage (out of a total exam score of 750) in total test scores in the mock examination, the female advantage declined to a mere 2 point advantage in the *Gaokao*. These differences translate to a 0.15 standard deviation decline in the difference in test scores between the *Gaokao* and mock examination for females relative to males.

While this is consistent with the earlier (mostly experimental) literature on that examines the gender gap under competitive and stressful environments, an alternative interpretation is that male students take the mock examination less seriously as it is a low-stakes test, but increase their preparation effort for the actual examination.[3] This behavioral difference could explain the observed gender gap in performance, and would not rely on the idea that women's performance suffers in a more competitive environment. It is worth pointing out that the mock examination is the only time that students can get a sense of their ranking in the province. The Department of Education reveals the full distribution of students' mock exam scores in the province. Given that university slots are allocated at the province level. One's standing in the overall test score distribution reveals important information about the type of college that students will likely qualify for, thus providing a strong incentive for students to take the mock examination seriously. To further examine the effort hypothesis, we use two newly-collected data sets based on a subsample of recent exam-takers. In the first dataset, we have information on student performance from two additional city-level mock examinations in May 2014 between the provincial mock examination in April 2014 and the *Gaokao* in June 2014 for the sample of exam-takers from the largest high-school in Anxi. We find that the gender gap in performance is similar among all the mock examinations. Second, we conduct a survey on students' exam preparation efforts, and do not find any statistically and economically significant gender differences in the rating of their *Gaokao* preparation such as hours of study in the six months preceding the *Gaokao*. Overall, these results do not support the hypothesis that there is gender difference in effort provision between the mock examination and *Gaokao*.

[2] In our setting, relative to the mock examination, the *Gaokao* is both higher-stakes and more competitive (students compete for a limited number of slots to qualify for their academic program of choice). We do not distinguish the role of exam stakes and the degree of competition and describe the *Gaokao* as "higher stakes", "more competitive" and entailing a greater degree of "competitive pressure" interchangeably.

[3] For example, Attali, Neeman and Schlosser (2011) find that men and whites tend to exert lower level of effort in a low-stakes GRE examination, and that this partially explains the larger performance differential across the low vs. high-stakes test for males and whites relative to females and other demographic groups.

We further examine the pressure explanation by utilizing two additional sources of variation in competitive pressure faced by students. First, we exploit the fact that the mock examination scores are used to calculate reference entry cutoffs for each of the four different tiers of universities to generate additional variation in the pressure involved in the examinations. Presumably, students who are closer to the entry thresholds for each university tier are more likely to feel greater pressure during the *Gaokao* relative to students who are further away from the entry thresholds. Therefore, our story would predict that the gender performance gap is likely to be accentuated among students who are close to the entry thresholds. Moreover, finding differential gender gaps in performance as a function of students' performance on the mock examination would also alleviate concerns that certain subgroups of students are simply not taking the mock examination seriously as this interpretation would imply that predicted entry cutoffs would have little bearing on the gap in students' relative performance. Consistent with the idea that females tend to underperform in stressful settings, we find striking evidence that the gender gap in performance is larger for students who are close to the university entry thresholds - for the group of students within three points of the entry cutoffs, females experience a 0.32 standard deviation larger decline in performance on the *Gaokao* relative to the mock examination compared to male students, whereas this difference is 0.14 and 0.1 standard deviations for the groups of students 6 to 10 points and 11 to 20 points away from the cutoff, respectively. Looking *within* gender, we provide additional evidence that these results are not entirely due to the increased effort of males, suggesting that the gender differential in performance on the *Gaokao* relative to the mock examination is likely to be driven, in part, by female underperformance in high-stakes settings.

Second, we exploit the fact that the individual subject components of the examinations are spread out over two days to explore whether there is a gender difference in the reaction to performance shocks in an earlier exam. The *Gaokao* is typically held over two consecutive days with the Chinese exam held on the morning of the first day, the Mathematics exam on the afternoon of the first day, followed by the combined subjects and English on the morning and afternoon of the second day, respectively.[4] We hypothesize that, relative to the others, students that experience a negative performance shock in the morning examination would be more likely to face greater pressure in the afternoon examination. To examine gender differences in the responsiveness to performance shocks, we examine how a student's relative performance in the morning examination of the *Gaokao* affects his/her relative performance in the afternoon examination and how this varies by gender. Relative performance is defined as the deviation of a student's performance in the *Gaokao* relative to his/her performance on the mock examination.

We find that a one standard deviation lower relative performance in the morning exam is associated with 0.11 standard deviations lower relative performance in the afternoon examination for males.

[4] In some provinces, the *Gaokao* is held over 3 days. Our sample is from Anxi county in Fujian province where the examination is held over two days.

Interestingly, this effect is significantly larger for females - a one standard deviation lower relative performance on the morning exam lowers their relative performance in the afternoon examination by 0.17 standard deviations. Moreover, we find that relative to males, females appear to be more responsive to negative shocks as compared to positive shocks. The gender differences in responsiveness to negative shocks are also most pronounced for students closest to the reference cutoffs. As the effort exerted in the mock examination and in the preparation during the last two months are completed before (and likely remains unchanged) during the two-day *Gaokao* period, the observed gender gap in the response to performance shocks is unlikely to be explained by gender difference in effort provision. Overall, these results suggest that female performance appears to be more detrimentally affected by negative shocks relative to males in high pressure settings, which could partially contribute to the observed gender gap in performance.

Our paper is closely related to a recent literature that examines whether men and women respond differently to competitive pressures and test stakes in real-world settings. Like our paper, most of these papers focus on academic performance and exploit differences in the nature of the test or test setting - for example, how competitive the tests are, the level of stakes involved and the grading scheme used, to examine whether the relative performance of females is affected by the nature of the test. For example, Ors, Palomino and Peyrache (2013) show that females outperform males in a less competitive national exam, but for the same cohort of students, males outperform females in the highly selective competitive entrance exam for admission to a top business school in France. Morin (2013) exploits a legislative change in Ontario that exogenously increased competition for university grades and documents that among students affected by the change, male performance improved relative to females. Azmat, Calsamiglia and Iriberri (2014) utilize variation in test stakes across different exams that Spanish high school students are required to sit for throughout the year and show that males tend to outperform females when test stakes are higher.

Our setting has a number of features that differentiate it from previous studies. First, the same examination board sets and implements both the high and low-stakes exams and the coverage of the test material is identical in both settings.[5] Second, as the *Gaokao* is the main requirement for admission into all colleges in China, there is limited sample selection of individuals into the actual high-stakes test based on their performance on the low-stakes test. This also ameliorates potential sample selection concerns that individuals with a greater distaste for competition may choose not to participate in the high-stakes examination. Third, we are able to exploit two different sources of variation in perceived pressure to shed some light on the mechanisms that lead to female underperformance in more competitive settings.

[5] In previous studies (e.g. Ors, Palomino and Peyrache, 2013, Azmat, Calsamiglia and Iriberri, 2014), the low-stakes and high-stakes settings considered typically involve different testing strategies, material and timing, which might conflate gender differences in the response to high vs. low-stakes with gender differences in the skills required in the high vs. low-stakes settings.

The rest of the paper proceeds as follows. The next section describes the institutional background of the *Gaokao* in China. Section 3 outlines the data used and the descriptive statistics. Section 4 reports the results on the gender gap in performance on the *Gaokao* relative to the mock examination and tests of the underlying mechanism. Section 5 concludes.

2 Institutional Background

The National College Entrance Examination (NCEE), commonly known as *Gaokao*, is an annual two or three day examination that is a pre-requisite for entrance into almost all institutions of higher education at the undergraduate level in China.[6] There are different tiers of universities in China, namely key universities (Tier 1), regular universities (sometimes further subdivided into two different tiers - Tier 2 and Tier 3), and technical colleges, and the differences among them are mostly based on ranking of the institutions and the duration of the programs (Davey, De Lian and Higgins, 2007).

The *Gaokao* is ultimately under the control of the Ministry of Education (MoE) and was once administered uniformly across the country. Starting in 2001, some provinces or direct-controlled municipalities arranged separate exam papers while others still adopted the national exam papers. The most commonly adopted examination system across the provinces is the "3+X" system - "3" represents the three compulsory subjects: Chinese, Mathematics and English (each accounting for 150/750 of the total score) and "X" represents the combined science subjects comprising physics, chemistry and biology for students on the science track, or combined arts subjects of history, geography and politics for students on the arts and social sciences track (accounting for 300/750 of the total score).[7] All students, regardless of stream, sit for the same Chinese and English exam. The coverage of the Math exam is different for students in the science vs. arts stream. The "3+X" system is typically held over two days in June in the following order: Chinese (Day 1, morning), Mathematics (Day 1, afternoon), Combined subjects (Day 2, morning) and English (Day 2, afternoon).

Two months prior to the *Gaokao*, a formal mock examination, administered by the province, is usually held to allow students to get a sense of the examination and their relative standing within the province. The mock examination results are released about one week after students sit for the exam and the Department of Education (DoE) in each province also releases the province-level distribution of test scores as well as a set of reference cutoffs for each of the four university tiers

[6] For example, in 2006, 9.5 million people applied for tertiary education entry in China, of which 93% were scheduled to take the national entrance exam. The remaining applicants were either exempted from the standardized exams (0.3%) or scheduled to take other types of standardized exams.

[7] Students choose to be in the science stream or arts and social sciences stream at the beginning of the second year of high school.

based on the proportion of students who were admitted into each university type in the previous year. Students usually sit for the examination in their last year of senior high school, although there has been no age restriction since 2001. In different provinces, students either apply for universities prior to the *Gaokao*, after the *Gaokao*, or after they have learnt about their estimated scores based on the mock *Gaokao* examination and their estimated rank in the province.

The *Gaokao* is highly competitive. It is commonly described as the "world's toughest exam" due to the intense pressure and competition that students are subject to. The *Gaokao* is virtually the only path for Chinese students to be admitted into universities and the number of exam takers typically exceed the available places for higher education. For example, in 2014, there were 9.39 million test takers vying for about 7 million college spots.[8] Furthermore, the 2000 or so universities in China are classified into four different tiers, with cutoff points to determine whether students can qualify for each tier of universities. Within each tier, each college also has a separate minimum exam score required for admission. It is estimated that less than 10% of candidates enroll into the top tier universities (key universities) and only less than 0.2% of exam takers will gain admittance into China's top five universities (Economist, 2012).[9] It is a national consensus that getting into a better university via the *Gaokao* greatly enhances an individual's chances to obtain a better job in China's fiercely competitive job market.

Due to its importance and competitiveness, the *Gaokao* imposes enormous pressure on test takers, as well as their parents and teachers. It is very common for students to spend hours studying for the *Gaokao* after returning home from ten hours of schooling, with little or no break on the weekends. Many schools dedicate the entire senior year of high school to preparing students for the exam. It is common to see astonishing new reports related to the *Gaokao* in the local and international media. For example, it was reported that some girls took contraceptives or received injections to prevent the onset of their menstrual cycle during the week of the exam.[10]

3 Data and Descriptive Statistics

Our data consists of test scores and demographic information for the universe of *Gaokao* test takers as well as the test scores for all first-time test takers who sat for the mock examination in Anxi, a county of Fujian Province, in 2008.[11] Appendix I provides some background information on the social and economic characteristics of Anxi county and Fujian province. The 2008 Provincial Mock

[8] See: http://www.businessweek.com/articles/2014-06-06/china-girds-for-high-stress-gaokao-weekend

[9] See: http://www.economist.com/blogs/analects/2012/06/university-entrance-exams

[10] http://www.nytimes.com/2009/06/13/world/asia/13exam.html

[11] We are only able to utilize data from a single year as prior to 2008, the mock exam data was unavailable. Post-2008, the dissemination of exam results was centralized at the province level, thus, we did not have access to the *Gaokao* data. 2008 was the only year that we were able to merge the individual-level mock exam data to the *Gaokao* data.

Examination of Fujian was held in mid-April and the *Gaokao* was held in mid-June. The province administered both the mock examination as well as the *Gaokao*, therefore, the mock examination is a good indication of the degree of difficulty and subject material covered by the actual examination. Furthermore, the DoE in Fujian uses the mock examination to determine the reference cutoff scores for each of the different tiers of universities based on the proportion of students eligible for each tier in the previous year.[12] Although *Gaokao* takers in Fujian typically apply to universities after they learn their actual scores and the actual cutoff points, the mock examination is taken seriously by students as a way to estimate their relative rank in the province and to ascertain the likely tier of university that they will qualify for.[13]

Our dataset was constructed by merging the mock examination scores to the *Gaokao* scores using each test-taker's name. Individuals with the same first and last name were dropped as they could not be uniquely identified. From the *Gaokao* sample, we further dropped a small number of individuals for whom we were not able to identify whether they were in the arts or science stream. Our final merged sample comprises 7,961 individuals - which is 98% of the universe of mock examination candidates and 94% of *Gaokao* candidates.[14] The summary statistics for the mock examination sample, the *Gaokao* sample and the merged sample are reported in Table 1. On average, the profile of students in the merged sample is very similar to that in the mock examination.[15] Candidates who sat for both examinations (in the merged sample) were of slightly higher ability than the overall *Gaokao* sample. Nevertheless, the qualitative differences in the actual test scores for the two samples are quite small, ranging from 0.4 to 1.3 points out of a 150 or 300 point test. There is also little observed difference in the demographic characteristics across the *Gaokao* and the merged sample. Overall, these results suggest that there is very little selection into the final sample based on an individual's performance on the mock examination and that the merged sample is broadly representative of the universe of first-time *Gaokao* test-takers in Anxi county.

4 Gender Gap in Performance on Mock Exam vs. *Gaokao*

Before turning to the formal econometric analysis, we present some suggestive graphical and descriptive evidence of the gender gap in performance on the high-stakes *Gaokao* vs. low-stakes mock examination. The top two panels of Figure 1 shows the distribution of total scores separately by

[12] Appendix Table A1 lists the score cutoffs in the 2007 and 2008 *Gaokao* in Fujian province as well as the proportion of students that meet the cutoffs for admission into each university tier.

[13] Our discussion in Section 4.2 highlights that the reference cutoffs are indeed informative about the fraction of students who are eventually eligible for each university tier based on the actual *Gaokao* scores.

[14] There were approximately 250 candidates who sat for the *Gaokao* and not the mock examination. Most of these students were either retaking the *Gaokao* or were private candidates (i.e. these students were not affiliated with a high school when they took the exam).

[15] Appendix I also includes a profile of candidates in Anxi county, Fujian province and China as a whole.

gender for the April mock examination and the June *Gaokao*. As observed in the figure, while the female test-score distribution is to the right of the male distribution in the mock examination, for the *Gaokao*, the male distribution appears to converge to that of the female distribution. Since we are presenting non-standardized scores in this section, it is important to distinguish between students in the Science and Arts stream as one of the two exam components, namely, math and the combined science/arts subject, differ across the two groups. The middle and bottom panels of Figure 1 graph the distributions separately for students in the Science stream and Arts stream. Among Science students, males appear to outperform females at almost all points of the test score distribution, and the male advantage becomes even more pronounced during the *Gaokao*. In contrast, among Arts students, we observe a strong female advantage in the mock examination at all points of the distribution. This advantage appears to be reduced in the *Gaokao*, particularly among students in the middle to upper-tail of the test score distribution.

Table 2 summarizes the means of the test-score distributions in Figure 1. Columns (1) to (3) report the mean scores for females, males and the gender gap (female-male), respectively for the mock examination. Columns (4) to (6) report similar statistics for the *Gaokao*. Column (7) reports the difference between the gender gaps reported in Columns (3) and (6). In the first three rows, we report the means for the total exam scores for all students as well as for students in the Science and Arts stream separately. On average, female and male students in the Science stream improve their scores on the *Gaokao* relative to the mock exam. Interestingly, the gender gap is narrower in the mock exam as compared to the *Gaokao* for science students. Male science students perform about 3 points better on the mock exam relative to females; in the *Gaokao*, the male test score advantage increases about three times to 9 points. For students in the Arts stream, while females test scores declined between the mock exam and *Gaokao*, male scores actually increased slightly. The gender gap in performance is also reduced significantly in the high-stakes setting for students in the Arts stream - the gender gap of 23 points in favor of females on the mock exam nearly halves to about 13 points on the *Gaokao*. In sum, the raw data shows strong evidence that the gender gap in performance is larger in high-stakes settings relative to low-stakes settings.

4.1 Empirical Strategy

Next, we turn to a formal empirical framework to more rigorously establish how gender impacts students' relative performance in high vs. low-stakes settings. Assume that an individual's performance on each subject test on the *Gaokao* is represented by the following equation:

$$P_{i,g}^{E,S} = \alpha_i^S + w_g^S + Z^{E,S} + Y_i^{E,S} + x_g^{E,S} + \epsilon_{i,g}^{E,S}$$

where E denotes the *Gaokao* Entrance Exam (M denotes the mock examination in the next equation), S denotes the type of student (science vs. arts stream), i denotes individuals, g denotes gender. α_i^S is the set of stream-specific individual characteristics[16] that do not change over the two exams, w_g^S represent gender-specific characteristics that do not change over the two tests, $Z^{E,S}$ represent the *Gaokao*-specific characteristics (such as the location, temperature) that do not vary across gender, $Y_i^{E,S}$ is the set of individual characteristics that may affect the two exams differently, $x_g^{E,S}$ captures the *Gaokao* factors that vary by gender and $\epsilon_{i,g}^{E,S}$ is the error term.

Correspondingly, an individual's performance on the mock examination is given by:

$$P_{i,g}^{M,S} = \alpha_i^S + w_g^S + Z^{M,S} + Y_i^{M,S} + x_g^{M,S} + \epsilon_{i,g}^{M,S}$$

For both equations, to remove the effect of the Z's or the exam-specific characteristics that are common to all individuals, we consider standardized test scores as the outcomes, that is, $\tilde{P}_{i,g}^{E,S} = \frac{P_{i,g}^{E,S} - \bar{P}_{i,g}^{E,S}}{\sigma_{P_{i,g}^{E,S}}}$ and $\tilde{P}_{i,g}^{M,S} = \frac{P_{i,g}^{M,S} - \bar{P}_{i,g}^{M,S}}{\sigma_{P_{i,g}^{M,S}}}$.

Taking the difference of the two resulting equations, we obtain:

$$\tilde{P}_{i,g}^{E,S} - \tilde{P}_{i,g}^{M,S} = (\tilde{x}_g^{E,S} - \tilde{x}_g^{M,S}) + (\tilde{Y}_i^{E,S} - \tilde{Y}_i^{M,S}) + (\tilde{\epsilon}_{i,g}^{E,S} - \tilde{\epsilon}_{i,g}^{M,S})$$

Notice that this first difference specification allows us to difference out the individual fixed effects that affect an individual's performance in both the *Gaokao* and mock examination (the α_i's) as well as the gender-specific characteristics that do not change across the two tests (the w_g's).

Empirically, we can directly measure $\tilde{P}_{i,g}^{E,S}$ and $\tilde{P}_{i,g}^{M,S}$ using our data on test scores. We use the female dummy to capture $(\tilde{x}_g^{E,S} - \tilde{x}_g^{M,S})$, and include school dummies, zip code dummies and student age to control for $(\tilde{Y}_i^{E,S} - \tilde{Y}_i^{M,S})$. The regression specification is given by:

$$\tilde{P}_{i,g}^{E,S} - \tilde{P}_{i,g}^{M,S} = \beta_0 + \beta_1 Female_i + Y_i\gamma + \epsilon_{i,g} \tag{1}$$

As the outcomes have been standardized using the type-specific (arts vs. science) mean and variance, both $\tilde{P}_{i,g}^{E,S}$ and $\tilde{P}_{i,g}^{M,S}$ have a mean of 0 and a variance of 1. For ease of interpretation, we further standardize the difference between $(\tilde{P}_{i,g}^{E,S} - \tilde{P}_{i,g}^{M,S})$ to have a mean of 0 and a variance of 1 so that the coefficient β_1 can be interpreted as the effect of being female on the difference in test scores between the *Gaokao* and mock examination in standard deviations.[17]

[16] This also captures individual selection into the arts/science stream as the selection into streams is individual-specific and does not change across the two exams. In our context, the choice of stream is chosen before students sit for either the mock exam or the *Gaokao*.

[17] The actual standard deviations for the difference in mock exam and *Gaokao* test scores ($\tilde{P}_{i,g}^{E,S}$ and $\tilde{P}_{i,g}^{M,S}$) are: Total (0.54), Chinese (0.91), Math (0.66), Combined Arts/Science (0.72), English (0.66).

4.2 Results

Table 3 reports the female coefficient estimate, β_1, from the estimation of equation (1) for the total score as well as for each of the individual subjects - Chinese, Mathematics, combined subjects and English. Panel (A) includes the full sample of students, while Panels (B) and (C) restrict the sample to students in the Science stream and Arts stream, respectively. Column (1) presents the raw gender difference in total test scores across the high-stakes *Gaokao* and low-stakes mock examination - on average, the difference in score between the *Gaokao* and mock exam among females is 0.16 standard deviations lower than that for males in the full sample. The corresponding estimates for students in the Science stream and Arts stream are -0.15 and -0.20, respectively. These differences are large and statistically significant and are virtually unaffected by the addition of controls for age, school fixed effects and zipcode fixed effects (Column (2)). Columns (3) to (10) report the estimates separately by subject. The results indicate that most of the effect is driven by a significantly worse relative performance by females on the combined subject test, as well as the English test. The distribution of raw standardized differences across the two tests by gender are shown in Appendix Figure 1. These results are consistent with the idea that females underperform relative to males on high-stakes vs. low-stakes settings.

To provide a sense of the magnitude of our estimates, Appendix Table A2 reports the gender difference in the likelihood of qualifying for a Tier 1 or Tier 2 university based on the reference cutoffs in the mock exam (Column (4)) and the actual cutoffs in the *Gaokao* (Column (8)). Based on their performance on the mock exam, females are 1.1 percentage points less likely to be eligible for a Tier 1 university. On the *Gaokao*, the gender difference nearly doubles to 2.1 percentage points. Columns (9) and (10) report the difference in the gender gap in Tier 1 eligibility across the *Gaokao* and mock exam with and without individual-level controls.[18] We find that females are significantly (0.8 percentage points) less likely than males to be eligible for a Tier 1 university based on their *Gaokao* scores relative to their performance on the mock exam. Given that, on average, 5.5% of *Gaokao* takers are eligible for Tier 1 universities, this works out to be a relative decline of about 15% (0.8/5.5=0.15). The second and third rows of Appendix Table A2 examine the gender gap in the likelihood of qualifying for Tier 2 universities and either Tier 1 or Tier 2 universities, respectively. The corresponding difference in the gender gap in high vs. low-stakes setting is 7% (1.7 percentage points) for Tier 2 eligibility and 8% (2.5 percentage points) for eligibility in either Tier 1 or Tier 2 universities.

Two potential mechanisms are consistent with the observed gender gap in performance on high-stakes exams. The first possibility is that males and females respond differently to competitive and stressful environments such as the *Gaokao*. This explanation is line with a large number of biological studies on gender differences in stress responses (Taylor et al., 2000, Lee and Harley, 2012).

[18] The set of controls are the same as those used in Table 3.

In addition, according to the "stress and gender" survey conducted by the American Psychological Society (2010), given similar self-reported stress levels, women are more likely than men to report that their stress levels are on the rise, and to report more physical and emotional symptoms of stress. This suggests that there could be a substantial gender difference when it comes to test anxiety and stress (Kirschbaum, Wust and Hellhammer, 1992).

The second possibility is that males could also have greater scope to differentially increase their preparation for the *Gaokao* in the two months between the mock examination and the *Gaokao*. Such behavior could generate the patterns of relative female underperformance in high-stakes settings that we observe in the data, for reasons that are potentially unrelated to female performance under pressure. Nevertheless, it is worth noting that while the mock exam scores have no bearing on the scores used by students to apply for university, the mock exam is the only time before the *Gaokao* that students participate in a province-level examination. Therefore, this is the only opportunity that students can get a real sense of their relative academic standing within the province and the types of universities that they are likely to qualify for. This is important as what ultimately matters for admission into different universities is a student's performance relative to his/her peers at the province-level.[19] As such, students have an incentive to put in their full effort on the mock exam.

In the next subsections, we conduct various analyses to evaluate these two potential mechanisms. First, we directly examine the preparation effort students put into preparing for the *Gaokao* during the intervening months from two newly-collected datasets. Second, we also provide some further empirical tests of our preferred explanation that female and male students react differently in response to competitive pressure.

4.3 Effort Provision in *Gaokao* Preparation

To examine the possibility that our observed empirical patterns are driven by gender differences in effort provision, we conduct two exercises using two newly-collected data sets. First, we obtain data on individual-level scores for several mock exams and the *Gaokao* from the largest high school in Anxi in 2014. Specifically, similar to that used in our aforementioned analyses, this dataset contains test score information from the mock exam administrated by the provincial Department of Education (on April 2014) and the *Gaokao* (on June 7-8, 2014). In addition, the dataset contains information on two additional mock exams that students took, one administered by the Xiamen City Department of Education in mid-May 2014, and the other administered by the Quzhou City Department of Education in late-May 2014. Apart from being administered by different layers of the government, the two additional mock exams are of similar nature to the provincial exam used in

[19] Prior to the provincial mock exam, students have the opportunity to take many practice exams, but these are typically at the school level.

the previous analysis.[20] This dataset not only allows us to verify whether our previous results about gender gap in performance on high-stakes exams can be replicated in another year and sample of students, but also provides us with the opportunity to examine whether there are changes in the gender gap in performance between the Gaokao and the April mock exam. More specifically, while the mock exams at different time points do not reflect changes in competitive pressures faced by students, they potentially capture changes in the preparation effort for the *Gaokao*.

The estimation results are reported in Table 4. Columns 1 and 2 report the difference between the April mock exam and the *Gaokao* using the same specification as our previous analysis. We replicate the negative and statistically significant effects, further confirming our findings in Table 3. Columns 3 and 4 and Columns 5 and 6 examine the difference between the April and mid-May mock exams, and the difference between the April and late-May mock exams, respectively. We find that all the estimates are statistically insignificant and the magnitude is close to zero. These results indicate that males do not appear to systematically improve in terms of their performance on the later mock exams, suggesting that the gender difference in performance in the high-stakes *Gaokao* relative to the low-stakes April mock exam documented in Table 3 and Columns (1) and (2) of Table 4 are unlikely to be driven by male students increasing their preparation efforts between the April mock exam and the actual Gaokao exam in June.

In the second exercise, we conducted a survey on students' preparation effort for the *Gaokao*. We surveyed students in a large middle school in Anxi, which has approximately 3,800 students and 200 full-time teachers in 2016. We surveyed graduating students in mid-May 2016, three weeks before the *Gaokao*. There were 323 students sitting for the *Gaokao* exam in June 2016, and we were able to obtain valid responses from 311 students.[21] The full questionnaire can be found in Appendix II.

Specifically, we asked students "How many hours, on average, did you spend studying each day?" for each month from December 2015 to May 2016. This question was also asked for each subject separately. The mean values for males, females and their differences are reported in Table 5A. As observed from the table, we do not find any statistically and economically significant gender differences in the reported time spent on exam preparation in the months leading up to the *Gaokao*. In addition, we asked students to rank "On a scale of 1 to 10, how prepared were you for the *Gaokao* exam? (1: unprepared, 10: very prepared)." Table 5B shows the mean values for males, females

[20] Another difference between the two additional mock exams conducted at the city level in May and the provincial level mock exam in April is that for the city level mock exams, reference entry cut-offs are not provided when the results are released (since the entry cutoff is determined by students' performance across the whole province).

[21] The high response rate was possible as teachers assisted in helping us to conduct the survey – students were required to independently finish the survey and the paper-based surveys were collected immediately upon completion by the teachers. The surveys were anonymous and students were informed that it was part of a research project on *Gaokao* preparation conducted by researchers from Xiamen University (which is regarded as the most prestigious university in the province).

and their differences. Similarly, we do not find any statistically and economically significant gender differences in the rating of their *Gaokao* preparation in the six months preceding the *Gaokao*. In Table 5C, we examine students' evaluation of the effectiveness of their study efforts on a scale of 1 to 10 with 1 for very ineffective and 10 for very effective. We continue to find no statistically and economically significant gender differences from December 2015 to May 2016. In Appendix Table A3, we further check whether there are any gender differences in sleeping hours, sleeping quality, sick days, and stress status in the six months preceding the *Gaokao*. Although there are some significant gender differences, the magnitude of the differences are small relative to the sample means.

Taken together, the results from both these exercises suggest that males do not appear to be differentially increasing their effort provision in the months leading up to the *Gaokao*. Therefore, it does not seem to be the case that our key findings that females underperform relative to males on the high-stakes Gaokao relative to the low-stakes mock exam reflects males' greater scope to differentially increase their preparation for the *Gaokao* in the intervening time between the mock exam and the *Gaokao*.

4.4 Gender Gap for Students Closer to Reference Thresholds

Next, we examine the hypothesis that males and females respond differentially to competitive pressure in greater detail. First, we exploit the fact that the mock examination scores are used to calculate reference entry cutoffs for different university tiers to generate additional variation in the pressure involved in the examinations. In particular, when the Department of Education in Fujian province releases the mock examination scores after the exam, they provide a list of the province-level test score distribution (in 10 point bins) as well as a set of reference entry cutoffs for entry into each of the four university tiers that are calculated based on the proportion of students eligible for each tier in the previous year. Table 6A lists the reference cutoffs for each stream in 2008. Table 6B and 6C report the fraction of students scoring above each of the reference cutoffs in Fujian province and Anxi county, respectively. While fewer students in Anxi are projected to qualify for Tier 1 universities relative to province-wide statistics, the fraction of students in Anxi who score above the reference cutoffs for Tier 1 and Tier 2 universities is similar to the fraction for the province as a whole. Appendix Table A1 lists the cutoffs for the *Gaokao* in 2007 and 2008 and the fraction of students in Fujian scoring above each of the reported cutoffs. Importantly, the fraction of students projected to be eligible for each of the different university tiers based on the reference cutoffs in the mock exam appears to be very similar to the fraction of students who were eligible for each tier based on the actual cutoffs in the 2007 and 2008 *Gaokao*.[22] This suggests that

[22]Note that the slight discrepancy in the fraction of students eligible for each Tier based on the reference cutoffs in the 2008 mock exam compared to the 2007 *Gaokao* cutoffs (see Appendix Table A1D) is likely to be due to the

the reference cutoffs are indeed informative about the types of universities that students are likely to qualify for.

We focus on the reference cutoffs for the top two university tiers (Tier 1 and Tier 2) as eligibility for these tiers is more selective - only about 22-28% of test-takers in Anxi or Fujian are projected to be eligible for entry into Tier 1 and Tier 2 universities.[23] As such, students who are close to the reference entry thresholds for Tier 1 and Tier 2 universities are more likely to face higher pressure on the *Gaokao* relative to students who are further away from these entry thresholds. This arises because students who are close to the entry thresholds are more likely to experience a larger change in the set of universities (in terms of quality and quantity) that they can apply to resulting from a change in relative performance as compared to students who are further away from the thresholds. This would predict that the gender gap in performance is likely to be larger among students who are close to the entry thresholds relative to students who are further from these thresholds. Moreover, finding differential gender gaps in performance as a function of students' performance on the mock examination would also alleviate concerns that certain subgroups of students are simply not taking the mock examination seriously as this interpretation would imply that the reference entry cutoffs should have little bearing on the gap in students' relative performance.

Table 7 reports the estimates of equation (1) for four different subgroups of students - those with mock examination scores within 3 points, 5 points, 6 to 10 points and 11 to 20 points of the reference cutoffs required for entrance into the top two university tiers. In Panel (A), we find that consistent with the idea that female students perform more poorly on the *Gaokao* relative to the mock examination when they face greater pressure, the gender gap in relative performance is largest among students within 3 points of the cutoffs and declines monotonically for students further away from the threshold. More specifically, compared to their male counterparts, female students perform 0.32 (0.25) standard deviations worse on the *Gaokao* relative to the mock examinations when they are within 3 (5) points of the cutoff. The gender performance gap is less than half as large at 0.14 and 0.10 for students 6 to 10 points and 11 to 20 points from the reference cutoffs. The estimates in Column (5) provide a formal test of significance of the difference across students within 3 points and students within 11 to 20 points of the reference cutoffs. The difference is -0.22 standard derivations with a standard error of 0.14. While the relatively large standard errors of our estimates imply that the difference is not statistically significant at conventional levels, the magnitude of the difference is economically large.[24] These qualitative patterns are similar for the

fact that the distribution of mock exam scores are reported in 10 point bins, hence the reference cutoffs are rounded to the nearest ten. The reference cutoffs chosen are in fact the mock exam score bins that most closely generate the same fractions of students eligible for each university tier as the 2007 *Gaokao* cutoffs.

[23] From Table 6B, admission into Tier 3 and the technical universities are a lot less competitive, with 40 to 50% of students likely eligible to enter at least a Tier 3 or better university and 80-90% of all test-takers eligible to enter at least a technical university.

[24] This is likely to be due to the relatively small sample size of students within 3 or 5 points of the reference cutoffs.

individual subjects (Panels (B) to (E)).

While this result provides additional evidence in support of the idea that women and men respond differently towards competitive pressures, it is also possible that men near the reference entry cutoffs are able to differentially increase their preparation for the *Gaokao* in the intervening two months, although as noted in the previous subsection, we do not find any evidence of clear gender differentials in effort provision in the months leading up to the *Gaokao*. Nonetheless, for this exercise, we can use the additional variation in stakes generated by distance from the reference cutoffs to look *within* gender to examine whether the observed patterns are driven by a decline in female performance or an improvement in male performance in the high-stakes setting. If part of the observed patterns are due to lower female performance when faced with more competitive pressure, this can help to alleviate concerns that our earlier results were largely driven by an increase in effort provision by males in the *Gaokao* relative to the mock exam.

4.5 Performance Differences by Gender

Table 8 examines how the performance gap of female and male students vary depending on how close their scores are to the reference entry cutoffs on the mock exam. We hypothesize that students who score just below (or close to) the reference entry cutoffs are likely to face higher pressure to perform well on the *Gaokao* relative to students with mock exam scores that are further away from the entry cutoffs. Utilizing this source of variation in pressure, we are able to examine the effect of competitive pressures on female and male performance, separately, without relying on a comparison of relative differences across genders.

The key dependent variable is the difference in standardized (within-gender) *Gaokao* and mock exam scores. We estimate separate regressions by gender of the performance gap between the *Gaokao* and mock exam on indicators of the distance from the predicted cutoffs based on the mock examination. As observed in Column (1), female students scored 0.22 standard deviations (standard error of 0.14) lower on the *Gaokao* relative to the mock exam when they are 1 to 3 points below the reference cutoffs, as compared to female students who are 11 to 20 points from the cutoffs. The difference in scores between the *Gaokao* and the mock exam is 0.05 (s.e. 0.137) and 0.03 (s.e. 0.065) standard deviations lower for female students who score 0 to 3 points above the cutoffs and within 4 to 10 points from the cutoffs, respectively, relative to female students who score 11 to 20 points from the cutoffs. These results indicate that female students tend to score worse on the high-stakes *Gaokao* when their mock exam scores are just below the reference thresholds for admission into Tier 1 or Tier 2 universities, relative to when their mock exam scores are further away from the thresholds. Interestingly, this pattern is reversed for male students - males who scores within 3 points of the reference thresholds appear to score higher on the *Gaokao* relative

to the mock exam as compared to males who are 11 to 20 points from the cutoffs (column (2)). While the within-gender differences are not significant at conventional levels, the magnitude of the estimates are quite large. Column (3) provides a statistical test of the difference in the estimates between the female and male sample. We find that the difference in the estimates for female and male students who score 1 to 3 points below the cutoff is statistically significant at the 5% level. In Columns (4) to (6), we show that the results are similar when the full set of controls are included.

In sum, these patterns provide suggestive evidence that the widening of the gender gap in performance in high-stakes vs. low-stakes settings is driven by a combination of a decline in females' performance coupled with an improvement in males' performance in high-stakes settings. This is consistent with the idea that the observed gender difference in performance in the *Gaokao* vs. the mock examination is unlikely to be entirely driven by an increase in effort provision by male students. Female and male students appear to react to high pressure environments quite differently - while female performance appears to suffer, male students appear to "up" their game when the stakes are higher.

4.6 Gender Gap in Response to (Negative) Performance Shocks

Next, we exploit the fact that the Gaokao is held over multiple subjects across a two-day period to explore whether males and females react differently to "shocks" in their performance on an earlier Gaokao subject exam. One advantage of exploiting variation in pressure (arising due to differences in performance on an earlier exam) during the two-day *Gaokao* exam is that we do not expect students to be able to adjust their preparation for the later exam within such a narrow time frame. We hypothesize that performance on an earlier subject exam may affect the pressure levels faced by individuals in later exams as university admission is based on the total score across all the subjects. More specifically, if a student performs worse on an early exam, he/she may face higher pressure to achieve better scores in the later exams to compensate for underperforming in the earlier exam. If males and females respond differently to competitive and stressful environments, they would react differently to "shocks" in their performance on an earlier *Gaokao* subject exam.

Focusing on the exams on the first day, we examine how a student's performance on the afternoon examination (mathematics) is affected by "shocks" to his/her performance on the morning examination (Chinese). To proxy for performance shocks, we use the deviation between an individual's score on the Gaokao and the mock examination on the morning Chinese exam. Our empirical strategy thus relates a student's relative performance on the morning Chinese test to his/her relative performance on the afternoon math test. The regression specification is as follows:

$$\tilde{M}_{i,g}^{E,S} - \tilde{M}_{i,g}^{M,S} = \beta_0 + \beta_1 Female_i \times (\tilde{C}_{i,g}^{E,S} - \tilde{C}_{i,g}^{M,S}) + \beta_2 (\tilde{C}_{i,g}^{E,S} - \tilde{C}_{i,g}^{M,S}) + \beta_3 Female_i + \mathbf{Y}_i \boldsymbol{\gamma} + \epsilon_{i,g} \quad (2)$$

where the outcome, $\tilde{M}_{i,g}^{E,S} - \tilde{M}_{i,g}^{M,S}$, is the difference in student i's mathematics score (afternoon test) on the Gaokao and mock examination and $\tilde{C}_{i,g}^{E,S} - \tilde{C}_{i,g}^{M,S}$ is the difference in student i's Chinese score (morning test) on the Gaokao and mock examination. The controls Y_i are identical to those in equation (1).

We are interested in the coefficient β_1 which measures how deviations from the mock examination score on the morning examination differentially affect the relative performance of females on the afternoon examination relative to males. Before turning to the regression estimates, Figure 2A presents the locally smoothed (lowess) graph of the unconditional relationship between the relative performance (Gaokao-mock exam) on the morning exam $\tilde{C}_{i,g}^{E,S} - \tilde{C}_{i,g}^{M,S}$ and the relative performance on the afternoon exam $M_{i,g}^{E,S} - \tilde{M}_{i,g}^{M,S}$ separately for males (solid line) and females (dashed line). From the figure, we can see that there is a positive relationship between a student's relative performance on the morning examination and his/her relative performance on the afternoon exam. Strikingly, this positive relationship appears a lot stronger for female students relative to male students, suggesting that females' performance on the afternoon exam is more strongly affected by their relative performance on the morning examination as compared to their male counterparts.

Column (1) of Table 9 reports the baseline coefficient estimates - the coefficient β_1 indicates that in response to a one standard deviation improvement in relative scores (Gaokao-mock exam score) in the morning exam, female relative performance on the afternoon exam is 0.06 standard deviations higher than that of males. This estimate is marginally significant at the 10% level. Column (2) includes a full set of controls that are interacted with the female dummy - β_1 increases slightly to 0.07 and is now significant at the 5% level. Notice that in both specifications, β_2 is also positive and statistically significant implying that the relative performance on the morning exam tends to be positively correlated with the relative performance on the afternoon exam for males as well. The striking finding from this regression is that relative performance on the morning exam tends to matter for performance on the afternoon exam significantly more for female students relative to male students.

Given that students closer to the reference entry cutoffs are more likely to face a greater degree of pressure, we test whether the gender differences in the response to initial performance shocks are more pronounced for groups of students who are closer to the thresholds - or in other words, have more to lose. Figure 2B depicts the locally smoothed unconditional relationship between the relative performance on the morning exam and the afternoon exam for subgroups of students within 3 points, 5 points, 6 to 10 points and 11 to 20 points of the reference cutoffs. We observe that for students close to the reference cutoffs (within 3 and 5 points), there is a clear positive relationship between the relative performance of the morning and afternoon exam for female students, particularly among female students who experienced a negative shock. In contrast, there appears to be virtually no relationship between the relative performance of male students in the morning exam and their

subsequent performance on the afternoon exam. Interestingly, there appears to be little evidence of a gender difference in the reaction to performance shocks for students further away from the reference thresholds.

Columns (3) to (6) of Table 9 report the coefficient estimates for students within 3 points, 5 points, 6 to 10 and 11 to 20 points of the cutoff, respectively. We find strong evidence that the gender gap in the response to initial performance shocks are largest among students closest to the entry thresholds. Among students within 3 to 5 points of the cutoff, a one-standard deviation increase in students' relative performance on the morning *Gaokao* is associated with a 0.34 to 0.19 standard deviation larger improvement in females' relative performance on the morning exam compared to their male counterparts. Interestingly, for this subgroup of students, there is virtually no relationship between the relative performance on the morning exam and afternoon performance for male students. For students further away from the threshold (6 to 20 points), there is little evidence of a gender difference in the responsiveness to initial performance. Overall, these results suggest that the gender differences in the responsiveness to performance shocks tend to be larger when the stakes are higher and when students are more likely to face greater competitive pressure. The fact that gender differences in the reaction to initial performance varies systematically across high(er) and low-stakes settings also suggest that our results are not entirely driven by unobservables that differentially affect the relative performance of females and males across different exams.[25]

To provide additional evidence that gender differences in the relationship between the relative performance on the afternoon test and morning test are indeed the consequence of performance shocks in the morning test, we look separately at the effect of (1) relative performance in the morning on students' performance on the *Gaokao* afternoon test and (2) relative performance in the morning on student's performance on the afternoon test on the mock examination. Consistent with a causal interpretation, the results in Appendix Table A4 indicate that the gender differences in the reaction to performance shocks documented in Table 9 are driven primarily by an improvement in female performance on the *Gaokao* afternoon examination (see the top panel). There is little evidence of gender differences in the effect of performance shocks on past performance on the afternoon test of the mock examination (see bottom panel). The latter result is reassuring as it suggests that there are no gender differences in the correlation between performance on the afternoon test in the mock examination and the incidence of performance shocks on the morning *Gaokao* exam. The fact that the performance shocks are only differentially correlated with the future performance of females and males, and are not differentially correlated with past performance provides a causal interpretation of our findings. The results for different subgroups of students based on their distance

[25] One concern might be that the presence of unobserved shocks that are correlated with performance on both exams that are magnified in high pressure settings could potentially generate these patterns. However, rather than being an alternative story, this possibility could be part of the mechanism that explains why females' subsequent performance is more affected by performance shocks.

to the entry cutoffs reported in Columns (2) to (5) of Appendix Table A4 are consistent with our previous findings. Appendix Table A5 shows that the results are similar if we consider the effects of performance shocks on the first three exams (Chinese, Math and Combined subjects) on the performance on the final English exam held in the afternoon of Day 2. This suggests that the results are not sensitive to the test subjects considered.[26]

Finally, in Table 10, we report estimates from a more flexible specification that looks separately at the effect of positive and negative shocks to relative performance. Interestingly, we find that relative to males, females appear to be more responsive to negative shocks as compared to positive shocks. The gender differences in responsiveness to negative shocks are also most pronounced for students closest to the reference cutoffs. Nevertheless, due to the large standard errors, with the exception of the subgroup of students within 5 points of the cutoff, we generally cannot reject that the magnitude of the difference in response to positive and negative shocks are significantly different. While there is a large behavioral literature that suggests that individuals are more responsive to losses than gains (e.g. Tversky and Kahneman, 1979), our results suggest that there may be important gender differences in the responsiveness to perceived losses relative to gains. These results are also broadly consistent with the empirical and experimental evidence that women may be more affected by negative feedback or performance (Roberts and Nolen-Hoeksema, 1989; Wozniak 2012; Goldin, 2013; Gill and Prowse, 2014; Buser, 2014).[27] In sum, in high pressure settings, female performance appears to be more detrimentally affected by negative shocks relative to males, which may contribute partially to the observed underperformance in the *Gaokao* relative to the mock examination.

5 Conclusion

We examine the gender gap in performance in response to competitive pressures and performance shocks in an important field setting - the *Gaokao* in China, an examination that is often touted as the world's most competitive. Using a unique dataset that links the examination records of the universe of candidates that sat for both the *Gaokao* and the mock examination held two months earlier in Anxi county in Fujian province, we study whether female and male students react differently to pressure by contrasting their performance in two (otherwise similar) settings where

[26] For the main analysis, we focus on the effects of Day 1 morning shocks on the performance on the Day 1 afternoon exam as this provides the cleanest empirical set up since students' performance on the first examination of the *Gaokao* cannot be affected by previous shocks. Also, examining the effects of performance shocks within the same day (vs. across Day 1 and Day 2 of the *Gaokao*) has the added advantage that students are less likely to have time to adjust to the earlier performance shock.

[27] As the feedback is noisy in our setting, another possible mechanism is that for the same degree of negative shocks, males may perceive them differently as they are overconfident about their performance (Barber and Odean, 2001). This overconfidence bolsters men against performance shocks.

the stakes vary considerably. We find that the gender gap in performance is significantly larger in the high-stakes *Gaokao* relative to the mock examination. These gender differences in exam performance across settings translate to a 15% decline in the likelihood that females are eligible for admission into a Tier 1 university in the *Gaokao* relative to the mock exam as compared to their male counterparts.

We find limited evidence that the empirical patterns are driven by male students choosing to exert less effort in the mock examination. In particular, we argue that there are strong incentives for both male and female students to take the mock exam seriously as it is the only chance that students can figure out their relative standing within the province and to get a real sense of the types of universities that they will likely qualify for. Moreover, we use variation in pressure induced by the reference cutoffs to look at how relative performance varies as a function of distance from the cutoffs within gender. These results indicate that the gender performance gap appears to be driven, in part, by a decline in female performance coupled with an improvement in male performance in higher pressure settings. In addition, we utilize the fact that the *Gaokao* has multiple subject components that are held in the morning and the afternoon over a two-day period to examine whether females and males respond differently to shocks to their performance on an earlier test. We find that relative to males, female relative performance on the first day's afternoon *Gaokao* is more strongly affected by performance shocks, especially the negative ones, on the morning exam, as measured by the deviation of the *Gaokao* morning exam score from the mock exam score. Consistent with the results found in the previous part of the paper, we also find that the gender gaps in the reaction to relative performance shocks are more pronounced for students who are close to the reference entry cutoffs for admission into the different university tiers, suggesting that females are more affected by performance shocks when competitive pressures are stronger. In sum, these results support the role of pressure for the observed performance gender gap.

Our study contributes to some recent literature exploring the role of stress in economic behavior. In an experimental setting, Ariely et al (2009) show that very high reward levels may have a psychological pressure, leading to a detrimental effect on performance. Chemin, De Laat and Haushofer (2013) observe that low levels of rain in the preceding year increased the level of the stress hormone of farmers in Ken. Moreover, Goh, Pfeffer and Zenios (2015) find that various types of stress in the workplace contribute substantially to healthcare cost and mortality in the United States. A number of recent studies investigate the role of stress measured by hormonse responses in competition (Apicella et al., 2011; Buser, Dreber, and Mollerstrom, 2015; Buckert et al., 2015; Zhong et al., 2015), while they find that gender difference in stress response does not contribute to gender difference in competitiveness. Relatedly, Goette et al. (2015) find that cortisol response to social stress correlates with self-confidence and that the effect depends on the level of trait anxiety. Our study suggests that gender differences in the response to stress may underpin gender differences in performance in high pressure settings.

Our results may have important implications for understanding the persistent underrepresentation of females in certain education fields and occupations that tend to be more competitive. To the extent that females tend to underperform in high pressure environments, this could potentially explain why women tend to opt out of educational and career tracks that are more competitive and where there is a large premium to performing under pressure (Buser, Niederle and Oosterbeek, 2014, Shurchkov, 2012 and Kleinjens, 2009). This idea that females are more likely to be easily "discouraged" when under pressure has important implications for policies that aim to increase academic diversity and to increase the representation of well-qualified women in more competitive, higher-paying fields and careers. The fact that the performance of males and females can vary dramatically depending on the testing environment suggests that the exclusive use of high-stakes testing (such as the *Gaokao*) as an ability screen and allocation mechanism into higher education works to the disadvantage of females and might lead to a relative paucity of females in top academic programs by virtue of the choice of testing mechanism. Our findings suggest that one way of achieving greater gender diversity could be to alter the stakes of admission examinations or to consider a wider range of admission pathways (for example using a combination of high-stakes testing and continual assessment).

Appendix I

Background Information on Anxi County

Anxi county is part of Quanzhou city in Fujian Province. Fujian has a total population of 36.9 million in 2010 and is a province on the southeast coast of mainland China. Quanzhou is the 12th largest Chinese extended metropolitan area (as of 2010) and is the largest prefecture-level city in Fujian. Quanzhou administers four districts, three county-level cities and four counties. Anxi is a mid-sized county in Quanzhou with a population of close to 1 million.

- County-level cities: Jinjian (1.97 mil), Nan'an (1.42 mil), Shishi (0.64 mil)
- Counties: Anxi (0.98 mil), Hui'an (0.72 mil), Yongchun (0.45 mil), Dehua (0.28 mil)

The following table provides some economic indicators of Anxi in relation to China as a whole, Fujian province and Quanzhou city in 2008.

In terms of participation in the *Gaokao*, the following table provides the gender breakdown and proportion of students in the Science and Arts stream in China, Fujian and Anxi in 2008.

	GDP per capita	Total (1000s) Population	% Urban	Annual Wages	Senior Sec Enrollment	Student-Teacher Ratio
China	22,698	132802	46	28,898	24,762,842	16.8
Fujian	30,123	3604	50	25,555	748,828	14.3
Quanzhou	34,840	779	50	22,225	160,614	14.9
Anxi	22,424	107.1	32	20,260	21,413	17.8

Note. The data is from the China Statistical Yearbook (2009) and the Fujian Statistical Yearbook (2009). All dollar values are in Yuan. 1 Yuan = 0.16 USD.

	China	Fujian	Anxi
Total Applicants	10,226,347	305,256	8432
Female	0.484	0.485	0.441
Science stream	0.534	0.597	0.535

Note. Figures for China and Fujian are from the Educational Statistics Yearbook of China (2009)

Appendix II

Gaokao Preparation Questionnaire

This is a survey on your behavior related your Gaokao. Please response to the questions truthfully. All the information provided will be kept confidentially. Thank you!

(A). Please respond to the following questions for each of the last 6 months before Gaokao.

	May	April	March	February	January	December
	2016	2016	2016	2016	2016	2015
A1. How many hours, on average, did you spend studying each day?						
A2. How many hours, on average, did you spend studying the material for the following subject each day?						
Subject: Chinese						
Subject: English						
Subject: Combined						
Subject: Math						
A3. On a scale of 1 to 10, how prepared were you for the Gaokao exam? (1for very unprepared, 10 for very prepared)						
A4. On a scale of 0% to 100%, how much of the material for the following subject have you prepared for?						
Subject: Chinese						
Subject: English						
Subject: Combined						
Subject: Math						
A5. How many hours of sleep, on average did you get each day?						
A6. On a scale of 1 to 10, what was the quality of your sleep? (1 for very poor, 10 for very good)						
A7. On a scale of 1 to 10, how much stress did you feel? (1 for very relaxed, 10 for very stressful)						
A8. On average, how many days were you sick each month?						
A9. Rating the effectiveness of your study efforts (1 for very inefficient, 10 for very efficient)						

(B). Please respond to the following questions regarding mock exam in each of the subject.

	Chinese	English	Combined	Math
B1. On a scale of 1 to 10, how much effort did you put in when taking <SUBJECT> in mock exam? (1: very little effort, 10: full effort)				
B2. On a scale of 1 to 10, how prepared did you feel when taking <SUBJECT> in mock exam? (1: not prepared at all, 10: very prepared)				
B3. How much stress did you feel when taking <SUBJECT> in mock exam? (1 very relaxed, 10 very stressful)				
B4. After taking the mock exam, how did you expect your performance to change on the Gaokao for <SUBJECT>? (1: Large improvement in performance, 2: Slight Improvement in performance 3: About the same performance, 4: Slight decline in performance, 5: Large decline in performance)				
B5. Based on your mock exam scores, what did you predict your Gaokao score would be for <SUBJECT>?				
B6. Were you sick when taking <SUBJECT> in mock exam? (Yes or No)				
B7. What was your mock exam score in <SUBJECT>?				

(C). Please respond to the following questions regarding Gaokao.

	Chinese	English	Combined	Math
C1. On a scale of 1 to 10, how much effort did you put in when taking <SUBJECT> in Gaokao? (1: very little effort, 10: full effort)				
C2. On a scale of 1 to 10, how prepared did you feel when taking <SUBJECT> in Gaokao? (1: not prepared at all, 10: very prepared)				
C3. When you sat for the Gaokao, would you say you were: 1: A lot less prepared, 2: Slightly less prepared, 3: As prepared, 4: Slightly more prepared 5: A lot more prepared than when you sat for the mock exam?				
C4. How much stress did you feel when taking <SUBJECT> in Gaokao? (1 very relaxed, 10 very stressful)				
C5. Were you sick when taking <SUBJECT> in Gaokao? (Yes or No)				
C6. Did your performance on <SUBJECT> in Gaokao meet your expectations? (1: Yes, 2: No)				
C7. What was your Gaokao score in <SUBJECT>?				

(D). Please respond to the following questions about your personal information.

D1. Name:

D2. Age:

D3. Gender:

D4. Stream:

D5. Tel:

D6. Email:

D7. Are you interested in participating further surveys in the future? 1. YES 2. NO

Thank you for your participation!

References

[1] American Psychological Society. 2010. “Stress in America Findings."

[2] Apicella, CL, A Dreber, PB Gray, M Hoffman, AC Little, and B Campbell, 2011. Androgens and competitiveness in men. Journal of Neuroscience, Psychology, and Economics 4, 54-62.

[3] Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. "Large stakes and big mistakes." The Review of Economic Studies 76, no. 2 (2009): 451-469.

[4] Attali Yigal, Zvika Neeman and Analia Schlosser. 2011. “Rise to the Challenge or Not Give a Damn: Differential Performance in High vs. low-stakes Tests." IZA Discussion Paper No. 5693.

[5] Azmat Ghazala, Caterina Calsamiglia and Nagore Iriberri. 2014. “Gender Difference in Response to Big Stakes." forthcoming Journal of the European Economic Association

[6] Barber, Bard M., and Terrance Odean. 2001 “Boys will be Boys: Gender, Overconfidence, and Common Stock Investment." Quarterly Journal of Economics. Vol 116(1): 261-292.

[7] Bertrand, Marianne. 2010. “New Perspectives on Gender." Handbook of Labor Economics, O. Ashenfelter and D. Card eds. Vol 4B: 1543-1590.

[8] Burgess, Simon, and Ellen Greaves. "Test scores, subjective assessment, and stereotyping of ethnic minorities." Journal of Labor Economics 31(3). (2013): 535-576.

[9] Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. “Gender, Competitiveness and Career Choices." Quarterly Journal of Economics 129(3). (2014). 1409-1447.

[10] Buser, Thomas. "The impact of losing in a competition on the willingness to seek further challenges." Tinbergen Discussion Paper (2014).

[11] Buser, T., Dreber, A., and Mollerstrom, J. (2015). Stress Reactions cannot explain the Gender Gap in Willingness to compete. *Working paper.*

[12] Buckert, M., C. Schwieren, B.M. Kudielka and C.J. Fiebach, 2015. How stressful are economic competitions in the lab? An investigation with physiological measures. *Working paper.*

[13] Chemin, Matthieu, Joost De Laat, and Johannes Haushofer. "Negative rainfall shocks increase levels of the stress hormone cortisol among poor farmers in Kenya." Available at SSRN 2294171 (2013).

[14] Cornwell, Christopher, David B. Mustard, and Jessica Van Parys. "Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school." Journal of Human Resources 48, no. 1 (2013): 236-264.

[15] Davey, G., De Lian, C., and Higgins, L. 2007. “The University Entrance Examination System in China." Journal of Further and Higher Education. Vol 31(4), 385-396.

[16] Gill, David, and Victoria Prowse. "Gender differences and dynamics in competition: The role of luck." Quantitative Economics 5.2 (2014): 351-376.

[17] Gneezy, Uri, Muriel Niederle and Aldo Rustichini. 2003. “Performance in Competitive Environments: Gender Differences." Quarterly Journal of Economics. Vol 118: 1049-1074.

[18] Gneezy, Uri, and Aldo Rustichini. “Gender and competition at a young age." American Economic Review (2004): 377-381.

[19] Goette, L., Bendahan, S., Thoresen, J., Hollis, F., & Sandi, C. (2015). Stress pulls us apart: Anxiety leads to differences in competitive confidence under stress. Psychoneuroendocrinology, 54, 115-123.

[20] Goh, Joel, Jeffrey Pfeffer, and Stefanos A. Zenios. "The Relationship Between Workplace Stressors and Mortality and Health Costs in the United States." *Management Science* (2015).

[21] Goldin, Claudia. 2013. “Can ‘Yellen Effect’ Attract Young Women to Economics?" Bloomberg View, `http://www.bloombergview.com/articles/2013-10-14/can-yellen-effect-attract-young-women-to-economics-`

[22] Jurajda Stepan and Daniel Munich. 2011. “Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities." American Economic Review Papers and Proceedings. Vol 101(3): 514-518.

[23] Kahneman, Daniel and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk." Econometrica. Vol 47(2): 263-292.

[24] Kirschbaum, C., Wust, S., and Hellhammer, D. 1992. “Consistent Sex Differences in Cortisol Responses to Psychological Stress." Psychosomatic Medicine. Vol 54(6): 648-657.

[25] Kleinjens, Kristin J. 2009. “Do Gender Differences in Preferences for Competition Matter for Occupational Expectations?" Journal of Economic Psychology. Vol 30: 701-710.

[26] Lavy, Victor. 2013. “Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-based Pay Tournaments among Teachers." Economic Journal. Vol 123: 540-573.

[27] Lavy, Victor. "Do gender stereotypes reduce girls’ or boys’ human capital outcomes? Evidence from a natural experiment." Journal of public Economics 92, no. 10 (2008): 2083-2105.

[28] Lee, Joohyung and Vincent R. Harley. 2012. “The Male Flight-Flight Response: A Result of SRY Regulation of Catecholamines?" BioEssays. Vol 34(6): 454-457.

[29] Morin, Louis-Philippe. 2013. “Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform." Journal of Labor Economics, forthcoming.

[30] Niederle, Muriel and Lise Vesterlund. 2007. “Do Women Shy Away from Competition? Do Men Compete Too Much?" Quarterly Journal of Economics. Vol 122(3): 1067-1101.

[31] Niederle, Muriel and Lise Vesterlund. 2011. “Gender and Competition." Annual Review of Economics. Vol 3(1): 601-630.

[32] Ors, Evren, Frédéric Palomino, and Eloic Peyrache. "Performance Gender Gap: Does Competition Matter?" Journal of Labor Economics 31.3 (2013): 443-499.

[33] Paserman, Daniel M. 2010. “Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players." Working Paper.

[34] Roberts Tomi-Ann and Susan Nolen-Hoeksema. 1989. “Sex Differences in Reactions to Evaluative Feedback." Sex Roles. Vol 21(11-12): 725-747.

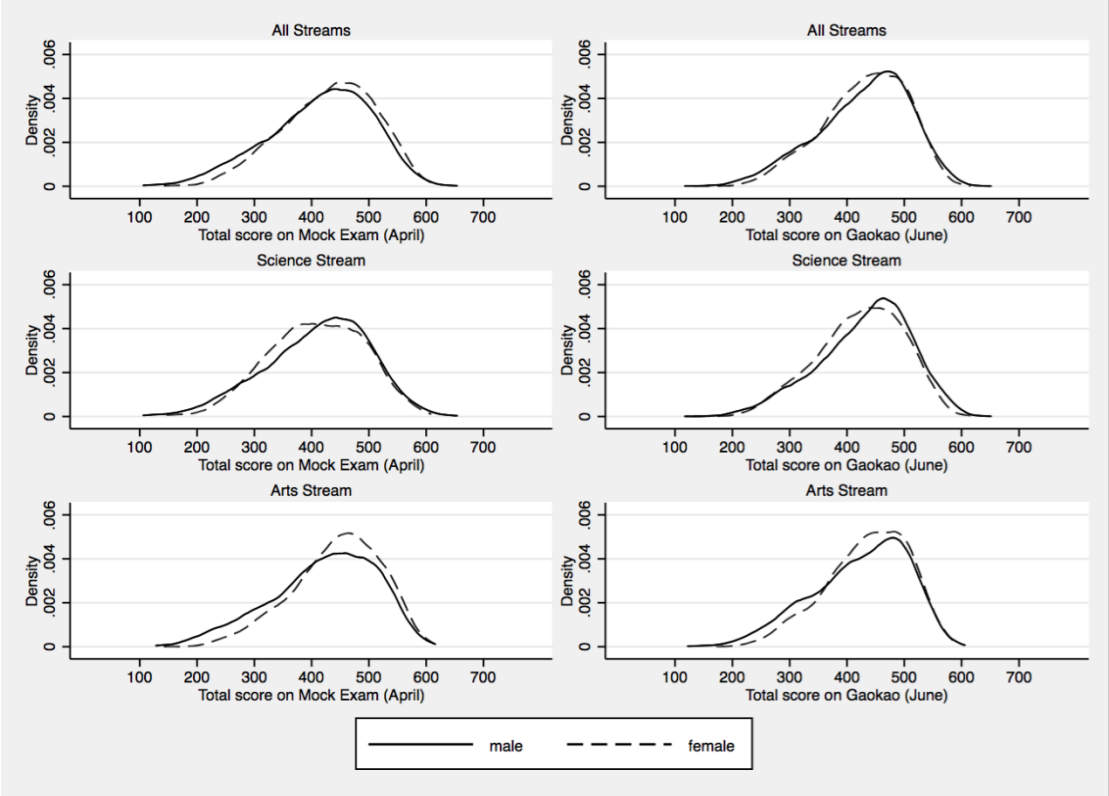
[35] Shurchkov, Olga. 2012. “Under Pressure: Gender Differences in Output Quality and Quantity under Competition ad Time Constraints." Journal of the European Economic Association. Vol 10(5): 1189-1213.

[36] Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A., and Updegraff, J. A. 2000. Biobehavioral Responses to Stress in Females: Tend-and-Befriend, not Fight-or-Flight. Psychological Review. Vol 107(3): 411.

[37] Wozniak, David. 2012. “Gender Differences in a Market with Relative Performance Feedback: Professional Tennis Players." Journal of Economic Behavior and Organization. Vol 83: 158-171.

[38] Zhong, Songfa, Idan Shalev, David Koh, Richard P. Ebstein, and Soo Hong Chew. "Competitiveness and Stress." Working Paper

Figure 1: Distributions of Female and Male Performance on the Mock Exam and Gaokao



Note. The sample includes students who sat for both the Mock Exam (April) and Gaokao (June). Each figure plots the distribution of total exam score for male and female students separately for the mock exam (left column) and actual Gaokao exam (right column). The top two graphs include all students, the middle two graphs include only students in the science stream and the bottom two graphs include only students in the arts stream.

Figure 2A: Relationship between Day 1 Afternoon Exam Score and Morning Exam Score by Gender

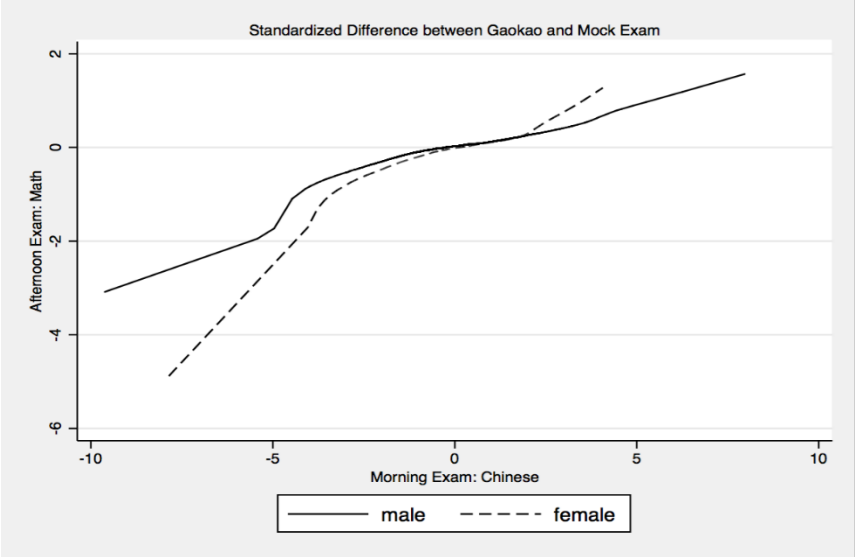
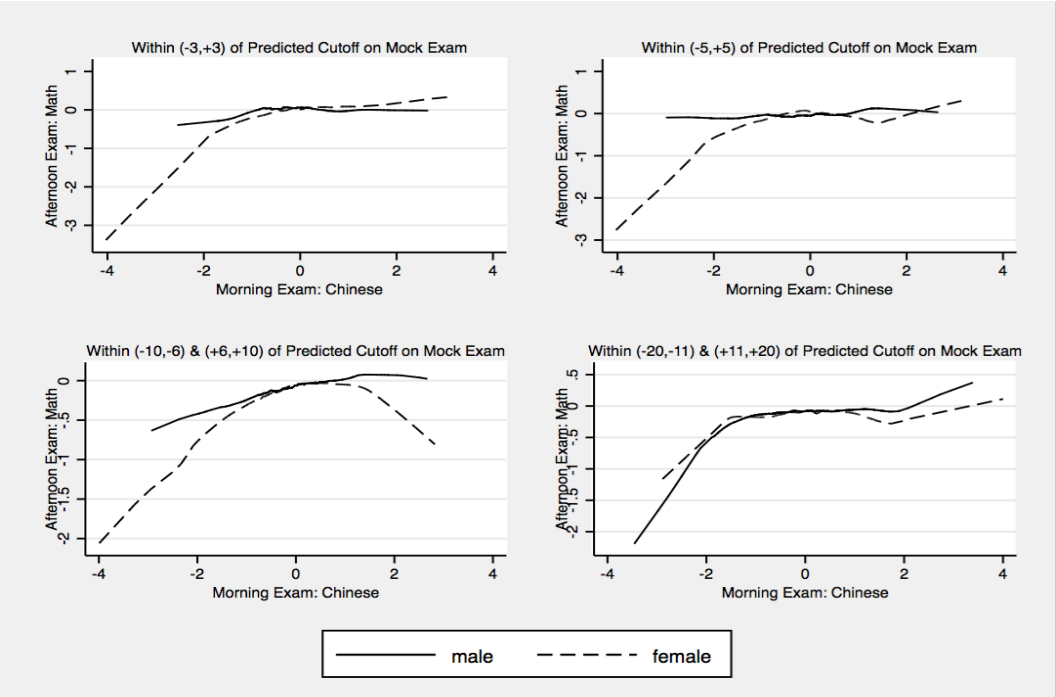


Figure 2B: Relationship between Day 1 Afternoon Exam Score and Morning Exam Score by Gender and Distance from Reference Cutoff



Note. The sample includes all students who sat for both the Gaokao and mock exam. The figures plot the relationship between the standardized difference between the Gaokao and the mock exam for the afternoon exam (Math) against the morning exam (Chinese) separately for male (solid line) and female students (dashed line). Figure 2A is for all students while each graph in Figure 2B is for different subsets of students depending on their performance on the mock exam relative to the reference cut-off.

Table 1: Summary Statistics

Sample	Mock Exam	Gaokao	Merged	Difference (Mock Exam - Merge)	Difference (Gaokao - Merge)
Observations	8164	8432	7961		
Female		0.44	0.45		-0.006
		(0.50)	(0.50)		
Science stream		0.54	0.54		-0.003
		(0.50)	(0.50)		
Age (in months)		227.24	227.15		0.089
		(10.15)	(10.10)		
Mock exam scores					
Total (out of 750)	420.78		421.17	-0.389	
	(87.12)		(86.78)		
Chinese (out of 150)	91.25		91.31	-0.053	
	(11.21)		(11.07)		
Math (out of 150)	94.27		94.37	-0.105	
	(25.32)		(25.24)		
Combined Science/Arts subjects (out of 300)	153.46		153.63	-0.169	
	(42.83)		(42.74)		
English (out of 150)	81.80		81.86	-0.063	
	(23.52)		(23.50)		
Gaokao scores					
Total (out of 750)		426.34	429.93		-3.595***
		(80.69)	(77.92)		
Chinese (out of 150)		96.78	97.21		-0.425***
		(10.15)	(9.65)		
Math (out of 150)		94.88	96.06		-1.176***
		(26.48)	(25.63)		
Combined Science/Arts subjects (out of 300)		151.25	152.52		-1.275**
		(35.88)	(35.07)		
English (out of 150)		83.42	84.14		-0.719**
		(21.02)	(20.57)		

Note. The mock exam sample includes all students who sat for all four papers in the mock examination in April 2008. The Gaokao sample includes all students who sat for all four papers in the actual examination in June 2008. The merged sample comprises students who could be identified in both the mock exam and Gaokao sample. The first three columns report the mean and standard deviation of the key variables in each of the samples. The last two columns report the difference in means between the mock exam and merged sample as well as the actual and merged sample. ***difference is significant at 1% level, **5%, *10%.

Table 2: Gender Gap in Mock Exam and Gaokao Scores

	Mock Exam (April)			Gaokao (June)			Diff-in-Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Female	Male	Difference	Female	Male	Difference	(6) - (3)
Total	429.91	414.08	15.830***	431.10	428.99	2.106	-13.724***
	(80.90)	(90.66)	[1.948]	(73.45)	(81.37)	[1.756]	[1.033]
Observations	3563	4398	7961	3563	4398	7961	7961
Total (Science stream only)	407.92	410.85	-2.932	423.26	432.11	-8.848***	-5.916***
	(82.00)	(90.37)	[2.876]	(75.23)	(80.16)	[2.576]	[1.525]
Observations	1370	2911	4281	1370	2911	4281	4,281
Total (Arts stream only)	443.65	420.41	23.240***	435.99	422.89	13.104***	-10.136***
	(77.10)	(90.92)	[2.787]	(71.89)	(83.36)	[2.578]	[1.429]
Observations	2193	1487	3680	2193	1487	3680	3,680

Note. The sample includes individuals who sat for both the mock (April) and Gaokao (June) examinations. Columns (3) and (6) report the gender difference in test scores for the mock and Gaokao, respectively. "Total (Science stream only)" and "Total (Arts stream only)" reports the total score for students in the science stream and arts stream, respectively. The last column reports the difference in the gender gap between the gaokao and the mock. Standard errors are reported in brackets. ***significant at 1% level, **5% level, *10% level.

Table 3: Regression Estimates of the Gender Gap in Performance

	Standardized Difference between Gaokao and Mock Examination									
	Total		Chinese		Math		Combined Science/Arts		English	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	A. Full Sample									
Female	-0.159***	-0.153***	-0.014	-0.004	-0.048**	-0.036	-0.167***	-0.170***	-0.122***	-0.119***
	[0.023]	[0.023]	[0.022]	[0.023]	[0.022]	[0.023]	[0.023]	[0.023]	[0.022]	[0.023]
Observations	7,961	7,961	7,961	7,961	7,961	7,961	7,961	7,961	7,961	7,961
R-squared	0.006	0.031	0.000	0.031	0.001	0.035	0.007	0.029	0.004	0.022
	B. Science Stream									
Female	-0.146***	-0.148***	-0.021	-0.015	-0.056	-0.038	-0.090***	-0.104***	-0.164***	-0.159***
	[0.034]	[0.035]	[0.033]	[0.034]	[0.036]	[0.036]	[0.031]	[0.032]	[0.032]	[0.033]
Observations	4,281	4,281	4,281	4,281	4,281	4,281	4,281	4,281	4,281	4,281
R-squared	0.004	0.043	0.000	0.045	0.001	0.057	0.002	0.039	0.006	0.035
	C. Arts Stream									
Female	-0.199***	-0.194***	-0.008	-0.006	-0.048	-0.037	-0.277***	-0.274***	-0.098***	-0.106***
	[0.033]	[0.034]	[0.033]	[0.033]	[0.031]	[0.032]	[0.035]	[0.036]	[0.034]	[0.035]
Observations	3,680	3,680	3,680	3,680	3,680	3,680	3,680	3,680	3,680	3,680
R-squared	0.010	0.033	0.000	0.043	0.001	0.038	0.017	0.042	0.002	0.027
Controls:										
Age		X		X		X		X		X
School FE		X		X		X		X		X
Zipcode FE		X		X		X		X		X

Note. Each column in each panel is a separate regression with the standardized difference between the Gaokao (June) score and the mock exam (April) score as the dependent variable for each of the subjects listed in the table. Panel A includes the full sample, Panel B includes only students in the Science stream and Panel C includes only students in the Arts stream. The total score is the sum of the scores across all four examination components. Both the Gaokao score and mock exam scores were standardized to have a mean of 0 and standard deviation of 1 by student type (i.e. Science vs. Arts and Social Science Stream). The difference between the standardized Gaokao and mock exam scores were re-standardized to have a mean of 0 and standard deviation of 1. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Table 4: Regression Estimates of the Gender Gap in Performance from the largest high school in Anxi in 2014

	(1)	(2)	(3)	(4)	(5)	(6)
	Gaokao - April Mock		Mid-May Mock - April Mock		Late-May Mock – April Mock	
Female	-0.115*	-0.111*	-0.036	-0.024	-0.010	-0.003
	[0.062]	[0.066]	[0.063]	[0.065]	[0.063]	[0.065]
Controls:						
zip code dummy		X		X		X
Age		X		X		X
Observations	1,016	1,016	1,016	1,016	1,016	1,016
R-squared	0.003	0.051	0	0.053	0	0.06

Note. Columns 1 and 2 report the difference between the April mock exam and the Gaokao. Columns 3 and 4 and Columns 5 and 6 examine the difference between the April and mid-May mock exams, and the difference between the April and late-May mock exams, respectively.Columns2, Columns 4 and Columns 6 include individual level controls such as age and zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Table 5a: Daily Study Hours

	(1)	(2)	(3)
	Male	Female	Difference
	A: Total		
May-16	7.511	7.296	0.216
Apr-16	7.33	7.039	0.29
Mar-16	7.063	6.841	0.221
Feb-16	6.585	6.502	0.083
Jan-16	6.563	6.456	0.107
Dec-15	6.537	6.306	0.231
	B: Chinese		
May-16	1.053	1.072	-0.019
Apr-16	1.015	1.039	-0.023
Mar-16	0.984	1.039	-0.055
Feb-16	0.948	0.977	-0.029
Jan-16	0.975	0.954	0.021
Dec-15	0.971	0.954	0.017
	C: Math		
May-16	2.085	2.177	-0.093
Apr-16	2.057	2.117	-0.06
Mar-16	1.984	2.054	-0.07
Feb-16	1.902	1.957	-0.055
Jan-16	1.892	1.971	-0.079
Dec-15	1.878	1.928	-0.05
	D: Integrate		
May-16	2.524	2.238	0.286
Apr-16	2.503	2.221	0.282
Mar-16	2.397	2.112	0.285
Feb-16	2.266	2.024	0.242
Jan-16	2.201	2.016	0.184
Dec-15	2.187	1.964	0.222
	E: English		
May-16	1.434	1.469	-0.035
Apr-16	1.431	1.431	-0.001
Mar-16	1.425	1.366	0.058
Feb-16	1.343	1.331	0.012
Jan-16	1.352	1.336	0.016
Dec-15	1.348	1.332	0.016

Table 5b: Gaokao Preparation

	(1)	(2)	(3)
	Male	Female	Difference
May-16	6.017	5.978	0.039
Apr-16	5.705	5.689	0.016
Mar-16	5.455	5.459	-0.005
Feb-16	5.102	4.919	0.184
Jan-16	5.034	5.044	-0.01
Dec-15	4.949	5.007	-0.059

Table 5c: Study Effectiveness

	(1)	(2)	(3)
	Male	Female	Difference
May-16	5.815	5.944	-0.129
Apr-16	5.582	5.77	-0.188
Mar-16	5.401	5.415	-0.014
Feb-16	5.02	5	0.02
Jan-16	4.901	5.119	-0.218
Dec-15	4.938	5.03	-0.092

Note. Table 5A: we asked students "How many hours, on average, did you spend studying each day?" for each month from December 2015 to May 2016. The results for each of the examination components are listed in the panels (A to E). Table 5B: we asked students to rank "On a scale of 1 to 10, how prepared were you for the Gaokao exam? (1: unprepared, 10: very prepared)." Table 5C: we examine students' evaluation of the effectiveness of their study efforts on a scale of 1 to 10 with 1 for very ineffective and 10 for very effective. The mean values for males, females and their differences are reported in each table.

Table 6A: Mock Exam Reference Cut-offs for each University Tier In 2008

	Science Stream	Arts Stream
Tier 1	540	570
Tier 2	470	500
Tier 3	420	460
Technical	300	340

Table 6B: Fraction of Students in Fujian Scoring above each of the Reference Cut-offs in the Mock Exam in 2008

Science Stream	% of sample	Arts Stream	% of sample
≥ 540	0.088	≥ 570	0.036
≥ 470	0.274	≥ 500	0.218
≥ 420	0.443	≥ 460	0.375
≥ 300	0.808	≥ 340	0.794

Table 6C: Fraction of Students in Anxi Scoring above each of the Reference Cut-offs in the Mock Exam in 2008

Science Stream	% of sample	Arts Stream	% of sample
≥ 540	0.050	≥ 570	0.022
≥ 470	0.275	≥ 500	0.237
≥ 420	0.494	≥ 460	0.427
≥ 300	0.881	≥ 340	0.857

Note. The data on the reference cut-offs are obtained from website of the Ministry of Education for Fujian Province. Table 4B: The proportion of students scoring above each of the reference cut-offs were calculated based on the distribution of mock exam test scores of all test-takers in Fujian published by the Department of Education. Table 4C: The proportion of students in Anxi scoring above each of the reference cut-offs is obtained from our merged dataset. The information can be found at the following link: http://www.qzzk.cn/wzyd.asp?NewsID=4543

Table 7: Is the Gender Gap in Performance Larger where it Matters More?

	Points from Cutoff based on Mock Exam Scores				
	(-3, +3)	(-5, +5)	(-10, -6) & (+6, +10)	(-20, -11) & (+11, +20)	Difference: Col (1) - Col (4)
	(1)	(2)	(3)	(4)	(5)
	A. Standardized Difference: Total				
Female	-0.319***	-0.254***	-0.144	-0.101*	-0.219
	[0.135]	[0.094]	[0.090]	[0.057]	[0.142]
R-squared	0.199	0.151	0.167	0.069	0.105
	B. Standardized Difference: Chinese				
Female	-0.092	-0.084	-0.080	0.051	-0.144
	[0.137]	[0.096]	[0.106]	[0.059]	[0.145]
R-squared	0.140	0.103	0.122	0.059	0.083
	C. Standardized Difference: Math				
Female	-0.126	-0.040	-0.030	-0.034	-0.092
	[0.114]	[0.087]	[0.088]	[0.058]	[0.125]
R-squared	0.191	0.116	0.139	0.062	0.090
	D. Standardized Difference: Combined Science/Arts				
Female	-0.320**	-0.283***	-0.163*	-0.128**	-0.192
	[0.127]	[0.091]	[0.090]	[0.060]	[0.136]
R-squared	0.196	0.157	0.122	0.057	0.094
	E. Standardized Difference: English				
Female	-0.241**	-0.168**	-0.092	-0.121**	-0.120
	[0.113]	[0.084]	[0.083]	[0.056]	[0.122]
R-squared	0.176	0.105	0.199	0.051	0.081
Controls:					
Age	X	X	X	X	X
School FE	X	X	X	X	X
Zipcode FE	X	X	X	X	X
Observations	296	486	444	963	1259

Note. Each cell is separate regression with the standardized difference between the Gaokao and mock exam score as the dependent variable, for each of the examination components listed in the panels (A to E), and for students at different points from the predicted cutoffs based on the mock exam scores. Column (1) restricts the sample to students within 3 points of the cutoff, Column (2) restricts the sample to students within 5 points of the cutoff and Columns (3) and (4) restrict the sample to students 6 to 10 points and 11 to 20 points from the cutoffs, respectively. Column (5) reports the difference in the estimates in Column (1) and (4). Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Table 8: Female Underperformance vs. Male Overperformance

	Standardized Scores within Own Gender Distribution: Total (Gaokao - Mock)					
	Female	Male	Female-Male	Female	Male	Female-Male
	(1)	(2)	(3)	(4)	(5)	(6)
Distance from cutoffs:						
(-3, -1)	-0.218	0.136	-0.354**	-0.197	0.117	-0.314*
	[0.144]	[0.108]	[0.180]	[0.147]	[0.113]	[0.185]
(0, 3)	-0.051	0.051	-0.102	-0.013	0.067	-0.080
	[0.137]	[0.086]	[0.162]	[0.121]	[0.089]	[0.150]
(-10, -4) & (+4, +10)	-0.032	0.013	-0.045	-0.030	0.014	-0.043
	[0.065]	[0.058]	[0.087]	[0.068]	[0.059]	[0.090]
(-20, -11) & (+11, +20)	Reference Group					
Controls	No	No	No	Yes	Yes	Yes
Observations	860	1,033	1,893	860	1,033	1,893
R-squared	0.004	0.002	0.003	0.114	0.076	0.094

Note. Each column is a separate regression of the standardized (within-gender) difference between the Gaokao and mock exam for females only (Columns (1) and (4)) and males only (Columns (2) and (5)) on indicators of the distance from the predicted cutoffs based on the mock examination. Columns (3) and (6) report the difference in coefficients for the female sample and male samples. (-3, -1) refers to a dummy variable indicating that a student scores 1-3 points below the predicted cutoff, (0, 3) refers to a dummy variable indicating a student scored 0 to 3 points above the predicted cutoff, and (-10, -4) & (+4, +10) refers to a dummy variable indicating that a student scored 4 to 10 points above or below the predicted cutoff. All the reported coefficients are relative to students who scored between 11 to 20 points from the predicted cutoff. Columns (4) to (6) include individual level controls such as age, school FE and zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Table 9: Gender Difference in Afternoon Performance in Response to Relative Performance on Morning Exam on Day 1

	Standardized Difference in Gaokao-Mock Exam Score on Afternoon Exam - Math					
			Points from Reference Cutoff based on Mock Exam Scores			
	All	All	(-3, +3)	(-5, +5)	(-10, -6) & (+6, +10)	(-20, -11) & (+11, +20)
	(1)	(2)	(3)	(4)	(5)	(6)
Relative performance on morning exam (Chinese)*Female	0.058*	0.069**	0.344**	0.187*	0.058	-0.039
	[0.030]	[0.031]	[0.137]	[0.104]	[0.112]	[0.085]
Relative performance on morning exam (Chinese)	0.114***	0.107***	-0.006	0.017	0.116*	0.086
	[0.018]	[0.018]	[0.067]	[0.056]	[0.061]	[0.065]
Observations	7,961	7,961	296	486	444	963
R-squared	0.054	0.062	0.352	0.202	0.223	0.093
Controls:						
Female dummy	X	X	X	X	X	X
Age FE	X	X	X	X	X	X
School FE	X	X	X	X	X	X
Zipcode FE	X	X	X	X	X	X
Age*Female		X	X	X	X	X
School*Female FE		X	X	X	X	X
Zipcode*Female FE		X	X	X	X	X

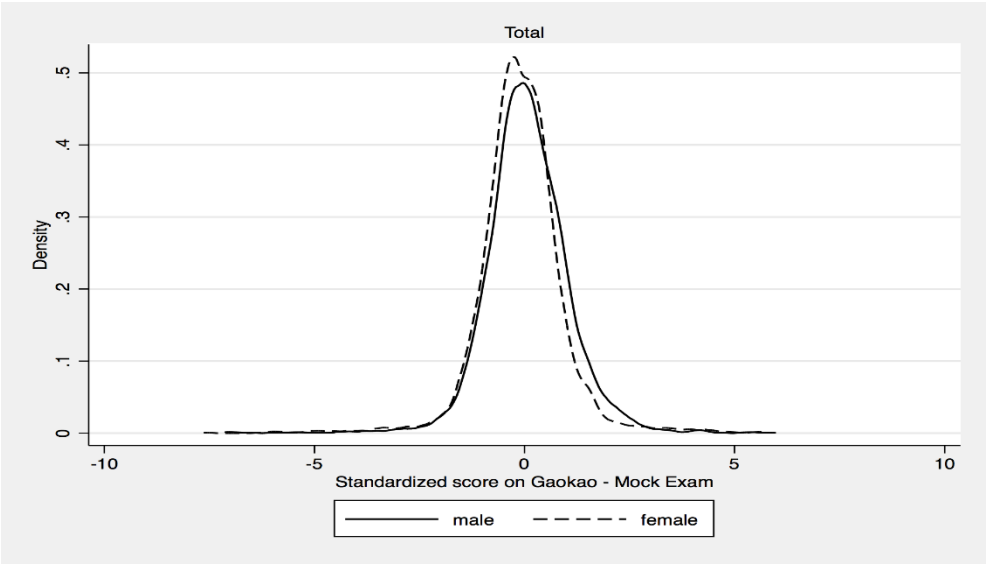
Note. Each column is a separate regression of the gender difference in the relationship between the relative performance on the Day 1 afternoon exam (Math) and the relative performance on the Day 1 morning exam (Chinese). The relative performance for both exams is measured using the standardized difference in Gaokao and mock exam scores. Columns (1) and (2) report the gender difference for the full sample while Columns (3) to (6) report the gender difference for subsets of students based on their performance on the mock examination relative to the reference cut-offs. Column (1) includes controls for a female dummy, age (in months) FE, school FE and zipcode FE. Columns (2) to (6) control for a set of fully interacted female*age (in months) FE, female*school FE and female*zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

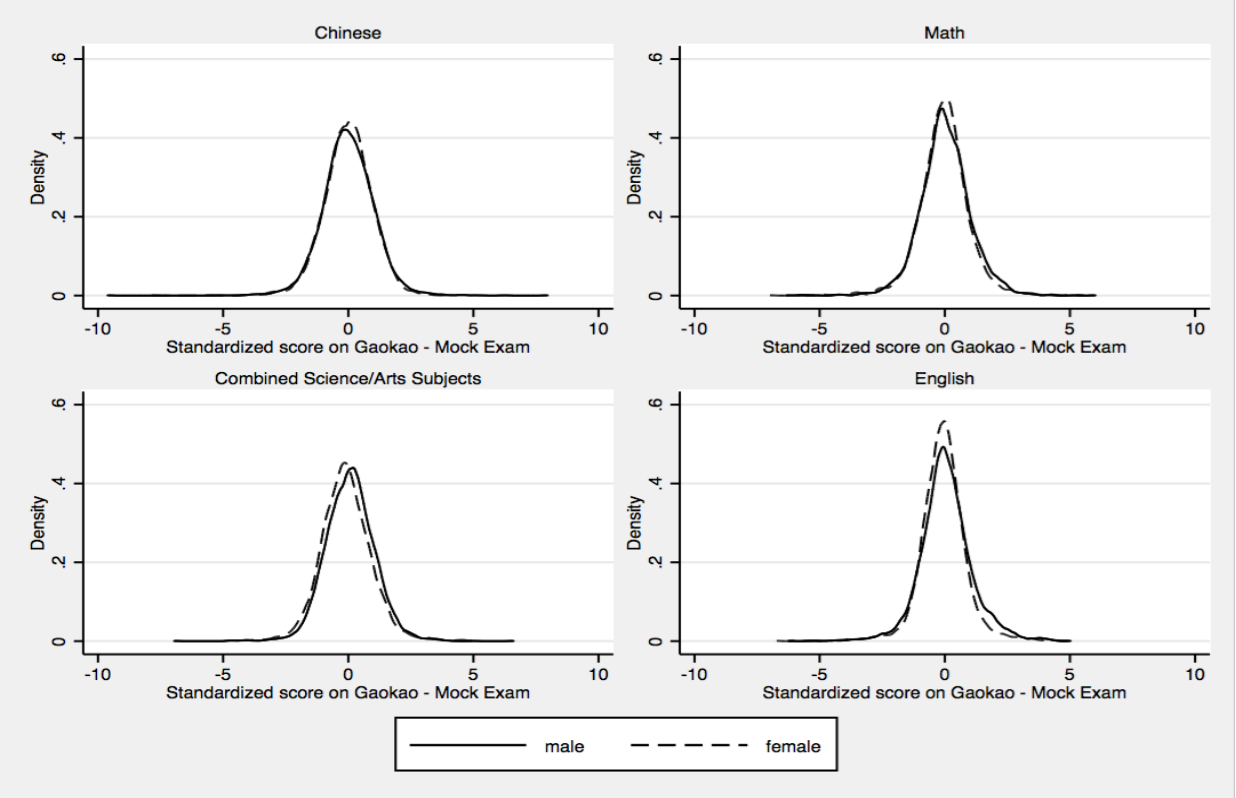
Table 10: Gender Difference in Afternoon Performance in Response to Relative Performance on Morning Exam on Day 1 - Positive vs. Negative Shocks

	Standardized Difference in Gaokao-Mock Score on Afternoon Exam (Math)				
		Points from Reference Cutoff based on Mock Exam Scores			
	All	(-3, +3)	(-5, +5)	(-10, -6) & (+6, +10)	(-20, -11) & (+11, +20)
	(1)	(2)	(3)	(4)	(5)
Positive relative performance on morning exam (Chinese)*Female	0.021	0.159	-0.207	-0.205	-0.051
	[0.054]	[0.212]	[0.187]	[0.169]	[0.130]
Negative relative performance on morning exam (Chinese)*Female	0.110*	0.430*	0.432**	0.238	-0.006
	[0.063]	[0.247]	[0.190]	[0.228]	[0.201]
Positive relative performance on morning exam (Chinese)	0.084***	-0.166	0.046	0.055	-0.062
	[0.030]	[0.141]	[0.105]	[0.108]	[0.097]
Negative relative performance on morning exam (Chinese)	0.130***	0.141	-0.006	0.180	0.223
	[0.037]	[0.152]	[0.104]	[0.124]	[0.158]
Observations	7,961	296	486	444	963
R-squared	0.063	0.376	0.219	0.241	0.102
p-value of F-test: Positive shock*Female = Negative shock*Female	0.379	0.495	0.051	0.204	0.878
Controls:					
Female dummy	X	X	X	X	X
Age FE	X	X	X	X	X
School FE	X	X	X	X	X
Zipcode FE	X	X	X	X	X
Age*Female	X	X	X	X	X
School*Female FE	X	X	X	X	X
Zipcode*Female FE	X	X	X	X	X

Note. Each column is a separate regression of the gender difference in the relationship between the relative performance on the Day 1 afternoon exam (Math) on the relative performance on the Day 1 morning exam (Chinese) allowing for the effects of relative performance on the morning exam to vary by "postive" or "negative" shocks. Positive (negative) shocks are defined as an improvement (worsening) in performance on the Day 1 morning Gaokao relative to the Day 1 morning mock exam. The relative performance for both exams is measured using the standardized difference in Gaokao and mock exam scores. Column (1) reports the gender difference for the full sample while Columns (2) to (5) report the gender difference for subsets of students based on their performance on the mock examination relative to the reference cut-offs. All regressions control for a set of fully interacted female*age (in months) FE, female*school FE and female*zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Appendix Figure 1: Distributions of Standardized Difference in Gaokao and Mock Exam Scores by Gender





Note. The sample includes students who sat for both the Mock Exam (April) and Gaokao (June). Each figure plots the distribution of standardized difference in test scores between the Gaokao and Mock Exam. Both the Gaokao score and mock exam scores were standardized to have a mean of 0 and standard deviation of 1 by student type (i.e. Science vs. Arts and Social Science Stream). The difference between the standardized Gaokao and mock exam scores were re-standardized to have a mean of 0 and standard deviation of 1. The top figure is for the total score, while the bottom figure is for each of the four subject components.

Appendix Table A1-A: Gaokao Cut-offs for each University Tier In 2008

	Science	Arts
Tier 1	534	547
Tier 2	471	487
Tier 3	428	452
Technical	320	332

Appendix Table A1-B: Fraction of Students in Fujian Scoring above each of the Cut-offs in the Gaokao in 2008

Science Stream	% of sample	Arts Stream	% of sample
≥ 534	0.087	≥ 547	0.036
≥ 471	0.288	≥ 487	0.228
≥ 428	0.449	≥ 452	0.380
≥ 320	0.786	≥ 332	0.810

Note. The data for the Gaokao cut-offs are from the following website: http://edu.qq.com/a/20080626/000135.htm. The proportion of students scoring above each of the reference cut-offs were calculated based on the distribution of Gaokao scores for all test-takers in Fujian published by the Department of Education.
See the following links for the data: Science stream (http://edu.people.com.cn/GB/116076/120173/7458397.html), Arts stream (http://edu.people.com.cn/GB/116076/120214/7458220.html).

Appendix Table A1-C: Gaokao Cut-offs for each University Tier In 2007

	Science	Arts
Tier 1	562	565
Tier 2	495	505
Tier 3	450	472
Technical	319	350

Appendix Table A1-D: Fraction of Students in Fujian Scoring above each of the Reference Cut-offs in the Gaokao in 2007

Science Stream	% of sample	Arts Stream	% of sample
≥ 562	0.087	≥ 565	0.032
≥ 495	0.290	≥ 505	0.219
≥ 450	0.447	≥ 472	0.360
≥ 319	0.798	≥ 350	0.797

Note. The data for the Gaokao cut-offs are from the following website: http://edu.qq.com/a/20080626/000135.htm. The proportion of students scoring above each of the reference cut-offs were calculated based on the distribution of Gaokao scores for all test-takers in Fujian published by the Department of Education.

See the following links for the data: Science stream (http://www.3773.com.cn/gaokao/Class149/267869.shtml), Arts stream (http://www.3773.com.cn/gaokao/Class149/267870.shtml).

Appendix Table A2: Gender Gap in the Probability of Qualifying for Tier 1 and Tier 2 Universities in the Mock Exam and Gaokao

	Mock Exam				Gaokao				Gaokao - Mock Exam	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Overall	Male	Female	Female-Male	Overall	Male	Female	Female-Male	Col (8) - (4)	Col (8) - (4)
Qualify for Tier 1	0.037	0.042	0.031	-0.011**	0.055	0.064	0.044	-0.021***	-0.010**	-0.008*
				[0.004]				[0.005]	[0.005]	[0.005]
Qualify for Tier 2	0.220	0.217	0.224	0.007	0.249	0.255	0.242	-0.013	-0.019**	-0.017*
				[0.009]				[0.010]	[0.009]	[0.009]
Qualify for Tier 1 or 2	0.257	0.259	0.255	-0.004	0.304	0.319	0.286	-0.033***	-0.029***	-0.025***
				[0.010]				[0.010]	[0.008]	[0.008]
Controls				No				No	No	Yes

Note. Columns (1) to (3) and (5) to (7) report the fraction of all, male and female test-takers that score above the thresholds for admission into each university tier for the mock exam (Cols (1) to (3)) and Gaokao (Cols (5) to (7)).Columns (4), (8), (9) and (10) are based on separate linear probability models with a dummy variable indicating that an individual met the cutoff for admission into Tier 1, Tier 2 and Tier 1 and Tier 2 universities, respectively, as the dependent variable. Column (4) and (8) report the gender difference for the mock exam and Gaokao, respectively. The last two columns report the difference in the gender gap in the probability of qualifying for each university tier for the Gaokao and mock exam without controls (Col (9)) and with controls (Col (10)). The cut-offs used to calculate the fraction of students who qualify for each university tier in the mock exam can be found in Table 4A. The cut-offs used for the Gaokao can be found in Appendix Table 1. The set of controls used in Col (10) are identical to those used in Table 3 (age, school FE and zipcode FE). Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Appendix Table A3-A: Gender Differences in Sleeping Hours

	Male	Female	Difference
May	6.849	6.685	0.164
Apr.	6.892	6.667	0.225**
Mar.	6.963	6.733	0.230**
Feb.	7.139	6.848	0.291**
Jan.	7.128	6.893	0.235**
Dec.	7.142	6.87	0.272**

Appendix Table A3-B: Gender Differences in Sleep Quality

	Male	Female	Difference
May	5.733	5.574	0.159
Apr.	5.724	5.567	0.158
Mar.	5.855	5.789	0.066
Feb.	5.989	6.189	-0.2
Jan.	6.097	6.2	-0.103
Dec.	6.278	6.341	-0.062

Appendix Table A3-C: Gender Differences in Stress Status

	Male	Female	Difference
May	7.273	7.111	0.162
Apr.	6.767	6.73	0.037
Mar.	6.256	6.193	0.063
Feb.	5.688	5.448	0.239
Jan.	5.307	5.141	0.166
Dec.	5.102	4.889	0.213

Appendix Table A3-D: Gender Differences in Number of Sick Days

	Male	Female	Difference
May	0.648	1.022	-0.374*
Apr.	0.966	1.348	-0.382
Mar.	0.824	1.481	-0.658**
Feb.	0.455	0.926	-0.472***
Jan.	0.341	0.622	-0.281*
Dec.	0.369	0.696	-0.327**

Note. The mean values for males, females and their differences are reported in each table.

Appendix Table A4: Gender Difference in Afternoon Performance in Response to Relative Performance on Morning Exam on Day 1

	A. Standardized Gaokao score on afternoon exam - Math				
		Points from Cutoff based on Mock Exam Scores			
	All	(-3, +3)	(-5, +5)	(-10, -6) & (+6, +10)	(-20, -11) & (+11, +20)
	(1)	(2)	(3)	(4)	(5)
Relative performance on morning exam (Chinese)*Female	0.059**	0.317***	0.128*	0.012	-0.067
	[0.024]	[0.102]	[0.072]	[0.075]	[0.050]
Relative performance on morning exam (Chinese)	0.004	-0.075	0.001	0.062	0.099***
	[0.017]	[0.047]	[0.038]	[0.042]	[0.037]
Observations	7,961	296	486	444	963
R-squared	0.220	0.377	0.234	0.231	0.137
	B. Standardized Mock Exam score on afternoon exam - Math				
Relative performance on morning exam (Chinese)*Female	0.013	0.088	0.004	-0.027	-0.041
	[0.023]	[0.069]	[0.049]	[0.047]	[0.035]
Relative performance on morning exam (Chinese)	-0.067***	-0.071	-0.010	-0.016	0.041*
	[0.016]	[0.045]	[0.034]	[0.037]	[0.022]
Observations	7,961	296	486	444	963
R-squared	0.202	0.273	0.161	0.148	0.143
Controls:					
Age*Female	X	X	X	X	X
School*Female FE	X	X	X	X	X
Zipcode*Female FE	X	X	X	X	X

Note. Each column in each panel is a separate regression of the gender difference in the relationship between the standardized Day 1 Math Gaokao scores (Panel A) or standardized Day 1 Math mock exam scores (Panel B) on the relative performance on the Day 1 morning exam (Chinese) interacted with the female dummy. The relative performance on the morning exam is measured using the standardized difference in Gaokao and mock exam scores. Column (1) reports the gender difference for the full sample while Columns (2) to (5) report the gender difference for subsets of students based on their performance on the mock examination relative to the reference cut-offs. All regressions control for a set of fully interacted female*age (in months) FE, female*school FE and female*Zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Appendix Table A5: Gender Difference in Last Exam (English) in Response to Cumulative Relative Performance on Previous 3 Exams (Chinese + Math + Combined subjects)

	Standardized exam score on final Gaokao exam - English				
		Points from Cutoff based on Mock Exam Scores			
	All	(-3, +3)	(-5, +5)	(-10, -6) & (+6,	(-20, -11) & (+11,
	(1)	(2)	(3)	(4)	(5)
Relative performance on Previous 3 Exams (Chinese + Math+ Combined)*Female	0.034	0.283**	0.176*	0.156*	-0.048
	[0.022]	[0.121]	[0.106]	[0.086]	[0.061]
Relative performance on Previous 3 Exams (Chinese + Math + Combined)	0.081***	0.115	0.140**	0.165***	0.232***
	[0.015]	[0.073]	[0.069]	[0.062]	[0.035]
Observations	7,961	296	486	444	963
R-squared	0.278	0.410	0.298	0.359	0.250
	Standardized exam score on final mock exam - English				
Relative performance on Previous 3 Exams (Chinese + Math + Integrate)*Female	-0.023	0.035	-0.016	-0.044	-0.114**
	[0.020]	[0.081]	[0.063]	[0.082]	[0.050]
Relative performance on Previous 3 Exams (Chinese + Math + Integrate)	-0.104***	-0.002	0.015	0.089	0.094***
	[0.014]	[0.065]	[0.050]	[0.062]	[0.034]
Observations	7,961	296	486	444	963
R-squared	0.296	0.333	0.271	0.236	0.217
Controls:					
Age*Female	X	X	X	X	X
School*Female FE	X	X	X	X	X
Zipcode*Female FE	X	X	X	X	X

Note. Each column in each panel is a separate regression of the standardized Day 2 English afternoon Gaokao scores (Panel A) or the standardized Day 2 English mock exam scores (Panel B) on the relative performance on the previous 3 examinations (Chinese, Math and Combined Subjects) interacted with the female dummy. The relative performance on the previous 3 examinations is measured using the standardized difference in Gaokao and mock exam scores. Column (1) reports the gender difference for the full sample while Columns (2) to (5) report the gender difference for subsets of students based on their performance on the mock examination relative to the reference cut-offs. All regressions control for a set of fully interacted female*age (in months) FE, female*school FE and female*zipcode FE. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.