Gender Gap under Pressure: Performance and Reaction to Shocks*

Xiqian Cai Yi Lu Jessica Pan Songfa Zhong

National University of Singapore

November, 2014

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. We also examine whether there are gender differences in the responsiveness to performance shocks by examining how relative performance on the morning exam of the *Gaokao* affects females' and males' performance on the afternoon exam. Consistent with the idea that females are more easily discouraged, we find that females are more responsive to a given change in relative performance on the morning exam as compared to males. This is particularly true for negative shocks. Notably, the gender difference in performance and reaction to shocks are more pronounced for subgroups of students where the stakes matter more.

*We thank seminar participants at the National University of Singapore and the Trans-Pacific Labor Seminar for helpful comments. Corresponding author: jesspan@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 women may simply prefer to opt-out from competition (for examples, see Bertrand, 2010, Gneezy, Niederle and Rustichini, 2003, Gneezy and Rustichini, 2004, Niederle and Vesterlund, 2007). These studies posit that gender differences in performance and attitudes in toward competition may explain an important part of the gender gap in educational choices and labor market outcomes (Buser, Niederle and Oosterbeek, 2012).

A growing line of research has attempted to assess whether such performance differences exist in real-world settings. Interestingly, the results are somewhat mixed. Earlier studies by Lavy (2008) 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 (Ors, Palomino and Peyrache, 2013, Jurajda and Munich, 2011 and Attali, Neeman and Schlosser, 2011).

This paper attempts to add to this literature by examining two related questions. We utilize a unique dataset from the National College Entrance Examination (*Gaokao*) and an arguably cleaner empirical setup to examine the gender gap in examination performance in response to competitive pressures. We are able to directly observe the performance of the same individual in a real high-stakes college entrance examination and an earlier low-stakes mock examination - by comparing the gender difference in performance in what is essentially the same examination in a high and low-stakes setting, we can provide a clean estimate of the effect of competitive pressures on the gender gap in performance in an important real world setting. Consistent with previous evidence, we find that males are significantly more likely to outperform females in a high pressure setting relative to a less competitive setting. Next, we extend the previous literature by analyzing whether there is a gender difference in the reaction to performance shocks. The latter is also an important question in the literature on gender gaps in performance as there is increasing evidence that women are less confident (and men tend to be overconfident) and may react more aversely to negative feedback than their male counterparts (Roberts and Nolen-Hoeksema, 1989 and Goldin, 2013). This lower degree of confidence and resistance to negative shocks may be another reason as to why women are more likely to opt-out of competitive fields (Niederle and Vesterlund, 2007). 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. To our knowledge, the latter question remains unexplored in the empirical literature.

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 exceed 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. In fact, the examination is so important that a couple of months prior to the actual examination, each province typically runs a mock examination 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.

It is worth noting that our setting has two key features that differentiate it from similar empirical tests that explore the same question. First, the same examination board sets and implements both the high and low-stakes and the coverage of the test material is identical in both settings.[1] 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.

We find strong evidence that females tend to do relatively poorer as compared to men on the high-stakes *Gaokao*, relative to their performance on the low-stakes mock examination. The gender differences are particularly striking for English and the Combined Science/Arts subjects. While

[1] In most of the previous studies (e.g. Ors, Palomino and Peyrache, 2013), the low-stakes and high-stakes settings considered typically involve different testing strategies and material, 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.

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. In terms of the individual subjects, the performance declines for females are -0.004, -0.036, -0.17 and -0.12 standard deviations for Chinese, Mathematics, Combined subjects and English, respectively. These patterns are consistent with females facing a greater degree of stress or test anxiety as compared to men.

An alternative interpretation of this gender differential in performance on the *Gaokao* relative to the mock examination is that male students take the mock examination less seriously as it is a low-stakes test, but put in their full effort on the actual examination.[2] 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. While we cannot fully dispel this alternative story, we provide some evidence that suggests that this mechanism is unlikely to be driving our results. 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 bring some empirical evidence to bear on this issue, first, we show that the female performance declines are observed across all tertiles of the test score distribution and are larger in the upper tail. To the extent that ability correlates with effort and that lower ability male students are less likely to exert effort on the test, we might have expected to find larger performance gaps among low ability students. Overall, the ubiquity of the gender difference in performance for both low and high ability students makes it less likely that the observed performance declines are driven exclusively by males putting in less effort on the low-stakes examination. Second, we exploit the fact that the

[2] 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 the larger performance differential across the low vs. high-stakes test for males and whites relative to females and other demographic groups.

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 "stakes" involved in the examinations. Students who are closer to the entry thresholds for each university tier are more likely to face higher stakes on 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 high(er) stakes 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 0.17 standard deviation larger decline in performance on the *Gaokao* relative to the mock examination compared to male students, whereas this difference is only 0.08 and 0.05 standard deviations for the groups of students 6 to 10 points and 11 to 20 points away from the cutoff, respectively. Finally, we use this additional variation in stakes to look at how performance varies for students as a function of their distance from the reference cutoffs *within* gender. We find some evidence that females who are close to the reference thresholds tend to perform relatively worse compared to those who are further from the entry thresholds. By contrast, we find that males tend to perform relatively better when they are closer to the reference threshold. Overall, these results suggest that the empirical patterns are unlikely to be driven entirely by lower effort provision of males in the mock exam. The results suggest that the gender differential in performance on the *Gaokao* relative to the mock examination is likely to be driven, in part, by female underperformance and males' improved performance in high-stakes settings.

Next, we exploit a unique feature of the *Gaokao* to test whether males and females respond differently to unanticipated shocks in performance. Evidence from social psychologists and, more recently, economists, suggest that women tend to be less confident than men and may be more easily discouraged by poor performance.[3] For example, in an editorial for Bloomberg, Claudia Goldin makes the observation that women appear to be less likely to major math-heavy fields such as economics (and engineering) as they are often discouraged by not getting sufficiently high grades in introductory courses.[4]

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.[5] 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 on 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 on the morning exam is associated with a 0.11 standard deviation lower relative performance on the afternoon examination for males. 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.

One concern with these estimates is whether they truly measure gender differences in the response to shocks in initial performance or whether they reflect unobserved factors that may be differentially correlated with the performance of female and male students. We provide a number of empirical tests that attempt to address this issue - first, we show that all our specifications are robust to the inclusion of a set of demographic controls (age, school fixed effects and zip code fixed effects) that are fully interacted with the female dummy. At the very least, this suggests that these observed characteristics are not differentially correlated with the relative performance of females and males. Second, we also find that the gender differences in the reaction to relative performance shocks are more pronounced for students who are close to the reference cutoffs (based on the mock examination scores) for admission into the different university tiers compared to students who are further away

[3] A non-technical summary of some of the studies can be found in this article: `http://www.theatlantic.com/features/archive/2014/04/the-confidence-gap/359815/`

[4] `http://www.bloombergview.com/articles/2013-10-14/can-yellen-effect-attract-young-women-to-economics-`

[5] 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.

from the cutoffs. This is consistent with the idea that female responses to positive/negative shocks are larger when the stakes larger. Third, we show that the relative performance shocks on the morning exam differentially affects female students' performance on the future *Gaokao* exam but is uncorrelated with the performance on the same subject in the (past) mock examination. Finally, we show that females are more affected by negative performance shocks relative to positive performance shocks. Overall, these results suggest that our findings are less likely to be accountable for by unobserved factors correlated with future performance that vary differentially by gender.

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. The results on the gender gap in the performance in the *Gaokao* relative to the mock examination are presented in Section 4. Section 5 examines the gender difference in the reaction to performance shocks. Section 6 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 intuitions and the duration of the programs (Davey et al, 2007).

The *Gaokao* is ultimately under the control of the Ministry of Education 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

[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.

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] 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).

A couple of 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 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 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.

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

[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

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 the universe of individuals who sat for the mock examination in Anxi, a county of Fujian Province, in 2008. Appendix I provides some background information on the social and economic characteristics of Anxi county and Fujian province. The 2008 Provincial Mock 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 was a good indication of the degree of difficulty and subject material covered by the actual examination. Furthermore, the Department of Education 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.[11] 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.[12]

Our dataset was constructed by merging the mock examination scores to the *Gaokao* scores using each test-taker's name.[13] From the *Gaokao* sample, we further dropped a small number (11) of

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

[11] Appendix Table 1 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.

[12] 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.

[13] Individuals with the same first name and last name were dropped as they could not be uniquely identified. This accounted for 94 mock examination candidates and 101 *Gaokao* candidates.

individuals for who 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. 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.[14] There were approximately 260 more candidates who sat for the *Gaokao* than the mock examination and, on average, the 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 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 *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 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 exam components, namely, the combined science/arts subject, differs 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 female students 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

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

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. In the remaining rows, we report the same statistics for each of the individual subject components - as all students, regardless of stream, sit for the same Chinese, Math and English exams, we report the means for these subjects for all students. The combined Science and Arts subjects are reported separately for students in the Science and Arts stream, respectively. 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.* Male students perform about 3 points better on the mock exam relative to females; on 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.* The relative decline in female performance across the two exam settings is observed for all subjects and is largest for the Combined Arts subject (-6 points) and English (-3 points). 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), 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[15] 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

[15] 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*.

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.[16]

4.2 Results

Table 3A 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. 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. This difference is large and statistically significant and is virtually unaffected by the addition of covariates (see Column (2)). As observed in Columns (3) to (10), most of this 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 2 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*

[16] 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).

and mock exam with and without individual-level controls.[17] 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 2 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.

Next, we evaluate two potential mechanisms that are consistent with the observed gender gap in performance on high-stakes exams. The first possibility is that, relative to females, males may take the mock examination less seriously and only put in their full effort in the *Gaokao* when the stakes actually matter (see, for example, Attali, Neeman and Scholsser, 2011). The second possibility is that males and females respond differently to competitive and stressful environments such as the *Gaokao*. We provide several pieces of evidence, both empirical and qualitative, that appear to support the latter mechanism.

Gender Differences in Effort Provision

If male students take the mock examination less seriously than female students, this 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. While we are unable to entirely rule out the possibility that the observed empirical patterns are driven by gender differences in effort provision given that we lack direct information on the effort exerted on the test (e.g. time taken to complete the examination, number of questions attempted), we provide some evidence that suggests that our results are unlikely to be due to lower effort by males in low-stakes settings.

First, 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 time that students can get

[17] The set of controls are the same as that used in Table 3A.

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.[18] As such, students have an incentive to put in their full effort on the mock exam.

To examine this issue empirically, we examine whether the gender gap in performance varies by students' ability. One might expect that if there are indeed gender differences in effort provision that low ability males might be the group that is most likely to reduce their full effort on the mock exam. If this is true, this would suggest that we would observe larger gender performance gaps among low ability students. To examine this possibility, we divide the sample into three groups based on students' overall performance on the mock exam. Table 3B reports similar estimates as Table 3A for students in the bottom third (Column (1)), middle third (Column (2)) and upper third (Column 3)) of the overall test score distribution on the mock exam. We find that, compared to males, females appear to underperform on the *Gaokao* relative to the mock exam in all three groups, with the relative performance declines more apparent among higher ability students. For example, among students in the bottom third based on mock exam scores, females performed -0.09 standard deviations worse on the *Gaokao* relative to the mock examination compared to males. Among students in the upper third, the female performance decline was about 50% larger at -0.13 standard deviations. These results indicate that the relative underperformance of females in high-stakes settings is observed for students of all abilities and tends to be more pronounced for higher ability students. This suggests that the overall patterns are unlikely to be driven by lower effort provision among male students.

Next, 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 "stakes" 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 4A lists the reference cutoffs for each stream in 2008. Table 4B and

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

4C 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 1 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*.[19] This suggests that 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.[20] 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 5 reports the estimates of equation (1) for four different subgroups of students - those with

[19] 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 1D) is likely to be due to 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.

[20] From Table 4B, 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.

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.[21] These qualitative patterns are similar for the individual subjects (Panels (B) to (E)). Overall, these results provide additional evidence in support of the idea that women underperform relative to men when subject to more competitive pressures.

Do Males and Females Respond Differently to Pressure?

Next, we turn to the possibility that men and women respond differently to competitive and stressful environments such as the *Gaokao*. This explanation is consistent with a large number of biological studies on gender differences in stress responses (Taylor et al., 2000, Lee and Harley, 2012). 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, 1992). Moreover, gender differences in stress response could also potentially reconcile why we only observe the widening of the gender gap in performance in response to competition for Day 2 subjects. To cope with challenging situations, the human body consumes a lot of energy (Sapolsky, 1994) - therefore, the gender difference in the response to stress could be

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

more pronounced as students are subject to prolonged stress. Consistent with this interpretation, we also find the largest gender performance gaps for the combined subject exam which is half an hour longer than the other subject tests.

On the basis of this explanation, there are two possible ways that men and women might differ in performance under pressure. First, female performance may be worse in more competitive settings as they are less able to deal with the challenges associated with high-stakes settings. Alternatively, males may be better able to rise to the challenge when faced with higher pressure and competition. The latter explanation is consistent with the experimental literature that has shown that men tend to perform better in more competitive situations (Gneezy, Niederle and Rustichini, 2003). To separate these two effects, we 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.

The first three columns of Panel A in Table 6 reports the mean of the total test scores for female and male students for subgroups of students classified based on the distance from the reference cutoff. For all three subgroups, female typically do worse on the *Gaokao* relative to the mock examination, with the score gap generally decreasing for female students further from the thresholds. Female students score an average for about 4 to 6 points lower on the *Gaokao* relative to the mock exam when they are within 3 points of the cutoffs compared to when they are 11 to 20 points from the cutoff. Strikingly, this pattern is reversed for male students - while males within 3 points of the threshold score about 1 point higher on the *Gaokao* relative to the mock examination, those who are 11 to 20 points from the cutoff score close close to 4 points lower on the *Gaokao*. Columns (4) and (5) report the difference of the estimates between Columns (3) and (1) without controls and with the full set of controls used in the previous tables. While the within gender differences are not significant at conventional levels, the female-male difference of the test score difference between the *Gaokao* and mock exam for students within 3 points and 11 to 20 points of the cutoffs is marginally significant at the 10% level without controls. Adding in the full set of controls reduces the coefficient and statistical significance, but the qualitative findings remain largely similar. Panel (B) reports similar estimates using standardized scores (by type) within each students' own gender distribution.

In sum, these patterns suggest 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. 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.

5 Gender Gap in Reaction to Shocks

In this section, 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. More specifically, 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:

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

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 affects the relative performance of females on the afternoon examination relative to males. Before turning to the regression estimates, Figure 2A presents a 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.

Before turning to the regression estimates, it is useful to note that we do not find any evidence of gender differences in the incidence of relative performance shocks on the Day 1 morning exam (Chinese). As shown in Table 3A, females do not appear to underperform relative to males in the Chinese *Gaokao* relative to the mock exam. This is important as it suggests that, on average, females and males appear to experience a similar change in relative performance on the morning exam. Column (1) of Table 7 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. These results are broadly consistent with the idea that women may be more affected by negative feedback or performance (Roberts and Nolen-Hoeksema, 1989 and Wozniak 2012).[22]

One potential concern with these estimates is whether they truly measure gender differences in the response to shocks in initial performance or whether they reflect unobserved factors that may be differentially correlated with the performance of female and male students. It is important to note that our results are not affected by common shocks (e.g. health shocks) that may potentially affect

[22] 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 bolster men against performance shocks.

performance on both components of the test since our empirical specification asks whether females and males who experience the same relative performance shock on the morning exam differed in their afternoon performance. For an unobserved factor to account for our results, it has to be the case that it differentially affects the performance of female students on the afternoon exam even among males and females with similar relative performance on the morning exam. One example of such a factor could be the menstrual cycle (see Wozniak, Harbaugh and Mayr, 2014 and Buser, 2012).

While it is impossible to fully dispel this concern due to our lack of exogenous variation in the initial performance shock, we provide a number of empirical observations that provide suggestive evidence that our results are not entirely driven by unobservables. First, as discussed above, we showed that the estimates are not sensitive to the addition of a full set of gender-interacted controls. This suggests that, at the very least, the set of observable characteristics (school and location fixed effects and age) are not differentially correlated with the relative performance of females and males.

Next, we use the distance from the reference cutoffs to generate additional variation in the extent to which performance on the *Gaokao* matters for different groups of students. Given that students closer to the threshold are more likely to be faced with 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 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 7 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 close 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.[23]

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 3 indicate that the gender differences in the reaction to performance shocks documented in Table 7 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

[23] 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.

of our findings. The results for different subgroups of students based on their distance to the entry cutoffs reported in Columns (2) to (5) of Appendix Table 3 are consistent with our previous findings and a causal interpretation.

Finally, in Table 8, 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. In a high pressure settings, female performance appears to be more detrimentally affected by negative shocks relative to males.

6 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 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 translates 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. Moreover, consistent with the idea that females tend to underperform when the stakes are higher, we show that the gender performance gap is largest among students whose mock examination scores are right around the reference thresholds that determine whether students are eligible to apply for admission into the top two tiers of universities.

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 show that the gender performance gap is largest among the high ability students - to the extent that high ability students are more likely to exert effort on the mock exams, this suggests that the observed empirical patterns are unlikely to be driven exclusively by lower effort provision among male students in the mock exam. Moreover, we use the variation in stakes 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 stakes settings. These findings are consistent with the findings in the experimental literature and a recent paper by Ors, Palomino and Peyrache (2012) that studies a similar question in the French higher education system.

In the second part of the paper, 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. Evidence from behavioral psychologists and, more recently, economists, suggest that women tend to be less confident than men and may be more easily discouraged by poor performance. Nevertheless, to our knowledge, this relationship has not been systematically tested in a field setting. Consistent with this line of reasoning, we find that relative to males, female relative performance on the first day's afternoon *Gaokao* is more strongly correlated with relative performance on the morning exam, as measured by the deviation of the *Gaokao* morning exam score from the mock exam score. In response to a given change in relative performance on the morning examination, female relative performance in the afternoon exam appears to be more affected than male relative performance. Consistent with the results found in the previous part of the paper, we also find that the gender gap 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. Furthermore, we provide some evidence that the gender difference in responsiveness to initial relative performance is more pronounced for negative performance shocks relative to positive shocks.

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 Kleinjans, 2009). This idea that females are more likely to be easily "discouraged" compared to males have 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.

	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.

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.

	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)

References

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

[2] 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.

[3] 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.

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

[5] Buser, Thomas. 2012. “The Impact of the Menstrual Cycle and Hormonal Contraceptives on Competitiveness.” Journal of Economic Behavior and Organization. Vol 83(1): 1-10.

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

[7] Croson, Rachel and Uri Gneezy. 2009. “Gender Differences in Preferences.” Journal of Economic Literature. Vol 47(2): 448-74.

[8] 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.

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

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

[11] 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-

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] Ors, Evren, Frdric Palomino, and Eloc Peyrache. ”Performance Gender Gap: Does Competition Matter?.” Journal of Labor Economics 31.3 (2013): 443-499.

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

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

[21] 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.

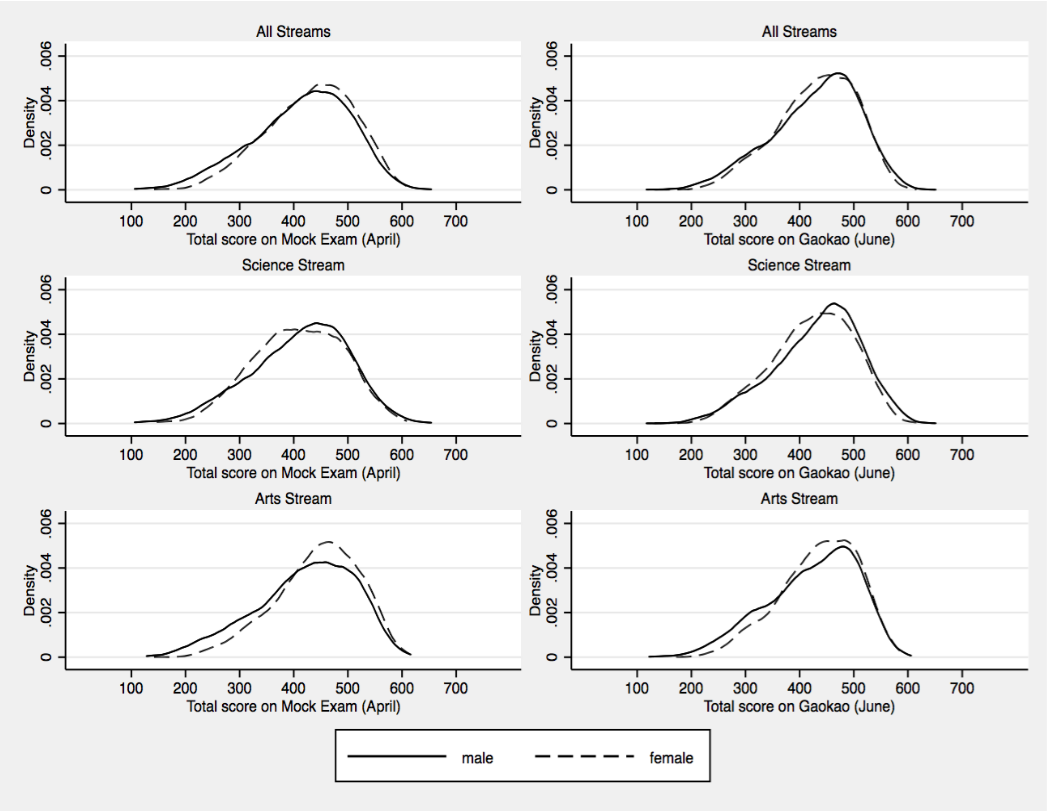
[22] Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A., and Updegraff, J. A. 2000. Biobehavioral Responses to Stress in Females: Tend-and-Befriend, not Fight-or-Flight. Psychological Review. Vol 107(3): 411.

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

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

[25] Wozniak, David, William T. Harbaugh and Ulrich Mayr. 2014. "The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices." Journal of Labor Economics. Vol 32(1): 161-198.

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

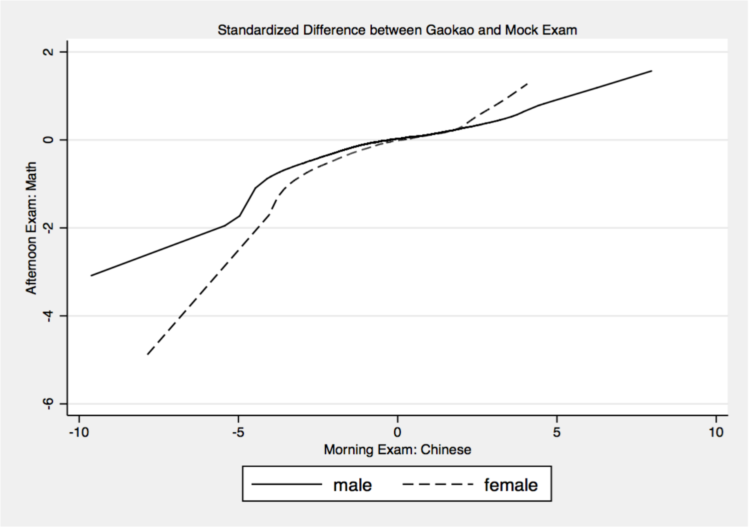
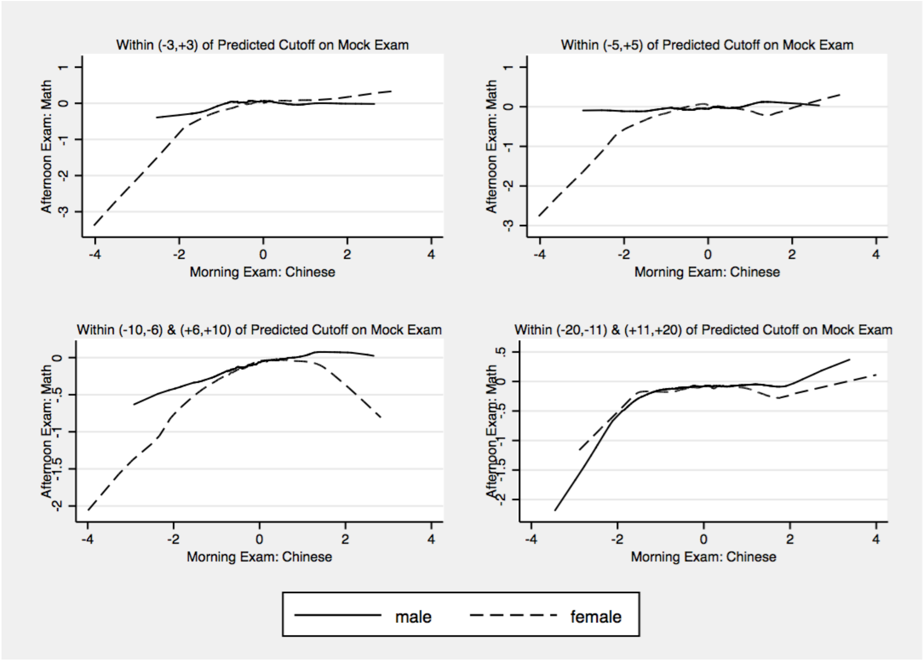


Figure 2B: Relationship between Day 1 Afternoon Exam Score and Morning Exam Score by Gender and Distance from Reference Cutoff



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)		
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]
Obs	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]
Obs	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]
Obs	2193	1487	3680	2193	1487	3680	3,680
Chinese	93.51	89.52	3.990***	99.07	95.70	3.367***	-0.622***
	(9.58)	(11.85)	[0.246]	(8.77)	(10.06)	[0.214]	[0.214]
Obs	3563	4398	7961	3563	4398	7961	7961
Math	93.77	94.86	-1.094*	95.22	96.74	-1.521***	-0.426
	(24.29)	(25.97)	[0.569]	(25.44)	(25.77)	[0.578]	[0.374]
Obs	3563	4398	7961	3563	4398	7961	7961
Combined Science	132.42	142.68	-10.264***	141.19	152.67	-11.485***	-1.221
	(42.02)	(45.77)	[1.461]	(36.64)	(39.72)	[1.270]	[0.973]
Obs	1370	2911	4281	1370	2911	4281	4,281
Combined Arts	170.71	169.40	1.307	154.74	159.42	-4.682***	-5.989***
	(31.35)	(35.74)	[1.115]	(27.49)	(31.22)	[0.976]	[0.792]
Obs	2193	1487	3680	2193	1487	3680	3,680
English	86.64	77.98	8.665***	87.28	81.60	5.686***	-2.979***
	(21.56)	(24.27)	[0.521]	(19.26)	(21.24)	[0.459]	[0.330]
Obs	3563	4398	7961	3563	4398	7961	7961

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. "Combined Science" and "Combined Arts" is the score on the combined examination 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 3A: 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)
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]
Controls:										
Age		X		X		X		X		X
School FE		X		X		X		X		X
Zipcode FE		X		X		X		X		X
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

Note. Each column is a separate regression of the standardized difference between the Gaokao (June) score and the mock exam (April) score for each of the subjects listed in the table. 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 3B: Regression Estimates of the Gender Gap in Performance by Initial Performance on Mock Exam

	Dependent Var: Standardized Difference between Gaokao and Mock Exam Score		
	Sample based on performance on mock exam:		
	Bottom Third	Middle Third	Upper Third
	(1)	(2)	(3)
Female	-0.085*	-0.121***	-0.131***
	[0.046]	[0.035]	[0.035]
Controls:			
Age	X	X	X
School FE	X	X	X
Zipcode FE	X	X	X
Observations	2,658	2,655	2,648
R-squared	0.065	0.049	0.049

Note. Each column is a separate regression of the standardized difference between the Gaokao (June) score and the mock (April) score for students in the bottom third (Column (1)), middle third (Column (2)) and upper third (Column (3)) of the distribution of total performance on the mock exam. 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 4A: 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 4B: 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 4C: 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 5: 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 of the standardized difference between the Gaokao and mock exam score for each of the examination components listed in the panels (A to E) for students at different points from the predicted cutoffs based on the mock exam scores. Column (1) is restricted to students within 3 points of the cutoff, Column (2) is restricted to students within 5 points of the cutoff and Columns (3) and (4) are restricted 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 6: Do Females Underperform where it Matters More?

	Points from Cutoff based on Mock Exam Scores			Difference between Col (1) and (3)	
		(-10, -4) &	(-20, -11) &		
	(-3, +3)	(+4, +10)	(+11, +20)	No Controls	With Controls
	(1)	(2)	(3)	(4)	(5)
	A. Raw Test Scores: Total (Gaokao - Mock)				
Female	-19.66	-16.16	-13.66	-6.000	-4.214
	(46.47)	(39.15)	(36.73)	[4.483]	[4.556]
Observations	126	289	445	571	571
Male	1.18	-2.01	-3.63	4.816	2.782
	(36.14)	(34.19)	(37.98)	[3.232]	[3.379]
Observations	170	345	518	688	688
Female-Male				-10.816*	-6.996
				[5.527]	[5.668]
Observations				1259	1259
	B. Standardized Scores within Own Gender Distribution: Total (Gaokao - Mock)				
Female	-0.350	-0.250	-0.218	-0.132	-0.087
	(1.08)	(0.90)	(0.82)	[0.103]	[0.105]
Observations	126	289	445	571	571
Male	-0.166	-0.234	-0.248	0.081	0.071
	(0.80)	(0.80)	(0.88)	[0.073]	[0.078]
Observations	170	345	518	688	688
Female-Male				-0.213*	-0.159
				[0.126]	[0.131]
Observations				1259	1259

Note. Columns (1) to (3) report the mean and (standard deviation) of the difference in test scores between the Gaokao and mock examination for students within 3 points, 4 to 10 points and 11 to 20 points of the predicted cutoffs. Panel A reports the raw test scores while Panel B reports the scores that are standardized by gender and by type (arts vs. science stream). Column (4) reports the regression estimates of the difference between Col (1) and Col (3) without controls. Column (5) reports a similar difference, controlling for the full set of controls in Table 5 interacted with the female dummy. For the estimates in Columns (4) and (5) standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Table 7: 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

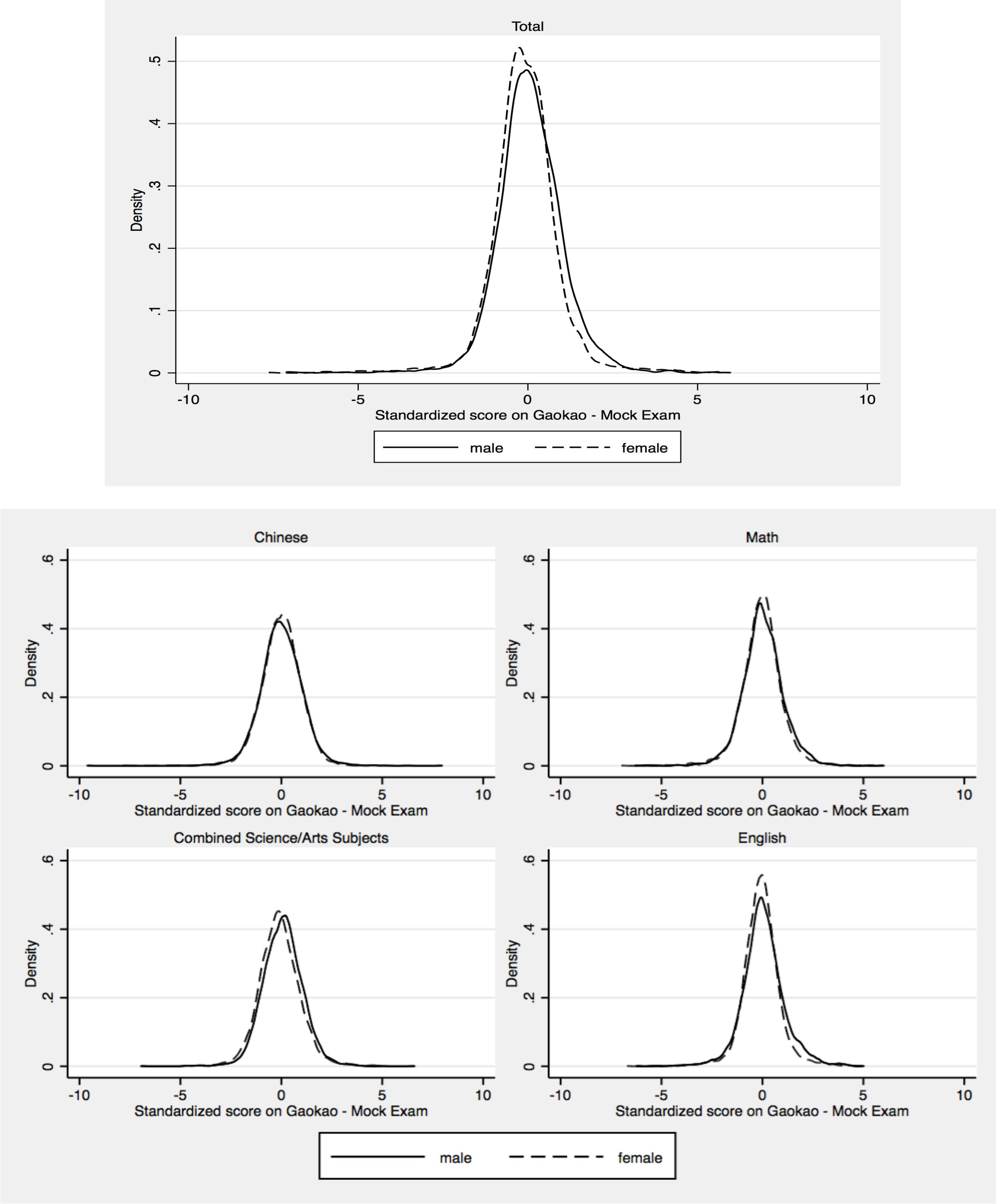
Note. Each column is a separate regression of the gender difference in the relationship between the relative performance on the Day 1 morning exam (Chinese) and the relative performance on the Day 1 afternoon exam (math). The relative performance for both exams is measured using the standardized difference 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 8: 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 morning exam (Chinese) and the relative performance on the Day 1 afternoon exam (math) 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 1A: 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 1B: 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 1C: 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 1D: 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 2: 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 3A. Robust standard errors are reported in brackets. ***significant at 1%, **5%, *10%.

Appendix Table 3: Gender Difference in Afternoon Performance in Response to Relative Performance on Morning Exam on Day 1

	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
	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 relative performance on the Day 1 morning exam (Chinese) and the standardized Day 1 math Gaokao (Panel A) and Mock Exam (Panel B) scores. 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%.