ISSN 2042-2695

CENTRE *for* ECONOMIC PERFORMANCE

CEP Discussion Paper No 1314

Revised, October 2015
(First published, November 2014)

Gender Differences in Response to Big Stakes

Ghazala Azmat, Caterina Calsamiglia and Nagore Iriberri

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract
It is commonly perceived that increasing incentives improves performance. However, the reaction to increased incentives might differ between men and women, leading to gender differences in performance. In a natural experiment, we study the gender difference in performance resulting from changes in stakes. We use detailed information on the performance of high-school students and exploit the variation in the stakes of tests, which range from 5% to 27% of the final grade. We find that female students outperform male students in all tests—but to a relatively larger degree when the stakes are low. The gender gap disappears in tests taken at the end of high school, which count for 50% of the university entry grade.

Keywords: Stakes, gender gaps, performance
JEL codes: D03; J16; I21; C30

This paper was produced as part of the Centre's Labour Markets Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Ghazala Azmat acknowledges financial support from ECO2011-30323-C03-02. The authors acknowledge support from the Institut Català de la Dona (U-8/10). Caterina Calsamiglia acknowledges financial support from the Ministerio de Economía y Competición (SEJ2005-01481, SEJ2005-01690, SEJ2006-002789-E and FEDER) and from the "Grupo Consolidado de tipo C" (ECO2008-04756), the Generalita de Catalunya (SGR2005-00626) and the Severo Ochoa program. Nagore Iriberri acknowledges financial support from Ministerio de Economía y Competición (ECO2012-31626), Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT869-13) y Ministerio de Ciencia e Innovación (ECO2011-25295).

Ghazala Azmat is an Associate at the Centre for Economic Performance, London School of Economics. She is also a Reader (Associate Professor) at the School of Economics and Finance, Queen Mary College, University of London. Caterina Calsamiglia is an Associate Professor at Universitat Automoma de Barcelona and an affiliated professor, Barcelona GSE. Nagore Iriberri is Ikerbasque Research Professor at the University of the Basque Country UPV/EHU, IKERBASQUE, Basque Foundation for Research.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1. Introduction

Pressure is a defining feature of many social and economic environments. Taking final exams in school, undergoing the last round of a job interview, giving a speech and answering questions at a press conference are some examples of situations with high stakes. In many instances, principals (evaluators), whether in a competitive or noncompetitive setting, use a one-shot process to gather information or evaluate the agent. This process is likely to induce pressure since agents understand that they will not have the opportunity to repeat the process—or that doing so will be costly. It is typically assumed that increased incentives lead to improved performance. However, men and women might react differently to increased incentives, engendering gender differences in performance.

In this paper, we provide empirical evidence showing that men and women react differently to increasing pressure, as defined by the size of the stakes at hand. We use detailed information from a high-performing private school in Barcelona, Spain, over a period of 12 years for several cohorts of high school students who take numerous tests with varying stakes. We follow several cohorts of students through their six years of high school (ages 12 to 18). For all subjects in each academic year (typically ten or 11 subjects), students undertake three types of tests with varying stakes (low, medium, and high). In particular, during the year, students undertake several low-stakes tests, two medium-stakes tests at the end of each semester, and then a high-stakes test at the end of the academic year. The test stakes vary from five percent to 27 percent of the final grade, which, for each subject, is the weighted average of all the tests taken throughout the academic year. At the end of high school, students undertake national-level (standardized) tests similar to the SAT in the United States, for which the stakes are very high as students' grade in the national-level test counts for 50 percent of the university entry grade.

The analysis shows that female students outperform male students by 0.18 standard deviations of the mean in low-stakes tests but by only 0.11 standard deviations in high-stakes tests. Moreover, in the national-level exams, the gap is reversed, such that male students outperform female students by 0.02 standard deviations, although this difference is not statistically significant. Our results persist over time, as well as within and between academic years.

To understand why gender differences in performance exist in the response to increased stakes, we conduct a small-scale field experiment, where the stakes for two ex-ante identical tests vary. The experiment allows us to determine whether the observed gender gap is due to a decrease in performance among female students as the stakes shift from low to high or an increase in performance among male students as the stakes shift from low to high—or a combination of both. Moreover, the experiment allows us to control for some features of the school evaluation system, such as the format of the tests and the amount of material. Overall, evidence suggests that male students do significantly better on the high-stakes exams, while female students do worse. However, this is only true partially, since the effects are statistically significant only for some subjects.

Overall, the described gender differences in performance are consistent with the existence of gender differences in response to pressure, resulting from different stakes; however, alternative hypotheses could also explain the main findings. First, since high-stakes exams are systematically at the end of the year, and differences in performance could result from the difference in test timing over the academic year. Second, the amount of material covered in the high-stakes tests is larger than that in other tests, which could affect students' performance differently. Third, teachers might grade male and female students differently depending on whether the exams are high or low stakes. Finally, students could allocate their effort differently when preparing for tests depending on the stakes associated with the tests. In Section 5 of the paper, we discuss these alternative hypotheses and use both administrative and experimental data to provide evidence that rules out these hypotheses as being the main driver of the observed effects.

There is an extensive literature that documents gender differences in labor-market outcomes (see Altonji and Blank, 1999; Bertrand, 2009). Understanding whether differential reactions to pressure can potentially explain part of this gap, using labor-market information is problematic. Once in the labor market, men and women will have made choices shaped by their professional environment and personal circumstances and, potentially, by their preference for and reaction to pressure. In our setting, we focus on a period in an individual's educational career when, for the last time, he or she is exposed to a homogeneous and compulsory procedure that will affect his or her future success in higher education and in the labor market.

Gender differences in academic attainment and achievement have been widely documented (see, for example, Goldin, Katz and Kuziemko, 2006). Differences in the nature of tests—in particular, their objectiveness, their competitive nature and the skills they measure—have been exploited to identify the potential channels that explain these gaps. Some noteworthy examples are Lavy (2008), Jurajda and Münich (2011), Cornwell, Mustard, and Van Parys (2013), and Örs, Palomino and Peyrache (2013).[1] Our setting uses a quasi-experimental design in the evaluation system to study gender differences in performance, for which the variation exploited is the test stakes, while other factors, such as the evaluators and the competitiveness of the environment, are held constant.

A recent literature shows that women and men differ in their competitiveness. Some papers have shown that women underperform compared to men in competitive environments (Gneezy, Niederle and Rustichini, 2003; Gneezy and Rustichini, 2004; Antonovics, Arcidiacono, and Walsh, 2009; Shurchkov, 2012; Iriberri and Rey-Biel, 2014), and others have shown that women shy away from competitive environments (Niederle and Vesterlund, 2007; Buser, Niederle, Oosterbek; 2013). Related to entry, Petrie and Segal (2014) show that if the rewards from competition are sufficiently large, women are willing to compete as much as men and that they win as many competitions as men. In our setting, because pressure is not defined as having a competitive nature, the rewards are independent of the performance of others.[2] Here, the gaps in performance result from the pressure that arises due to variation in the size of the test stakes. Moreover, because all tests are compulsory, there is no possibility to shy away from tests of varying stakes.

[1] Lavy (2008) compares blind and non-blind scores on the college matriculation exams of male and female students and finds evidence of gender stereotyping and discrimination against male students by teachers. In their setting, the size of the stakes is held constant for both types of test, and only the observability of the students' identity changes. Similarly, Cornwell, Mustard, and Van Parys (2011) show that the apparent advantage that female students had in grades provided by a teacher versus those provided by an external evaluation disappears once noncognitive traits are controlled for. Jurajda and Münich (2011) find that female applicants to Czech university programs perform less well than male applicants when admission rates are low but perform as well as male applicants when admission rates are high. Similarly, Örs, Palomino and Peyrache (2013) compare gender differences in test performance under direct competition, when relative performance determines college entry, with test performance in school and college; the authors argue that the payoffs resemble a piece-rate scheme. They show that male students perform better in the tournament setting than predicted by their previous high school grades and that woman perform worse.

[2] This does not exclude the possibility that there may be some type of competition for relative standing, such that, even when payments are piece-rate, individuals might still be concerned with their relative position among their peers (see Azmat and Iriberri, 2010 and 2014). The extent of the competitiveness across the tests, however, is somewhat constant.

The psychology literature shows that increasing the stakes beyond a certain level can lead to a decline in performance, commonly referred to as "choking under pressure" (Baumeister, 1984). The most relevant sources of pressure seem to be the presence of an audience, competition with others, personal traits, and one's own ego-relevant threat (see Ariely et al., 2009). Research in psychology has shown that an emphasis on the importance of the process can harm the individual's capacity to exhibit her "true" capability (Beilock, 2011).[3] Our results show that the strain of pressure, as defined as an increase of exam stakes, is heterogenous across gender. Moreover, we find some, although only suggestive, evidence of females "choking" and males "excelling" under pressure.

In experimental research on examining how stakes affect performance through effort choice, stakes are defined as monetary incentives. Camerer and Hogarth (1999) review the experimental results and conclude that the size of the monetary payoff has little effect on effort choice.[4] Ariely et al. (2009) provide experimental evidence that choking under pressure becomes relevant as the size of the stakes is increased. They suggest that there is an optimal amount of performance-contingent incentives (pressure) that leads to a maximal performance, and deviation from this will reduce performance. Using three different experiments, they show that the optimal pressure depends on the task at hand. In two experiments, they vary the monetary stakes, and in the final experiment, the monetary stakes are low and held constant, but increased pressure is induced by performing in front of an audience. In the final experiment, they also explore whether there are gender differences. They find that choking under pressure occurs in all experiments. However, it occurs earlier for cognitive tasks compared with physical effort tasks. In the third experiment, they show that performing in front of an audience reduces performance; however, there are no gender differences. This paper focuses on

[3] In her book "Choke: What the Secrets of the Brain Reveal about Getting it Right When You Have To," Beilock (2011) summarizes the existing literature in social psychology and emphasizes the similarity between students, athletes and business people who choke when the stakes are high. The mechanisms that could produce such a reaction include increased arousal, narrowed attention, and preoccupation with the reward itself. It has also been identified that consciously thinking about a task that is usually done automatically can be detrimental to performance. Increased pressure can induce a shift from this "automated" to a "controlled" procedure. See Yerkes and Dodson, 1908; Langer and Imber, 1979; Camerer, Lowenstein and Prelen, 2005 for details on this literature.

[4] Lavy and Angrist (2009), in an educational setting, show that providing monetary incentives to students in low-achieving schools improves the matriculation rates of girls but has no effect on boys.

cognitive tasks, for which pressure is induced only by varying the weights of different exams on the final grade.

The rest of the paper is organized as follows. In Section 2, we describe the performance data and the evaluation system that generated it. In Section 3, we present the main results. Section 4 describes the design of the experiment and presents the results. Section 5 discusses and tests for the alternative hypotheses for the main results. Section 6 concludes by discussing the implications of our findings.

2. Study Design

Evaluation System

The school performance data are from a private school in Barcelona, Spain that offers four years of Compulsory Secondary Education (*ESO*) for ages 12 to 15 and *Bachillerato,* which comprises the two years prior to university for ages 16 to 18. There are, therefore, six levels, which we refer to as Levels 1 to 6. At the end of Level 6, students who plan to pursue a university degree take externally designed and graded national exams (*Selectividad*). Compared with other schools in the region, both private and public, student performance on the national exams is substantially higher and attracts students from highly educated families. Average performance in the national exams in the school in 2010 was 7.73 (out of 10), compared to 7.49 in other private schools and 6.3 in the remaining public schools.

During each academic year, students take several exams in each of their subjects, and the final score for each subject is determined as follows. In each trimester, students take bi-monthly tests and an end-of-trimester test, except in trimester three, when they have an end-of-year test. For each subject, all tests are graded by the same teachers and follow a similar format although the amount of material covered differs. The final grade is determined by a weighted average of the first, second, and third trimester tests and the end-of-year test. The weights on the bi-monthly tests in each trimester are approximately 2.5 percent. The end-of-trimester test in trimester one and two is worth 11 percent, and the end-of-year test is worth 27 percent.[5] We define the bi-monthly,

[5] The weights described are the reduced-form weights resulting from the following compounded formula. Every trimester, a grade is constructed as follows. In the first two trimesters, the trimester-grade is constructed giving 60 percent of the weight to the (average) bi-monthly exams and 40 percent to the end-of-trimester exam. In the third trimester, the grade is determined only by the bi-monthly exams. The final grade is determined by giving 27 percent of the weight to trimester one and two, 18 percent to trimester three and 27 percent to the end-of-the-year exam.

end-of-term, and end-of-year tests as "low," "medium," and "high" stakes, respectively. The evaluation system is summarized in Figure 1.

Students in Level 6 take national-level exams, which, together with their high-school test scores in Levels 5 and 6, determine their college entry grade.[6] The weight assigned to the Level 5 and Level 6 end-of-year grade is 25 percent each, and 50 percent is assigned to the national-level exam grade. We therefore define the national exams as "super-high" stakes tests. The subject material evaluated in the national exams is the same as that covered in Level 6. However, not all subjects are tested in the national exam, which allows for some interesting variation, which we will exploit later.

Data Description

We have panel data on student test scores for all subjects, for levels 1 to 6, between the academic years 2000 to 2012, giving a total sample of 1,404 students. For each subject per academic year, we observe eight measures of performance in the low-stakes tests, two in the medium-stakes tests and one in the high-stakes tests.[7] All test scores are standardized to a distribution with zero mean and a unit standard deviation. The standardization is performed by academic year, level, subject and test type. The sample size varies by subject, including the majority of students in compulsory subjects but far fewer in elective subjects. In addition, we have information on the scores for the national exams (super-high stakes). We also have information on subject teachers for some but not all years. Table 1 presents the descriptive statistics separately for each level.

From Table 1, we can see that there are an equal number of female and male students. Each year, students take approximately eight subjects.[8] In Levels 1 to 4, all subjects are common and compulsory. High school in Spain is compulsory until age 16 (the end of Level 4). From the table, we see that approximately 20 percent of students leave at the end of Level 4. While some students may choose not to pursue Levels 5 and 6 and, thus, the national-level exams, others might choose to do so in another school. Although fewer in number, some students leave at the end of the other levels. There are

[6] College admission in Spain is administered through a centralized system. Applicants submit a single application with a list of up to eight major-university options. Students are then ordered according to their grades and are assigned to their preferred option following that order.

[7] In the data, the test scores for the low-stakes tests are typically provided for each month and are an average of the bi-monthly tests.

[8] Table A.1 in the Appendix summarizes all the subjects that students take per level.

very few students who leave during the academic year (less than two percent). Students performing very well in the low-stakes and medium-stakes tests, such that they obtain the highest possible grade throughout the course, may be given an exemption from the high-stakes test. Table 1 indicates that approximately five percent of students are exempt from the test. However, most importantly, there is no gender difference in terms of those who leave or those who are exempt from tests. This holds for students in all levels.[9]

3. Analysis and Results

3.1. Baseline Regressions

We start by estimating gender differences in school performance on different types of tests:

$$(1)\ Y_{ilsy} = \alpha + \beta Female_i + \varepsilon_{ilsy},$$

where the outcome variable Y_{ilsy} is the standardized score for student *i*, in level *l*, subject *s*, and academic year *y*. *Female* takes the value one if the student is female and zero otherwise. Table 2 displays the estimation results.

Panel A, column 1 presents the results for the gender difference in the final end-of-year test score. Columns 2, 3, 4 and 5 present the results for tests with low stakes, medium stakes, high stakes, and super high stakes, respectively. In Panel B, we restrict the analysis to final-year students (Level 6) who take the super-high-stakes test (*Selectividad*).

Overall, we see that, in school, female students outperform men by 0.16 standard deviations of the mean. Looking at different types of tests, we find that the gender gap in performance falls as the stakes of the test increase. Outperformance is highest in low-stakes tests, that is, 0.18 standard deviations, and lowest in high-stakes tests, that is, 0.11 standard deviations. Moreover, in super-high-stakes tests (Panel B), the sign of the coefficient reverses (-0.03), such that male students outperform female students. This difference, however, is not significant. Panel B indicates that the same patterns persist when we restrict the analysis to Level 6. Female students perform relatively better in

[9] We will discuss this further in Section 3.

low-stakes tests—0.18 standard deviations—than in high-stakes tests—0.08 standard deviations.

In the following analysis, we study student relative performance on different types of tests. We estimate the following regression:

$$(1)\ Y_{ilsy} = \alpha + \beta Female_i + \gamma Low\_Stake\_Test_{ilsy} + \delta Female_i * Low\_Stake\_Test_{ilsy} + \varepsilon_{ilsyt},$$

where the outcome variable, Y_{ilsy} is the standardized score for student *i*, in level *l*, subject *s*, and academic year *y*; *Female* refers, as before, to a dummy for female students; and *Low_Stake_Test* is a dummy variable that takes the value one when the test score refers to that of a low-stakes test and zero otherwise.[10]

Table 3 presents the estimated parameters for equation (2). Columns 1 and 2 compare the low-stakes versus high-stakes tests; columns 2 and 3 compare the low-stakes versus medium-stakes tests; columns 5 and 6 compare the medium-stakes versus high-stakes tests; and, finally, columns 7 to 10 compare all three types of tests. The estimates are performed with and without the inclusion of student fixed effects. In all regressions, the coefficient of interest is the interaction between the female and the lower-stakes test type.

Consistent with Table 2, we find that female students perform significantly better on low-stakes than on high-stakes tests by 0.06 standard deviations, and compared to medium-stakes tests, by 0.05 standard deviations. From columns 5 and 6, we see that there is no significant gender difference when comparing medium- and high-stakes tests. From columns 7 to 10, however, we do see monotonicity in the effect of increasing stakes on gender differences in performance. In particular, compared with high-stakes tests, the coefficient is positive for both the interaction between low-stakes tests and being female and between medium-stakes tests and being female. However, the coefficient on the interaction is larger (and significant) for low-stakes than medium-stakes tests. Thus, in the analysis that follows, we will focus only on comparisons between low-stakes and high-stakes tests. Furthermore, none of the results changes when using individual fixed-effect estimation. Hence, from now on, we will show the results using an OLS regression.

[10] Similarly, for the comparison of medium stakes and high stakes, the dummy we define is *Medium-Stakes Test*, which takes the value of 1 when the test is that of medium stakes.

To understand whether gender differences change during the student academic career, we explore the gaps for each academic level separately. Table 4 presents the estimation results separately for Levels 1 to 6. Overall, we see similar patterns across all levels, with the exception of Level 5, in which female students outperform male students in low-stakes relative to high-stakes tests, but the difference is not significant. The magnitudes do, however, vary. In Levels 1 to 3, the difference is 0.05 standard deviations, while, in Levels 4 and 6, the effect is double (0.10 standard deviations). In Level 4, students decide whether to stay at the same school for the last two years prior to entering university. Since the grades obtained in the last two years count toward the university entry grade, this decision is important, as it will affect student access to university. In Level 6, the students sit for their Selectividad exams, and the weight on their high-school grades is sizeable.

Students undertake several subjects, typically ten or 11, per academic year. In Table 5, we disaggregate the analysis by subject type, classifying subjects as either Arts or Science.[11] Women and men have traditionally exhibited performance differences that depend on the type of subject, e.g., following the stereotype, women are, on average, more likely to outperform men in the Arts rather than in Science.

Table 5 shows the estimation results for gender difference for science and arts subjects separately for the final grade (columns 1 and 2), for low-stakes tests (columns 3 and 4), and for high-stakes tests (columns 5 and 6). In columns 7 and 8, we bring the specifications together by interacting gender with the test stakes for science and art subjects, respectively. First, according to the final scores, female students outperform male students in Arts subjects, while there is no significant gender difference in Science subjects. We clearly see that in low-stakes tests, female students outperform males in Arts subjects but exhibit no performance difference in Science subjects. In high-stakes tests, we find that, while female students outperform male students in Arts subjects (0.25 standard deviations), they significantly underperform relative to male students in Science subjects (-0.10 standard deviations). The gender differences depend on the stakes of the tests, as well as on the type of subject. The interactions in columns 7 and 8 indicate that gender differentials regarding the stakes are significant and positive in both Arts and Science subjects. Table A.2 in the Appendix separately addresses some of the main subjects. Overall, the results are similar across the subjects.

[11] Table A.1 in the Appendix classifies subjects between science and arts subjects.

3.2. Robustness of the Baseline Regressions

We perform three robustness checks with respect to the baseline results. The results are shown in Table 6. For ease of exposition, column 1 replicates the overall gender difference in tests with different stakes (column 1 in Table 3). In column 2, we include the teacher fixed effects to see whether the teacher information explains any of the gender differences for different stakes.[12] In column 3, we restrict the analysis to a balanced sample of students who complete all six levels. Finally, in column 4, we exclude student-subjects who have been exempt from high-stakes tests. The last two robustness checks rule out the possibility that the gender differential effect is driven by students who leave school and who are exempt from taking high-stakes tests.

From column 2, we find that including teacher fixed effects does not change the size or the significance of the gender difference in low-stakes versus high-stakes test performance. This suggests that the effect is not driven by some teachers, as the size and significance of the coefficient does not change by adding teacher fixed effects. Similarly, in column 3, we find that restricting the analysis to students who stay in school has a quantitatively small effect. We again see that the interaction effect is significant and positive. Finally, students performing very well in low-stakes and medium-stakes tests, such that they are obtaining the highest possible grade throughout the course, may be given an exemption from the high-stakes test. Although this is typically a very small number of students per subject, sometimes none, we check the robustness of our results from excluding them. In column 4, we see the same result, both quantitatively and qualitatively, thus showing the robustness of the gender differential effect that depends on the size of the stakes for each test.

In Appendix Tables A.3 and A.4, we complement some of this analysis by looking at the likelihood of students to leave the school and to be exempt from a high-stakes exam, respectively. Based on Table A.3, we see that there is no gender difference in leaving school. Moreover, although high-performing students, as characterized by the average final test score over all subjects, are more likely to stay at the school, there is no gender differential in this. Table A.4 shows that, conditional on the low-stakes test score, female and male students have a similar exemption probability. In Level 1, the interaction is negative and significant, suggesting that female students have a smaller

[12] We have teacher information for the academic years 2000 to 2009.

chance of exemption. However, across levels, there is no systematic pattern in the coefficient, which with similar magnitudes is positive, although insignificant, in some levels (Levels 3, 4 and 5).

In summary, the results suggest that the difference in performance between male and female students is larger for the small stakes exams than for high stakes exams. The difference in the gaps depending on the exam stake could be because female students perform worse as the exam stakes increase, or because male students perform better, or both. The experiment in the next section helps disentangle the two effects, and further verify the results in an environment where we can control for important features in the evaluation system at the school.

4. Experiment in the School

4.1. Description of the Design

The experiment was conducted using all students in Level 6 for two subjects: Mathematics and Catalan (the language spoken in the region). One is an art subject (Catalan) and the other a science subject (Mathematics). The design of the experiment included the following features.

First, we asked teachers in these subjects to prepare two of the higher stakes exams (medium stakes), one of which was used for an actual higher stakes exam and the other for a lower stakes exam (low stakes). They randomly determined which test would be used for higher and lower stakes. This guaranteed that the two tests were comparable in terms of the format of the exam, the duration of the exam, and the amount, and even the content, of the covered material.

Second, in terms of timing, the exams were taken one week apart: the higher stakes exam was taken in the first week, and the lower stakes exam was taken in the second week. Since these exams were given at the end of the first term, these tests correspond to the regular medium- and low-stakes tests that the students would take anyway. [13] More important, the teachers did not provide any feedback on the result of the first, higher stakes exam before students took the unannounced second, lower stakes exam. In addition, the second exam took place before any new material was covered.

[13] Note that the end of term does not coincide with the holiday period. The term ends three weeks before the holiday period; thus, the next term starts before the holiday period. This allows us to compare the two tests without holidays in between.

Therefore, neither female nor male students should have devoted additional time to studying before the lower stake exam. The preparation time for the higher stakes exam should therefore be similar to the preparation time for the lower stakes exam.

The design allows us to compare the test scores directly, without the need to standardize them in order to determine how differently students reacted to the change in exam stakes. Moreover, we can verify the robustness of the effect by controlling for the type, format and content of the test and the effort in preparation for the different types of exam.

4.2. Results from the Experiment

Figure 2 shows the mean test scores obtained by male and female students in the exams depending on their stakes. Panel A displays the overall difference, and Panels B and C display the results separated for Catalan and Math, respectively. In all three Panels, female students perform better in both types of exams compared with male students, which is consistent with the overall finding regarding gender differences in school performance, shown in Table 2. We also see that female students perform better in the lower stakes exams than in the higher stakes exams, both in Catalan and in Math. For male students, the reverse happens, but only on the Catalan exam. Panel A indicates that male students perform better in higher stakes exams than in lower stakes exams. As indicated by Panels B and C, however, this is the case only for Catalan and not for Math.

Table 7 presents the results. In all columns, the outcome variable is the test score, and we control for gender, exam stakes and an interaction term between gender and exam stakes. Columns 1 to 2 present the OLS and fixed-effect results for the overall exam scores, columns 3 to 4 present the results for Catalan, and columns 5 to 6 present the results for Math. We find that, in line with the figures, there is a gender difference in students' response to exams with different stakes. Female students perform relatively better in the lower stakes exams than the higher stakes exams, while male students show the opposite pattern. This difference is not significant in the overall results but highly significant for Catalan. For Math, in columns 5 and 6, the effects show the opposite pattern, but the difference is very small and not significant. Interestingly, across subjects, we see very similar patterns and levels of significance for Math and Catalan in the main analysis (see Table A.2).

The results found in the field experiment, shown in Table 7, partially suggest that female students perform worse when the stakes increase and that male students perform better. This suggests that female students choke under pressure and that male students excel under pressure, engendering a higher gender gap when the stakes are low rather than high. However, this result is only significant for one of the subjects, Catalan—not Math.

5. Alternative Hypotheses for the Main Results

In this section, we propose alternative interpretations of our results. Our main hypothesis is that male and female students react differently to the pressure inherent in different stakes. Female students perform better overall than male students, but this gender gap is larger when the stakes are low. The results from the field experiment partially suggest that female performance decreases as the exam stakes increase and that male performance increases. In this section, we provide evidence to help rule out alternative hypotheses that might drive the main results. The alternative hypotheses include 1) the timing of the exams in the school evaluation system; 2) the different exam formats and volume of material covered; 3) teachers' differential grading; and 4) students' different levels of preparation depending on the stakes.

5.1. The Relevance of the Timing of Exams in the Evaluation System

The schedules for the low-stakes and high-stakes tests differ. Low-stakes tests are administered throughout the academic year, while high-stakes tests are administered only at the end of the academic year (see Figure 1 for the timing of exams). When comparing low-stakes tests with medium-stakes tests, we also find that medium-stakes tests come after the low-stakes tests. One hypothesis we consider is that rather than the pressure inherent in stakes, these estimates capture a gender difference in exam timing: male students slack off during the academic year and only study hard at the end of the academic year, while female student have more consistent study habits throughout the academic year.

To understand whether this is the case, we look at the performance in each of the eight low-stakes exams relative to the high-stakes exams. Table 8 reports the results. If the gender difference arises because the high-stakes test occurs later in the academic year, we should expect to find that the gender difference in the low-stakes test relative

to the high-stakes test will become smaller as the academic year progresses. Note that Low-Stakes Test 8 takes place only one week before the high-stakes exam. In Table 8, we see that all of the interactions are positive, although not all coefficients are significant. Moreover, there is no systematic pattern in terms of the gender differences in response to the low-stake tests as the students reach the end-of-year exam in terms of either the size of the coefficient or the level of significance. For Term 1 and Term 3, the coefficient on the interaction term is smaller in the last low-stakes test (*Low-Stakes Test 3* and *Low-Stakes Test 8*); however, for Term 2, this is not the case (*Low-Stakes Test 6*). This helps to rule out the alternative explanation that the different timings of the low- and high-stakes tests drive the identified gender difference. We do not observe a gender difference in time allocation in the low-stakes test. That is, it does not seem that male students wait until the end of the year to study because they do not exhibit a larger increase than girls in their test scores in the later low-stakes tests.

5.2. The Relevance of the Format and Amount of Material

The material covered in the high-stakes exams is substantially greater than that covered in the low-stakes exams, as the high-stakes exams cover the material through the academic year. It might be that male and female students perform differently depending on the amount of material.[14] We can provide two pieces of evidence that suggest that this is not the only driver of our results.

First, the field experiment included two tests that were comparable and randomly allocated into higher and lower stakes exams. Therefore, the format and amount of material covered are also comparable. Based on the field experiment, as shown in Table 7, we find that female students do significantly better when the stakes are lower compared with higher for Catalan (although no effect was found for Math), consistent with our interpretation of the baseline regressions.

Second, the structure of the Level 6 test system provides us with an alternative way to check that it is the stakes of high-stakes tests that explain the gender differences and not the difference in the type of exam and the amount of material covered. We exploit variation in the implicit stakes of different subjects for different students, which naturally occurs in Level 6. At this level, as with all the other levels, students take approximately ten or 11 subjects. In high school, all subjects are examined in the same

[14] Spencer, Steele, and Quinn (1999) show that the level of difficulty of exams could affect the gender difference in performance.

way as with all other levels. However, one important difference is that the national-level examination (i.e., the Selectividad) at the end of Level 6 tests students on only approximately five subjects, which differ for different individuals depending on their choice of track at the beginning of Level 5. Thus, for students in a given class, the subjects for which the grades matter to enter university differ. The relevance of a given subject for students in the same class hence varies depending on their choice of track, and teachers do not necessarily know for whom the subject is relevant or not. Therefore, the format and material covered in the exam for a subject is the same for all students, but for a subset of them, the subject is also relevant for Selectividad, while for others it is not. Recall that the high school Level 6 test score (for all ten or 11 subjects) counts for 25 percent of the university entry grade, but the super-high-stakes test scores of the Selectividad (for a smaller set of subjects) count for 50 percent. We can therefore classify subjects as those that "count" toward Selectividad (i.e., Selectividad subjects) and those that "do not count" (i.e., non-Selectividad subjects). Whether the subject counts will have implications for their implicit stakes. Thus, we use the variation in whether the exam counts to determine whether female and male students perform differently on their high school tests.

Table 9 shows the results for test scores in subjects taken for Selectividad and non-Selectividad subjects. Three results are noteworthy. First, the first columns show that female students outperform male students in all types of tests, regardless of whether the material is a Selectividad subject. Second, students perform relatively better on subjects that count toward a Selectividad exam than on those that do not, regardless of the exam's weight (high or low stakes) in the final grade. Finally, from column 2, looking at the interaction between gender and low stakes, we clearly see that, as in the baseline results, female students perform relatively better in low-stakes exams. In column 3, which also includes an interaction term between gender and Selectividad subject, we find that, while female students continue to perform better in the low-stakes exams, the gap between female and male students closes in Selectividad subjects, as indicated by the coefficient for the interaction, which is negative and significant.

This finding indicates the increased importance of the exam stakes, although the exams for all subjects, whether they are included in the Selectividad, count the same for the high school final grade and are taught and graded by the same teacher. The difference in performance suggests that students perform better in the Selectividad subjects because the national exams carry greater weight. However, the gender

difference, especially in high-stakes exams, indicates that this increased pressure, in terms of what the exam will eventually be relevant for, generates a different reaction among female and male students. Note that we are exploiting the variation in stakes, which arises not because of an increase in the weight of the grade of the exam but because of an increase in the weight of the *subject* owing to its inclusion or not in the national exam, Selectividad. Female students underperform relative to male students when the subject is included in the national exam and thus when the pressure is higher. This result further confirms that male and female students respond differently to different stakes and that our results are not merely driven by the amount of material and type of test, as these two factors are held constant.

5.3. Teachers Grading Differently

Teachers might grade male and female students differently (Lavy, 2008; Cornwell, Mustard, and Van Parys, 2013). In our setting, exams are graded by the teachers within the school, and they are therefore not blind. One hypothesis we investigate is that the observed gender difference is driven not by students' reaction to stakes but by the teachers' differential grading depending on the stakes, such that teachers always discriminate in favor of female students but discriminate less in higher stakes tests. We can provide two pieces of evidence that help to rule out this alternative explanation.

First, in the robustness tests of the baseline regressions, shown in Table 6, we have included teacher fixed effects, and the results remain unchanged. This rules out the alternative explanation that some specific teachers, with certain characteristics, are driving the observed gender difference in performance.

Second, as in the previous section, we can also exploit the variation in the implicit stakes of different subjects for different students, which naturally occurs in Level 6. Again, the variation that we exploit is not in the weight of different tests but in the inclusion of the subject in the individual's national exam. The variation in stakes that results from whether the subject is a Selectividad subject is indeed blind to the teachers. In other words, when grading, teachers do not know whether that particular subject is a Selectividad subject for a given student. In Table 9, we observe that the gender difference in performance also depends on whether the subject is a Selectividad subject. This provides convincing evidence that the main results are unlikely to be

driven by the teachers grading exams with different stakes differentially for male and female students.

5.4. Gender Difference in Preparation Time

The final hypothesis that we consider is that the observed gender difference is explained by differential reactions in preparation for tests with different stakes. In other words, male students may study and prepare more (less) when the stakes are higher (lower) than female students do.

The field experiment allows us to directly test this alternative hypothesis. In the experiment, the tests were only one week apart, there was no feedback between them, and more important, the second test was unannounced. This helps to ensure that students did not considerably change their level of preparation between the first and the second test. As shown in Table 7, we find evidence for gender differences in students' reaction to different stakes—but only for Catalan and not for Mathematics.

6. Discussion and Implications

We find gender differences in performance in response to different stakes. In particular, overperformance by female students compared with male students in school is reduced or even negated as the stakes are increased substantially.

It is important to understand the consequences of gender differences in response to increased stakes. Here, we discuss the potential labor-market implications of our findings that men and women have different levels of pressure tolerance.

If selection into jobs is unaffected by individual pressure tolerance and if pressure tolerance is payoff relevant to the worker through his or her productivity, we would expect that, all else being equal, female workers would earn a lower wage than male workers. In particular, for jobs in which pressure is high, we may expect female workers to underperform relative to their male counterparts.

Prior to entering the labor market, however, individuals must typically go through a selection process. Individuals self-select into occupations and firms, and employers in those firms select those whom they want to hire. Pressure in the selection process may lead companies to disregard individuals with lower pressure tolerance at the cost of other potentially productive skills. A firm will interview workers to measure their skills, and candidates, when interviewed, provide a signal of their skills that may be affected

by their tolerance for pressure. The level of pressure in the signaling process may impose a tradeoff in the productivity of the selected candidates. That is, increased pressure in the selection process will lead to a workforce with a higher tolerance for pressure, but this might come at the expense of other relevant skills. This type of scenario is discussed in Gneezy and List (2013).

In anticipation of the importance of pressure in the signaling process and in the firm valuation of pressure in the workplace, individuals concerned with their labor-market success may self-select into certain occupations and firms. For instance, an individual with low pressure tolerance will avoid firms that reward pressure tolerance, either in the interview stage or on the job, unless other skills can compensate for this low tolerance. This might be one explanation for occupational segregation, whereby there is an over- (under-) representation of women in some professions but not in others. Moreover, recent evidence highlights that women, unless they are top performing, abandon certain college majors, unlike their male counterparts.[15] The concern is that these majors tend to lead to high-paying jobs and, again, that they might explain gender wage gaps. In line with our reasoning, if these jobs involve high pressure, lower performing female students rationally switch majors.

In this paper, we find that the increased pressure arising from an increase in the stakes induces different reactions in female and male students. So far, we have denoted pressure as a single-dimension variable. However, as the social psychology literature has acknowledged, there are different sources of pressure, and the reaction to each form may be different for different individuals (see Beilock, 2011). For instance, the economics literature has also emphasized the asymmetric reactions of men and women to competitive environments, even when the stakes are not particularly high. Understanding the different sources of pressure and how they affect different groups in society is important to understanding the potential inefficiencies in the labor market.

[15] A recent article in the *Washington Post*, "Women should embrace the B's in college to make more later," by C. Rampell, cites P. Arcidiacono: "STEM majors, as with economics, begin with few women enrolling and end with even fewer graduating. This *leaky pipeline* has been somewhat puzzling because women enter college just as prepared as men in math and science." Similarly, work by C. Goldin (2013) is cited showing that the fraction of women graduating in economics decreases substantially with their grade in introductory economics in the first year, which is not true for male students.

References

Altonji, J. and R. Blank (1999), "Race and Gender in the Labor Market" in O. Ashenfelter and D. Card, (eds), Handbook of Labor Economics, Volume 3c, Elsevier Science B.V.: 3144-3259.

Angrist, J. and V. Lavy (2009), "The Effects of High Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial", *The American Economic Review*, 99 (4): 1384-1414.

Antonovics, K., P. Arcidiacono and R. Walsh (2009), "The Effects of Gender Interactions in the Lab and in the Field", *The Review of Economics and Statistics,* 91 (1): 152-162.

Ariely, D., U. Gneezy, G. Loewenstein and N. Mazar (2009), "Large Stakes and Big Mistakes", *Review of Economic Studies*, 76 (2): 451-469.

Azmat, G. and Iriberri, N. (2010) "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students", *The Journal of Public Economics*, 2010, vol. 94 (7-8), pp. 435-452.

Azmat, G. and Iriberri, N. (2014) "The Provision of Relative Performance Feedback: An Analysis of Performance and Satisfaction," *Journal of Economics & Management Strategy*, forthcoming.

Baumeister, R. F. (1984), "Chocking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance", *Journal of Personality and Social Psychology,* 46 (3): 610-20.

Beilock, S. (2011), *Choke: What the Secrets of the Brain Reveal About Getting it Right When You have To,* by Simon and Schuster, Free Press.

Bertrand, M. (2010), "New Perspectives on Gender", in O. Ashenfelter and D. Card, (eds), Handbook of Labor Economics, Volume 4b, Elsevier Science B.V.: 1545-1592.

Buser, T., M. Niederle and H. Oosterbeek (2014), "Gender, Competitiveness and Career Choices", *The Quarterly Journal of Economics,* 129 (3).

Camerer, C. F. and R. Hogarth (1999), "The Effect of Financial Incentives in Experiments: a Review and Capital-Labor-Production Framework", *Journal of Risk and Uncertainty,* 19 (1), 7-42.

Camerer, C.F., G. Lowenstein and D. Prelec (2005), "Neuroeconomics: How Neuroscience Can Inform Economics", *Journal of Economic Literature,* 43 (1): 9-64.

Cornwell, C., D. Mustard and J. Van Parys, (2013), "Non-cognitive Skills and Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School", *Journal of Human Resources,* 48 (1): 236-264.

Gneezy, Uri and J. List (2013), "The Why Axis: Hidden Motives and the Undiscovered Economics of Everyday Life", *New York: Public Affairs*.

Gneezy, U., M. Niederle and A. Rustichini (2003), "Performance in Competitive Environments: Gender Differences", *The Quarterly Journal of Economics*, 118 (3), 1049-1074.

Gneezy, U. and A. Rustichini (2004), "Gender and Competition at a Young Age", *American Economic Review P&P*, 94(2), 377-381.

Goldin, C., L. Katz, and I. Kuziemko, (2006) "The Homecoming of American College Women: The Reversal of the Gender Gap in College", *Journal of Economic Perspectives*, 20:133-156.

Iriberri N. and P. Rey-Biel (2011) "On Women's Underperformance in Competitive Environments: When and Why", UPF working paper 1288.

Jurajda, Stepan, and Daniel Münich (2011), "Gender Gap in Performance under Competitive Pressure", *American Economic Review: Papers and Proceedings*, vol. 101(3), pp. 514-518.

Langer, E. J. and L.G. Imber (1979), "When Practice Makes Imperfect: The Debilitating Effects of Overlearning", *Journal of Personality and Social Psychology*, 37 (11), 2014-2024.

Lavy, V. (2008) "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment", *Journal of Public Economics*, 92 (10-11): 2083-2105.

Niederle, M. and L. Vesterlund (2007), "Do Woman Shy Away from Competition? Do Men Compete too Much?" *The Quarterly Journal of Economics*, 122 (3): 1067-1101.

Örs, E., F. Palomino and E. Peyrache (2013), "Performance Gender-Gap: Does Competition Matter?" *Journal of Labor Economics*, 31 (3): 443-449.

Petrie, R. and C. Segal (2014), "Gender Differences in Competitiveness: The Role of Prizes", mimeo.

Shurchkov, O. (2012), "Under Pressure: Gender Differences in Output Quality and Quantity Under Competition and Time Constraints," *Journal of the European Economic Association* 10(5): 1189–1213.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999), "Stereotype threat and women's math performance," *Journal of Experimental Social Psychology*, 35, 4–28.

Yerkes, R.M. and J. D. Dodson (1908), “The Relationship of Strength of Stimulus to Rapidity of Habit-Formation,” *Journal of Comparative Neurology of Psychology,* 18 (5): 459-482.

Tables and Figures

Table 1: Descriptive Statistics

	OVERALL			Level 1			Level 2			Level 3			Level 4			Level 5			Level 6		
Variable	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.
No. Low-Stakes Tests	38146	6.78	1.19	8994	7.01	0.97	7974	7.03	0.97	6904	6.83	0.99	6435	6.82	1.02	4626	6.6	1.62	3213	5.59	1.47
No. Medium-Stakes Tests	38146	1.98	0.16	8994	2.00	0.07	7974	2.00	0.04	6904	1.97	0.16	6435	2.00	0.05	4626	1.97	0.23	3213	1.93	0.37
No. High-Stakes Tests	38146	1.00	0.06	8994	1.00	0	7974	1.00	0	6904	1.00	0.00	6435	1.00	0.00	4626	0.99	0.11	3213	0.97	0.17
No. Students	38146	309.53	97.53	8994	85.12	6.49	7974	78.17	5.37	6904	69.74	5.12	6435	59.46	7.67	4626	41.7	13.7	3213	35.49	17.06
Prop. Female Students	38146	0.50	0.06	8994	0.49	0.07	7974	0.50	0.05	6904	0.51	0.06	6435	0.52	0.05	4626	0.51	0.07	3213	0.51	0.06
No. Subjects	38146	8.11	1.09	8994	7.96	0.25	7974	7.91	0.31	6904	7.83	1.02	6435	8.76	1.54	4626	8.86	0.99	3120	7.22	1.56
Prop. Leavers	38146	0.10	0.03		--	--	7974	0.03	0.04	6904	0.07	0.05	6435	0.1	0.05	4626	0.21	0.04	3213	0.1	0.05
Prop. Students Exempt	38146	0.05	0.08	8555	0.04	0.07	7636	0.04	0.06	6553	0.05	0.07	5753	0.09	0.10	4334	0.05	0.10	3157	0.00	0.02

Notes: The descriptive statistics are calculated over the years 2000 to 2012. Each observation refers to a student-subject. *No. Low-Stakes Tests* is the number of low-stakes tests taken by students for a given subject in a given academic year. Students typically take low-stakes tests bi-monthly (per subject), such that they take approximately six low-stakes tests in terms 1 and 2 and four in term 3. We have information only on the average monthly low-stakes test scores. *No. Medium-Stakes Tests* is the number of medium-stakes tests taken by students for a given subject in a given academic year. *No. High-Stakes Tests* is the number of high-stakes tests taken by students for a given subject in a given academic year. *No. Students* is the number of students in a given academic year. *Prop. Female Students* is the proportion of female students in a given academic year. *No. Subjects* is the number of subjects that students take in a given academic year. *Prop. Leavers* is the proportion of students who leave at the end of the previous academic year. *Prop. Students Exempt* is the proportion of students given an exemption from a high-stakes test in one (or more) of the subjects in a given academic year.

Table 2: Performance under Different Stakes

A: Performance under Different Stakes: Overall					
	Final Test Score	Low-Stakes Test Score	Medium-Stakes Test Score	High-Stakes Test Score	Super-High-Stakes Test Score
Female	0.159***	0.175***	0.126***	0.118***	-0.0301
	[0.0419]	[0.0438]	[0.0407]	[0.0374]	[0.0382]
Constant	-0.0799***	-0.0884***	-0.0636**	-0.0595**	0.016
	[0.0294]	[0.0309]	[0.0283]	[0.0263]	[0.0272]
Observations	38,146	38,637	38,247	38,857	2,598
R-squared	0.006	0.008	0.004	0.004	0.001

B: Performance under Different Stakes: Level 6 only					
	Final Test Score	Low-Stakes Test Score	Medium-Stakes Test Score	High-Stakes Test Score	Super-–High-Stakes Test Score
Female	0.169***	0.185***	0.0939***	0.0803**	-0.0301
	[0.0344]	[0.0336]	[0.0345]	[0.0332]	[0.0382]
Constant	-0.0820***	-0.0937***	-0.0467*	-0.0407*	0.016
	[0.0244]	[0.0239]	[0.0246]	[0.0236]	[0.0272]
Observations	3,213	3,372	3,207	3,473	2,598
R-squared	0.008	0.009	0.002	0.002	0.001

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. *Final Test Score* is the students' accumulated test score at the end of the academic year. *Low-Stakes Test Score* is the students' test score in the low-stakes tests; *Medium-Stakes Test Score* is the students' test score in the medium-stakes tests; *High-Stakes Test Score* is the students' test score in the high-stakes tests; and *Super-High-Stakes Test Score* is the students' test score in the national exam, Selectividad, taken in Level 6. All test scores are standardized to a distribution with zero mean and a unit standard deviation. The standardization is performed by academic year, level, subject and test type.

Table 3: Performance under Different Stakes with Interactions

	Low-Stakes vs. High-Stakes		Low-Stakes vs. Med.-Stakes		Med.-Stakes vs. High-Stakes		Low-Stakes vs. (Med.- and High-Stakes)			
	Test Score	Test Score	Test Score	Test Score	Test Score	Test Score	Test Score	Test Score	Test Score	Test Score
Female	0.118***		0.126***		0.118***		0.122***		0.118***	
	[0.0374]		[0.0407]		[0.0374]		[0.0387]		[0.0374]	
Low-Stakes Test	-0.0290***	-0.0277***	-0.0248***	-0.0253***			-0.0269***	-0.0265***	-0.0290***	-0.0281***
	[0.00897]	[0.00899]	[0.00730]	[0.00718]			[0.00735]	[0.00732]	[0.00897]	[0.00902]
Female*Low-Stakes Test	0.0574***	0.0579***	0.0490***	0.0513***			0.0532***	0.0546***	0.0574***	0.0585***
	[0.0124]	[0.0124]	[0.00988]	[0.00979]			[0.00996]	[0.00995]	[0.0124]	[0.0124]
Med.-Stakes Test					-0.00418				-0.00418	-0.00319
					[0.00722]				[0.00722]	[0.00721]
Female*Med.-Stakes Test					0.00845	0.0048			0.00845	0.008
					[0.0102]	[0.00719]			[0.0102]	[0.0102]
Constant	-0.0595**	-0.000747	-0.0636**	-0.00026	-0.0595**	-0.00116	-0.0615**	-0.000332	-0.0595**	-0.000753
	[0.0263]	[0.00309]	[0.0283]	[0.00246]	[0.0263]	[0.00180]	[0.0270]	[0.00166]	[0.0263]	[0.00339]
Observations	77,494	77,494	76,884	76,884	77,104	77,104	115,741	115,741	115,741	115,741
R-squared	0.006		0.006		0.004		0.005		0.005	
No. of students		1,404		1,404		1,404		1,404		1,404
Student FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. *Low-Stakes vs. High-Stakes* compares low-stakes test scores with high-stakes test scores. *Low-Stakes vs. Med.-Stakes* compares low-stakes test scores with medium-stakes test scores. *Med.-Stakes vs. High-Stakes* compares medium-stakes test scores with high-stakes test scores. *Low-Stakes Test* takes the value one if the test is a low-stakes test and zero otherwise. *Med.-Stakes Test* takes the value one if the test is a medium-stakes test and zero otherwise. All test scores are standardized.

Table 4: Performance under Different Stakes with Interactions by Level

	Level 1 Test Score	Level 2 Test Score	Level 3 Test Score	Level 4 Test Score	Level 5 Test Score	Level 6 Test Score
Female	0.210***	0.127***	0.119**	0.026	0.0796	0.0803
	[0.0436]	[0.0459]	[0.0486]	[0.0532]	[0.0566]	[0.0595]
Low-Stakes Test	-0.0233*	-0.0247	-0.0244*	-0.0492***	-0.00791	-0.0529**
	[0.0130]	[0.0159]	[0.0148]	[0.0153]	[0.0202]	[0.0218]
Female*Low-Stakes Test	0.0472**	0.0489**	0.0482**	0.0957***	0.0155	0.105***
	[0.0189]	[0.0219]	[0.0215]	[0.0210]	[0.0270]	[0.0332]
Constant	-0.104***	-0.0639**	-0.0601*	-0.0134	-0.0406	-0.0407
	[0.0303]	[0.0324]	[0.0327]	[0.0375]	[0.0421]	[0.0432]
Observations	18,112	16,026	14,156	13,108	9,247	6,845
R-squared	0.014	0.006	0.005	0.002	0.002	0.005

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. The regressions compare low-stakes test scores with high-stakes test scores. All test scores are standardized. *Low-Stakes Test* takes the value one if the test is a low-stakes test and zero otherwise.

Table 5: Performance under Different Stakes with and without Interactions for Science versus Arts Subjects

	Final Test Score		Low-Stakes Test Score		High-Stakes Test Score		Low vs. High Test Score	
	Science	Arts	Science	Arts	Science	Arts	Science	Arts
Female	-0.0689	0.298***	-0.0499	0.312***	-0.101***	0.250***	-0.101***	0.250***
	[0.0452]	[0.0438]	[0.0468]	[0.0458]	[0.0388]	[0.0407]	[0.0388]	[0.0407]
Low-Stakes Test							-0.0252*	-0.0316***
							[0.0134]	[0.00902]
Female*Low-Stakes Test							0.0507***	0.0622***
							[0.0176]	[0.0127]
Constant	0.0358	-0.152***	0.0249	-0.159***	0.0501*	-0.127***	0.0501*	-0.127***
	[0.0322]	[0.0309]	[0.0342]	[0.0319]	[0.0275]	[0.0288]	[0.0275]	[0.0288]
Observations	14,405	23,741	14,589	24,048	14,623	24,234	29,212	48,282
R-squared	0.001	0.023	0.001	0.025	0.003	0.016	0.002	0.02

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. *Final Test Score* is the students' accumulated test score at the end of the academic year. *Low-Stakes Test Score* is the students' score on the low-stakes tests, *High-Stakes Test Score* is the students' score on the high-stakes tests. *Low-Stakes vs. High-Stakes* compares low-stakes test scores with high-stakes test scores. Science subjects are subjects that are classified as science subjects. Art subjects are subjects that are classified as humanity subjects. A full list of subjects is given in Table A.1.

Table 6: Robustness checks: Teacher Fixed Effects, Staying Students and Students Exempt from the High-Stakes Test

	Test Scores	Test Scores	Test Scores	Test Scores
Female	0.118***	0.158*	0.0887	0.125***
	[0.0374]	[0.0887]	[0.0553]	[0.0349]
Low-Stakes Test	-0.0290***	-0.0254**	0.0216	0.0330***
	[0.00897]	[0.0106]	[0.0145]	[0.00906]
Female*Low-Stakes Test	0.0574***	0.0502***	0.0458**	0.0655***
	[0.0124]	[0.0140]	[0.0191]	[0.0125]
Constant	-0.0595**	-0.086	0.194***	-0.0619**
	[0.0263]	[0.0670]	[0.0391]	[0.0241]
Observations	77,494	54,520	37,020	72,774
R-squared	0.006	0.021	0.004	0.007
Teacher FE	No	Yes	No	No
Students in sample for all 6 levels	No	No	Yes	No
Student-subjects not exempt	No	No	No	Yes

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. The regressions compare low-stakes test scores with high-stakes test scores. All test scores are standardized. *Low-Stakes Test* takes the value one if the test is a low-stakes test and zero otherwise. Column 2 includes teacher fixed effects. Teacher identity is available only for years 2000 to 2009. Column 3 is restricted to cohorts of students who are in the sample for all six levels. Column 4 excludes test scores for when students are exempt from sitting the high-stakes exams.

Table 7: Results in the Field Experiment

	Lower-Stakes vs. Higher-Stakes Test Score					
	Overall		Catalan		Math	
Female	2.392		0.829		3.955	
	[3.015]		[2.677]		[5.030]	
Lower-Stakes Test	-0.305	-1.164	-1.796*	-2.800***	1.186	0.68
	[1.643]	[1.471]	[1.046]	[0.830]	[3.023]	[2.672]
Female*Lower-Stakes Test	1.112	2.186	3.112**	4.116***	-0.867	-0.118
	[1.962]	[1.768]	[1.417]	[1.261]	[3.826]	[3.425]
Constant	50.36***	51.87***	48.28***	48.89***	49.41***	51.07***
	[2.922]	[1.272]	[1.772]	[0.299]	[3.678]	[0.844]
Observations	181	181	92	92	89	89
R-squared	0.026	0.055	0.025	0.24	0.013	0.003
Student FE	No	Yes	No	Yes	No	Yes

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. The regressions compare lower-stakes test scores with higher-stakes test scores in two subjects (for *Overall* and for *Catalan* and *Math*). The regression for the *Overall* includes a dummy for *Catalan*. Test scores are in levels.

Table 8: Exploiting the different Timings of Low-Stakes Tests

	High-Stake vs. Low-Stake T1	High-Stake vs. Low-Stake T2	High-Stake vs. Low-Stake T3	High-Stake vs. Low-Stake T4	High-Stake vs. Low-Stake T5	High-Stake vs. Low-Stake T6	High-Stake vs. Low-Stake T7	High-Stake vs. Low-Stake T8
Female	0.162***	0.145***	0.139***	0.156***	0.145***	0.144***	0.158***	0.155***
	[0.0336]	[0.0364]	[0.0351]	[0.0338]	[0.0346]	[0.0365]	[0.0352]	[0.0354]
Low-Stakes Test	-0.013	-0.0185**	-0.00732	-0.00966	-0.00816	0.00338	0.0303**	0.0446***
	[0.00932]	[0.00851]	[0.00810]	[0.00903]	[0.00850]	[0.00741]	[0.0136]	[0.0124]
Female*Low-Stakes	0.0309**	0.0381***	0.0197*	0.0228*	0.018	0.0342***	0.0317*	0.0164
	[0.0134]	[0.0118]	[0.0113]	[0.0128]	[0.0118]	[0.0109]	[0.0186]	[0.0171]
Constant	-0.191***	-0.188***	-0.187***	-0.191***	-0.189***	-0.193***	-0.197***	-0.199***
	[0.0240]	[0.0253]	[0.0247]	[0.0241]	[0.0245]	[0.0253]	[0.0247]	[0.0247]
Observations	76,990	77,277	76,830	77,018	77,186	74,054	58,467	55,602
R-Squared	0.005	0.005	0.004	0.004	0.004	0.005	0.004	0.003

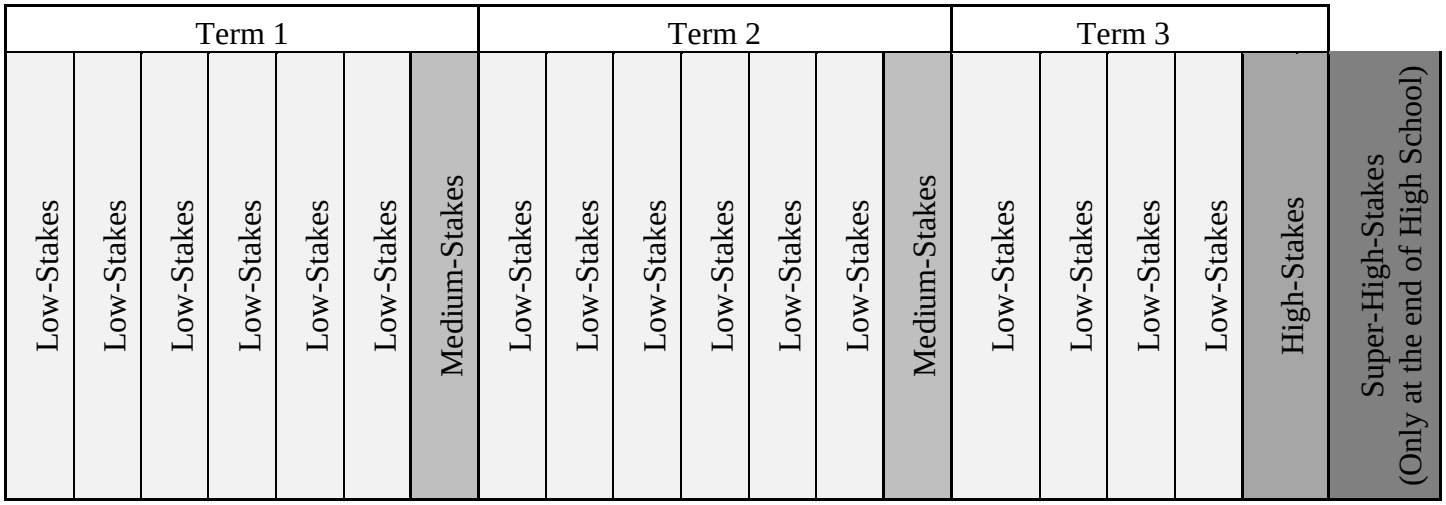
Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. The regressions compare test scores on low-stakes tests taken throughout the year. *Low Stake 1* to *Low Stake 8* are the first to last low-stakes tests in an academic year. The high-stakes exams are compared to each of these low-stakes exams separately in each column. All test scores are standardized.

Table 9: Selectividad Subject versus Non-Selectividad Subject (Level 6 only)

	Test Score	Test Score	Test Score
Female	0.135***	0.0828**	-0.0681**
	[0.0242]	[0.0342]	[0.0345]
Low-Stakes Test	-0.0176	-0.0705**	0.170***
	[0.0242]	[0.0344]	[0.0523]
Selectividad Subject	0.130***	0.130***	0.106**
	[0.0288]	[0.0288]	[0.0485]
Female*Low-Stakes Test		0.105**	0.155***
		[0.0484]	[0.0390]
Female*Selectividad Subject			-0.121**
			[0.0544]
Constant	-0.154***	-0.127***	-0.143***
	[0.0307]	[0.0330]	[0.0376]
Observations	6,426	6,426	6,426
R-squared	0.008	0.009	0.008

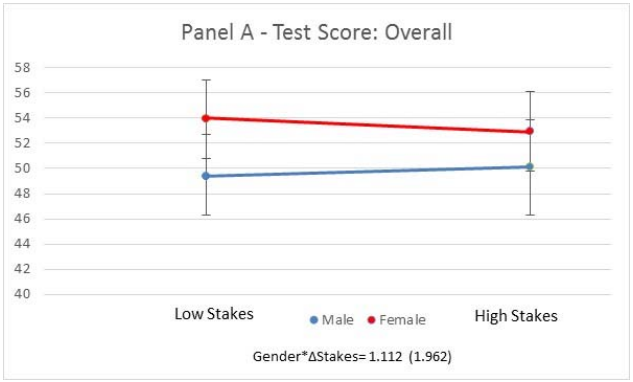
Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. The regressions compare low-stakes test scores with high-stakes test scores. All test scores are standardized. *Low-Stakes Test Score* is the students' test score on the low-stakes tests, *High-Stakes Test Score* is the students' test score on the high-stakes tests. *Selectividad Subject* indicates the subjects in which a student will do the national exams at the end of Level 6.

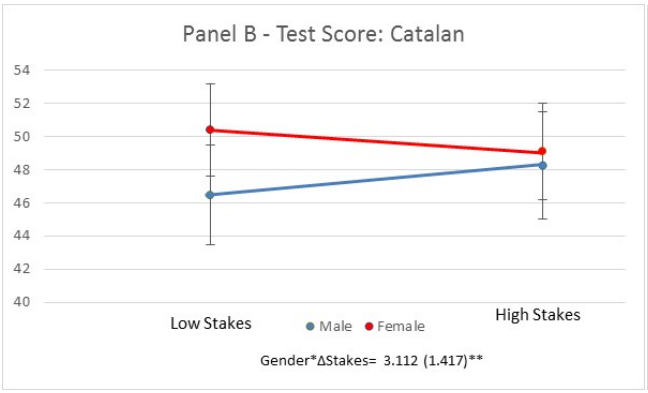
Figure 1: Evaluation System in the School

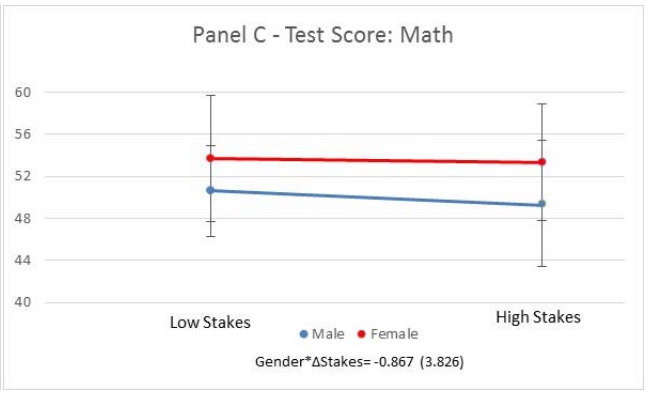


Notes: This is the evaluation system used in each subject. *Low-Stakes* is the test that counts for approximately 2.5 percent of the final grade. *Medium-Stakes* is the test that counts for approximately 11 percent of the final grade. *High-Stakes* is the test that counts for approximately 27 percent of the final grade. *Super-High-Stakes* is the national exam, Selectividad, taken at the end of Level 6, which counts for 50 percent of the university entry test score.

Figure 2: Test Scores in Experiment







Notes: Panels A, B and C present the mean test scores and standard error bands (90% confidence level) in lower- and higher-stakes exams for male and female students in the experiment, together with the interactive coefficient (and standard errors) for gender and stakes from the OLS specification.

Tables and Figures in the Appendix

Table A.1: List of Subjects and their Classification as a Science or Arts Subject

List of Subjects	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Arts-Science
Spanish	X	X	X	X	X	X	Arts
Catalan	X	X	X	X	X	X	Arts
English	X	X	X	X	X	X	Arts
French	X	X	X	X	X	X	Arts
Math	X	X	X	X	X	X	Science
Biology	X	X	X	X	X	X	Science
History-Geography	X	X	X	X	X	X	Arts
IT	X	X	X	X	X		Science
Chemistry-Physics			X	X	X	X	Science
Latin			X	X	X		Arts
Technical Drawing			X		X	X	Science
Art History					X	X	Arts
Contemporary Sciences					X		Arts
Economics					X	X	Science
Math Applied to Social Sciences					X	X	Arts
Philosophy					X	X	Arts
Audiovisual Culture						X	Arts
History Philosophy						X	Arts

Table A.2: Performance under Different Stakes for Main Subjects

	Low vs. High Test Score							
	Biology	Chemistry/Physics	Math	Catalan	Spanish	English	French	History/Geography
Female	0.0354	-0.165***	-0.154***	0.291***	0.294***	0.225***	0.404***	0.038
	[0.0450]	[0.0608]	[0.0446]	[0.0469]	[0.0440]	[0.0495]	[0.0506]	[0.0475]
Low-Stakes Test	-0.0292	-0.0116	-0.0208	-0.0440***	-0.0340**	-0.0363**	0.000979	-0.0336
	[0.0213]	[0.0272]	[0.0191]	[0.0139]	[0.0163]	[0.0143]	[0.0143]	[0.0210]
Female*Low-Stakes	0.0579**	0.0236	0.0415	0.0871***	0.0673***	0.0733***	-0.00263	0.0657**
	[0.0293]	[0.0391]	[0.0257]	[0.0207]	[0.0228]	[0.0203]	[0.0211]	[0.0292]
Constant	-0.0179	0.0810*	0.0772**	-0.147***	-0.149***	-0.113***	-0.206***	-0.0194
	[0.0317]	[0.0441]	[0.0315]	[0.0333]	[0.0309]	[0.0357]	[0.0354]	[0.0330]
Observations	7,116	3,060	9,368	9,578	9,584	9,070	9,097	6,658
R-squared	0.001	0.006	0.005	0.029	0.028	0.018	0.041	0.002

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. Standard errors are clustered at the student level. *Low-Stakes Test Score* is the students' score on the low-stakes tests; *High-Stakes Test Score* is the students' score on the high-stakes tests. *Low-Stakes vs. High-Stakes* compares low-stakes test scores with high-stakes test scores. Each column presents the results for a given subject (*Biology*, *Chemistry/Physics*, *Math*, *Catalan*, *Spanish*, *English*, *French*, *History/Geography*).

Table A.3: Probability of Leaving School

	Level 1 Pr.(Leaving)	Level 2 Pr.(Leaving)	Level 3 Pr.(Leaving)	Level 4 Pr.(Leaving)	Level 5 Pr.(Leaving)	Level 6 Pr.(Staying)
Female	-0.21	-0.4	0.00874	0.221	-0.0683	0.388
	[0.316]	[0.308]	[0.0925]	[0.167]	[0.0650]	[0.382]
Av. Final Test Score	-0.0802***	-0.101***	-0.0769***	-0.0919**	-0.0229**	0.542***
	[0.00803]	[0.0234]	[0.00898]	[0.0405]	[0.00994]	[0.0833]
Female*Av. Final TS	0.026	0.0496*	0.00258	-0.0288	0.0129	-0.0774
	[0.0227]	[0.0291]	[0.0148]	[0.0253]	[0.00979]	[0.0707]
Observations	622	707	771	825	875	875
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. *Pr. (Staying)* is the probability that the student will stay in school for all six levels. *Pr. (Leaving)* is the probability that the student will leave at the end of the academic year. *Av. Final Test Score* is the average final test score over all subjects at the end of the academic year. Controls include year and subject dummies.

Table A.4: Probability of being Exempt from High-Stakes Test

	Overall Pr.(Exempt)	Level 1 Pr.(Exempt)	Level 2 Pr.(Exempt)	Level 3 Pr.(Exempt)	Level 4 Pr.(Exempt)	Level 5 Pr.(Exempt)	Level 6 Pr.(Exempt)
Female	0.00268	0.138**	0.0931*	-0.133	-0.111	-0.0928	0.0482
	[0.0351]	[0.0542]	[0.0556]	[0.0934]	[0.0776]	[0.0958]	[0.0426]
Score_LS_Test	0.0646***	0.0806***	0.0718***	0.0567***	0.107***	0.0558***	-0.00128
	[0.00401]	[0.00665]	[0.00576]	[0.0111]	[0.00824]	[0.0120]	[0.00328]
Female* Score_LS_Test	0.00103	-0.0214**	-0.0114	0.0227	0.0185	0.014	-0.00581
	[0.00532]	[0.00841]	[0.00866]	[0.0138]	[0.0117]	[0.0136]	[0.00557]
Observations	38,409	8,824	8,013	7,079	6,555	4,565	3,373
R-squared	0.144	0.201	0.175	0.229	0.237	0.235	0.198
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% level, and *** denotes significance at the 1% level. *Pr. (Exempt)* is the probability that the student is exempt from a subject's high-stakes test. *Score_LS_Test* is the average test score in the low-stakes tests in the subject from which the student is exempt. Controls include year and subject dummies.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1313	Saul Estrin Ute Stephan Sunčica Vujić	Do Women Earn Less Even as Social Entrepreneurs?
1312	Nicholas Bloom Renata Lemos Raffaella Sadun John Van Reenen	Does Management Matter in Schools?
1311	Erling Barth Alex Bryson James C. Davis Richard Freeman	It's Where You Work: Increases in Earnings Dispersion across Establishments and Individuals in the US
1310	Christos Genakos Pantelis Koutroumpis Mario Pagliero	The Impact of Maximum Markup Regulation on Prices
1309	Gianmarco I.P. Ottaviano Filipe Lage de Sousa	Relaxing Credit Constraints in Emerging Economies: The Impact of Public Loans on the Performance of Brazilian Manufacturers
1308	William Fuchs Luis Garicano Luis Rayo	Optimal Contracting and the Organization of Knowledge
1307	Alex Bryson Richard B. Freeman	Employee Stock Purchase Plans – Gift or Incentive? Evidence from a Multinational Corporation
1306	Andrew E. Clark Sarah Flèche Claudia Senik	Economic Growth Evens-Out Happiness: Evidence from Six Surveys
1305	Jorge De la Roca Gianmarco I.P. Ottaviano Diego Puga	City of Dreams

1304	Jan-Emmanuel De Neve George W. Ward Femke De Keulenaer Bert Van Landeghem Georgios Kavetsos Michael I. Norton	Individual Experience of Positive and Negative Growth is Asymmetric: Evidence from Subjective Well-being Data
1303	Holger Breinlich Anson Soderbery Greg C. Wright	From Selling Goods to Selling Services: Firm Responses to Trade Liberalization
1302	Esteban Aucejo Teresa Foy Romano	Assessing the Effect of School Days and Absences on Test Score Performance
1301	Gianmarco I.P. Ottaviano	European Integration and the Gains from Trade
1300	Antoine Dechezleprêtre Ralf Martin Myra Mohnen	Knowledge Spillovers from Clean and Dirty Technologies
1299	Stephen Machin Richard Murphy	Paying Out and Crowding Out? The Globalisation of Higher Education
1298	Iain Cockburn Jean O. Lanjouw Mark Schankerman	Patents and the Global Diffusion of New Drugs
1297	David W. Johnston Grace Lordan Michael A. Shields Agne Suziedelyte	Education and Health Knowledge: Evidence from UK Compulsory Schooling Reforms
1296	Fabio Pinna Stephan Seiler	Consumer Search: Evidence from Path-tracking Data
1295	Matt Dickson Paul Gregg Harriet Robinson	Early, Late or Never? When Does Parental Education Impact Child Outcomes?

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7404 0612
Email info@cep.lse.ac.uk Web site http://cep.lse.ac.uk