Journal of Economic Behavior and Organization 191 (2021) 1065–1090

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Gender differences under test pressure and their impact on academic performance: A quasi-experimental design☆

Daniel Montolio[a,b,a,*], Pere A. Taberner[a,b,c]

[a] Department of Economics, Universitat de Barcelona, Av. Diagonal 690, Barcelona 08034, Spain
[b] Institut d'Economia de Barcelona (IEB), c/ John M. Keynes 1-11, Barcelona 08034, Spain.
[c] Knowlegde Sharing Network (KSNET), c/ Mallorca 100, Barcelona 08029, Spain

ARTICLE INFO

Article history:
Received 30 November 2020
Revised 31 August 2021
Accepted 16 September 2021
Available online 24 October 2021

JEL classification:
A22
I24
J16

Keywords:
Gender differences
Stakes
Pressure
Academic performance
Higher education
Field data

ABSTRACT

Student performance at university is a strong determinant of individual decisions and future outcomes, most notably labour opportunities. Although published studies have found gender differences in student performance in response to pressure, little is known about such differences when university students respond to different levels of pressure, resulting from different stakes. Based on field data, this study aims to examine gender differences in student performance in response to different stakes when sitting multiple choice tests, a frequently employed exam format at university. To do so, the introduction of continuous assessment in the evaluation system of a university course allows us to exploit a unique quasi-experimental set up in which the same students take similar tests throughout the course but under different levels of pressure, i.e. facing different stakes. Exploiting individual student data in a panel data framework, we find that male students outperform their female counterparts when under high pressure. However, as the stakes faced decrease, the gender gap shrinks and even reverses in favour of female students at the lowest pressure scenario. We also analyse possible mechanisms responsible for the observed gender gap by studying whether students excel or choke under pressure depending on their gender, and by studying gender differences when omitting questions on multiple choice tests.

1. Introduction

Student performance is a strong determinant of individual decisions and future outcomes. In many countries, university choice depends on the grades obtained throughout high school and on university entrance examination results. Moreover, university final Grade Point Average (GPA) may play an important role when applying to postgraduate degrees and certain types of jobs. All this means that student performance both at high school and at university are strong determinants of

☆ We would like to thank the Editor and three anonymous referees for helpful insights and comments. Our gratitude also to Pedro Rey-Biel, José Montalbán, Andreu Arenas, Estíbaliz Royuela-Colomer, participants to the *27th Encuentro de Economía Pública* 2020 and participants to *GEEZ Seminars* for very useful comments and feedback. Special thanks to all the members of GIDEI (Grup d'Innovació Docent en Economia dels Impostos) for the data and their insightful comments. Thanks to the Faculty of Economics and Business of the University of Barcelona for the data, and especially to José Luís Andújar. Daniel Montolio acknowledges financial support from project 2017SGR796 (Generalitat de Catalunya). All remaining errors are our own.

* Corresponding author.
E-mail addresses: montolio@ub.edu (D. Montolio), peretaberner@gmail.com (P.A. Taberner).

https://doi.org/10.1016/j.jebo.2021.09.021

future labour outcomes (Rose, 2006; Bulman, 2017). Indeed, the well-documented gender gap in labour outcomes (see, for example, Lavy, 2013; Goldin et al., 2017; Blau and Kahn, 2017; Kunze, 2018) highlights the need to understand the potential determinants of gender differences in academic performance. This would allow to mitigate differences across gender and to promote increased equality in education and, hence, in labour opportunities.

Students at higher education sit a significant number of examinations exposing them to different levels of pressure, depending on the number of credits at stake, the weight of the test in the final grade, the specific rules of assessment which might require students to perform better, the length of the test,[1] the difficulty of the content, the type of examination or the simple fact of having to sit an exam. In this context, we focus on the pressure resulting from the stakes, mainly determined by the weight of the test on the final grade and the specific rules students face when sitting exams, and how it can differently affect performance depending on their gender.

The goal of this paper is, therefore, to study gender differences in academic performance under different levels of pressure which are determined by the stakes at hand. Here, the introduction of continuous assessment in the evaluation system of a university course allows us to exploit a unique quasi-experimental setting in which the same students take similar tests but exposed to different stakes. The main sources of pressure on students, i.e. the stakes they face, are determined by both the greater weight attached to a test on the final grade and a specific rule of evaluation (an eliminatory rule) that require a better performance from the students. The use of unique student administrative data provides us with a rich individual-based data set to control for personal and group characteristics that might also determine student performance and affect gender differences. Panel data results show that male students perform better than female students do when under greater pressure, but that this gender gap narrows as pressure decreases until mitigated and even reversed in favour of female students.

Few studies have analyzed the role of pressure in educational performance across gender using real-world settings; that is, little is known about how pressure affects student performance across gender at the university level.[2] Moreover, to the best of our knowledge, only one economic study to date has attempted to analyse pressure as the weight attached to an exam, albeit in this instance at high school level (Azmat et al., 2016). The strength of our quasi-experimental set up is based on five specific characteristics: (i) the tests are computer corrected so require no subjective bias correction, (ii) the tests present an almost identical format (multiple choice), employing the same questions, with the same level of difficulty and a very similar structure, (iii) same cohorts of students sit these tests in scenarios characterised by different levels of pressure, (iv) the pressure students are under is analysed in a real world environment, i.e. sitting their university exams. Moreover, our data and empirical strategy allow us to explore the possible mechanisms responsible for the results obtained, that is, we are able to disentangle the main drivers of the gender differences observed.

The paper is structured as follows. Section 2 reviews the most recent and relevant literature on the topic. Section 3 describes the quasi-experimental setting, explaining the specific evaluation system used on the course and the sources of pressure. Section 4 explains the data used and presents the main descriptive statistics. Section 5 outlines the empirical strategy and Section 6 reports the results and identifies the possible mechanisms responsible for the results obtained. Finally, the last section discusses the results and presents the conclusions.

2. Literature review

Numerous studies have attempted to explain the determinants of gender differences in educational achievement. In the literature, the "*nurture vs nature*" heated debate has been widely explored to determine the main roots of such gender differences (González De San Román and De La Rica, 2016; Dee, 2007).[3] The economic literature, in general, has tried to identify these environmental mechanisms;[4] however, the consensus is that the gender gap is multifactorial and it is difficult to narrow the determinants down to just a few factors. One of these factors studied is the pressure that students face when sitting exams, induced by different kind of sources such as competition, time constraints or stakes.

[1] In January 2018, the University of Oxford had decided to extend the time available to computer science and maths students to complete their exams. According to Telegraph (2018), the decision was taken for two reasons: the gender grade gap between students being awarded first-class degrees in these subjects and evidence suggesting that females underperform when under time pressure. Accordingly, the board of examiners decided to lengthen exams from 90 to 105 min (Mail, 2018).

[2] Studies analysing competitive pressure (understood as the set up in which a student's performance depends on the efforts and results of other students) use data from schools, national contests, or university entrance exams. For instance, Örs et al. (2013) examine student performance on entrance exams to a master's program (their competitive setting), but they then compare this gender gap with performance during the master's (their non-competitive setting) due to the increasing importance attached to that test.

[3] Advocates of biological determinants (nature) argue that innate differences determine the gender gap in student achievement, while the proponents of environmental determinants (nurture) claim that the main drivers are social and cultural norms.

[4] The most frequently examined causes to date have been family interactions (e.g. González De San Román and De La Rica, 2016; Farré and Vella, 2013; Rodríguez-Planas and Nollenberger, 2018), teacher-student interactions (e.g. Hoffmann and Oreopoulos, 2009; Muralidharan and Sheth, 2016; Lim and Meer, 2017; 2020), competitive pressure (e.g. Örs et al., 2013; Jurajda and Münich, 2011; Pekkarinen, 2015), and self-beliefs (Contini et al., 2017; Lubienski et al., 2013), among others.

2.1. Psychology literature

In a seminal study, Baumeister (1984) defines the term "*choking under pressure*" as a decline in performance when an individual faces a situation that specifically calls for an improved performance, with pressure being considered as "any factor or combination of factors that increases the importance of performing well on a particular occasion" (Baumeister, 1984, p.610). He shows, by way of three experiments, that people's performance decreases under different pressure scenarios: namely, implicit competition, cash incentives and audience-induced pressure. Similarly, Beilock (2011) suggests that a stressful situation can result in an individual performing below their true potential. She reports that students might choke when rewards are high and that high-skilled students are more likely to choke under pressure in important exams. She suggests that the mechanisms behind the process could be too much concern for the reward itself or gender stereotypes, among others.

A large body of the psychological literature has focused on the causes of test anxiety, on its links with academic achievement and on gender differences in relation to this type of anxiety. Putwain et al. (2010) report a relationship between test anxiety and parental/teacher pressure, achievement goals and the importance of good performance. Likewise, Chin et al. (2017) find that test anxiety is attributable to worry and the social repercussions of failing. von der Embse et al. (2018) perform a meta-analysis gathering 30 years of psychological research on test anxiety and its predictors (from 1988 to present). They review 238 published studies with the goal of examining the relationship between test anxiety and educational performance and the influence of individual characteristics on test anxiety. The results show a negative relationship between test anxiety and educational achievement in terms of students' average grades, outcomes of university entrance examinations and standardized tests. Additionally, self-esteem, test difficulty, importance and consequences of an exam are associated with higher test anxiety. Of relevance to us here, several psychological studies provide evidence that female students report higher levels of test anxiety in higher education (see, for example, Núñez Peña et al., 2016; Eman et al., 2012; Backović et al., 2012). Núñez Peña et al. (2016) propose two possible explanations for gender differences in relation to test anxiety, both related to social gender roles. One explanation is that women are under greater social pressure to perform well at university and so are more worried by exams. The other is that men are less likely to admit their real anxiety as they consider they should show themselves to be emotionally strong.

2.2. Economic literature

Few studies in the economic literature focus on the role of pressure on gender differences in scholastic achievement using field data. These works have been looking at the following sources of pressure: competition, time and the stakes involved, being competitive pressure the most studied. There is only one paper, to the best of our knowledge, which defines pressure as we do: according to the weight (stake) of an exam within the overall grade for the year (Azmat et al., 2016). They analyse gender differences in school performance and university entrance exams when placed under different levels of pressure on primary and secondary level. They use data from the six years of high school and the entrance exams to university. The midterms during the year are classified as low-stake, the final exam at the end of each term as medium-stake, the final exam at the end of the year as high-stake, and entrance exams as super-high-stake. Their results suggest that females outperform males on all tests, but to a relatively higher degree when the stakes are low. However, this gender gap disappears in the case of university entrance tests.

Studies analysing the gender gap when under competitive pressure have been undertaken primarily within secondary and high schools, and in relation to university entrance examinations or national contests. Jurajda and Münich (2011) analyse the gender gap in university admission rates in the Czech Republic under competitive pressure. Using secondary school results and data from student performance in the entrance examination, they group each university according to that year's admission rate by quartiles. In this way, they can define the competitiveness of the university according to the admission rate. Their results suggest that boys outperform girls (i.e. they are more likely to be admitted) when applying in a competitive scenario, but there is no gender gap at less selective universities. The authors control for both skills and subject of study. Likewise, Örs et al. (2013) use data from HEC Paris, the prestigious business school. They focus on the performance of students applying to the MSc in Management, one of the best master's programs in business in Europe and one offering great job opportunities. They find that men outperform women in the HEC entrance exam (competitive setting), while the same women outperform men in the national *baccalauréat* exam and the first year of the master's program (non-competitive settings). In the same vein, Cai et al. (2019) analyse gender differences on the Chinese university entrance examination, defined as a highly competitive and high-stakes situation. They also find that male students outperform female students in this exam, while this gender gap is narrower in a previous low-stakes examination. Arenas and Calsamiglia (2020) exploit a policy reform in the Spanish university entrance exams in which stakes at hand increased due to the variation of the weight of those exams on the university entrance grade. They find that the implementation of this policy has a negative effect on the performance of female students and this effect was larger among the top students. This lead a decrease on the proportion of female students in the most wanted undergraduate studies.

On these previous studies, the exam format is not the main focus. In fact, Azmat et al. (2016) and Örs et al. (2013) analyse examinations with different formats such as multiple choice, open-question and oral. However, some authors have specifically focused on the gender gap under competitive pressure in multiple choice tests. Pekkarinen (2015) studies the gender gap in performance in the Finnish university entrance exams, scenario which is defined as highly competitive, and

Iriberri and Rey-Biel (2019) analyse the gender gap in secondary students participating in a two-stage maths competition in Madrid in which only the best performers in the first stage continue to the second one. Both studies find the same results: male students outperform female students on these multiple-choice examinations and the mechanism behind this gap is because girls omit more items. In addition, Iriberri and Rey-Biel (2019) find that the gender gap is greater in the second stage (higher competitive environment) than in the first (lower competitive pressure), while there are no gender differences in maths grades at school. They argue that the decision to omit a test item might be driven by gender differences on lower levels of confidence or risk aversion (see also, Balart et al., 2020; Karle et al., 2020). Therefore, these results highlight gender differences in answering multiple choice questions, as other studies have already shown (Akyol et al., 2016; Riener and Wagner, 2017; Baldiga, 2014; Espinosa and Gardeazabal, 2020).

Finally, De Paola and Gioia (2016) analyse gender differences under time pressure on student performance. They ran a field experiment in a course at the University of Calabria (Italy) where students were given the opportunity to choose between two evaluation schemes at the beginning of a course: the traditional system or their experimental alternative. Students opting for the latter were randomly selected into one of two groups. In one group, they sat the first midterm test with no time pressure and the second under time pressure, while in the other group the time pressure conditions were reversed. They find that overall students perform worse under time pressure, but that this effect is due specifically to the underperformance of female students. In the same vein, Shurchkov (2012) run a laboratory experiment among university students to analyse gender differences under competitive and time pressure when completing math and verbal tasks. They find that females underperform males in the math tasks under high pressure but outperform males in verbal tasks under low pressure. Concretely, they find that this increasing performance in verbal tasks is due to the more time given to the participants. Apparently, women use this extra time in improving the quality of their performance, while men use it to increase the quantity of work leading to a higher proportion of mistakes.

3. Quasi-experimental setting

The present paper takes advantage of the evaluation system employed on a course entitled *Principles of Taxation* taught at the University of Barcelona, which allows us to exploit a quasi-experimental design to address our research question. Our empirical set up is made possible by (i) the implementation of a continuous assessment (hereinafter, CA) on the course, this is a evaluation system in line with the European Higher Education Area (EHEA) guidelines, also known as Bologna Process[5]; and, (ii) the specific design adopted for the evaluation system and its evolution during the academic years analysed; both resulting in different stakes faced by students when sitting their exams.

The Faculty of Economics and Business (University of Barcelona) implemented the Bologna Process in its Bachelor's Degree in Business Administration in the 2011/12 academic year. This meant that teachers had to redefine the courses on the Bachelor program, rethink the teaching process and introduce CA (Gallardo et al., 2010). The CA means an evaluation system conducted over the teaching term, and not only with a final examination at the end of the term. Therefore, students must pass through several evaluation tasks such as midterms, assignments or papers, before the final exam. *Principles of Taxation* was no exception, and prior to this date, the course coordinators opted for the gradual introduction of CA. Thus, two years before the official implementation of the Bologna Process, in the 2009/10 academic year, professors introduced a pilot CA system as an alternative to that of single assessment (SA).

The adoption of the CA divides the course's evaluation system in two periods: first, the CA over the term and, second, a final exam at the end of the term.[6] In the first period, the CA is conducted at the same time as lectures are delivered, and it is based on various multiple choice midterms during the term covering the content introduced between each midterm. In the second period, the final exam is sat at the end of the term, after lectures are finished and over the evaluation weeks. The final exam contains two parts: students, first, take a multiple choice test and, second, they complete an open-question test. For comparability reasons, we make use only of the multiple choice part of the final exam.[7] We use seven academic years of the course – from 2009/10 to 2015/16 – since the multiple choice element of the final exam was eliminated in 2016/17.

We exploit the comparability between CA midterms and the multiple choice part of the final exam, since they contain the same kind of questions and share a similar structure. All the questions are designed in the same way by the same teachers with the same level of difficulty. Each multiple choice item comprises four options of which only one is correct and three incorrect. Students score 1 point for each correct question and lose 0.25 of a point for each wrong question, while omitting the item altogether has no effect on their score. However, while the midterms are computer-based, the final exam is completed on paper.[8] Students have 30 min in each midterm to answer ten multiple choice questions plus two

[5] For further information about EHEA and Bologna Process: http://www.ehea.info.

[6] Given the panel structure used in our empirical set up we have decided to call æperiodsg the two broad moments of the evaluation that the students have to face. Note, according to Faculty rules, students can, however, opt out of CA and take the SA, i.e. a single final exam, constituting the sole form of student assessment for a given course.

[7] Hereinafter, and for the sake of clarity, when we use in the text the term æfinal examg we are referring to the multiple choice questions of the final examination.

[8] The literature is inconclusive whether there are gender differences in answering computer-based or paper-based exams. The only economic study, Marcenaro-Gutiérrez and López-Agudo (2016), finds gender differences in favour of females answering on computer, but this gap varies according the

or three small exercises in which they need to solve a problem by entering a number into a box. Similarly, students have 30 min to answer 20 multiple choice questions of the first part of the final exam. We are able to check whether these small differences are a threat to our estimates by performing an heterogeneity analysis. As explained below, we compare results across CA midterms (and across final exams) with varying degrees of pressure, hence, cancelling any possible difference between midterms and final exam's multiple choice questionnaires.

Table 1 presents a summary of the evaluation system and the main sources of pressure we analyse.[9] Our two direct measures of the stakes the students face are given by, on the one hand, the weight of the CA or the final exam in the overall course grade and, on the other hand, on the fact that the multiple choice part of the final exam is eliminatory during four years. This is, students have to fulfil at least one of two requirements: (i) score 4 or more on this multiple choice part of the final exam[10] or, (ii) obtain an average CA grade of 5 or more. If neither requirement is met, the open-question part of the final examination is not marked and the student automatically fails the whole course.

We define the CA part of the course – the first period – as low stakes or low pressure scenario given the weight of midterms in the overall course grade compared with that of the final exam. In contrast, the final exam – second period – is defined as a high pressure scenario due to the greater weight attached to it in the overall course grade. This setting can be further defined as a sequential game of two periods in which same students first sit low-stakes midterms and then sit a high-stakes final exam.

Moreover, the eliminatory rule included within the evaluation system also further determines the stakes to which students are subject and adds some relevant heterogeneity over the timespan analysed. Table 1 shows that during the first four academic years, the eliminatory nature of the multiple choice part of the final exam makes the pressure even higher. This is, in case of not fulfilling the requirements previously presented, the open-question part is not marked and the student automatically fails the whole course. To account for this additional source of pressure, we divide the timespan into two intervals of years, indicated by the dashed line in Table 1, with more homogeneous stakes faced by students over the academic years within those intervals. Thus, we compare different cohorts of students (in the 2 intervals) exposed to different levels of pressure according to the specific eliminatory rule which requires students to perform better.[11]

Table 1
Summary evaluation system for *Principles of Taxation*.

Academic year	Stakes from the % on course grade		Stakes from the evaluation design	
	1st Period CA midterms	2nd Period Final exam	Multiple choice part eliminatory in final exam	
2009/10	10%	90% (30%)	Yes	1st Interval
2010/11	20%	80% (27.7%)	Yes	
2011/12	30%	70% (23.3%)	Yes	
2012/13	40%	60% (20%)	Yes	
2013/14	40%	60% (20%)	No	2nd Interval
2014/15	40%	60% (20%)	No	
2015/16	40%	60% (20%)	No	

Note: In parenthesis, we report the weight of the final multiple choice test on the overall course grade. As explained in more detail in Appendix A (see Table A.1), students sit a different number of CA midterms over academic years. Moreover, in some of these years, they could discard one midterm. This implies a midterm weight varying between 3.33 and 20% of the overall course grade.

Therefore, we define the first interval, from 2009/10 to 2012/13, as high pressure and the second, from 2013/14 to 2015/16, as low pressure due to this eliminatory rule of the multiple choice part of the final exam. Table 2 summarises the four scenarios of pressure resulting when interacting the stakes from the periods (the % of the CA and final exam on the final course grade) with the stakes shaped by the eliminatory rule (which pushes students to perform better, either in the CA and in the final exam). This rule also provides higher pressure over the CA midterms in the first interval than in the second one because students do not need to score a minimum of four on the final exam in case of passing the CA. Thus, independently of the weight on the final course grade, students have incentives to perform well and pass the CA during this first interval. Additionally, the CA is associated with lower pressure in the second interval due to the formula to obtain the

competences assessed or the previous training or use of ICT. However, studies from other fields find mixed results on gender differences in this topic, either there is no differences or differences are in favour of female students (Wallace and Clariana, 2005; Kies et al., 2006).

[9] Further details on the evaluation system can be found in Appendix A and Table A.1.

[10] The Spanish system usually defines grades between 0 and 10.

[11] The suppression of the eliminatory rule by the teaching staff of Principles of Taxation was not induced by the aim of mitigating gender differences in performance across students. It was suppressed because it was one of the primary sources of students' complaints, precisely for the pressure that the rule exerted on them.

overall course grade.[12] It is computed as the maximum between (i) the whole final exam grade (including the two parts) and, (ii) the weighted average grade of the CA grade and the final exam grade (see Table A.1 in Appendix A for more details). Therefore, a poor CA performance in the second interval can be rectified with a good performance solely in the final examination.

Table 2
Stake scenarios coming from the evaluation system.

Periods	Intervals	
	2009/10–2012/13 (High stakes)	2013/14–2015/16 (Low stakes)
CA midterms (Low stakes)	Low / High	Low / Low
Final exam (High stakes)	High / High	High / Low

Note: each cell presents the combination of the stakes faced due to both the type of evaluation (CA midterms vs final exam) and the interval considered (2009/10–2012/13 vs 2013/14–2015/16).

On top of that, our setting is characterized by a number of factors that allow us to effectively investigate gender differences in academic performance when students are exposed to different levels of pressure. First, multiple choice tests are corrected by machine/computer, and not by teachers. This allows us to avoid any possible teacher gender-bias or bias towards specific subgroups of students, as some authors have suggested (see, for exemple, Falch and Naper, 2013; Goldin and Rouse, 2000). Second, students sit similar exams: that is, multiple choice format, comprising the same kind of questions with a similar level of difficulty. Thus, the effects that arise from comparing completely different types of test or degrees of test difficulty are not a concern here, though we are aware of small differences between them. Third, we focus essentially on students who complete both the CA component and the final exam. This group of students might have different characteristics compared to students that, for instance, sits only the final exam. Given this situation, we analyse the potential bias of self-selection using the Heckman (1979) procedure.

4. Data and descriptive statistics

This study uses data from two sources: administrative and course data. First, the administrative data was provided by the University of Barcelona's Faculty of Economics and Business. They contain full demographic and academic information for all students enrolled on the course *Principles of Taxation* over the seven academic years. Second, the course data were provided by the *Economics of Taxation Teaching Innovation Group* (GIDEI, from the acronym in Catalan). It contains a full set of information about the grades and groups.

4.1. Administrative and course data

The administrative data comprises two sorts of student information: demographic and academic. The demographic information includes student gender, date of birth, country of birth, province of birth, city of birth, nationality and student ID. The academic data contains general information about the whole undergraduate program and specific information for the year that the student is enrolled on *Principles of Taxation*. The general information includes the student's access path to the degree, university access grade, the year of starting the degree and the GPA for the whole undergraduate program. The specific information includes the academic year in which the student took *Principles of Taxation*, the number of cumulative credits passed – including those passed in that year, the courses the student is enrolled on that year (course title, group, term and final grade) and whether that year the student win a scholarship. The Faculty's administrative data allow us to compute rich vectors of control variables for individual and group characteristics.

The course data comprises two sorts of information: grades and group information.[13] The grade information includes student ID, academic year, grade of each midterm, average CA grade, final exam grade, detailed information on the final multiple choice test (number of questions answered correctly, incorrectly and omitted), the grade for each question in the open-question part, and the overall course grade. The group information also contains details about the group teaching schedule, the teachers assigned to each group, teacher gender and language of instruction (English being employed with some groups). The course data provides us with all the student grades plus details for computing the control variables for group characteristics.

[12] Students know at the beginning of the course all the evaluation rules.

[13] In our set up a group implies that the same number of students (around 90–100 per group) is taught by the same instructor(s) for the whole term (maximum of two different instructors per group). That is, a group is a similar concept of a class at primary or secondary school.

Table 3
Descriptive statistics for the CA and final exam grades - balanced sample.

	CA Grade (Low-stakes)			Final Exam Grade (High-stakes)		
	Male	Female	Statistic	Male	Female	Statistic
1st Interval - High-stakes ($N = 1449$)						
N	684	765		684	765	
Percentage (%)	47.20	52.80		47.20	52.80	
$H_o^a : Pr(F_i) = 0.5$			2.13**			2.13**
$H_o^b : Pr(F_i) = Pr(M_i)$			3.01***			3.01***
Mean Test	5.34	5.44	−1.02	4.42	4.35	0.68
Median Test	5.44	5.50	0.20	4.38	4.38	0.00
KS Test			0.05			0.03
10th percentile	2.82	3.04		1.88	1.88	
25th percentile	4.18	4.44		3.00	3.00	
75th percentile	6.63	6.71		5.75	5.63	
90th percentile	7.58	7.63		7.13	6.88	
2nd Interval - Low-stakes ($N = 2463$)						
N	1337	1126		1337	1126	
Percentage (%)	54.28	45.72		54.28	45.72	
$H_o^a : Pr(F_i) = 0.5$			−4.25***			−4.25***
$H_o^b : Pr(F_i) = Pr(M_i)$			−6.01***			−6.01***
Mean Test	4.74	5.13	−4.09***	5.23	5.15	1.01
Median Test	4.88	5.44	15.91***	5.25	5.13	1.79
KS Test			0.09***			0.03
10th percentile	1.32	1.57		2.88	2.90	
25th percentile	2.82	3.63		3.90	4.00	
75th percentile	6.63	6.88		6.50	6.38	
90th percentile	7.82	8.00		7.60	7.40	

Note: In line with the definition of a balanced sample, students opting for CA and sitting the final exam are the same. Therefore, the number of males and females, their percentages and tests A and B are the same for CA and the multiple choice part of the final exam. The null hypothesis for test A (H_o^a) is that the proportion of females (F_i) is equal to 50% and for test B (H_o^b) that the proportion of males (M_i) and females (F_i) are equal, where i denotes the CA or final exam sample. Z-statistic for Test A and B. The null hypothesis for the Mean Test is equal mean grades across the gender (unequal variances), t-statistic. The Median Test is a non-parametric 2-sample test in which the null hypothesis is equal medians across gender, chi-squared test statistic with continuity correction. The KS Test is the Two-sample Kolmogorov-Smirnov (KS) Test in which the null hypothesis is equal grades distribution (CA or final exam, respectively) across gender, D-statistic. *** denotes significance at 1% level, ** the 5% level and * the 10% level.

4.2. Sample and descriptive statistics

The database covers the timespan between 2009/10 and 2015/16, i.e. seven academic years. According to the administrative data, 5464 students enroll on the first term course, *Principles of Taxation*, during this time. Students enrolled in groups taught in English were removed given the very different format of their evaluation system compared to that of the Catalan/Spanish-taught groups. The first two academic years, 2009/10 and 2010/11, correspond to the pre-Bologna program. Those students who at that time are in the final year of their degree or have registered for all the credits to obtain the degree, are removed from the sample, since for these students, the eliminatory rule applied to the multiple choice part of the final exam do not affect them. This would mean their adopting a different strategic behaviour and having to face completely different levels of pressure on the CA component and final exam. Therefore, we are left with 5013 students, i.e. 92% of all students enrolled on *Principles of Taxation* in the seven-year time span. Figure B.1 in Appendix B shows the number of students who opt for CA or SA, the number of students who sit the midterms but do not sit the final exam and the number of students who sit neither midterms nor the final exam.

Table 3 shows the main descriptive statistics for the balanced sample (students opting for CA and sitting the final exam) in each interval explained in Section 3 and depicted at Table 2. The first interval – high-stakes – comprises 1449 students, and the second interval – low-stakes – 2463 students. On the one hand, in the first interval, the percentage of male students is 47.2%, which is significantly different from the percentage of female students. Mean and median for the CA grade are higher for female students but not statistically significantly different from male students. However, the mean of the final exam is higher for male students, not statistically significantly different from female students, and the median is the same. In terms of percentiles, the gender grade gap in favor of female students falls at higher percentiles in the case of CA. However, in the final exam, while there are no differences across gender in the first decile and the 25th percentile, a gender gap favoring male students emerges in the 75th percentile and increases in the last decile.

On the other hand, in the second interval, the percentage of male students is 54.3%, significantly different from the percentage of female students. Mean and median for the CA grade are 0.39 and 0.56 higher for female students, respectively, statistically significantly different at 1% level. However, the final exam grade's mean and median are higher for male students but not statistically significantly different. In terms of percentiles, the gender grade gap favoring female students falls at higher percentiles in the case of CA, the same as in the first interval. However, in the final exam, while the gender gap favors female students in the first decile and the 25th percentile, it favors males in the 75th percentile and increases in

the last decile. Therefore, there are differences across the intervals, highlighting the importance of taking this into account in the empirical analysis. In Appendix B, Fig. B.2 shows the kernel distributions of the CA grade and final exam grade by gender, and Fig. B.3 shows the kernel distributions of the CA grade and final exam grade by gender in each interval.

5. Empirical strategy

The identification strategy involves analysing the gender gap in student performance when exposed to different stakes while enrolled on the course *Principles of Taxation*. Given the nature of our data set and the structure of the evaluation system we can rely on a panel data strategy in which we follow same students over our two periods of evaluation.[14] This is, students complete the CA in $t = 1$, i.e. first period, and then, they take the final exam in $t = 2$, i.e. second period. Notice that academic year is not the time variable in this setting, rather it is a student characteristic. This two-period panel data model that we develop is similar to that presented in Iriberri and Rey-Biel (2019), but we introduce a number of modifications to strengthen the identification strategy.[15] We introduce a set of individual variables, we control for group characteristics and we have information for seven cohorts of students. Furthermore, we perform a Quantile Regression Panel Data (QRPD) with period fixed effects (FE, hereafter). This allows us to analyse, not only the potential average gender difference, but also gender differences along the student performance distribution. Given the vast evidence on differences along the student distribution (see, for example, Örs et al., 2013; Rask and Tiefenthaler, 2008; Meghir and Rivkin, 2011), we cannot ignore this in our setting.

We first estimate test grades on student gender controlling for individual, group and year characteristics by OLS with period FE (Pooled OLS) and random effects (RE) to estimate average gender differences. The baseline econometric specification can be expressed as follows:

$$Grade_{igyt} = \alpha_0 + \alpha_1 \cdot Female_{igy} + \alpha_2 \cdot Final\_Exam_t + \alpha_3 \cdot Female_{igy} \cdot Final\_Exam_t + \alpha_4 \cdot X_{igy} + \mu_{gy} + \varepsilon_{igyt} \quad (1)$$

where the dependent variable $Grade_{igyt}$ denotes the grade obtained by student i in group g in academic year y and during period t, where $t = 1$ refers to CA and $t = 2$ to the final exam, $Female_{igy}$ is a dummy variable which takes value 1 if the student is female and 0 if male, $Final\_Exam_t$ is the period FE, i.e. a dummy variable which takes a value of 1 if it is the second period or 0 if the first, X_{igy} is the vector of individual controls,[16] μ_{gy} is year-group FE[17] and ε_{igyt} the error term.[18]

The dummy variable $Female_{igy}$ identifies gender differences in the 1st period, this is when stakes at hand are low and $Female * Final\_Exam$ reveals the difference in the gender gap between the CA and the final exam grades. Therefore, $Female + Female \cdot Final\_Exam$ identifies gender differences in the 2nd period, this is when stakes at hand are high. Following Machado and Santos Silva (2019), we also estimate all versions of Eq. (1) by using a QRPD model for the 10th, 25th, 50th, 75th, and 90th percentiles to account for differences along the grade distribution.

Second, we perform an heterogeneity analysis that seeks to examine gender differences within academic years with more homogeneous stakes and pressure, as previously explained. This setting allows us to check that the results found for Eq. (1) are not driven by differences between the CA and the final exam, other than the stakes faced. We thus estimate the following baseline econometric specification by Pooled OLS, RE and QRPD:

$$\begin{aligned} Grade_{igyt} = {} & \alpha_0 + \alpha_1 \cdot Female_{igy} + \alpha_2 \cdot Final\_Exam_t + \alpha_3 \cdot Female_{igy} \cdot Final\_Exam_t + +\alpha_4 \cdot 1st\_Interval_y \\ & + \alpha_5 \cdot 1st\_Interval_y \cdot Female_{igy} + \alpha_6 \cdot X_{igy} + \mu_{gy} + \varepsilon_{igyt} \end{aligned} \quad (2)$$

where $1st\_Interval_y$ is the dummy variable that takes a value of 1 if the student belongs to the first interval (high pressure in academic years 2009/10–2012/13) and 0 to the second (low pressure in academic years 2013/14–2015/16). The interaction $Female * Final\_Exam$ is interpreted as before and is the difference in the gender gap associated with the CA component and the final exam. $Female * 1st\_Interval$ is the difference in the gender gap between the first and second intervals (for

[14] Even if defining our empirical set up in a panel data framework is superior in terms of identification, we have also performed all the estimations in a pooled cross-sectional manner (see Wooldridge, 2012), that is, the CA grade and final exam grade are treated as different outcome variables as in Örs et al. (2013). The results, reported in Appendix D (Table D.1 and D.2), are pretty much in line with those obtained in our more restrictive panel data set up.

[15] In the setting described by Iriberri and Rey-Biel (2019), school students face elimination at the end of the first stage depending on their performance and are, as such, subject to competitive pressure. However, in our setting, students are free to decide to participate in the second period (the final exam), with high incentives to do so. These students are not subject to competitive pressure and do not face the pressure of being eliminated during the first period; rather, they face the pressure of either passing or failing the course and satisfying their own grade goal, i.e. they are subject to test pressure. Moreover, we focus on those students who take both the CA and the final exam, i.e. 78% of our sample.

[16] Given the invariant nature of our main interest variable, student gender, we cannot run regressions using individual FE; this is the reason why we control for a large set of individuals characteristics. However, in Appendix C (see Tables C.1 and C.2) we report results using individual FE for the parameters of interest that can be estimated, that is, the interaction term between female student and the final exam.

[17] Even if the model that includes the year-group fixed effect is the most restrictive specification, allowing us to control for peer and teacher effects and year effects at the same time, in a less restrictive version, we also estimate the model introducing year and group fixed effects separately and using group variables instead of group FE.

[18] Group FE are important. Note that the allocation of students across groups is not random. When students enrol on their courses at the beginning of the academic year, they choose the group they wish to join. Priority in the enrolment process is determined by the student's GPA: those with the highest averages having first choice. For this reason, we need to control for peer-group and teacher effects. Indeed, to ensure robustness, and thanks to the rich data set we have, we replace, as robustness check, the group FE (μ_{gy}) with a list control variables of group characteristics (Z_{gt}).

grades in general, both periods). Hence, the gender gap in the same first period during the second interval (low stakes) is given by *Female* and the gender gap in the first period during the first interval is given by $Female + Female * 1st\_Interval$. Additionally, $Female + Female * Final\_Exam + Female * 1st\_Interval$ shows the gender gap in the second period in the first interval and $Female + Female * Final\_Exam$ the gender gap in the second period in the second interval (recall Table 2).

The control variables for individual and group characteristics used in the empirical models are defined in Table B.1 in Appendix B. The individual variables are age, age squared, nationality, university access grade,[19] dummy variable which takes a value of 1 if student enrolled between four and six courses on that term, dummy variable which takes a value of 1 if student enrolled seven or eight courses on that term, dummy variable which takes a value of 1 if student enrolled nine or more courses on that term,[20] average grade obtained that first term without *Principles of Taxation* grade and this average grade squared, status as scholarship holder and if the student is retaking the subject. The access grade proxies (direct and indirectly) student ability and family background and so allows us to control for the students' unobserved academic abilities.[21] Moreover, the number of courses enrolled on that term and the average grade obtained allow us to proxy extra information for that term: for instance, whether the student is working hard, subject to an extra effort by taking on more courses, their personal circumstances during that term or whether enrolled on a double degree program. In short, any circumstance that might lead a student to perform better or worse. The group variables are morning/afternoon group, percentage of female students in the group, gender of the teacher, average age of group, group's average access grade, average number of courses enrolled on by group and the average grade obtained in that term's courses without *Principles of Taxation* grade by group.

Finally, and given that we use a balanced sample, we also analyse the possibility of sample selection in our set up. Self-selection may arise from those students who have not taken either the CA or the final exam, and from those students who have not taken any exam at all. In order to do so we perform a Heckman panel data to correct for sample selection (Kyriazidou, 1997; Wooldridge, 2010), and we estimate an unbalanced panel data with the full sample.

6. Results

Table 4 shows gender differences over the first period – low stakes – and the second period – high stakes –. Estimates from the table are computed as explained in the previous section, so we already show the gender gap coefficient for each period. The RE and Pooled OLS estimates, i.e. the average difference, show that male students outperform female students in the second period (final exam) when the stakes faced are high by around 0.081 standard deviations (s.d.), statistically significant at the 1% level.[22] However, there are slightly significant differences (females outperforming males) on the CA grade, this is when stakes are low. In Appendix C, we present results for the full econometric specification (Table C.3). Importantly for us, the results hold with student FE (Table C.1). Moreover, the results are also aligned when using only those students that failed or passed the CA (Table C.4) to compare students with the same type of motivation when facing the final test.[23]

The QRPD estimates allow us to examine differences in the gender gap across the grade distribution in each period. Results indicate that the gender gap in favour of male students over the second period is lower at the bottom of the distribution and it widens as we move up along the distribution, from −0.02 s.d. at the bottom decile to −0.14 s.d. at the top decile. These differences are statistically significant for the 50th, 75th and 90th percentiles, while they are not for the 10th and 25th percentiles.[24]

Note that these results might be biased due to self-selection into the exams, i.e. the characteristics that lead students not to sit midterms or the final exam could be correlated with their gender. The obtained results hold with the full sample (Table C.6 in Appendix C), i.e. with all the students,[25] and we perform a two-step Heckman procedure to correct for sample selection in panel data. In this regard, Table 5 shows that, despite some selection to the first period across gender, the main results are robust to self-selection and our main estimates are not biased. Our findings using the Heckman procedure show that female students are more likely than their male counterparts to complete the CA component, while there are no gender differences regarding sitting the final exam, but even so, our results remain. Furthermore, Heckman results show that female students outperform male students in the first period (CA) when the stakes faced are low by around 0.066 s.d., statistically significant at the 5% level. Finally, we perform the Heckman technique with the pooled-cross sectional data, and results also hold (see Table D.3 in Appendix D).

[19] This information is not available for a small number of students (190 out of 3.912, that is, 4.86%) as they accessed via a different path.

[20] The reference base is students with three or less courses enrolled on in the term.

[21] For example, Iriberri and Rey-Biel (2019) proxy a student's mathematical ability with their maths grades at school, Örs et al. (2013) proxy student ability with the ranking of the preparation school and Jurajda and Münich (2011) control student ability with test scores on entrance examinations.

[22] Each grade variable is standardized with mean 0 and standard deviation 1 at the year level.

[23] In Appendix C, we also provide, as a robustness exercise, further results using each CA evaluation moment separately (see Table C.5) in a panel data framework. Results are consistent with those reported in the main text and, as expected, the results for the CA are mainly driven by the first CA, that is, the CA test that usually more students take as the number of observations show and that is, in principle, more comparable across academic years.

[24] We also estimate the gender differences for the CA grade and the final exam grade separately using a pooled cross-sectional structure. Results are available along Appendix D and they indicate very similar findings than those obtained with the panel data structure.

[25] Table B.2 in Appendix B shows descriptive statistics as in Table 3 but for the full and balanced samples. The percentage of female students is similar in both samples and grades by gender are also similar, but in general higher in the balanced sample.

Table 4
Gender gap in the CA grade and final exam - balanced sample.

	1st Period: CA - Low-stakes			2nd Period: Final exam - High-stakes		
	(1)	(2)	(3)	(4)	(5)	(6)
Pooled OLS	0.054*	0.054*	0.052*	−0.081***	−0.081***	−0.083***
	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
RE	0.054**	0.054*	0.052*	−0.081***	−0.081***	−0.083***
	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
$Q(\tau = 0.10)$	0.093*	0.101*	0.091*	−0.021	−0.017	−0.024
	(0.054)	(0.054)	(0.054)	(0.058)	(0.057)	(0.057)
$Q(\tau = 0.25)$	0.074**	0.078**	0.073*	−0.050	−0.048	−0.052
	(0.037)	(0.037)	(0.037)	(0.039)	(0.039)	(0.039)
$Q(\tau = 0.50)$	0.053*	0.052*	0.051*	−0.082***	−0.083***	−0.084***
	(0.027)	(0.027)	(0.027)	(0.029)	(0.029)	(0.029)
$Q(\tau = 0.75)$	0.033	0.028	0.031	−0.113***	−0.116***	−0.115***
	(0.035)	(0.035)	(0.035)	(0.037)	(0.037)	(0.037)
$Q(\tau = 0.90)$	0.017	0.008	0.016	−0.137***	−0.143***	−0.139***
	(0.049)	(0.049)	(0.049)	(0.052)	(0.052)	(0.052)
Observations	7410	7410	7410	7410	7410	7410
Number of ind.	3705	3705	3705	3705	3705	3705
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: The dependent variable measures student performance over the two periods: the CA grade (average multiple choice midterms) in the first period and final exam grade (multiple choice part) in the second. The dependent variable is standardised with mean 0 and standard deviation 1 at each period and year level. The *Female* dummy variable takes a value of 1 if the student is female and 0 otherwise. *Final_Exam* takes a value of 1 if the grade refers to the second period (final exam grade) and 0 if it refers to the first period (CA grade). The gender gap of the first period comes from the variable *Female* in Eq. (1) and the gender gap in the second period is computed as $Female + Female * Final\_Exam$. Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. **Individual variables:** age, age^2, nationality, university access grade, dummy variables for the number of courses enrolled on in the first term (further details in Table B.1 in Appendix B), average grade obtained that first term (excluding *Principles of Taxation*) and this variable squared, status as scholarship holder and repeat student. **Group variables:** morning/afternoon group, percentage of female students in the group, gender of the teacher, average age of group, group's average access grade, average number of courses enrolled on by group and the average grade obtained in that term's courses by group (excluding *Principles of Taxation*). The sample size difference observed with Table 3 is due to missing values in three control variables among 207 students.

Table 5
Gender gap for the two-period panel data – Heckman procedure .

	1st Period: CA - Low-stakes		2nd Period: Final exam - High-stakes	
	(1)	(2)	(3)	(4)
Main equation	0.067**	0.066**	−0.085***	−0.085***
	(0.028)	(0.028)	(0.026)	(0.027)
Selection	0.210***	0.217***	0.003	0.006
	(0.071)	(0.070)	(0.078)	(0.078)
Observations	9244	9244	9244	9244
Number of ind.	4622	4622	4622	4622
Individual controls	Yes	Yes	Yes	Yes
Group controls	Yes	No	Yes	No
Group FE	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes

Note: The dependent variable measures student performance over the two periods: the CA grade in the first period and final exam grade in the second. The dependent variable is standardised with mean 0 and standard deviation 1 at each period and year level. The *Female* dummy variable takes a value of 1 if the student is female and 0 otherwise. *Final_Exam* takes a value of 1 if the grade refers to the second period (final exam grade) and 0 if it refers to the first period (CA grade). Main and selection equations contain the same individual and group variables than in Table 4. Moreover, since selection equation must have at least one different independent variable to those included in the second step equation, we include the university GPA as independent variable in the selection equation. The Heckman procedure in a panel data framework did not converge when using year-group FE.

6.1. Heterogeneity analysis

So far, we have studied the stakes, defined as the weight of a test in the overall course grade. Now we turn to analyse the second source of pressure, that is, specific rules in the system of evaluation that increases the students' need to perform well (see Tables 1 and 2). As explained in Section 3, this timespan can be divided into two intervals characterized by fairly homogenous stakes and, hence, pressure: the first, from 2009/10 to 2013/14, of high stakes, and the second, from 2013/14 to 2015/16, of low stakes.

This heterogeneity analysis aims to test the hypothesis that as the stakes are increased, the wider the gender gap is in favour of male students. Moreover, we can compare gender gap within the same period (CA or final exam) but with different stakes at hand due to the changes mentioned over the years. Therefore, this heterogeneity analysis allows us to remove a great deal of potential threats to our interpretation of the results. One may still argue that an alternative explanation to our results could be related with, for instance, the differences in the amount of material to be covered in each evaluation. Each midterm in the CA (computer based) covers two or three topics, while the final exam (paper based) covers all the contents of the entire course. This would be a threat if males and females react and study differently depending on the amount of material to prepare for the examination, or the preparation time devoted to each exam. Both contents and preparation time can influence the strategy adopted by students to face the CA and the final exam irrespective of the pressure faced by the weight of the particular exam on the final grade.

Table 6 reports the results of the Pooled OLS, RE and QRPD estimates from Eq. (2). As stated, results are ranked from the lowest stake scenario, column (1), to the highest stake scenario, column (4). On the one hand, in the lowest pressure scenario the gender gap is 0.08 s.d. in favour of female students in the CA over the second interval, and this gender gap is mitigated over the first interval, column (2), with an increase on the stakes. On the other hand, in column (3), male students above the 75% of the grade distribution outperform female students by around 0.094 and 0.123 s.d. in the final exam over the second interval. In the highest-pressure scenario, column (4), the gender gap in favour of male students widens, on average, to 0.132 s.d., statistically significant at the 1% level. Our results indicate the presence of gender differences in the lowest stake scenario in favour of female students, the first period over the second interval of years, and it shrinks until it reversed in favour of male students in the highest scenario, the second period over the first interval of years. These results can be easily interpreted in Fig. 1 where we plot these average gender gaps in each pressure scenario, from the lowest to the highest.

Table 6
Gender gap in the CA grade and final exam: heterogeneity results - balanced sample.

	Lowest	Pressure Scenario		Highest
	CA - Low	CA - High	Final Exam - Low	Final Exam - High
	(1)	(2)	(3)	(4)
Pooled OLS	0.080**	0.003	−0.055*	−0.132***
	(0.036)	(0.026)	(0.032)	(0.035)
RE	0.080**	0.003	−0.055*	−0.132***
	(0.036)	(0.026)	(0.032)	(0.035)
$Q(\tau = 0.10)$	0.126**	0.020	0.014	−0.091
	(0.061)	(0.073)	(0.064)	(0.075)
$Q(\tau = 0.25)$	0.104**	0.012	−0.018	−0.111**
	(0.043)	(0.051)	(0.045)	(0.052)
$Q(\tau = 0.50)$	0.078**	0.002	−0.057*	−0.133***
	(0.031)	(0.037)	(0.033)	(0.038)
$Q(\tau = 0.75)$	0.054	−0.007	−0.094**	−0.155***
	(0.040)	(0.048)	(0.042)	(0.049)
$Q(\tau = 0.90)$	0.035	−0.014	−0.123**	−0.172**
	(0.056)	(0.067)	(0.059)	(0.069)
Observations	7410	7410	7410	7410
Number of ind.	3705	3705	3705	3705
Individual controls	Yes	Yes	Yes	Yes
Year-group FE	Yes	Yes	Yes	Yes

Note: Same definition of the dependent, *Female* and *Final_Exam* variables than in Table 4. The 1*st_Interval* takes value 1 if academic years from 2009/10 to 2012/13 which is defined as high stakes, and 0 otherwise. The gender gap of the first period and 2nd Interval is defined by the variable *Female* in Eq. (2), the gender gap of the first period and 1st Interval is defined by *Female* + *Female* ∗ 1*st_Interval*, the gender gap in the second period and 2nd Interval is computed as *Female* + *Female* ∗ *Final_Exam* and the gender gap in the second period and 1st Interval *Female* + *Female* ∗ *Final_Exam* + *Female* ∗ 1*st_Interval*. Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Same individual and group variables than in Table 4.

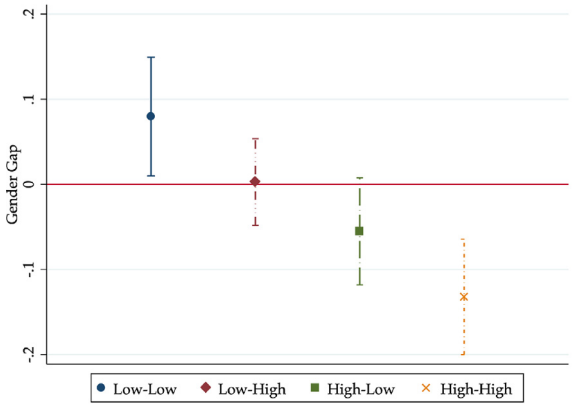

Fig. 1. Gender gap ordered from the lowest to the highest pressure scenario. *Note:* The first level of pressure refers to CA-Final Exam and the second one to first-second interval. Therefore, Low-Low = CA - 2nd Interval, Low-High = CA - 1st Interval, High-Low = Final Exam - 2nd Interval, High-High = Final Exam - 1st Interval. Coefficients obtained from four different estimations. Positive gender gap is in favour of female students and negative gender gap in favour of male students.

Quantile results in Table 6 show that the gender gap is different along the grade distribution for each pressure scenario defined, but they always follow the same pattern. Given a pressure scenario, the gender gap shrinks along the distribution in favour of male students, i.e. gender differences are more in favour of female students at the bottom than at the top of the distribution. Note that in the first column of results in Table 6, the lowest pressure scenario, the gender gap is 0.126 s.d. in favour of female students at the bottom decile. This gender gap shrinks to 0.104 s.d. at the 25th percentile and to the 0.078 s.d. at the median. Conversely, in the last column, the highest pressure scenario, the gender gap is 0.111 s.d. in favour of male students at the 25th percentile, and this gender gap widens along all the distribution. This gender gap is 0.133 s.d. at the median, 0.155 s.d. at the 75th percentile and 0.172 s.d. at the top decile.[26]

Therefore, with this heterogeneity analysis, we observe that as the stakes increase, the gender gap grows in favour of male students. However, as stakes fall, the gender gap is mitigated, and is even reversed in favour of female students. A similar pattern occurs along the grade distribution, the gender gap varies in favour of male students from the bottom to the top of the distribution. This means that if the gender gap is in favour of females in the bottom, the gender gap shrinks. However, if the gender gap is in favour of males in the bottom, the gender gap widens. On top of that, we show that, given the same period of the evaluation system but different stakes at hand, the gender gap varies according to our hypotheses. Therefore, differences between first and second period, i.e. CA and final exam, could be seen as orthogonal to the gender of students and are not a threat in our setting.

6.2. Potential mechanisms: disentangling the gender gap

After the results showed so far is important to ask: do female students choke under pressure and, as result, perform worse? Or, are male students motivated by pressure and, so, perform better? Or, are these two mechanisms operating simultaneously? To address these questions, we estimate the same two-period panel data model (Eq. (1)) as in Table 4, obtaining now the estimates of the difference between CA (low-pressure) and the final exam (high pressure) separately for males and females. More precisely, for male students, the estimated parameter is now the coefficient of the variable *Final_Exam*. Similarly, the difference between CA (low-pressure) and the final exam (high pressure) for female students is given by $Final\_Exam + Female * Final\_Exam$. Hence, Table 7 reports these results by Pooled OLS, RE and QRPD regressions.

Both the Pooled OLS and the RE results show that female students perform worse on the final exam (high stakes) than on the CA component (low stakes), by 0.153 s.d. (statistically significant at the 1% level). However, the performance of male students on the CA component and final exam are largely the same. These results indicate that female students tend to choke when stakes are high, and their performance suffers when they feel higher pressure. However, male students appear not to be affected by the stakes at hand and their performance remains unchanged. From the quantile regression estimates for females, there is a pattern along the distribution: the difference is greater for the bottom decile 0.21–0.22 s.d. (statistically significant at the 1% level) and this gap shrinks as we move to the top of the distribution, with a 0.09 s.d. difference at the top decile.

We repeat this exercise for the heterogeneity setting; we estimate the same specifications as in Table 7 adding the dummy variable *1st_Interval*. Since we have collinearity between the dummy variable *1st_Interval* and the year-group FE, we estimate the regressions without the year FE to overcome this issue and with group FE. Fig. 2 shows the results which

[26] We also compute Eq. (2) with the full sample obtaining very similar results than in the balanced sample (Table C.7 in Appendix C).

Table 7
Differences across periods by gender - balanced sample.

	Male			Female		
	(*Final_Exam*)			(*Final_Exam* + *Female* ∗ *Final_Exam*)		
	(1)	(2)	(3)	(4)	(5)	(6)
Pooled OLS	−0.018	−0.018	−0.018	−0.153***	−0.153***	−0.153***
	(0.038)	(0.038)	(0.038)	(0.036)	(0.036)	(0.036)
RE	−0.018	−0.018	−0.018	−0.153***	−0.153***	−0.153***
	(0.038)	(0.038)	(0.038)	(0.036)	(0.036)	(0.036)
$Q(\tau = 0.10)$	−0.101*	−0.098*	−0.095*	−0.215***	−0.216***	−0.210***
	(0.056)	(0.056)	(0.055)	(0.055)	(0.054)	(0.054)
$Q(\tau = 0.25)$	−0.060	−0.059	−0.059	−0.185***	−0.185***	−0.183***
	(0.038)	(0.038)	(0.038)	(0.037)	(0.037)	(0.037)
$Q(\tau = 0.50)$	−0.016	−0.016	−0.016	−0.151***	−0.151***	−0.151***
	(0.028)	(0.028)	(0.028)	(0.027)	(0.027)	(0.027)
$Q(\tau = 0.75)$	0.027	0.026	0.025	−0.119***	−0.118***	−0.121***
	(0.036)	(0.036)	(0.036)	(0.035)	(0.035)	(0.035)
$Q(\tau = 0.90)$	0.061	0.058	0.056	−0.093*	−0.092*	−0.098**
	(0.050)	(0.050)	(0.050)	(0.049)	(0.049)	(0.049)
Observations	7410	7410	7410	7410	7410	7410
Number of ind.	3705	3705	3705	3705	3705	3705
No. males/females	1909	1909	1909	1796	1796	1796
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: Same definition of the dependent, *Female* and *Final_Exam* variables than in Table 4. Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Same individual and group variables than in Table 4.

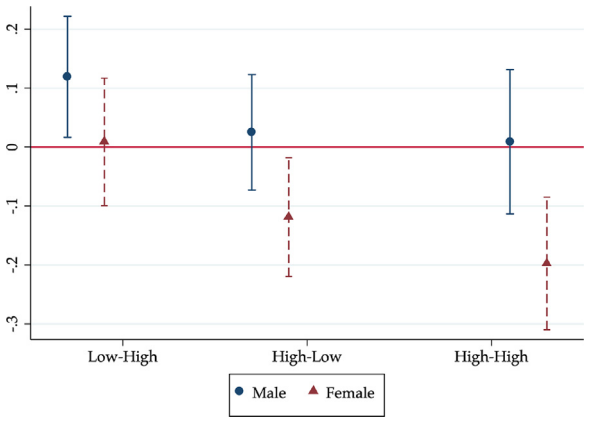

Fig. 2. Student performance differences across gender compared with the lowest pressure scenario. *Note:* the reference scenario is the Low-Low = CA - 2nd Interval. Then, Low-High = CA - 1st Interval, High-Low = Final Exam - 2nd Interval, High-High = Final Exam - 1st Interval. Ordered by pressure scenario (from low to high). Coefficients obtained from two different estimations: one for male students and one for female students.

come from two independent regressions: one with the subsample of only male students (solid line) and a second one with only female students (dashed line). Thus, these regressions do not contain neither the female variable nor any interaction with it. Note that the reference pressure scenario is the lowest (Low-Low scenario), this is the CA over the second interval. Therefore, the results, i.e. the estimates from the other three scenarios, are based on this reference scenario.

Male students in the first interval (high-stakes) increase their CA performance compared with their counterparts in the second interval (low-stakes), while female students do not change their performance due to the change in the stakes faced in both intervals and obtain similar results in the CA component of their final grade. This translates into an increase in the gender gap of the CA, as we observe in the Fig. 1. When we compare students' performance between CA and final exam over the first interval (High-Low scenario), we observe that male students do not change much their performance, while female students' performance significantly decreases under the increasing pressure faced when performing the final exam. Finally, the highest-pressure scenario, the final exam over the first interval (High-High scenario), male students do

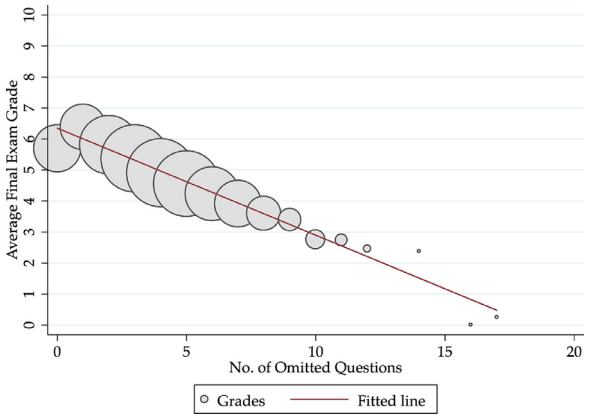

Fig. 3. Average final exam grade by number of omitted questions on the final multiple choice test. *Note:* The multiple choice part of the final exam comprises 20 questions and the grade ranges from 0 to 10 (observations 3746). The size of the circle indicates the number of observations.

Table 8
Gender differences in the omission of questions on the multiple choice test of the final exam.

	OLS	Quantile Regressions				
	(1)	(2) $\tau = 0.10$	(3) $\tau = 0.25$	(4) $\tau = 0.50$	(5) $\tau = 0.75$	(6) $\tau = 0.90$
Female	0.243***	0.206***	0.267***	0.308***	0.279***	0.181*
	(0.050)	(0.066)	(0.055)	(0.056)	(0.068)	(0.090)
Female · 1*st_Interval*	0.069	0.299***	0.113	0.046	−0.059	0.037
	(0.078)	(0.119)	(0.111)	(0.092)	(0.104)	(0.128)
Observations	3562	3562	3562	3562	3562	3562
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Year-group FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: The dependent variable measures the number of omitted questions on the final exam. It is standardized with mean 0 and standard deviation 1 at year level. The *Female* dummy variable takes a value of 1 if the student is female and 0 otherwise. 1*st_Interval* dummy variable takes a value of 1 if it refers to the first interval (high pressure), and 0 to the second interval (low pressure). Standard errors, clustered at year-group level, for the OLS and bootstrapped standard errors with 1000 replications for the quantile regressions, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. The sample includes 3562 observations because information was missing for some students. All the regressions include the same individual variables than in Table 4, and the CA grade (from 0 to 10) as a control variable.

not change significantly their performance, while female students underperform their counterparts in the lowest pressure scenario (reference). In this highest pressure scenario, the difference between male and female is even bigger, leading to a higher gender gap as that observed in Fig. 1. Therefore, we conclude that the increase in the gender gap in favour of male students as the stakes increase is mainly driven by the decrease of female students' performance, i.e. choke under pressure.

These gender differences in response to stakes might be given by different behaviour changes across gender. Specially, our previous results suggest that female students are those who might change their behaviour when they feel more pressure. In this vein, both economic and educational literature have analysed student behaviour when taking multiple choice tests and, more specifically, several studies have found gender differences (Pekkarinen, 2015; Iriberri and Rey-Biel, 2019; Akyol et al., 2016; Riener and Wagner, 2017; Baldiga, 2014). They focus on the number of test items omitted due to the evidence showing that the number of questions omitted is a good predictor of the student's grade. While in our case this information is not available for the midterms, it is for the final exam. Thus, Fig. 3 shows the average final exam grade by the number of omitted questions on this test. This figure indicates a negative relation between the grade scored and the number of omitted questions: as the number of omitted test items increases, the lower the average final exam grade is.

Finally, we analyse whether female and male students follow different strategies when omitting items over the multiple choice test in the final exam. Table 8 reports the results obtained for estimations with year-group FE using OLS and quantile regressions with a pooled cross-sectional data structure.[27] Female students omit more questions than their male counterparts on the final exam in the second interval – low stakes - by around 0.243 s.d. (*Female* coefficient), which is statistically

[27] Given that we only have omitted questions for the multiple choice test in the final exam we have to rely on a pooled cross-section framework. Moreover, we should stress that the quantile results identify the distribution of the number of omitted items, and not the grade distribution as in the previous tables.

significant at the 1% level. The gender gap on the omitted-question distribution in this second interval widens from the first decile to the median, before narrowing in the 75th percentile and the last decile of the distribution. Then, we focus our attention on the interaction *Female* · 1*st_Interval*, which identifies the difference in the gender gap between the first (high-stakes) and second interval (low-stakes). This interaction is positive and significant at the 1% level in the first decile which corresponds to that part of the distribution in which students omit fewer items. When stakes increase the gender gap on omitting items also increases among the high-skilled students. Moreover, results using group variables and year FE, and group FE together with year FE, shown in Table D.4 in Appendix D, show that the gender gap is also statistically significant for students at the 25th percentile. According to Fig. 3, this part of the omitted-question distribution concentrates students with the highest final exam grades, i.e. the best students. This result implies that female students with the highest grades are impacted more strongly by the stakes at hand and omit more items. Therefore, among the best students, females omit even more questions than males on a high pressure scenario.

7. Discussion and conclusion

This paper has analysed gender differences in academic performance in response to different stakes at hand, measured as the weight of a specific test in the overall course grade and in relation to rules in the system of evaluation that increase the importance of a good performance. The specific design of the evaluation system of a course taught at the University of Barcelona allows us to exploit a unique quasi-experimental set up. In this setting, students first take midterm multiple choice tests as part of the CA component (designated low pressure), and, at the end of the term, they sit the final exam which includes a multiple choice test (designated high pressure). Our primary goal has been to analyse, in a panel data framework, the performance of those students who complete both the CA component and the final exam and, in so doing, ensure that our results are not biased by individuals who drop out of the course during the term or by those who opt solely to sit the final exam.

Our main findings suggest that there are indeed gender differences in student responses to pressure: male students are found to outperform female students when sitting high-stake exams (0.132 s.d.). However, as the stakes at hand decreases, the gender gap in favour of male students is narrowed until it is mitigated and ultimately, in the lowest stake scenario, reversed in favour of female students (0.08 s.d.). These results are robust to the use of the full sample of students and to the possible presence of sample selection in relation to completing the CA component and sitting the final exam. We have also examined the potential mechanisms that might account for these gender gaps. We find that this widen gender gap in favour of male students as stakes increase is due to female students decrease their performance. This is, male students perform at the same level, while female counterparts decrease their performance as pressure increases. Moreover, there is suggestive evidence that the top female students omit more items than male students on the final exam as test pressure rises.

The findings reported herein concur with those studies that provide evidence of a gender gap in academic performance attributable to pressure. Thus, with studies in which male students outperform female students when faced with high levels of pressure (Örs et al., 2013; Iriberri and Rey-Biel, 2019; Jurajda and Münich, 2011; Pekkarinen, 2015), as well as reports that either find no gender differences or evidence indicating that girls outperform boys when faced with low levels of pressure (Örs et al., 2013; Jurajda and Münich, 2011). In contrast, but at the school level, Azmat et al. (2016) find that girls outperform boys and that this gender gap narrows as pressure increases. However, when these authors turn their attention to university entrance exams – a very high stakes test – they find that the gender gap disappears. Yet, it should be borne firmly in mind that these last findings are drawn in relation to very different types of courses and exams, and that the age range of their sample is from 12 to 18. Evidence also indicates that in situations of high pressure the gender gap is increasing over the students' grade distribution (Örs et al., 2013; Jurajda and Münich, 2011; Ellison and Swanson, 2010). For example, Jurajda and Münich (2011) find significant gender differences in the 50th and 75th percentiles as do Örs et al. (2013) in the 25th, 50th, 75th and 90th percentiles. Here, in high pressure scenarios, we have found significant gender differences in the 50th, 75th and 90th percentiles. Finally, female students appear to omit more test items than do male students in multiple choice tests (Pekkarinen, 2015; Espinosa and Gardeazabal, 2010; 2020). Here, moreover, we find that the gender gap in relation to the omission of test items increases in high pressure scenarios for the best students (consistent with Funk and Perrone, 2016; Coffman and Klinowski, 2020), while Iriberri and Rey-Biel (2019) make the same finding but for all students.

This study contributes to a greater understanding of gender differences in academic performance at the university level by providing additional evidence about the existence of gender differences in the taking of multiple choice tests as a result of the stakes faced by students. We confirm that the design itself of the evaluation system, that may entail different pressure scenarios, may induce a gender gap that punishes female students if stakes are particularly high in a given test. Moreover, we confirm that also the strategies employed on multiple choice tests vary with gender and that female students are more likely to omit test items. Some studies have shown that multiple choice tests benefit certain subgroups, above all male students, and so call into question the suitability of this type of exam (Riener and Wagner, 2017; Espinosa and Gardeazabal, 2010). Clearly, education systems should be properly designed to assess student knowledge and abilities, and not student reaction to pressure or test-taking strategies – differences in grades should be driven by differences in knowledge and abilities. Finally, our results indicate that the evaluation changes promoted within the EHEA process (the Bologna Process) favour gender equality with respect to higher education outcomes.

Appendix A. Evaluation system

The course analysed is a mandatory course on the Bachelor's Degree in Business Administration taught at the Faculty of Economics and Business (University of Barcelona). According to the program, students have to take this course in the first term of their third year.[28] The course is an introduction to economics of taxation where students are given an initial grounding in the Spanish tax system, i.e. the role of the public sector, general principles of taxation and specific tax theory. It combines theoretical and practical content, that is, taxation theory with numerical exercises.

Over the seven academic years considered, the course coordinators introduced some changes to the evaluation system. Table A.1 summarises the specific rules and the changes made to them. The table is divided into three main sections: rules applied over the midterms (CA), rules governing the final exam and rules for computing the overall course grade. Note that the number of midterms fell over the years from four to two. During the first five years – 2009/10 to 2013/14 –, students could discard the midterm with the worst grade, as long as they had opted to sit them all. In the last two years, with just two midterms, this possibility was eliminated. The CA grade, therefore, was computed as the average grade of all midterms, each with the same weight.

The existence of a minimum grade required on the whole final exam (first and second part) over the last three years represented pressure to students for the whole final exam in the same direction of the eliminatory ruled analysed in the main text (reinforcing the stakes faced by students across intervals). A comparison shows that the eliminatory rule was a source of greater pressure on the final exam than the minimum grade requirement.[29]

Finally, the use of two homogenous intervals (2009/10 - 2012/13 vs 2013/14 - 2015/16) of pressure is not perfect, but it offers a close approximation. The number of midterms and the weight of the CA grade are not exactly the same in the first interval but, despite this, the incentives to perform well in the CA component remain high and constant given the fact that students do not need to score a minimum of four on the multiple choice part of the final exam in case of passing the CA. Given that the weight of the final exam in this interval is high, not being entitled to have the second part of the final exam marked means automatically failing the whole course. In the second interval, the system employed in the first year presents some differences with the other two, but it is more similar to the system employed in the last two years than it is to those in the first interval. In any case, the use of year fixed effects should help eradicate these small differences.

Table A.1
Evaluation system for *Principles of Taxation*.

	CA		Final Exam	Overall Course Grade		
Academic Year	No. Midterms	Average Grade	Multiple choice eliminatory	% CA	Min. grade Final Exam*	Overall grade
2009/10	4	3 highest	Yes	10%	No	CA & F.Exam
2010/11	4	3 highest	Yes	20%	No	CA & F.Exam
2011/12	3	2 highest	Yes	30%	No	CA & F.Exam
2012/13	3	2 highest	Yes	40%	No	CA & F.Exam
2013/14	3	2 highest	No	40%	Yes	CA & F.Exam
2014/15	2	the 2	No	40%	Yes	max{F.Exam, CA & F.Exam}
2015/16	2	the 2	No	40%	Yes	max{F.Exam, CA & F.Exam}

* Students had to obtain a minimum final exam grade of 4 (out of 10) for the CA grade to be taken into account for the overall course grade.

[28] In Spain, with few exceptions, bachelor degrees comprise four academic years.

[29] Student surveys over the years have identified the "eliminatory rule" as a key source of pressure and have expressed their rejection of it. Yet, the "minimum grade rule" for the CA grade to be taken into account has not been perceived by students to be a major source of pressure, given their perception that they could pass the course by performing well on the final exam.

Appendix B. Data

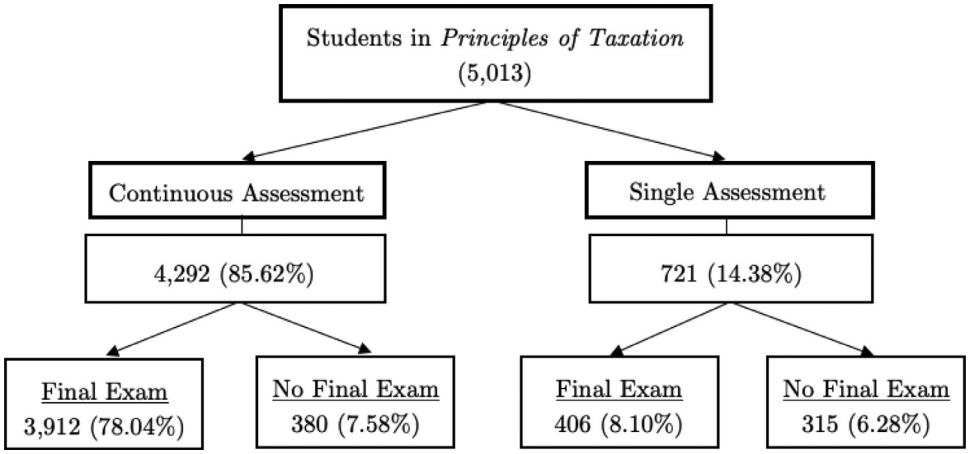

Fig. B.1. Subgroups of students by system of evaluation on the *Principles of Taxation* course (number and percentage). *Note:* All percentage calculations made for the whole sample (i.e. 5013 students).

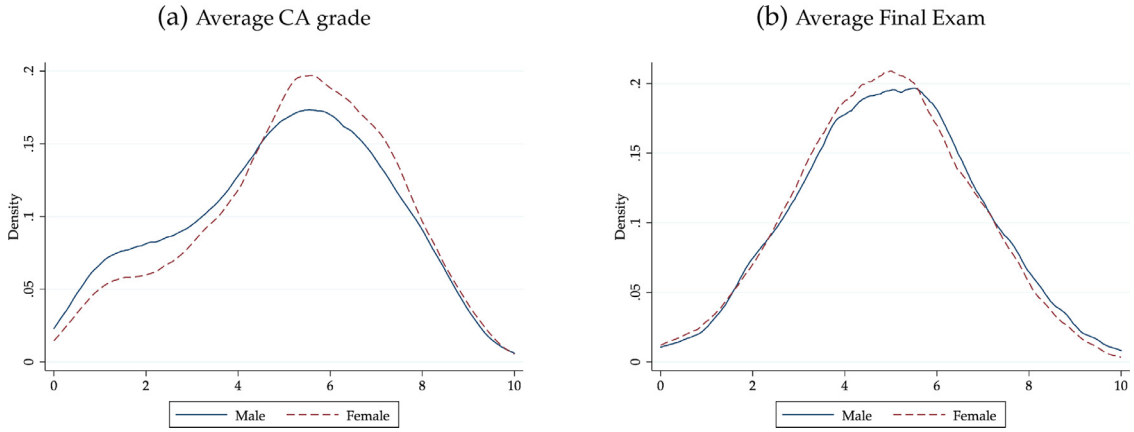

Fig. B.2. Kernel density estimations for the whole timespan - balanced sample.

Table B.1
Overview of all the variables included in the regressions.

Variable	Description
1. Dependent variables	
Grade	The final grade of the CA is standardised with mean 0 and standard deviation 1 at year level (1st period) and the grade of the multiple choice part of the final exam, standardised with mean 0 and standard deviation 1 at year level (2nd period).
Number of omitted questions	The number of omitted items on the multiple choice part of the final exam (out of 20 questions), standardised with mean 0 and standard deviation 1 at year level.
2. Individual characteristics	
Age	The age variable is the difference between 1st September of the year in which student is enrolled on *Principles of Taxation* and her date of birth.
Age^2	*Age* squared.
Nationality	Dummy variable which takes a value of 1 if Spanish nationality and 0 if other nationality.
University access grade	The grade ranges between 1 and 10. The grade comprises 60% from the two years of high school grade *(Bachillerato)* and 40% from the university entrance examination *(Selectividad)*.
Dummies: no. of courses enrolled	Defined as courses enrolled on in the first term of the academic year in question, *Principles of Taxation* included, three dummy variables are included: (i) between 4 and 6 courses, (ii) 7 and 8 courses, (iii) 9 or more courses. Thus, 3 or less courses is the reference base.
Average grade of the courses	Average grade of all the courses that the student enrolled the first term of the academic year, *Principles of Taxation* excluded.
[Average grade of the courses]^2	*Average grade of the courses* squared.
Repeat Student	Dummy variable which takes a value of 1 if the student has failed the course *Principles of Taxation* in a previous year, 0 otherwise.
Scholarship	Dummy variable which takes a value of 1 if the student has been awarded a scholarship that year, 0 otherwise.
3. Group characteristics	
Morning group	Dummy variable which takes a value of 1 if the student is enrolled in a morning group, 0 otherwise.
% of female classmates	Percentage of female students in that group (taking into account all the students enrolled in the group).
Age	The average age of the whole group (taking into account all the students enrolled in the group).
University access grade	The average university access grade in the group.
No of courses enrolled	The average number of courses that students in this group enrolled on in the first term of the academic year in question, *Principles of Taxation* included.
Average grade of the courses	The average grade of the group from all the courses that the students enrolled on in the first term of the academic year in question, *Principles of Taxation* excluded.
Female teacher	Variable which ranges from 0 to 1. This is the proportion of female professors teaching that group, in ECTS terms.

Table B.2
Descriptive statistics for the CA and final exam grades.

	CA Grade (Low-stakes)			Final Exam Grade (High-stakes)		
	Male	Female	Statistic	Male	Female	Statistic
Full Sample ($N = 5013$)						
N	2207	2085		2253	2065	
Percentage (%)	51.42	48.58		52.18	47.82	
$H_o^a : Pr(F_i) = 0.5$			-1.86^{*}			-2.86^{***}
$H_o^b : Pr(F_i) = Pr(M_i)$			-2.63^{***}			-4.05^{***}
Mean Test	4.73	5.03	-4.38^{***}	4.89	4.78	1.92^{*}
Median Test	4.94	5.26	12.64^{***}	4.90	4.88	2.65
KS Test			0.07^{***}			0.04
10th percentile	1.38	1.69		2.25	2.38	
25th percentile	2.88	3.51		3.50	3.50	
75th percentile	6.50	6.69		6.25	6.00	
90th percentile	7.69	7.71		7.50	7.25	
Balanced Sample ($N = 3912$)						
N	2021	1891		2021	1891	
Percentage (%)	51.66	48.34		51.66	48.34	
$H_o^a : Pr(F_i) = 0.5$			-2.08^{**}			-2.08^{**}
$H_o^b : Pr(F_i) = Pr(M_i)$			-2.94^{***}			-2.94^{***}
Mean Test	4.94	5.25	-4.47^{***}	4.95	4.83	2.07^{**}
Median Test	5.19	5.44	13.96^{***}	5.00	4.88	1.83
KS Test			0.08^{***}			0.04
10th percentile	1.63	2.13		2.38	2.38	
25th percentile	3.38	3.94		3.63	3.60	
75th percentile	6.63	6.82		6.25	6.13	
90th percentile	7.81	7.82		7.50	7.25	

Note: see notes in Table 3.

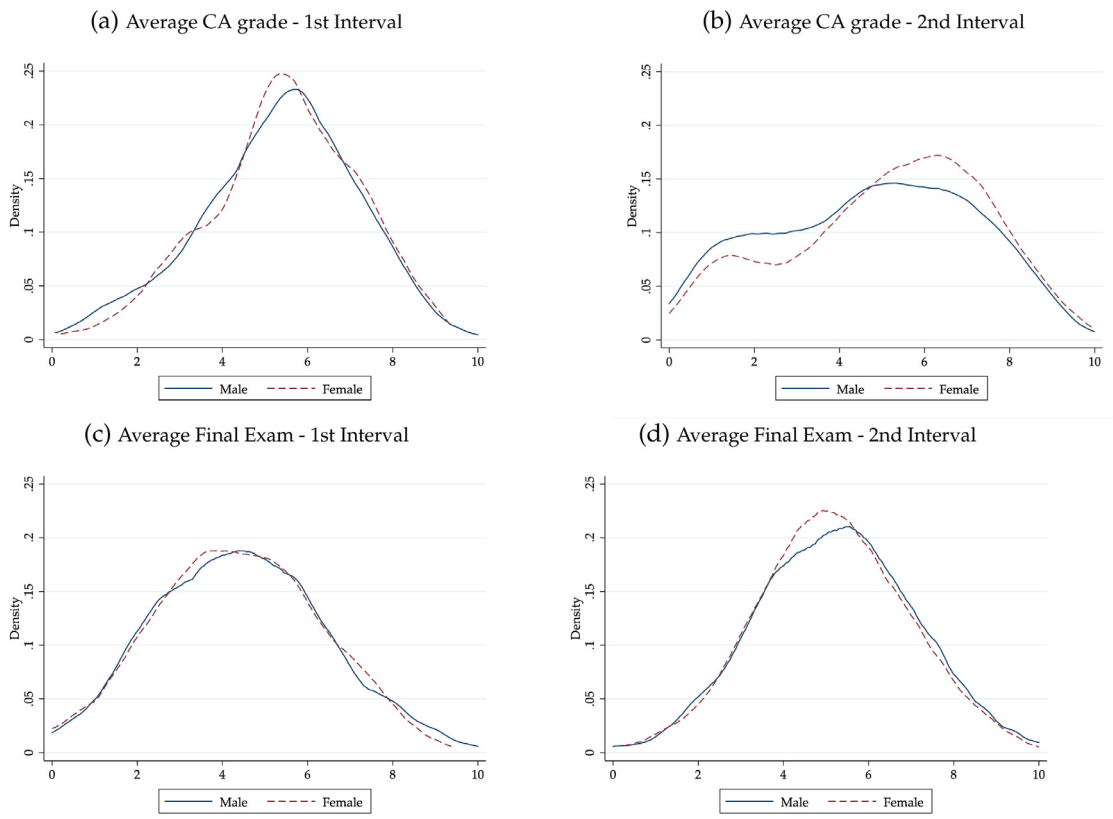

Fig. B.3. Kernel density estimations by intervals - balanced sample.

Appendix C. Additional Results

Table C.1
Estimated coefficients - balanced sample with individual FE.

	Pooled OLS (1)	RE (2)	FE (3)
Female	0.052*	0.052*	
	(0.028)	(0.028)	
Final_Exam	−0.018	−0.018	−0.018
	(0.038)	(0.038)	(0.038)
Female · Final_Exam	−0.135***	−0.135***	−0.135***
	(0.037)	(0.037)	(0.037)
Observations	7410	7410	7410
Number of ind.	3705	3705	3705
Individual controls	Yes	Yes	No
Individual FE	No	No	Yes
Year-group FE	Yes	Yes	No

Note: see notes in Table 4. The FE model has been estimated with the same sample than the pooled OLS and the RE models.

Table C.2
Gender differences: mechanisms – balanced sample with individual FE.

	Male (*Final_Exam*)			Female (*Final_Exam* + *Female* ∗ *Final_Exam*)		
	Pooled OLS (1)	RE (2)	FE (3)	Pooled OLS (4)	RE (5)	FE (6)
Difference between CA - Final Exam	−0.018 (0.038)	−0.018 (0.038)	−0.018 (0.038)	−0.153*** (0.036)	−0.153*** (0.036)	−0.153*** (0.036)
Observations	7410	7410	7410	7410	7410	7410
Number of ind.	3705	3705	3705	3705	3705	3705
Individual controls	Yes	Yes	No	Yes	Yes	No
Individual FE	No	No	Yes	No	No	Yes
Year-group FE	Yes	Yes	No	Yes	Yes	No

Note: see notes in Table 7. The FE model has been estimated with the same sample than the pooled OLS and the RE models.

Table C.3
Full set of estimated coefficients of the two-period panel data - balanced sample.

	(1)	(2)	(3)
Female	0.054**	0.054*	0.052*
	(0.027)	(0.028)	(0.028)
Final_Exam	−0.018	−0.018	−0.018
	(0.038)	(0.038)	(0.038)
Female · Final_Exam	−0.135***	−0.135***	−0.135***
	(0.037)	(0.037)	(0.037)
Age	−0.010	−0.023	−0.026
	(0.040)	(0.039)	(0.039)
Age^2	−0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)
Spanish nationality	0.023	0.025	0.013
	(0.045)	(0.046)	(0.049)
University entrance grade	0.116***	0.119***	0.117***
	(0.015)	(0.015)	(0.015)
Dummy: number of courses = 4 - 6	−0.123**	−0.122**	−0.126**
	(0.057)	(0.057)	(0.058)
Dummy: number of courses = 7 - 8	−0.308***	−0.311***	−0.303***
	(0.073)	(0.073)	(0.075)
Dummy: number of courses = 9 or +	−0.256**	−0.271***	−0.269**
	(0.099)	(0.101)	(0.109)
Average grade of courses	−0.223***	−0.233***	−0.230***
	(0.070)	(0.070)	(0.070)
Average grade of courses^2	0.047***	0.048***	0.047***
	(0.006)	(0.006)	(0.006)
Repeat student	−0.224***	−0.234***	−0.220***
	(0.038)	(0.039)	(0.039)
Scholarship	0.102***	0.104***	0.101***
	(0.028)	(0.027)	(0.028)
Morning group	−0.046		
	(0.068)		
% female students	0.041		
	(0.257)		
Average university entrance grade	0.088		

(*continued on next page*)

Table C.3 (*continued*)

	(1)	(2)	(3)
	(0.073)		
Average grade of the courses	−0.007		
	(0.059)		
Average number of courses enrolled	−0.082		
	(0.073)		
% female teachers	−0.044		
	(0.052)		
Average age	−0.050		
	(0.039)		
Constant	0.713	−0.343	−0.377
	(1.581)	(0.546)	(0.529)
Observations	7410	7410	7410
Number of ind.	3705	3705	3705
Individual controls	Yes	Yes	Yes
Group controls	Yes	No	No
Year FE	Yes	Yes	No
Group FE	No	Yes	No
Year-group FE	No	No	Yes

Note: see notes in Table 4.

Table C.4
Gender gap in the final exam conditional on pass or fail the CA - balanced sample.

	CA - Low-stakes	Final exam - High-stakes		
	All students	All students	Students: pass CA	Students: fail CA
	(1)	(2)	(3)	(4)
Pooled OLS	0.052*	−0.083***	−0.050	−0.100***
	(0.028)	(0.028)	(0.036)	(0.037)
RE	0.052*	−0.083***	−0.050	−0.100***
	(0.028)	(0.028)	(0.036)	(0.037)
$Q(\tau = 0.10)$	0.091*	−0.024	0.047	−0.052
	(0.054)	(0.057)	(0.070)	(0.085)
$Q(\tau = 0.25)$	0.073*	−0.052	0.003	−0.075
	(0.037)	(0.039)	(0.049)	(0.059)
$Q(\tau = 0.50)$	0.051*	−0.084***	−0.049	−0.101**
	(0.027)	(0.029)	(0.038)	(0.044)
$Q(\tau = 0.75)$	0.031	−0.115***	−0.104**	−0.127**
	(0.035)	(0.037)	(0.050)	(0.057)
$Q(\tau = 0.90)$	0.016	−0.139***	−0.148**	−0.147*
	(0.049)	(0.052)	(0.071)	(0.079)
Observations	7410	7410	4218	3192
Number of ind.	3705	3705	2109	1596
Individual controls	Yes	Yes	Yes	Yes
Group controls	No	No	No	No
Period FE	Yes	Yes	Yes	Yes
Group FE	No	No	No	No
Year-group FE	Yes	Yes	Yes	Yes

Note: see notes in Table 4. Reported results in column (1) and (2) corresponds to the results reported in Table 4, column (3) and (6), respectively.

Table C.5
Gender gap over all the midterms and the final exam - balanced sample.

	1st CA (1)	2nd CA (2)	3rd CA (3)	4th CA (4)	*Final_Examen* (5)
Pooled OLS	0.075**	−0.002	−0.005	0.169	−0.079***
	(0.028)	(0.036)	(0.040)	(0.156)	(0.028)
RE	0.074***	0.002	0.002	0.191	−0.080***
	(0.028)	(0.036)	(0.042)	(0.154)	(0.029)
$Q(\tau = 0.10)$	0.084	0.038	−0.051	0.371	−0.006
	(0.059)	(0.064)	(0.091)	(0.265)	(0.056)
$Q(\tau = 0.25)$	0.080*	0.019	−0.029	0.276	−0.041
	(0.041)	(0.044)	(0.063)	(0.183)	(0.039)
$Q(\tau = 0.50)$	0.075**	−0.003	−0.004	0.166	−0.080***
	(0.030)	(0.033)	(0.046)	(0.135)	(0.029)
$Q(\tau = 0.75)$	0.070*	−0.024	0.020	0.059	−0.118***
	(0.039)	(0.043)	(0.061)	(0.176)	(0.038)
$Q(\tau = 0.90)$	0.067	−0.040	0.039	−0.022	−0.148***
	(0.055)	(0.059)	(0.084)	(0.244)	(0.052)
Observations	12,179	12,179	12,179	12,179	12,179
No. of students by period	3525	3083	1640	226	3705

Note: The dependent variable measures student performance over five periods: four midterms in the CA and the final exam grade (only the multiple choice part). The dependent variable is standardised with mean 0 and standard deviation 1 at each period and year level. Note this is an unbalanced panel data since the number of midterms is not the same across academic years (see Table A.1 in Appendix A). The *Female* dummy variable takes a value of 1 if the student is female and 0 otherwise and periods fixed effects are included (2*nd_CA*, 3*rd_CA*, 4*th_CA*, and *Final_Examen*, where 1*st_CA* period is the reference base). The estimated model also includes individual controls and year-group fixed effects. For the academic year 2009/10, we only have the average grade of the CA and not the grade of each midterm. Therefore, in this table, students from this year have missing values in the midterms' grades. The gender gap of the first period, column (1), comes from the variable *Female*. Then, the gender gap in the second period, column (2), is computed as *Female* + *Female* ∗ 2*nd_CA*, in the third period, column (3), as *Female* + *Female* ∗ 3rd_*CA*, in the fourth period, column (4), as *Female* + *Female* ∗ 4th_*CA*, and in the last period, the final exam, column (5), the gender gap is computed as *Female* + *Female* ∗ *Final_Examen*. Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level.

Table C.6
Gender gap in the CA grade and final exam - full sample.

	1st Period: CA - Low-stakes			2nd Period: Final exam - High-stakes		
	(1)	(2)	(3)	(4)	(5)	(6)
Pooled OLS	0.054**	0.053*	0.051*	−0.087***	−0.087***	−0.089***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.026)
RE	0.052**	0.052*	0.050*	−0.087***	−0.087***	−0.089***
	(0.026)	(0.027)	(0.027)	(0.026)	(0.026)	(0.026)
$Q(\tau = 0.10)$	0.086	0.093*	0.085	−0.026	−0.020	−0.026
	(0.054)	(0.053)	(0.053)	(0.055)	(0.055)	(0.055)
$Q(\tau = 0.25)$	0.070*	0.074**	0.069*	−0.055	−0.052	−0.056
	(0.037)	(0.037)	(0.037)	(0.038)	(0.038)	(0.038)
$Q(\tau = 0.50)$	0.053**	0.052*	0.050*	−0.088***	−0.088***	−0.090***
	(0.027)	(0.027)	(0.027)	(0.028)	(0.028)	(0.028)
$Q(\tau = 0.75)$	0.036	0.032	0.032	−0.119***	−0.123***	−0.123***
	(0.035)	(0.035)	(0.035)	(0.036)	(0.036)	(0.036)
$Q(\tau = 0.90)$	0.023	0.016	0.019	−0.144***	−0.150***	−0.148***
	(0.048)	(0.048)	(0.048)	(0.050)	(0.050)	(0.050)
Observations	8111	8111	8111	8111	8111	8111
Number of ind.	4406	4406	4406	4406	4406	4406
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: see notes in Table 4.

Table C.7
Gender gap in the CA grade and final exam: heterogeneity results - full sample.

	Lowest	Pressure Scenario		Highest
	CA - Low	CA - High	Final Exam - Low	Final Exam - High
	(1)	(2)	(3)	(4)
Pooled OLS	0.063*	0.032	−0.077**	−0.108***
	(0.035)	(0.025)	(0.032)	(0.030)
RE	0.057*	0.037	−0.082***	−0.103***
	(0.034)	(0.025)	(0.031)	(0.029)
$Q(\tau = 0.10)$	0.095	0.067	−0.016	−0.044
	(0.061)	(0.072)	(0.062)	(0.074)
$Q(\tau = 0.25)$	0.079*	0.050	−0.046	−0.075
	(0.042)	(0.049)	(0.043)	(0.051)
$Q(\tau = 0.50)$	0.062**	0.031	−0.079**	−0.110***
	(0.031)	(0.036)	(0.031)	(0.037)
$Q(\tau = 0.75)$	0.045	0.012	−0.111***	−0.143***
	(0.040)	(0.047)	(0.041)	(0.048)
$Q(\tau = 0.90)$	0.032	−0.002	−0.135**	−0.169**
	(0.055)	(0.065)	(0.056)	(0.067)
Observations	8111	8111	8111	8111
Number of ind.	4406	4406	4406	4406
Individual controls	Yes	Yes	Yes	Yes
Year-group FE	Yes	Yes	Yes	Yes

Note: see notes in Table 6.

Appendix D. Robustness check: pooled cross-section estimates

Table D.1
Gender gap for the pooled cross-section data - balanced sample.

	1st Period: CA - Low-stakes			2nd Period: Final Exam - High-stakes		
	(1)	(2)	(3)	(4)	(5)	(6)
OLS	0.046	0.048	0.043	−0.073***	−0.076***	−0.074***
	(0.028)	(0.029)	(0.029)	(0.027)	(0.027)	(0.027)
$Q(\tau = 0.10)$	0.021	0.027	0.003	−0.069	−0.058	−0.070
	(0.048)	(0.046)	(0.047)	(0.057)	(0.057)	(0.060)
$Q(\tau = 0.25)$	0.077*	0.065	0.072*	−0.003	−0.034	−0.033
	(0.041)	(0.044)	(0.042)	(0.046)	(0.043)	(0.042)
$Q(\tau = 0.50)$	0.062*	0.066*	0.066*	−0.089**	−0.079**	−0.088**
	(0.036)	(0.036)	(0.035)	(0.037)	(0.036)	(0.039)
$Q(\tau = 0.75)$	0.039	0.044	0.044	−0.113**	−0.099**	−0.113***
	(0.035)	(0.038)	(0.039)	(0.042)	(0.045)	(0.045)
$Q(\tau = 0.90)$	−0.013	−0.032	−0.041	−0.157***	−0.173***	−0.152***
	(0.045)	(0.043)	(0.041)	(0.045)	(0.043)	(0.046)
Observations	3705	3705	3705	3705	3705	3705
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: The dependent variable measures CA grade in columns (1) to (3) and final exam grade in columns (4) to (6). Each dependent grade variable is standardized with mean 0 and standard deviation 1 at year level. The coefficients shown are the *Female* dummy variable (1 if female student) for the OLS and QR. Standard errors, clustered at year-group level, for the OLS and bootstrapped standard errors with 1000 replications for the QR. Standard errors are in parentheses and *** denotes significance at 1% level, ** at 5% level and * at 10% level. Same individual and group control variables than in Table 4.

Table D.2
Gender gap for the pooled cross-section data - full sample.

	1st Period: CA - Low-stakes			2nd Period: Final Exam - High-stakes		
	(1)	(2)	(3)	(4)	(5)	(6)
OLS	0.045	0.048*	0.042	−0.079***	−0.082***	−0.082***
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.025)
$Q(\tau = 0.10)$	0.017	−0.004	0.015	−0.082	−0.060	−0.068
	(0.048)	(0.047)	(0.044)	(0.052)	(0.052)	(0.052)
$Q(\tau = 0.25)$	0.061	0.074*	0.088**	0.003	−0.044	−0.035
	(0.043)	(0.044)	(0.042)	(0.042)	(0.041)	(0.039)
$Q(\tau = 0.50)$	0.056	0.077**	0.068*	−0.092***	−0.079**	−0.099***
	(0.038)	(0.036)	(0.036)	(0.036)	(0.035)	(0.037)
$Q(\tau = 0.75)$	0.029	0.037	0.041	−0.110***	−0.102**	−0.121***
	(0.035)	(0.037)	(0.037)	(0.041)	(0.040)	(0.042)
$Q(\tau = 0.90)$	−0.013	−0.036	−0.033	−0.177***	−0.173***	−0.157***
	(0.041)	(0.039)	(0.039)	(0.042)	(0.039)	(0.043)
Observations	4047	4047	4047	4064	4064	4064
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: see notes in Table D.1.

Table D.3
Gender gap for the pooled cross-section data – Heckman procedure.

	1st Period: CA - Low-stakes			2nd Period: Final Exam - High-stakes		
	(1)	(2)	(3)	(4)	(5)	(6)
Main equation	0.079**	0.077**	0.070**	−0.081**	−0.084***	−0.084***
	(0.031)	(0.030)	(0.031)	(0.033)	(0.031)	(0.031)
Selection	0.199***	0.207***	0.214***	−0.015	−0.015	−0.018
	(0.056)	(0.056)	(0.057)	(0.054)	(0.053)	(0.054)
Observations	4622	4622	4622	4622	4622	4622
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Group controls	Yes	No	No	Yes	No	No
Year FE	Yes	Yes	No	Yes	Yes	No
Group FE	No	Yes	No	No	Yes	No
Year-group FE	No	No	Yes	No	No	Yes

Note: The dependent variable in the second step (Main equation) measures the CA grade in columns (1) to (3) and the final exam grade in columns (4) to (6). The dependent variable in the first step (Selection Equation) is a dummy variable which takes a value of 1 if the student takes the CA component, columns (1) to (3), or the final exam, columns (4) to (6). Each dependent grade variable is standardised with mean 0 and standard deviation 1 at year level. The coefficients shown are the *Female* dummy variable (1 if a female student) from the selection equation (first step) and the main equation (second step). The explanatory variables of the selection equation are the same as those in each main equation, but we add the university GPA to the CA grade, and the university GPA and a dummy variable capturing whether the student completed the CA component or not to the final exam grade. This is because the selection equation must have at least one different variable to those included in the second step equation. Moreover, in the final exam grade, we control for the student having previously completed, or otherwise, the CA component. The other individual and group variables are the same than in Table 4. Heckman standard errors are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level.

Table D.4
Gender gap mechanisms: omitted questions - robustness check.

	OLS	$Q(\tau = 0.10)$	$Q(\tau = 0.25)$	$Q(\tau = 0.50)$	$Q(\tau = 0.75)$	$Q(\tau = 0.90)$
	(1)	(2)	(3)	(4)	(5)	(6)
Equation (1) with group variables and year FE						
Female	0.242***	0.185***	0.249***	0.329***	0.261***	0.161*
	(0.050)	(0.058)	(0.050)	(0.056)	(0.067)	(0.083)
Female · 1*st_Interval*	0.083	0.334***	0.228**	0.026	0.015	0.050
	(0.079)	(0.118)	(0.101)	(0.092)	(0.104)	(0.128)
Equation (1) with group FE and year FE						
Female	0.242***	0.171***	0.243***	0.305***	0.258***	0.197**
	(0.050)	(0.059)	(0.051)	(0.056)	(0.072)	(0.082)
Female · 1*st_Interval*	0.081	0.360***	0.202**	0.091	0.046	0.060
	(0.079)	(0.115)	(0.104)	(0.094)	(0.109)	(0.121)
Observations	3562	3562	3562	3562	3562	3562
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes

Note: see notes in Table 8.

References

Akyol, P., Key, J., Krishna, K., 2016. Hit or Miss? Test Taking Behavior in Multiple Choice Exams. NBER Working Paper No. 22401 doi:10.3386/w22401. http://www.nber.org/papers/w22401.pdf.

Arenas, A., Calsamiglia, C., 2020. Gender Differences in High-Stakes Performance and College Admission Policies. Unpublished results. Mimeo.

Azmat, G., Calsamiglia, C., Iriberri, N., 2016. Gender differences in response to big stakes. J. Eur. Econ. Assoc. 14 (6), 1372–1400. doi:10.1111/jeea.12180.

Backović, D.V., Živojinović, J.I., Maksimović, J., Maksimović, M., 2012. Gender differences in academic stress and burnout among medical students in final years of education. Psychiatr. Danub. 24 (2), 175–181.

Balart, P., Ezquerra, L., Hernandez-Arenaz, I.n., 2020. Framing effects on risk-taking behavior: evidence from a field experiment. SSRN Electron. J. doi:10.2139/ssrn.3556710.

Baldiga, K., 2014. Gender differences in willingness to guess. Manage. Sci. 60 (2), 434–448. http://proceedings.aom.org/cgi/doi/10.5465/AMBPP.1991.4976539.

Baumeister, R.F., 1984. Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. J. Pers. Soc. Psychol. 46 (3), 610–620. doi:10.1037/0022-3514.46.3.610.

Beilock, S., 2011. Choke: What the Secrets of the Brain Reveal About Getting It Right When You Have To. Simon and Schuster. https://books.google.es/books?id=yNkes7-8pUMC.

Blau, F.D., Kahn, L.M., 2017. The gender wage gap: extent, trends, and explanations. J. Econ. Lit. 55 (3), 789–865. doi:10.1257/jel.20160995.

Bulman, G., 2017. Weighting recent performance to improve college and labor market outcomes. J. Public Econ. 146, 97–108. doi:10.1016/j.jpubeco.2016.12.002.

Cai, X., Lu, Y., Pan, J., Zhong, S., 2019. Gender gap under pressure: evidence from China's national college entrance examination. Rev. Econ. Stat. 101 (2), 249–263. doi:10.1162/rest_a_00749.

Chin, E.C.H., Williams, M.W., Taylor, J.E., Harvey, S.T., 2017. The influence of negative affect on test anxiety and academic performance: an examination of the tripartite model of emotions. Learn. Individ. Differ. 54, 1–8. doi:10.1016/j.lindif.2017.01.002. http://linkinghub.elsevier.com/retrieve/pii/S104160801730002X.

Coffman, K.B., Klinowski, D., 2020. The impact of penalties for wrong answers on the gender gap in test scores. Proc. Natl. Acad. Sci. U.S.A. 117 (16), 8794–8803. doi:10.1073/pnas.1920945117.

Contini, D., Tommaso, M.L.D., Mendolia, S., 2017. The gender gap in mathematics achievement: evidence from Italian data. Econ. Educ. Rev. 58, 32–42. doi:10.1016/j.econedurev.2017.03.001. http://linkinghub.elsevier.com/retrieve/pii/S0272775716303466.

De Paola, M., Gioia, F., 2016. Who performs better under time pressure? Results from a field experiment. J. Econ. Psychol. 53, 37–53. doi:10.1016/j.joep.2015.12.002.

Dee, T.S., 2007. Teachers and the gender gaps in student achievement. J. Hum. Resour. XLII (3), 528–554. doi:10.3368/jhr.XLII.3.528.

Ellison, G., Swanson, A., 2010. The gender gap in secondary school mathematics at high achievement levels: evidence from the american mathematics competitions. J. Econ. Perspect. 24 (2), 109–128. doi:10.1257/jep.24.2.109.

Eman, S., Dogar, I.A., Khalid, M., Haider, N., 2012. Gender differences in test anxiety and examination stress. J. Pakistan Psychiatric Soc. 9 (2), 80–85.

von der Embse, N., Jester, D., Roy, D., Post, J., 2018. Test anxiety effects, predictors, and correlates: a 30-year meta-analytic review. J. Affect. Disord. 227, 483–493. doi:10.1016/j.jad.2017.11.048. http://linkinghub.elsevier.com/retrieve/pii/S0165032717303683.

Espinosa, M.P., Gardeazabal, J., 2010. Optimal correction for guessing in multiple-choice tests. J. Math. Psychol. 54 (5), 415–425. doi:10.1016/j.jmp.2010.06.001. http://linkinghub.elsevier.com/retrieve/pii/S0022249610000623.

Espinosa, M.P., Gardeazabal, J., 2020. The gender-bias effect of test scoring and framing: a concern for personnel selection and college admission. B.E. J. Econ. Anal. Policy 20 (3). doi:10.1515/bejeap-2019-0316.

Falch, T., Naper, L.R., 2013. Educational evaluation schemes and gender gaps in student achievement. Econ. Educ. Rev. 36, 12–25. doi:10.1016/j.econedurev.2013.05.002. http://linkinghub.elsevier.com/retrieve/pii/S0272775713000782.

Farré, L., Vella, F., 2013. The intergenerational transmission of gender role attitudes and its implications for female labour force participation. Economica 80 (318), 219–247. doi:10.1111/ecca.12008. http://ideas.repec.org/p/iza/izadps/dp2802.html.

Funk, P., Perrone, H., 2016. Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams. CEPR Discussion Paper No. DP11716.

Gallardo, E., Montolio, D., Camós, M., 2010. The European higher education area at work: lights and shadows defining continuous assessment. Revista d'Innovació Docent Universitària 2, 10–22. doi:10.1344/105.000001524.

Goldin, C., Kerr, S.P., Olivetti, C., Barth, E., 2017. The expanding gender earnings gap: evidence from the LEHD-2000 census. Am. Econ. Rev. 107 (5), 110–114. doi:10.1257/aer.p20171065.

Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. Am. Econ. Rev. 90 (4), 715–741.

González De San Román, A., De La Rica, S., 2016. Gender gaps in PISA test scores: the impact of social norms and the mother's transmission of role attitudes. Estudios de Economía Aplicada 34, 79–108.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Hoffmann, F., Oreopoulos, P., 2009. A professor like me: the influence of instructor gender on college achievement. J. Hum. Resour. 44 (2), 479–494. doi:10.1353/jhr.2009.0024.

Iriberri, N., Rey-Biel, P., 2019. Competitive pressure widens the gender gap in performance: evidence from a two-stage competition in mathematics. Econ. J. 129 (620), 1863–1893. doi:10.1111/ecoj.12617. https://academic.oup.com/ej/article/129/620/1863/5473520.

Jurajda, v., Münich, D., 2011. Gender gap in performance under competitive pressure: admissions to Czech universities. Am. Econ. Rev. 101 (3), 514–518. doi:10.1257/aer.101.3.514.

Karle, H., Engelmann, D., Peitz, M., 2020. Student Performance and Loss Aversion. CRC TR 224 Discussion Paper Series. University of Bonn and University of Mannheim, Germany. https://econpapers.repec.org/RePEc:bon:boncrc:crctr224_2020_150.

Kies, S.M., Williams, B.D., Freund, G.G., et al., 2006. Gender plays no role in student ability to perform on computer-based examinations. BMC Med. Educ. 6 (1), 57. doi:10.1186/1472-6920-6-57.

Kunze, A., 2018. The gender wage gap in developed countries. In: Averett, S.L., Argys, L.M., Hoffman, S.D. (Eds.), The Oxford Handbook of Women and The Economy. No. 10826. Oxford University Press, pp. 368–394. doi:10.1093/oxfordhb/9780190628963.013.11.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica 65 (6), 1335. doi:10.2307/2171739. https://www.jstor.org/stable/2171739?origin=crossref.

Lavy, V., 2013. Gender differences in market competitiveness in a real workplace: evidence from performance-Based pay tournaments among teachers. Econ. J. 123 (569), 540–573.

Lim, J., Meer, J., 2017. The impact of teacher-student gender matches. J. Hum. Resour. 52 (4), 979–997. doi:10.3368/jhr.52.4.1215-7585R1.

Lim, J., Meer, J., 2020. Persistent effects of teacher–student gender matches. J. Hum. Resour. 55 (3), 809–835. doi:10.3368/jhr.55.3.0218-9314R4.

Lubienski, S.T., Robinson, J.P., Crane, C.C., Ganley, C.M., 2013. Girls' and Boys' mathematics achievement, affect, and experiences: findings from ECLS-K. J. Res. Math. Educ. 44 (4), 634. doi:10.5951/jresematheduc.44.4.0634.

Machado, J.A.F., Santos Silva, J.M.C., 2019. Quantiles via moments. J. Econom. 213 (1), 145–173. doi:10.1016/j.jeconom.2019.04.009. http://www.sciencedirect.com/science/article/pii/S0304407619300648 https://linkinghub.elsevier.com/retrieve/pii/S0304407619300648.

Mail, D. (2018). Oxford University extends time for maths and computer science exams in bid to help women get better grades. Retrieved on 9th June 2018 from http://www.dailymail.co.uk/news/article-5294031/Oxford-University-extends-time-maths-help-women.html#ixzz550SPGjEv,

Marcenaro-Gutiérrez, O., López-Agudo, L.A., 2016. Mind the gap: analysing the factors behind the gap in students' performance between pencil and computer based assessment methods. Revista de Economia Aplicada 24 (71), 93–120.

Meghir, C., Rivkin, S., 2011. Econometric Methods for Research in Education, vol. 3. Elsevier, pp. 1–87. doi:10.1016/B978-0-444-53429-3.00001-6.
Muralidharan, K., Sheth, K., 2016. Bridging education gender gaps in developing countries: the role of female teachers. J. Hum. Resour. 51 (2), 269–297. doi:10.3368/jhr.51.2.0813-5901R1.
Örs, E., Palomino, F., Peyrache, E., 2013. Performance gender gap: does competition matter? J. Labor Econ. 31 (3), 443–499. doi:10.1086/669331.
Núñez Peña, M.I., Suárez-Pellicioni, M., Bono, R., 2016. Gender differences in test anxiety and their impact on higher education Students' academic achievement. Procedia Social Behav. Sci. 228, 154–160. doi:10.1016/j.sbspro.2016.07.023. http://linkinghub.elsevier.com/retrieve/pii/S1877042816309491.
Pekkarinen, T., 2015. Gender differences in behaviour under competitive pressure: evidence on omission patterns in university entrance examinations. J. Econ. Behav. Organ. 115, 94–110. doi:10.1016/j.jebo.2014.08.007. http://linkinghub.elsevier.com/retrieve/pii/S0167268114002261.
Putwain, D.W., Woods, K.A., Symes, W., 2010. Personal and situational predictors of test anxiety of students in post-compulsory education. Br. J. Educ. Psychol. 80 (1), 137–160. doi:10.1348/000709909X466082.
Rask, K., Tiefenthaler, J., 2008. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. Econ. Educ. Rev. 27 (6), 676–687. doi:10.1016/j.econedurev.2007.09.010.
Riener, G., Wagner, V., 2017. Shying away from demanding tasks? Experimental evidence on gender differences in answering multiple-choice questions. Econ. Educ. Rev. 59, 43–62. doi:10.1016/j.econedurev.2017.06.005. http://linkinghub.elsevier.com/retrieve/pii/S0272775716306847.
Rodríguez-Planas, N., Nollenberger, N., 2018. Let the girls learn! It is not only about math... it's about gender social norms. Econ. Educ. Rev. 62, 230–253. doi:10.1016/j.econedurev.2017.11.006. https://linkinghub.elsevier.com/retrieve/pii/S0272775717304041.
Rose, H., 2006. Do gains in test scores explain labor market outcomes? Econ. Educ. Rev. 25, 430–446. j.econedurev.2005.07.005
Shurchkov, O., 2012. Under pressure: gender differences in output quality and quantity under competition and time constraints. J. Eur. Econ. Assoc. 10 (5), 1189–1213. doi:10.1111/j.1542-4774.2012.01084.x.
Telegraph, T. (2018). Oxford University extends exam times for women's benefit. Retrieved on 9th June 2018 from https://www.telegraph.co.uk/education/2018/02/01/oxford-university-extends-exam-times-womens-benefit/.
Wallace, P., Clariana, R.B., 2005. Gender differences in computer-administered versus paper-based tests. Int. J. Instr. Media 32 (2), 171.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press.
Wooldridge, J.M., 2012. Introductory Econometrics: A Modern Approach. South-Western, Cengage Learning.