IEB
Institut
d'Economia
de Barcelona

IEB Working Paper 2018/21

GENDER DIFFERENCES UNDER TEST PRESSURE AND THEIR IMPACT ON ACADEMIC PERFORMANCE: A QUASI-EXPERIMENTAL DESIGN

Daniel Montolio, Pere A. Taberner

Public Policies

IEB Working Paper 2018/21

GENDER DIFFERENCES UNDER TEST PRESSURE AND THEIR IMPACT ON ACADEMIC PERFORMANCE: A QUASI-EXPERIMENTAL DESIGN

Daniel Montolio, Pere A. Taberner

Postal Address:
Institut d'Economia de Barcelona
Facultat d'Economia i Empresa
Universitat de Barcelona
C/ John M. Keynes, 1-11
(08034) Barcelona, Spain
Tel.: + 34 93 403 46 46
ieb@ub.edu
http://www.ieb.ub.edu

IEB Working Paper 2018/21

GENDER DIFFERENCES UNDER TEST PRESSURE AND THEIR IMPACT ON ACADEMIC PERFORMANCE: A QUASI-EXPERIMENTAL DESIGN [*]

Daniel Montolio, Pere A. Taberner

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 are exposed to test pressure. Based on field data, this study aims to examine gender differences in student academic performance in response to different levels of pressure 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. Exploiting two data structures—namely, pooled cross-sections and panel data—we find that male students outperform their female counterparts when under high pressure. However, in low test pressure scenarios the gender gap is narrowed and even reversed in favour of female students. Finally, we analyse the mechanisms responsible for the gender gap by studying how each gender responds to test pressure, and by studying gender differences when omitting test items on multiple choice formats.

JEL Codes: A22, I24, J16
Keywords: Gender differences, test pressure, academic performance, field data, higher education

Daniel Montolio
Universitat de Barcelona &
Institut d'Economia de Barcelona (IEB)
E-mail: montolio@ub.edu

Pere A. Taberner
KSNET
E-mail: peretaberner@gmail.com

[*] We would like to thank 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 Jose Luis Andujar. All remaining errors are our own.

1 Introduction

In January 2018, most of the news media reported that the University of Oxford had decided to extend the time available to computer science and maths students to complete their exams. According to The 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 minutes (Daily Mail, 2018). The social response was divided: some argued that the move was sexist, while others believed that it would serve to provide greater gender equality in education.[1]

Student performance is a strong determinant of individual decisions and future outcomes. First, university choice depends on the grades obtained throughout high school and on university entrance examination results. Second, a university student's accumulated average grade is an important criterion when applying for undergraduate grants or scholarships. And, third, an undergraduate student's final grade point average (GPA) is crucial when applying for postgraduate degrees or jobs. For example, admission to many master's programs or applications for graduate scholarships depend, in part, on a student's GPA. Likewise, many companies request a student's grade transcript and take this into account in their job selection procedures. Moreover, in many countries, many students sit public examinations to become public servants and in some disciplines there is the need to pass an exam to practice the profession. All this means that student performance both at university and in these public exams are strong determinants of future labour outcomes. However, the well-documented gender gap in labour outcomes (see, for example, Blau and Kahn, 2000; Wolfers, 2006; Goldin et al., 2017; Blau and Kahn, 2017) highlights the need to understand the potential determinants of gender differences in academic performance so as to mitigate them and to promote increased gender equality in education and, hence, in labour opportunities.

In recent years, economic researchers have become increasingly interested in gender differences in performance when under pressure (see, for example, Gneezy et al., 2003; Niederle and Vesterlund, 2010; Paserman, 2007; Shurchkov, 2012). However, few studies have analysed the role of pressure in educational performance across gender using real-world data. While these papers focus mainly on competitive pressure, their principal data are drawn from schools, national contests or university entrance exams. Little is known about how test pressure affects student performance across gender at the university level.[2] While in higher education, students sit a significant number of exams exposing them to different levels of pressure, depending on the number of credits at stake, the weight of the exam in the final grade, the specific rules of assessment which might require students to perform better, the difficulty of the exam content, the type of exam or the simple fact of having to sit an exam. To the best of our knowledge, only one economic study to date has attempted to analyse pressure as the weight attached to an

[1] This was not the first occasion that the University of Oxford had modified rules governing its examinations to reduce the gender grade gap. In 2017, history students were given the opportunity to take exams at home in an attempt at increasing the number of women obtaining top grades (The Times, 2017).

[2] Ors 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).

exam, albeit in this instance at high school level (Azmat et al., 2016). Additionally, we are aware that the multiple choice test is the most frequently adopted format to assess student knowledge at university, in entrance exams and public examinations. Here, also, published studies have documented gender differences in the results obtained on multiple choice tests and in the respective answering behaviour by gender. However, little is known about how this gender gap varies according to the level of pressure faced, highlighting the need for more studies that can shed light on gender differences attributable to pressure on this test format.

The goal of this paper, therefore, is to study gender differences in academic performance under different levels of pressure when sitting multiple choice tests. 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 levels of pressure. This setting allows us to structure the dataset in two ways and to exploit each to the full: independently pooled cross-sections and panel data. Moreover, the use of unique student administrative data provides us with a rich individual-based dataset to control for personal and group characteristics that might also determine student performance and affect gender differences.

We define test pressure as the increased need students feel to perform well on an examination due to the increasing importance attached to that test. The principal source of test pressure on students in our main setting is the greater weight attached to a test in the final grade awarded for the course. Moreover, because in our setting some of the rules of evaluation change over the years, modifying the pressure students are under when sitting these exams, we perform a heterogeneity analysis. In this case, the secondary source of test pressure is identified as the rules of evaluation which require a better performance from the students. The strength of our quasi-experimental set up is based on five specific characteristics: (i) the tests are machine/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) the same cohort 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, and (v) 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 empirical results show that, after controlling for individual and group characteristics, male students perform better than female students when under greater test pressure, but that this gender gap narrows as test pressure decreases until mitigated and even reversed in favour of female students. Our pooled cross-sectional estimates allow us to obtain the main results for each performance separately, to analyse differences along each grade distribution and to assess self-selection bias related to the decision to take or not take each exam. Our panel data enable us to reshape the data set and to analyse the setting as a sequential game of two periods, allowing us to confirm our main results and, moreover, to test for heterogeneity effects in relation to exam pressure. Finally, we examine the mechanisms underlying the gender gap and study whether the gap reflects a fall in female performance or a rise in male performance under pressure.

The paper is structured in six sections. Section 2 reviews the most recent and relevant

literature on the topic. Section 3 describes the institutional setting, explaining the specific evaluation system used on the course. 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, especially in relation to mathematics. In the literature, the *"nurture vs nature"* heated debate has been widely explored (González De San Román and De La Rica, 2016; Dee, 2005) with the advocates of biological determinants arguing that innate differences determine the gender gap in student achievement, while the proponents of environmental determinants claim that the main drivers are social and cultural norms. The economic literature, in general, has tried to identify these environmental mechanisms, but has not found it easy to disentangle all the possible factors responsible for this gender gap (Bedard and Cho, 2010; Contini et al., 2017). 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; Holmlund and Sund, 2008; Escardíbul and Mora, 2013), competitive pressure (e.g. Ors et al., 2013; Jurajda and Münich, 2011; Pekkarinen, 2015), and self-beliefs (Contini et al., 2017; Lubienski et al., 2013), among others. As such, the consensus is that the gender gap is multifactorial and it is difficult to narrow the determinants down to just a few factors. However, as discussed in the introduction, test pressure has been the object of few studies in economic research and, therefore, constitutes the main focus of this paper.

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 difference 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

2.2.1 Gender gap on academic performance under pressure

Few studies in the economic literature focus on the role of pressure on gender differences in scholastic achievement using field data. However, the few that have been undertaken look at the following sources of pressure: competition, time and the stakes involved, with competitive pressure – "the stress that one feels when competing" (Iriberri and Rey-Biel, 2018, p.2) – being the most analysed. Azmat et al. (2016) define the pressure a student is under according to the weight attached to a particular exam within the overall grade for the year, while time pressure can be defined as the stress students feel as a result of the time constraints they are under to complete the exam (De Paola and Gioia, 2016).

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, Ors 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). Similarly, Pekkarinen (2015) also studies the gender gap in performance in the Finnish university entrance exams (which use a multiple-choice format)

and finds the same results as the previous authors: male students outperform female students on the entrance exams and fewer girls tend to be admitted to university. He also analyses gender differences in responses to the multiple choice entrance exams and finds that the gap in performance is because girls omit more items on exams of this type. Finally, in a very recent study, Iriberri and Rey-Biel (2018) analyse the gender gap in secondary students participating in a maths contest in Madrid, Spain. The competition comprises two multiple choice exams but only the best performers in the first stage continue to the second. They find that while there are no gender differences in maths grades at school, boys outperform girls in the maths contest. Moreover, the gender gap is greater in the second stage (higher competitive environment) than in the first (lower competitive pressure).

Empirical studies of the impact of time pressure on educational performance are scare. De Paola and Gioia (2016) examine how time pressure influences performance at the University of Calabria when students are 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 sit 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.

Azmat et al. (2016) analyse gender differences in school performance and university entrance exams when placed under different levels of pressure. Despite focusing primarily on school performance, the importance of this study lies in its definition of pressure, which is the same as that employed herein. As discussed above, they define pressure according to the weight (stake) of an exam within the overall grade for the year. 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.

2.2.2 Multiple choice format

The multiple choice test is a frequently employed exam format at university. It provides a rapid system for the evaluation of student knowledge in large class sizes, an objective grading system, and easy correction (Akyol et al., 2014) while offering high marker reliability (Espinosa and Gardeazabal, 2010). On the down side, it has been noted that students might adopt a guessing strategy when answering questions, and that student knowledge is not evaluated properly (Akyol et al., 2014; Espinosa and Gardeazabal, 2010). To overcome guessing, wrong answers are often penalized, but this can give rise to other problems. According to Akyol et al. (2014), such a system is fair if the student either knows the answer or she does not. However, usually, students doubt among various options and have to decide whether to answer the question or to skip it. This means that the decision to omit a test item might reflect behavioural differences such as lower levels of confidence or risk aversion (Iriberri and Rey-Biel, 2018), the latter being

the subject of various studies (Niederle and Vesterlund, 2007; Baldiga, 2014; Espinosa and Gardeazabal, 2010). This particular characteristic of the multiple choice format benefits students that are less averse to risk and penalizes those that are more averse to taking risks (Espinosa and Gardeazabal, 2010).

A number of studies have examined gender differences in answering multiple choice questions. They suggest that male students normally outperform their female counterparts in this exam format (Pekkarinen, 2015; Iriberri and Rey-Biel, 2018)), while the latter tend to omit more items than males (Riener and Wagner, 2017; Baldiga, 2014; Iriberri and Rey-Biel, 2018). Moreover, the students that omit most items tend to perform worse (Baldiga, 2014; Pekkarinen, 2015) and to be more risk averse (Espinosa and Gardeazabal, 2013; Akyol et al., 2014).

3 Institutional Setting

The present paper takes advantage of the evaluation system employed on a course entitled *Principles of Taxation* (*Fonaments de la Fiscalitat* in Catalan), which allows us to exploit a quasi-experimental design to address our research question. It 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.[3] The course is an introduction to taxation during which 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.[4] It combines theoretical and practical content, that is, taxation theory with numerical exercises. Our empirical set-up is made possible by (i) the implementation of a system of continuous assessment (hereinafter, CA) on the course, in line with the Bologna Process; and, (ii) the specific design adopted for the evaluation system and its evolution during the academic years analysed.

3.1 Course evaluation system

The Faculty of Economics and Business implemented the Bologna Process in its Bachelor's Degree in Business Administration in the 2011/12 academic year. This meant teachers had to redefine the courses on the Bachelor program, introducing CA and rethinking the teaching process (Gallardo et al., 2010). *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 implementation of the Bologna Process, in the 2009/10 academic year, they introduced a pilot CA system as an alternative to that of single assessment (SA).

With the adoption of the CA, the course's evaluation system comprises two periods: first, the CA and, second, a final exam at the end of the term.[5] In the first period, the CA is conducted at the same time as lectures are delivered. Assessment is based on various (see below for details) multiple choice tests (30 minutes) taken on a computer during the term covering

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

[4] For further information about the course and its content (academic year 2017/18), see `https://goo.gl/Z6z1wp`

[5] 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.

the content introduced between each midterm. In the second period, the final exam is sat at the end of the term and after lectures have finished. The final exam is divided into two parts: students, first, take a multiple choice test (30 minutes) and, second, complete an open-question test (90 minutes). The multiple choice midterm and final tests are largely similar, employing a similar structure and questions with the same level of difficulty. The small differences that exist are explained in section 3.2. This system of evaluation, which allows us to compare student performance when completing very similar multiple choice tests, was in operation until the 2016/17 academic year. In that year the structure of the final exam was changed and the multiple choice element eliminated. Therefore, the timespan of interest for us comprises the seven academic years (from 2009/10 to 2015/16) in which the midterm CA test and first part of the final exam employed very similar multiple choice tests.

Hereinafter, and for the sake of clarity, we refer to the CA multiple choice midterm tests as *midterms* and the average grade awarded during this period of CA as the *CA grade*. The final exam, that is, the multiple choice questions plus the open-question test, is referred to as the *final exam*, where the multiple choice questions on this exam are referred to as the *final test* and the grade awarded on that test as the *final test grade*.[6] The grade awarded to the student for the whole of the final exam is denoted as *final exam grade* while the whole course grade is denoted as the *overall course grade*.

Over these seven academic years, the course coordinators have introduced minor changes to the evaluation system. To understand fully the pressures faced by students, we summarise the specific rules and the changes made to them in Table 1. The Table 1 is divided into three main sections: rules applied during the period of 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 (if they had opted to sit them all). In the last two years, with just two midterms, this possibility was eliminated. The CA grade, therefore, is computed as the average grade of all midterms, each with the same weight.

The final test was eliminatory for the first four years of our study. This means that students had to fulfil at least one of two requirements: (i) score 4 or more on the final test or, (ii) have a CA grade of 5 or more. If neither requirement was met, the second part of the final exam, the open questions, was not marked. As Table 1 shows, this rule was eliminated in the 2013/14 academic year and the final test was no longer eliminatory. Finally, certain changes have been made in computing the overall course grade. First, the weight of the CA grade has been progressively increased from 10 to 40%. This reflects the decision to gradually introduce CA, allowing the coordinators to experiment with the course design in readiness for the official implementation of the Bologna Process. Later, in the 2013/14 academic year, a minimum final exam grade of 4 was required for the CA grade to be taken into account in the overall course grade. Students scoring less than 4 were deemed to have failed the course. Finally, the overall course grade was computed with the weighted average between CA grade and the final exam grade, in line with the rules outlined above. This rule was changed in the 2014/15 academic year and the overall course grade was computed as the maximum between the final exam grade,

[6] We do not specifically refer to the second part of the final exam (the open-question section) as it plays no role in our study.

on the one hand, and the weighted average between the CA grade and the final exam grade, on the other.

Table 1: Evaluation system for *Principles of Taxation*

Academic Year	CA		Final Exam	Overall Course Grade		
	No. Midterms	Average Grade	Test eliminatory	% CA	Minimum grade F.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 for the CA grade to be taken into account for the overall course grade.

3.2 Quasi-experimental setting

The specific evaluation system employed by the course *Principles of Taxation* over these seven years allows us to exploit a quasi-experimental situation with regards to the test pressure faced by students, where test pressure is defined as the increased necessity of performing well on a test due to the increasing weight of that examination within the overall course grade.[7] In other words, a good performance on an exam that has more impact on the overall course grade (high stakes) is more important than an exam of lesser impact (low stakes). Therefore, students feel under greater pressure to perform well on a high-stakes exam given the graver consequences of obtaining a poor grade or of failing. Based on this definition, we define two levels of pressure in our main setting: the low pressure scenario corresponds to the midterms (CA) while the high pressure scenario corresponds to the final test. Recall that the weight of the CA grade is lower than that of the final exam in the overall course grade (see Table 1). Depending on the academic year students sit a different number of midterms, the weight varying between 3.33 and 20%. However, the final test (multiple choice) accounts for 33% of the final exam grade (i.e. between 20 and 30% of the overall course grade). Thus, the pressure associated with each of the midterms is lower than the pressure associated with the final test. Additionally, students are given two hours to complete the two parts of the final exam, which means the pressure associated with the final test (the first part of this exam) is high, being an important part of this evaluation stage.

[7]There is some debate as to what constitutes a quasi-experiment and what methodology should be employed. In our main setting, our quasi-experiment involves students taking very similar tests in the same year, but exposed to different levels of pressure. We do not, therefore, define either treatment or control groups, as the same students are exposed to the different (high-low) pressure scenarios. This rich setting allows us to compile panel data with all those students sitting the tests over the term. Therefore, we use the term quasi-experiment as this natural setting allows us to isolate the role played by pressure in the students' academic performance, while ensuring no effects are attributable to exam format, content or test difficulty or to different groups of students. As such, we can isolate gender differences in test performance under varying degrees of pressure in a similar way to a designed experiment.

Furthermore, it should be borne in mind that a number of specific rules included within the evaluation system determined the pressure to which students were subject. There are three additional characteristics that help to further reduce the pressure faced in the CA part of the course (the first period). First, the fact that students sit several midterms has the effect of lowering the level of risk associated with the CA grade. Second, in five of the seven years studied, students could discard their worst midterm grade (assuming they sat them all). Third, in the last two years, the overall course grade is computed as the maximum between the final exam grade, on the one hand, and the weighted average between the CA grade and the final exam grade, on the other. This means that a poor CA performance can be rectified in the final exam. These three characteristics mean that the associated risk can be diversified over the CA and so students face less pressure when sitting the midterms. In contrast, the final exam (the second period) is defined as a high pressure scenario due to the greater weight attached to it in the overall course grade. Moreover, during the first four academic years, the eliminatory nature of the final test made the pressure even higher, given that the second part of the final exam was not marked and the student would automatically fail the whole course. In the following three years, the minimum grade required on the final exam represented pressure to students for the whole final exam. A comparison of the two rules suggests the former (elimination) was a source of greater pressure on the final test than the minimum grade requirement.[8]

These modifications to the system of evaluation represent differences over the academic years in the pressure students face in completing the CA and when sitting the final test. For this reason, we divided the timespan into two intervals of what we consider to be more homogeneous levels of pressure: 2009/10-2012/13 and 2013/14-2015/16 (as indicated by the dashed line in Table 1). By so doing, we are able to examine test pressure attributable to specific rules of evaluation and, as such, this represents our secondary source of test pressure. Moreover, it also allows us to examine the main setting which controls for differences in the pressure faced by students over successive academic years. We define the first interval, 2009/10-2012/13, as being of high pressure and the second, 2013/14-2015/16, as being of low pressure. On the one hand, the incentives to perform well in the CA component are higher in the first interval than in the second due to the eliminatory nature of the final test in that first period (greater pressure) and because of the formula employed to compute the overall course grade in the second (less pressure). On the other hand, the incentives to perform well in the final test are higher in the first interval than in the second due to the eliminatory nature of this test and also to the greater weight given to the final exam grade in the overall course grade.[9]

To sum up, we first define our main setting as the pressure attributable to the weight of a

[8]Student questionnaires 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 can pass the course by performing well on the final exam.

[9]We are aware that this division in two homogenous intervals 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 final test in case of passing the CA. The weight of the final exam in this interval is high, given that not being entitled to have the second part 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.

test in the overall course grade. Thus, we define the CA as low pressure, due to the lower weight of the midterms, and the final test as high pressure, due to its higher weight. This setting can be further defined as a sequential game of two periods in which students first sit low pressure midterms and then sit a high pressure final test. Then, we define a secondary (heterogeneous) setting in which the test pressure varies according to specific rules of evaluation which require students to perform better. Here, we define the first interval as high pressure and the second as low pressure. We should stress that in the main setting we compare the same students exposed to different levels of pressure, i.e. the same student first completes the CA and then goes on to sit the final test. In the secondary setting we compare different cohorts of students exposed to different levels of pressure.

Midterms and the final test are comparable, insofar as they contain the same kind of questions and share a similar structure. 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. Item difficulty on the midterms and final test is the same given that the questions are designed in the same way by the same teachers. However, while the midterms are computer-based, the final test is completed on paper. Moreover, on the midterms, students have 30 minutes to answer ten multiple choice questions plus two or three small exercises in which they must solve a problem by entering a number into a box. However, in the final exam, students have 30 minutes to answer 20 multiple choice questions.

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, tests are corrected by machine, not by teachers. More specifically, midterms are corrected by computer with students taking the exam on the university web page in a computer classroom while the paper-based exams are corrected by machine. This allows us to avoid any possible teacher gender-bias or bias towards specific subgroups of students, as some authors have suggested (see, for example, 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 the CA component and that sit the final exam. This allows us to compare how the behaviour and performance of the same individuals change under different levels of pressure. However, we also examine our main setting with all students enrolled on the course, that is, we also include students opting for single assessment (SA) and students who while completing the CA component decide not to sit the final exam. Additionally, we also examine this main setting including those students who complete neither the CA component nor the final exam. We should stress that the group of students who complete the CA component is not exactly the same as the group that sits the final exam. Given this situation, we analyse the possible issue of self-selection in each period separately.

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 data

The administrative data comprise 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 was 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 Grade Point Average (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 was enrolled on that year (name, group, term and final grade) and whether that year the student won a scholarship. The Faculty's administrative data allow us to compute rich vectors of control variables for individual and group characteristics.

4.2 Course data

The course data comprise two sorts of information: grades and group information. The grade information includes student ID, academic year, grade of each midterm, average CA grade, final test grade, detailed information on the final test (number of questions answered correctly, incorrectly and omitted), the grade for each question in the open-question part, the grade for the final exam 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 provide us with all the student grades plus details for computing the control variables for group characteristics.

4.3 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, 5,464 students enrolled 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 were in the final year of their degree or had registered for all the credits to obtain the degree, were removed from the sample, since for these students, the eliminatory rule applied to the multiple choice test did not affect them. This

would have meant their adopting a different strategic behaviour and having to face completely different levels of pressure on the CA component and final test. Therefore, we were left with 5,013 students, i.e. 92% of all students enrolled on *Principles of Taxation*in the seven-year period. Figure A.1 in Appendix B shows the number of students who opted for CA or SA, the number of students who sat the midterms but did not sit the final exam and the number of students who sat neither midterms nor the final exam.

Table 2: Descriptive statistics for the CA and final test grades

	CA Grade (Low Pressure)			Final Test Grade (High Pressure)		
	Male	Female	Statistic	Male	Female	Statistic
Full Sample (N=5,013)						
N	2,207	2,085		2,253	2,065	
Percentage (%)	51.4	48.6		52.2	47.8	
$H_o^a : Pr(M_i) = 0.5$			-1.862***			-2.861***
$H_o^b : Pr(M_i) = Pr(F_i)$			-2.634***			-4.046***
Mean Test	4.73	5.03	-4.37***	4.89	4.78	1.92*
Median Test	4.94	5.26	12.64***	4.9	4.88	2.65
KS Test			0.071***			0.037
10th percentile	1.38	1.69		2.25	2.38	
25th percentile	2.88	3.51		3.5	3.5	
50th percentile	4.94	5.26		4.9	4.88	
75th percentile	6.5	6.69		6.25	6	
90th percentile	7.69	7.71		7.5	7.25	
Balanced Sample (N=3,912)						
N	2,021	1,891				
Percentage (%)	51.7	48.3				
$H_o^a : Pr(M_i) = 0.5$			-2.079**			
$H_o^b : Pr(M_i) = Pr(F_i)$			-2.94***			
Mean Test	4.94	5.25	-4.46***	4.96	4.83	2.07**
Median Test	5.19	5.44	13.96***	5	4.88	1.83
KS Test			0.076***			0.038
10th percentile	1.63	2.13		2.38	2.38	
25th percentile	3.38	3.94		3.63	3.6	
50th percentile	5.19	5.44		5	4.88	
75th percentile	6.63	6.82		6.25	6.13	
90th percentile	7.81	7.82		7.5	7.25	

Notes: Full sample: 4,292 students opted for CA and 4,318 sat the final exam. In line with the definition of a balanced sample, students opting for CA and final test are the same. Therefore, the number of males and females, their percentages and Tests *a* and *b* are the same for CA and the final test. The null hypothesis for Test *a* (H_o^a) is that the proportion of males (M_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 test 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 test, respectively) across gender, D-statistic. *** denotes significance at the 1% level, ** the 5% level and * the 10% level.

Table 2 shows the main descriptive statistics for the full sample and the balanced sample.[10] The full sample comprises 5,013 students, while the balanced sample comprises all students who opt for CA and sit the final test, i.e. 3,912 students. In the case of the full sample, 85.6% of the students opt for CA and 86.1% take the final exam. The percentage of male students opting for CA is 51.4%, which is significantly different from the percentage of female students. Likewise, the percentage of male students sitting the final exam is 52.2%, again significantly different from the percentage of female students. Female students score 0.3 points (out of ten) more on the CA grade than male students (significant at 1% level), but score 0.11 points (out of ten) less on the final test (significant at 10% level). In terms of percentiles, the gender grade gap in favour of female students falls at higher percentiles in the case of CA. However, in the final test, while the gender gap is in favour of female students in the first decile it is zero in the 25th percentile. From the median upward, a gender gap in favour of male students emerges and increases at higher percentiles. In the case of the balanced sample, the descriptive statistics are quite similar to those for the full sample, but grades are, in general, higher. Female students score 0.31 points more on the CA grade (significant at 1% level), while male students score 0.13 points more on the final test (significant at the 5% level). Table A.1 in Appendix A shows the same descriptive statistics for the first interval (2009/10 - 2012/13, high pressure) and second interval (2013/14 - 2015/16, low pressure) in the case of the balanced sample. The main finding that male students perform worse in comparison with female students in the CA component, but perform better in the final test remains the same. However, there are differences across the intervals which highlight the importance of taking into account this in the empirical analysis. In Appendix B, Figure A.2 shows the kernel distributions of the CA grade and final test grade by gender and Figure A.3 shows the kernel distributions of the CA grade and final test grade by gender in each interval.

5 Empirical Strategy

The identification strategy involves analysing the gender gap in student performance when exposed to different levels of test pressure while enrolled on the course *Principles of Taxation*. For this purpose, our strategy can be broken down into two main parts. In each part, we give our database a different data structure: an independently pooled cross-sectional structure and a panel data structure. Following the definitions of Wooldridge (2012), in the first part, we have a pooled cross-sectional dataset in which we observe the CA grade and the final test grade over seven academic years. It is important to highlight that the students observed each year are different and that the CA grade and final test grade are different variables in the dataset. This means we observe grades obtained at different points of time (years) from different cohorts of students. The main advantage of this data structure is that observations across years are independently distributed (Wooldridge, 2012). In the second part, we reshape the database giving it a panel data structure and include all students who take both the CA and the final test. This means we follow the same student across the two periods. This sequential setting of two periods (students first complete the CA and then sit the final test) allows us to build a two-period panel

[10] Following the notation used by Iriberri and Rey-Biel (2018), the balanced sample comprises students opting both for CA and sitting the final test.

data with the balanced sample. In this data structure, the time variable is each period, i.e. the CA and the final test. Students complete the CA in $p = 1$ (i.e. first period) and, then, they take the final test in $p = 2$ (i.e. second period). Notice that academic year is not the time variable in this data structure, rather it is a student characteristic. For this reason, we control for year-group fixed effects. Finally, note that the CA and final test grades now form part of the same dependent variable denoted *grade*.[11]

We exploit the advantages of these two data structures. On the one hand, the pooled cross-sectional structure allows us to analyse the two grades separately, i.e. the difference between a low (CA) and high pressure scenario (final test). Second, this allows us to examine differences in the gender gap across the distribution of grades in each situation. Third, we can analyse the possibility of sample selection in each case. Self-selection may arise from those students who have not taken either the CA or the exam, and from those students who have not taken any exam at all. On the other hand, the two-period panel data structure allows us to estimate the same gender gaps in the balanced sample, but here we use another methodology. This allows us, moreover, to analyse whether the gender gaps in the CA and the final test are statistically different. Due to the strength of this second setting, a heterogeneous analysis is provided to examine differences between the two homogenous academic year intervals defined in section 3.2. This heterogeneity analysis seeks to examine the second source of test pressure: evaluation rules that increase test pressure.

The whole empirical strategy aims at identifying the causal effect between a student's gender and their grade – controlling for individual characteristics, group characteristics and year fixed effects – and how this causal effect varies according to the level of pressure faced. 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, we employ three strategies to control for these group and year effects.

Pooled Cross-Section Model. We follow the methodology outlined in Ors et al. (2013) introducing a number of modifications to strengthen the model. In line with their two-stage admission procedure, we estimate CA and final test grades separately by quantile regressions (QR) for the 10th, 25th, 50th, 75th, and 90th percentiles with bootstrapped standard errors with 1,000 replications and we compare the gender gap between the two main scenarios. We take the same percentiles as these authors so as to (i) compare the median coefficient with the ordinary least squares (OLS) coefficient, and (ii) estimate the coefficients in the top and bottom deciles. Our contributions include the fact that we also estimate the regressions by OLS, we analyse self-selection into CA and the final test using the Heckman technique and we have more control variables.[12] The baseline empirical model estimates the grade on student gender and on the control variables of individual characteristics. Note that here we regress the CA grade and

[11] Recall we use the average continuous assessment grade, the CA grade, as the grade for the first period.

[12] Note that Ors et al. (2013) analyse the entrance exams for a Master's program at an expensive, prestigious business school – a special situation with a high payoff. Students are under very high competitive pressure in order to go through to the second stage of the examination and then to be accepted on to the program. However, in our setting, we deal with a compulsory course on an undergraduate degree taught at a public university. Hence, the scenario is one that all university students face many times during higher education.

final test grade separately, so they constitute two different dependent variables. The econometric specification can be expressed as follows:

$$y_{igt} = \beta_0 + \beta_1 \cdot Female_{igt} + \beta_2 \cdot X_{igt} + \varepsilon_{igt} \tag{1}$$

where the dependent variable y_{igt} denotes the grade (either $y = CA\ grade$ or $y = final\ test\ grade$) obtained by student i in group g in the academic year t, $Female_{igt}$ is a dummy variable which takes a value of 1 if the student is female and 0 if male, X_{igt} is the vector of individual controls and ε_{igt} is the error term. We estimate the baseline model using three different strategies to control for year and group fixed effects. We first regress Eq.(1) with control variables of group characteristics (Z_{gt}) and year fixed effects (μ_t); second, we regress the baseline model with group fixed effects (μ_g) and year fixed effects (μ_t); and, third, we estimate Eq.(1) by adding year-group fixed effect (μ_{gt}).[13]

Two-Period Panel Data Baseline Model. This setting is similar to that presented in Iriberri and Rey-Biel (2018), but again we introduce a number of modifications to strengthen the model.[14] We estimate the final test grade on student gender by OLS with period fixed effect (pooled OLS), random effects (RE) and student fixed effects (FE). The econometric specification estimated by pooled OLS and RE can be expressed as follows:

$$Grade_{igtp} = \alpha_0 + \alpha_1 \cdot Female_{igt} + \alpha_2 \cdot Final\_Test_p + \alpha_3 \cdot Female_{igt} \cdot Final\_Test_p + \alpha_4 \cdot X_{igt} + \mu_{gt} + \varepsilon_{igtp} \tag{2}$$

where the dependent variable $Grade_{igtp}$ denotes the grade obtained by student i in group g in academic year t and during period p, where $p = 1$ refers to CA and $p = 2$ to the final test and $Final\_Test_p$ is a dummy variable which takes a value of 1 if it is the second period (final test) or 0 if the first (CA).

The econometric specification estimated by student FE can be expressed:

$$Grade_{igtp} = \alpha_0 + \alpha_2 \cdot Final\_Test_p + \alpha_3 \cdot Female_{itg} \cdot Final\_Test_p + \mu_{it} + \mu_{gt} + \varepsilon_{itgp} \tag{3}$$

where μ_{it} is the student fixed effect. In the heterogeneity analysis, we estimate the following specification by pooled OLS and RE:

$$\begin{aligned} Grade_{igtp} = {} & \alpha_0 + \alpha_1 \cdot Female_{igt} + \alpha_2 \cdot Final\_Test_p + \alpha_3 \cdot Female_{igt} \cdot Final\_Test_p + \\ & + \alpha_4 \cdot 1st\_Interval_{igt} + \alpha_5 \cdot 1st\_Interval_{igt} \cdot Female_{igt} + \alpha_6 \cdot X_{igt} + \mu_{gt} + \varepsilon_{igtp} \end{aligned} \tag{4}$$

[13] 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.

[14] In the setting described by Iriberri and Rey-Biel (2018), 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 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 test, i.e. 78% of our sample. In addition, Iriberri and Rey-Biel (2018) only employ the students' school characteristics and maths grades at school. In our case, we employ individual and group characteristics plus seven years of information.

where $1st\_Interval_{igt}$ 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 control variables for individual and group characteristics used in all the empirical models are detailed in Table A.2 in Appendix A. The individual variables are age, nationality, university access grade [15], number of courses enrolled on in the first term, average grade obtained that first term, and status as scholarship holder and repeat student. The access grade proxies (direct and indirectly) student ability and family background and so allows us to control for the students' unobserved academic abilities.[16] 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 by group.

6 Results

The results are presented in three subsections. The first refers to the pooled cross-sectional structure, in which we present the main results and the sample selection analysis. The second refers to the two-period panel data structure, in which we present the main results and the heterogeneity analysis. The third subsection presents the possible mechanisms underpinning the results.

6.1 Pooled cross-sectional structure

6.1.1 Main results

Table 3 reports the results of the main setting in which the dependent variable is either the CA grade (low pressure) or the final test grade (high pressure). Each grade variable is standardized with mean 0 and standard deviation 1 at year level. The columns are the three augmented baseline models in each period and the rows are the coefficients of the variable *Female* from the OLS (first row) and for each regression from the QR (next five rows).

[15] 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.

[16] For example, Iriberri and Rey-Biel (2018) proxy a student's mathematical ability with their maths grades at school, Ors 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.

Table 3: Gender gap in the CA grade and final test - balanced sample

	CA Grade (Low Pressure)			Final Test Grade (High Pressure)		
	(1)	(2)	(3)	(4)	(5)	(6)
Ordinary Least Square						
Female	0.034	0.037	0.032	-0.082***	-0.084***	-0.081***
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
Quantile Regressions						
$\tau = 0.10$	-0.008	0.008	-0.012	-0.023	-0.0053	-0.013
	(0.052)	(0.050)	(0.050)	(0.055)	(0.052)	(0.054)
$\tau = 0.25$	0.056	0.062*	0.075**	-0.024	-0.059	-0.028
	(0.036)	(0.036)	(0.038)	(0.039)	(0.037)	(0.037)
$\tau = 0.50$	0.044	0.067**	0.037	-0.124***	-0.112***	-0.119***
	(0.035)	(0.033)	(0.035)	(0.036)	(0.035)	(0.038)
$\tau = 0.75$	0.039	0.034	0.026	-0.116***	-0.112***	-0.091**
	(0.032)	(0.034)	(0.035)	(0.039)	(0.039)	(0.042)
$\tau = 0.90$	0.0017	-0.025	-0.030	-0.170***	-0.166***	-0.148***
	(0.037)	(0.037)	(0.040)	(0.043)	(0.042)	(0.044)
N	3,722	3,722	3,722	3,722	3,722	3,722
$Adj.\ R^2_{OLS}$	0.331	0.333	0.342	0.307	0.310	0.321
Individual var.	Yes	Yes	Yes	Yes	Yes	Yes
Group var.	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

Notes: The dependent variable measures CA grade in columns (1) to (3) and final test 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 1,000 replications for the QR. Standard errors are in parentheses and *** denotes significance at 1% level, ** at 5% level and * at 10% level. 190 observations were discarded because they have no university access grade (students access degree via another path) or there is no average grade for the first term (students who take no courses despite enrolling). Individual variables: age, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and 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.

The OLS results show that female students outperform male students by around 0.03 standard deviations in the CA grade (first period - low pressure), though this difference is not statistically significant. However, when the same students sit the final test, male students outperform female students by around 0.08 standard deviations (second period - high pressure), which is statistically significant at the 1% level. Thus, male and female students perform at a similar level when subject to the low pressure associated with a lower stake in the overall course grade. However, a gender gap emerges in favour of male students when subject to the high pressure associated with a higher stake in the overall course grade. We also estimate these specifications when the dependent variable is the grade of the first midterm in the CA (see results in Panel A, Table B.1 in Appendix A). The importance of this midterm lies in the fact

that its test content is similar across the years and is common to all seven years. Moreover, as it is the first test, it is more likely that almost all the students will sit it. In this case, female students outperform male students by around 0.06 standard deviations in the OLS estimates, statistically significant at the 10% level, and around 0.08–0.09 standard deviations in the 25th, 50th and 75th percentiles in the QR estimates, statistically significant at levels between 1 and 10%.

The QR estimates allow us to examine differences in the gender gap across the grade distribution in each period. Over the CA grade distribution (first period - low pressure), in columns (1) to (3), we find no gender grade gap across percentiles, albeit with a few exceptions. In the case of the final test (second period - high pressure), in columns (4) to (6), male students outperform female students in the median, 75th percentile and 90th percentile, being statistically significant at the 1% level, while there are no significant differences in the first decile and the 25th percentile. Coefficients are similar across the three augmented specifications. Moreover, the gender gap is higher for the top students (last decile) by around 0.16–0.17 standard deviations. Comparing the average effect with the median effect, the gender gap in the median is around 0.10 –0.12 standard deviations vs. the 0.08 in the OLS. In addition, we also estimate these specifications with the final test grade as our dependent variable, including the CA grade as an independent variable (see results in Panel B, Table B.1 in Appendix A). Since it is a two-period sequence, the CA grade gives more information about student performance on the final test. The results and interpretation are the same, and no significant changes emerge.[17]

6.1.2 Sample selection analysis

In this section we analyse whether the characteristics that lead students not to sit midterms or the final test are correlated with their gender. First, we estimate the same regressions as in the previous section but using the full sample. Second, we take into account those students who do not sit either the CA or the final test (sample selection approach).

Table 4 reports the estimates using the same specifications as those in Table 3 above, but for the full sample (that is, with all the students completing the CA component and all the students who sit the final test). The results are very similar to those obtained with the balanced sample. Female students outperform male students in the CA by around 0.03 standard deviations (which is not statistically significant), and male students outperform female students in the final test by around 0.08 standard deviations, statistically significant at the 1% level. The results from the QR are very similar to those obtained in the balanced sample, but with a notable difference. In the CA grade, the median estimates are statistically significant at the 10% level. Therefore, taking into account all the students who complete the CA, a gender gap emerges in favour of female students in the median (around 0.06 standard deviation), while there are no differences in the OLS estimates.

[17]The pooled cross-sectional structure uses the estimates without CA grade as our independent variable, because in the sample selection this variable cannot be added due to the nature of the analysis.

Table 4: Gender gap in the CA and final test grades - full sample

	CA Grade (Low Pressure)			Final Test Grade (High Pressure)		
	(1)	(2)	(3)	(4)	(5)	(6)
Ordinary Least Square						
Female	0.034	0.037	0.032	-0.079***	-0.083***	-0.081***
	(0.026)	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)
Quantile Regressions						
$\tau = 0,1$	-0.040	-0.035	-0.005	-0.025	-0.033	-0.013
	(0.051)	(0.050)	(0.047)	(0.051)	(0.052)	(0.053)
$\tau = 0,25$	0.038	0.045	0.066*	-0.032	-0.0572	-0.036
	(0.040)	(0.038)	(0.038)	(0.038)	(0.0368)	(0.037)
$\tau = 0,5$	0.062*	0.069**	0.054*	-0.116***	-0.112***	-0.114***
	(0.034)	(0.033)	(0.032)	(0.034)	(0.035)	(0.037)
$\tau = 0,75$	0.048	0.048	0.027	-0.103***	-0.098***	-0.089**
	(0.031)	(0.033)	(0.033)	(0.037)	(0.037)	(0.039)
$\tau = 0,9$	-0.008	-0.025	-0.010	-0.176***	-0.177***	-0.163***
	(0.034)	(0.035)	(0.035)	(0.040)	(0.041)	(0.044)
N	4,070	4,070	4,070	4,100	4,100	4,100
$Adj.\ R^2_{OLS}$	0.355	0.357	0.364	0.303	0.305	0.316
Individual var.	Yes	Yes	Yes	Yes	Yes	Yes
Group var.	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

Notes: The dependent variable measures the CA grade in columns (1) to (3) and the final test grade in 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) for the OLS and QR. Standard errors, clustered at year-group level, for the OLS and bootstrapped standard errors with 1.000 replications for the QR. Standard errors are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. 4,292 students take the CA component, 222 observations being eliminated because the students had no university access grade (having accessed the degree via another path) or no average grade for the first term (students who enrol but then fail to take any course). 4,318 take the final exam, 218 observations being eliminated for the same reasons as above. Individual variables: age, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and 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.

We then analyse the gender gap in the CA grade taking into account those students who do not complete the CA component and the gender gap in the final test grade taking into account those students who do not sit the final test. To do so, we conduct the two-step procedure technique developed by Heckman (1979) to correct for sample selection. The first step estimates the probability of completing the CA (or final test), and the second step regresses the specifications defined in the empirical strategy, but correcting for this sample selection. We denote the first step 'selection equation' and the second as 'augmented model'. Table 5 reports the *Female*

estimates from the two Heckman steps for both the CA and final test grades. In the CA grade, the selection equation shows that female students are more likely to complete the CA than male students; therefore, there is sample selection over the first period of the term. Once this selection bias is corrected, the gender gap in the case of CA is around 0.10 standard deviations in favour of female students, which is statistically significant at the 1% level. However, the selection equation for the final test grade does not show any gender differences in the probability of taking the final test. Therefore, there is no selection bias in the final test, and the gender gap in favour of male students estimated with the balanced and full sample remains unchanged at around 0.08 standard deviations.

Table 5: Gender gap in the CA grade and final test - Heckman

	CA Grade (Low Pressure)			Final Test Grade (High Pressure)		
	(1)	(2)	(3)	(4)	(5)	(6)
Augmented Model						
Female	0.099***	0.102***	0.090***	-0.081***	-0.085***	-0.082***
	(0.034)	(0.029)	(0.034)	(0.027)	(0.027)	(0.027)
Selection Equation						
Female	0.172***	0.181***	0.167***	-0.061	-0.061	-0.066
	(0.054)	(0.047)	(0.048)	(0.055)	(0.055)	(0.056)
N	4,668	4,668	4,668	4,668	4,668	4,668
Individual var.	Yes	Yes	Yes	Yes	Yes	Yes
Group var.	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

Notes: The dependent variable in the second step (Augmented B. Model) measures the CA grade in columns (1) to (3) and the final test 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 (or the final test) and 0 otherwise. 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 augmented baseline model (second step). The explanatory variables of the selection equation are the same as those in each augmented baseline model, 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 test grade. This is because the selection equation must have at least one different variable to those included in the 2nd step equation. Moreover, in the final test grade, we control for the student having previously completed, or otherwise, the CA component. Heckman standard errors are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. The full sample comprises 5,013 students, 345 observations being eliminated because the students had no university access grade (having accessed the degree via another path) or no average grade for the first term (students who enrol but then fail to take any course). Of the 4,668 students, 598 do not complete the CA component and 568 do not sit the final exam.

6.2 Two-period panel data structure

6.2.1 Main Results

Table 6 shows the estimates for the main setting when the data are structured as two-period panel data. With this structure, the dummy variable *Female* identifies the gender gap in the CA grade (1st period - low pressure) and $Female \cdot Final\_Test$ reveals the difference in the gender gap between the CA and the final test grades. Thus, $Female + Female \cdot Final\_Test$

identifies the gender gap in the final test grade (2nd period - high pressure). For the sake of simplicity, we only show results with year-group FE, the most restrictive manner of controlling for peer-teacher and year effects. The pooled OLS and RE estimates, columns (1) and (2), show that female students outperform male students in the first period by around 0.043 standard deviations, which is not statistically significant. However, the gender gap is reversed in the second period (final test) by 0.134 standard deviations, and is statistically significant at the 1% level. This means that male students outperform female students in the final test (second period) by around 0.091 standard deviations (0.043 - 0.134).[18] The FE estimates, column (3), show exactly the same coefficients and significance for the interaction $Female \cdot Final\_Test$.[19] In the FE model, we cannot compute the gender gap in each period since the variables do not vary over time (i.e. over the two periods), such as $Female$, are controlled for by the individual FE. Nevertheless, we compute this model as it is the most restrictive one due to the individual FE. This allows us to control for possible omitted variables such as individual, group or year variables and to verify the coefficients $Final\_Test$ and $Female \cdot Final\_Test$.

These estimates confirm the previous results (section 6.1.1) from the pooled cross-sectional structure. Table 7 compares the results obtained with the two data structures used. On the one hand, the pooled cross-sectional structure estimates a gender gap of around 0.032 standard deviations for the CA grade, while the two-period panel estimates a 0.043 standard deviation, neither of which is statistically significant. On the other hand, the gender gap estimated is -0.086 and -0.091 standard deviations for the final test, respectively, both statistically significant at the 1% level. This indicates that the results from the two-period panel data are completely aligned with those from the pooled cross-sectional structure. Moreover, with the two-period panel model, both pooled OLS and RE present very similar estimates and, therefore, a very similar gender gap.

[18] We estimate the same specifications, but instead we define the dummy period variable as CA, which takes a value of 1 if it refers to the first period (CA) and 0 if it refers to the second period (final test). In this case, the dummy variable $Female$ reveals the gender gap in the final test grade (2nd period - high pressure) and $Female \cdot CA$ reveals the difference in the gender gap between the CA and the final test. Thus, $Female$ plus $Female \cdot CA$ identifies the gender gap in the CA grade (1st period - low pressure). Table B.2 in Appendix A shows the results. The coefficient Female is -0.091, and statistically significant at the 1% level, i.e. the gender gap in the final test. CA and $Female \cdot CA$ are the same as in Table 6 but, as expected, they present the opposite sign.

[19] Note that individual and group variables and year-group FE are constant through the two periods, thus they are omitted due to individual FE.

Table 6: Gender Gap - Main Setting

	Pooled OLS	RE	FE
	(1)	(2)	(3)
$Female$	0.0427	0.0427	
	(0.0273)	(0.0273)	
$Final\_Test$	-0.0189	-0.0189	-0.0189
	(0.0377)	(0.0377)	(0.0375)
$Female \cdot Final\_Test$	-0.1337***	-0.1337***	-0.1337***
	(0.0374)	(0.0374)	(0.0372)
N	7,444	7,444	7,444
Number of Ind.	3,722	3,722	3,722
Individual var.	Yes	Yes	No
Individual FE	No	No	Yes
Period FE	Yes	Yes	Yes
Year-Group FE	Yes	Yes	No

Notes: The dependent variable measures student performance over the two periods: the CA grade in the first period and final test 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. $Final\_Test$ takes a value of 1 if the grade refers to the second period (final test grade) and 0 if it refers to the first period (CA grade). Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Balanced sample.

Table 7: Gender gap - pooled cross-section vs two-period panel data

	Pooled Cross-Section		Two-Period Panel Data	
	CA Grade	Final Test Grade	Pooled OLS	RE
	(1)	(2)	(3)	(4)
CA Gender Gap (low pressure)	0.032	-	0.0427	0.0427
Final Test Gender Gap (high pressure)	-	-0.086***	-0.091***	-0.091***

Notes: Columns (1) and (2) refer to the pooled cross-sectional structure presented in section 6.1.1 while columns (3) and (4) refer to the two-period panel data structure. The *CA Gender Gap* in column (1) is taken from the coefficient *Female* in Table 3, column (3), OLS estimate and the *Final Test Gender Gap* in column (2) is taken from the coefficient *Female* in Table B.1 in Appendix A, column (6), OLS estimate. The *CA Gender Gap* in columns (3) and (4) is taken from the coefficient *Female* in Table 6, columns (1) and (2), respectively and the *Final Test Gender Gap* in columns (3) and (4) is taken from the coefficient *Female* in Table B.2 in Appendix A, columns (1) and (2), respectively.

6.2.2 Heterogeneity analysis

Having estimated the gender gap in the main setting and checked the robustness of our results, we now analyse the heterogeneous effects on the gender gap attributable to the different levels of pressure experienced by students across the seven academic years when completing the CA component and sitting the final test. In the main setting, we have studied the test pressure, defined as the weight of a test in the overall course grade. In this heterogeneity analysis, we analyse the second source of test pressure, that is, specific rules in the system of evaluation that increases the students' need to perform well. Recall that over the seven-year timespan the course

coordinators introduced various changes to the evaluation system that determined the level of pressure associated with the CA component and the final test. As explained in section 3.2, this timespan can be divided into two intervals characterized by fairly homogenous test pressure: the first, from 2009/10 to 2013/14, of high pressure, and the second, from 2013/14 to 2015/16, of low pressure. This heterogeneity analysis allows us to test the hypothesis that as the test pressure is increased, the wider the gender gap is in favour of male students.

For this purpose, we estimate Eq.(4) which contains two more variables than Eq.(3). We introduce the dummy variable $1st\_Interval$ which takes a value of 1 for those years of higher test pressure (first interval) and 0 for the years with lower test pressure (second interval). Moreover, and more importantly, we introduce the interaction between this dummy variable and the student gender variable: $1st\_Interval \cdot Female$. This interaction is our variable of interest, and shows the additional gender gap that emerges in the years of higher test pressure. Table 8 reports the estimates of the pooled OLS and RE with these two new variables. Since the time variable in this data structure is the two periods (i.e., the CA and the final test), the two new variables included in the specification are constant over the panel data. For this reason, the FE estimator is not estimated since we would obtain the same results as those reported in column (3) in Table 6.

Table 8: Gender gap - Heterogenous Effects

	Pooled OLS	RE	Pooled OLS	RE
	(1)	(2)	(3)	(4)
$Female$	0.0650*	0.0650*	0.0650*	0.0650*
	(0.0350)	(0.0350)	(0.0350)	(0.0350)
$Final\_Test$	-0.0189	-0.0189	-0.0189	-0.0189
	(0.0377)	(0.0377)	(0.0377)	(0.0377)
$Female \cdot Final\_Test$	-0.1337***	-0.1337***	-0.1337***	-0.1337***
	(0.0374)	(0.0374)	(0.0374)	(0.0374)
$1st\_Interval$	0.191***	0.349***		
	(0.0227)	(0.0273)		
$Female \cdot 1st\_Interval$	-0.0633*	-0.0633*	-0.0633*	-0.0633*
	(0.0372)	(0.0372)	(0.0372)	(0.0372)
N	7,444	7,444	7,444	7,444
Number of Ind.	3,722	3,722	3,722	3,722
Individual var.	Yes	Yes	Yes	Yes
Individual FE	No	No	No	No
Period FE	Yes	Yes	Yes	Yes
Year-Group FE	Yes	Yes	Yes	Yes

Notes: The dependent variable measures the performance of students over the two periods: the CA grade in the first and the final test grade in the second. The dependent variable is standardized 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. The $Final\_Test$ takes a value of 1 if the grade refers to the second period (final test grade) and 0 to the first period (CA grade). The $1st\_Interval$ dummy variable takes a value of 1 if it refers to the first interval (high pressure) and 0 to the second (low pressure). Standard errors, clustered at year-group level, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Balanced sample.

Now, the interpretation of the dummy variable $Female$ refers to the gender gap in the CA

component during the second interval (low pressure). The interaction $Female \cdot Final\_Test$ is interpreted as before, and is the difference in the gender gap associated with the CA component and the final test. $Female \cdot 1st\_Interval$ is the difference in the gender gap between the first and second intervals (for grades in general, both periods). The results in Table 8 show that in the second low pressure interval, female students outperform male students on the CA component by 0.065 standard deviations, being statistically significant at the 10% level ($Female$ coefficient). The additional gender gap between the first period (CA) and the second period (final test), $Female \cdot Final\_Test$, remains equal than before, -0.134 standard deviations, and is statistically significant at the 1% level. However, male students outperform female students more in the first interval than they do in the second by around 0.063 standard deviations, a result that is statistically significant at the 10% level.

Interpreting the sign and magnitude of the dummy variable $1st\_Interval$ is not straightforward as it gives rise to collinearity with the year-group FE. This means that any interpretation of this variable depends on the constant and the two year-group FE omitted. However, we are interested in the additional gender gap that emerges in the high pressure interval, that is, the interaction $Female \cdot 1st\_Interval$, and not in the general differences between the two intervals. In light of this, we estimate the same specification without the dummy variable $1st\_Interval$ to check whether the coefficient of the interaction might present issues of collinearity. The results in columns (3) and (4) are unchanged from those in columns (1) and (2), and the interaction coefficient remains the same. This is because year-group FEs control perfectly for the differences between the two intervals. As mentioned, we cannot measure the general effect of this pressure heterogeneity, our interest lying solely in the interaction itself.

In order to understand the gender gap in each period and interval, the coefficients are interpreted as follows. The gender gap in the CA component in the first interval is given by $Female + Female \cdot 1st\_Interval$ and the gender gap in the CA component in the second interval is given by $Female$. Additionally, $Female + Female \cdot Final\_Test + Female \cdot 1st\_Interval$ shows the gender gap in the final test in the first interval and $Female + Female \cdot Final\_Test$ the gender gap in the final test in the second interval. Table 9 shows the computation of these four gender gaps. In the CA component (low pressure), the gender gap is almost 0 in the first interval (high pressure) and 0.065 standard deviations in favour of female students in the second interval (low pressure). In the final test (high pressure), male students outperform female students in the final test by around 0.132 standard deviations in the first interval (high pressure). However, in the second interval (low pressure), the gender gap in favour of male students narrows to 0.069 standard deviations.[20] Therefore, when we analyse for the two sources of test pressure, we observe that as the test pressure increases, the gender gap grows in favour of male students. However, as test pressure falls, the gender gap is mitigated, and is even reversed in favour of female students. These important results can be easily interpreted in Figure 1.

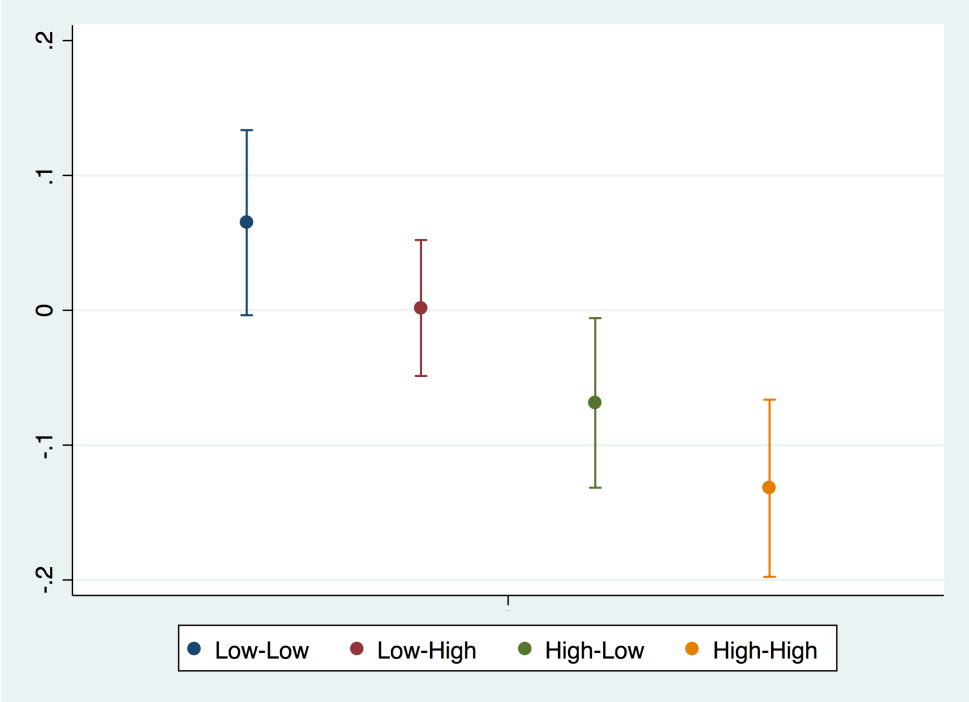
[20] As a robustness exercise, we estimate the three augmented models using the pooled cross-sectional structure for each grade and interval. OLS estimates in Table B.3 in Appendix A also show four similar gender gaps to those shown in Table 9. In addition, Table B.4 shows the similarity between the estimates obtained using both data structures.

Table 9: Gender gap - 1st Interval and 2nd Interval

	CA Grade (Low Pressure)		Final Test Grade (High Pressure)	
	1st Int. (High)	2nd Int. (Low)	1st Int. (High)	2nd Int. (Low)
	(1)	(2)	(3)	(4)
Female	0.0650*	0.0650*	0.0650*	0.0650*
Female · *Final_Test*			-0.1337***	-0.1337***
Female · 1*st_Interval*	-0.0633*		-0.0633*	
Gender Gap	0.0017	0.0650	-0.1320	-0.0687

Notes: Coefficients taken from Table 8 column (2). *Gender Gap* derived from own calculations with the estimated coefficients.

Figure 1: Gender gap ordered from the lowest to the highest pressure scenario



Notes: The first level of pressure refers to CA-Final Test and the second one to first-second interval. Therefore, Low-Low = CA - 2nd Interval, Low-High = CA - 1st Interval, High-Low = Final Test - 2nd Interval, High-High = Final Test - 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.

6.3 Potential mechanisms

The aim of this section is to analyse the possible mechanisms that might explain the gender differences identified above in response to test pressure. First, we seek to disentangle the gender gap and analyse whether female students choke under pressure or whether male students are more highly motivated when placed under pressure. Second, we examine gender differences in the way students answering multiple choice tests. More specifically, we seek to find differences in item omission strategies in the final test.

6.3.1 Disentangling the gender gap

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 two-period panel data separately for female and male students. Thus, the dummy variable $Female$ and the interaction $Female \cdot Final\_Test$ are not included in these specifications. The variable of interest as we seek to identify differences across gender is the dummy variable $Final\_Test$ and its significance, sign and magnitude. Table 10 reports the results for each gender using RE and FE.

Table 10: Gender differences between the CA and final test grades

	RE		FE	
	(1)	(2)	(3)	(4)
	Female	Male	Female	Male
$Final\_Test$	-0.1525***	-0.0189	-0.1525***	-0.0189
	(0.0358)	(0.0378)	(0.0355)	(0.0375)
N	3,604	3,840	3,604	3,840
Number of Ind.	1,802	1,920	1,802	1,920
Individual var.	Yes	Yes	No	No
Individual FE	No	No	Yes	Yes
Year-Group FE	Yes	Yes	No	No

Notes: The dependent variable measures performance (the CA grade in the first period; the final test grade in the second period). The dependent variable is standardised with mean 0 and standard deviation 1 at each period and year level. The dummy variable $Final\_Test$ takes a value of 1 in the second period (final test, i.e. high pressure) and 0 in the first period (CA, i.e. low pressure). 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, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and status as scholarship holder and repeat student.

The dummy variable $Final\_Test$ identifies differences in performance between the CA and the final test for female and male students, respectively. The RE and FE results are very similar, and show that female students perform worse on the final test than on the CA component, by around 0.15 standard deviations (statistically significant at the 1% level). However, the performance of male students on the CA component and final test are largely the same, their performance decreasing by around 0.02 standard deviations (which is not statistically significant). On the one hand, these results indicate that female students tend to choke under pressure, and that their performance suffers when they face a high pressure test environment. On the other hand, male students appear not to be affected by higher test pressure and their performance remains unchanged. To sum up, when the weight attached to a test in the overall grade is high, female students may choke under pressure, while male students appear unaffected.

We repeat this exercise for the heterogeneity setting, that is, when the test pressure is attributable to the rules of the evaluation system. We estimate the same specifications as in Table 11 but add the dummy variable $1st\_Interval$. Now, we are interested in interpreting this variable: that is, in identifying its significance, sign and magnitude. Since we have collinearity between

the dummy variable $1st\_Interval$ and the year-group FE, we estimate the regressions without the year-group FE to overcome this issue. Table 11 shows the results from these estimates controlling with group variables and group FE.

Table 11: Gender differences between the CA and final test grades - heterogeneity analysis

	RE			
	(1)	(2)	(3)	(4)
	Female	Male	Female	Male
$Final\_Test$	-0.153***	-0.019	-0.153***	-0.019
	(0.036)	(0.038)	(0.036)	(0.038)
$1st\_Interval$	0.056	0.164***	0.039	0.116***
	(0.036)	(0.038)	(0.036)	(0.038)
N	3,604	3,840	3,604	3,840
Number of Ind.	1,802	1,920	1,802	1,920
Group FE	No	No	Yes	Yes
Group var.	Yes	Yes	No	No

Notes: The dependent variable measures performance (the CA grade in the first period; the final test grade in the second period). The dependent variable is standardised with mean 0 and standard deviation 1 at each period and year level. The dummy variable $Final\_Test$ takes a value of 1 in the second period (final test, i.e. high pressure) and 0 in the first period (CA, i.e. low pressure). 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, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and 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.

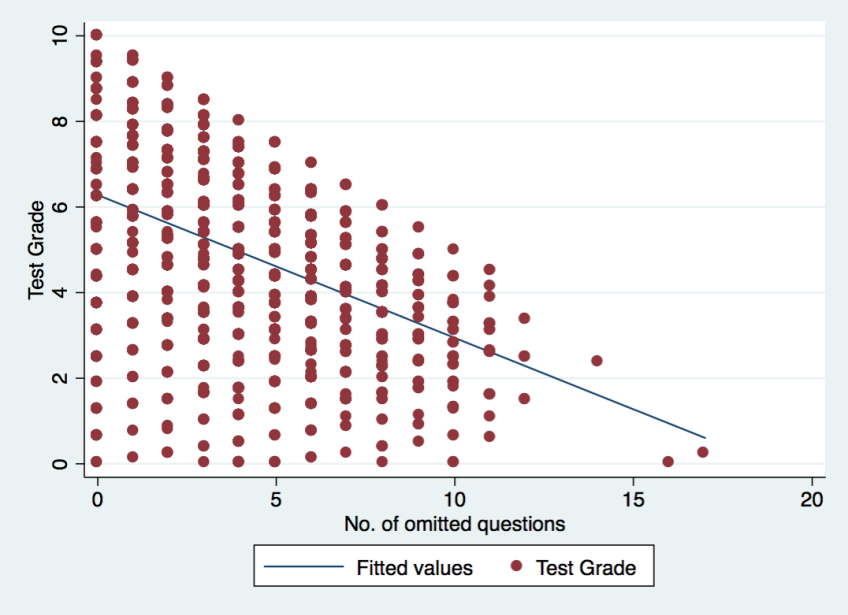
The $Final\_Test$ coefficient is unchanged, and is not affected by either the new dummy variable or the elimination of the year-group FE. Male students perform better (the result being statistically significant at the 1% level) during the first interval of years (high pressure) than they do during the second (low pressure). However, there is no statistically significant difference between the two intervals for the female students. Despite some differences in the $1st\_Interval$ coefficient depending on the way we control for peer-group effects, the sign and significance remain the same. Therefore, behaviour under pressure across gender changes when the source of test pressure changes. In this case, male students perform better under pressure, while the performance of female students remains unaffected. Thus, the additional gender gap estimated in section 6.2.2 is attributable to an improvement in the performance of male students. The second source of test pressure (i.e. rules of evaluation) is sensitive to the definition of the rules themselves, while the first source of test pressure (i.e. the weight of an exam) is more general and applicable.

6.3.2 Omitted questions

Both the economic and educational literature have analysed student behaviour when taking multiple choice tests and, more specifically, several studies have examined gender differences when responding to this test format. The typical focus is on the number of test items omitted,

it has been shown that the number of questions omitted is a good predictor of the student's grade. While this information is not available for the midterms, it is for the final test. This being the case, we use the pooled cross-sectional data structure to analyse gender differences in omitting final test items. Here, we should stress that the QR identifies the distribution of the number of omitted items, and not the grade distribution. Table 12 and Figure 2 show the relation between the number of omitted questions on the final test and the grade scored. The table reports the average final test grade in each part of the omitted-question distribution showing that as the number of omitted test item increases, the lower the final test grade. This negative relation is depicted in Figure 2.

Figure 2: Scatter plot of final test grade and number of omitted questions on the final test



Note: The final test comprises 20 questions and the grade ranges from 0 to 10.

Table 12: Mean final test grade by quantiles of omitted-question distribution

	All	0-10%	10-25%	25-50%	50-75%	75-90%	90-100%
Mean Final Test Grade	4.99	6.15	5.94	5.26	4.67	4.15	3.50
N	3,579	363	591	899	840	567	319

Table 13 reports the results obtained for the augmented baseline model with year-group FE using OLS and QR. Female students omit more questions than their male counterparts on the final test by around 0.24 standard deviations, which is statistically significant at the 1% level. The gender gap on the omitted-question distribution widens from the first decile to the median, before narrowing in the 75th percentile and the last decile of the distribution. For this reason, we focus our attention on the interaction $Female \cdot 1st\_Interval$, i.e. the first four years of our study which are characterized by a higher degree of test pressure due to the rules of the evaluation system. The OLS estimate shows that female students omit more questions than their male counterpart in this first interval, but the coefficient is not statistically significant. However,

the gender gap is significant at the 10% level in the first decile and the 25th percentile, which corresponds to that part of the distribution in which students omit fewer items. According to Table 12, this part of the omitted-question distribution concentrates students with the highest final test grades, i.e. the best students. This result implies that students with the highest grades are impacted more strongly by the test pressure attributable to specific evaluation rules than are students with lower grades. In other words, among the best students, females omit more questions than males on the final test.

Table 13: Gender differences in the omission of questions on the final test

	OLS	Quantile Regressions				
	(1)	(2)	(3)	(4)	(5)	(6)
		$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
$Female$	0.237***	0.203***	0.225***	0.285***	0.225***	0.169*
	(0.050)	(0.064)	(0.054)	(0.053)	(0.064)	(0.087)
$1st\_Interval$	-0.092*	-0.187	-0.134	0.074	-0.0282	-0.347
	(0.052)	(0.334)	(0.222)	(0.228)	(0.285)	(0.451)
$Female * 1st\_Interval$	0.071	0.242*	0.196*	0.076	-0.019	0.024
	(0.078)	(0.124)	(0.109)	(0.092)	(0.099)	(0.125)
N	3,579	3,579	3,579	3,579	3,579	3,579

Notes: The dependent variable measures the number of omitted question on the final test. 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. $Final\_Test$ takes a value of 1 if the grade refers to the second period (final test grade), and 0 to the first period (CA grade). $1st\_Interval$ 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 1,000 replications for the QR, are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. The sample includes 3,579 students as opposed to 3,722, as information was missing for 143 students. Individual variables: age, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and status as scholarship holder, repeat student and CA grade. All the regressions include individual variables and year-group FE.

7 Discussion and Conclusion

This paper has analysed gender differences in academic performance in response to different levels of test pressure, measured as the weight of a specific test in the overall course grade and in relation to a number of 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 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 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. This focus allows us to construct a panel data structure. Our main findings suggest that there are indeed gender differences in student responses to test pressure.

In this balanced sample, male students are found to outperform female students under high test pressure (by between 0.065 and 0.13 standard deviations). However, as the test pressure decreases, the gender gap in favour of male students is narrowed until it is mitigated (0.002 standard deviations) and ultimately, in the lowest pressure scenario, reversed in favour of female students (0.065 standard deviations). We analysed the possible presence of sample selection in relation to completing the CA component and sitting the final exam. Our findings show that female students are more likely than their male counterparts to complete the CA component, while there are no gender differences with regards to sitting the final exam. Therefore, our results hold true for the majority of students (balanced sample – 78%) and, when generalizing the findings to all students, the Heckman technique reinforces these findings.

We have also examined the potential mechanisms that might account for these gender gaps and find that they are likely to differ depending on the source of test pressure. When we analyse the pressure attributable to the weight of the test, female students appear more likely to choke under pressure, while male students maintain their level of performance. Thus, the gender gap results from a fall in the performance of female students. However, our analysis of the test pressure attributable to the rules of the evaluation system shows that male students are more likely to excel while the performance of female students remains unchanged. Thus, in this instance, the gender gap results from a hike in the performance of male students. Moreover, there is suggestive evidence that the top female students omit more items than male students on the final test 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, we find studies in which male students outperform female students when faced with high levels of pressure (Ors et al., 2013; Iriberri and Rey-Biel, 2018; 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 (Ors 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 (Ors et al., 2013; Jurajda and Münich, 2011). For example, Jurajda and Münich (2011) find significant gender differences in the 50th and 75th percentiles as do Ors 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). 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, while Iriberri and Rey-Biel (2018) make the same finding but for all students.

This study makes various contributions to the literature. First, a unique quasi-experimental set up is exploited to analyse gender differences in response to test pressure at university. Using

novel real-world data, we provide evidence of how university students respond to test pressure and we highlight differences across gender in a common situation in higher education. Second, the only study we are aware of that defines pressure as the weight of a specific exam uses field data at the school level. Moreover, we incorporate a new source of test pressure that depends on the specific rules applied in the course's evaluation system. As such, this study contributes to a greater understanding of gender differences in academic performance at the university level. Third, the setting offers the following strengths: (i) there is no gender bias in test correction as this is performed by machine or computer, (ii) the exams compared are of the same type (multiple choice) and, moreover, made up of similar questions with a similar level of difficulty, and (iii) the main setting analyses the same students when they find themselves in different scenarios of pressure. Fourth, the fact that students take multiple choice tests allows us to examine how the gender gap in terms of item omission varies with test pressure. Fifth, our results provide further evidence in the current debate about the adequacy of the multiple choice format of testing.

This study provides additional evidence about gender differences in the taking of multiple choice tests. We confirm that the strategies employed on such 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 assessing student knowledge and abilities, and not student test-taking strategies – differences in grades should be driven by differences in abilities (Riener and Wagner, 2017). Yet, university groups are large and teachers have to mark many exam scripts in a short period of time. Given these circumstances and the low level of resource provision for teaching at university, it is far from easy to offer university teachers a feasible alternative to the multiple choice test format.

References

Akyol, P., Key, J., and Krishna, K. (2014). Hit or Miss? Test Taking Behavior in Multiple Choice Exams. (May 5, 2014) Retrieved from http://econfin.massey.ac.nz/school/seminar%20papers/albany/2014/key.pdf.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender Differences in Response To Big Stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

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

Baldiga, K. (2014). Gender Differences in Willingness To Guess. *Management Science*, 60(2):434–448.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3):610–620.

Bedard, K. and Cho, I. (2010). Early gender test score gaps across OECD countries. *Economics of Education Review*, 29(3):348–363.

Beilock, S. (2011). *Choke: What the Secrets of the Brain Reveal About Getting It Right When You Have To*. Simon and Schuster.

Blau, F. D. and Kahn, L. M. (2000). Gender Differences in Pay. *Journal of Economic Perspectives*, 14(4):75–99.

Blau, F. D. and Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, 55(3):789–865.

Chin, E. C., Williams, M. W., Taylor, J. E., and Harvey, S. T. (2017). The influence of negative affect on test anxiety and academic performance: An examination of the tripartite model of emotions. *Learning and Individual Differences*, 54:1–8.

Contini, D., Tommaso, M. L. D., and Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, 58:32–42.

Daily Mail (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.

De Paola, M. and Gioia, F. (2016). Who performs better under time pressure? Results from a field experiment. *Journal of Economic Psychology*, 53:37–53.

Dee, T. S. (2005). Teachers and the Gender Gaps in Student Achievement. *NBER Working Paper No. 11660*.

Eman, S., Dogar, I. A., Khalid, M., and Haider, N. (2012). Gender Differences in Test Anxiety and Examination Stress. *Journal of Pakistan Psychiatric Society*, 9(2):80–85.

Escardíbul, J.-O. and Mora, T. (2013). Teacher Gender and Student Performance in Mathematics: Evidence from Catalonia. *Document de Treball de l'IEB 2013/7*.

Espinosa, M. P. and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5):415–425.

Espinosa, M. P. and Gardeazabal, J. (2013). Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment. *Journal of Economics and Management*, 9(2):107–135.

Falch, T. and Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36:12–25.

Farré, L. and Vella, F. (2013). The Intergenerational Transmission of Gender Role Attitudes and its Implications for Female Labour Force Participation. *Economica*, 80(318):219–247.

Gallardo, E., Montolio, D., and Camós, M. (2010). The European Higher Education Area at work: Lights and shadows defining Continuous Assessment. *Revista d'Innovació Docent Universitària*, 2:10–22.

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

Goldin, C., Kerr, S. P., Olivetti, C., and Barth, E. (2017). The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census. *American Economic Review*, 107(5):110–114.

Goldin, C. and Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, 90(4):715–741.

González De San Román, A. and 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. and Oreopoulos, P. (2009). A Professor Like Me: The Influence of Instructor Gender on College Achievement. *Journal of Human Resources*, 44(2):479–494.

Holmlund, H. and Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, 15(1):37–53.

Iriberri, N. and Rey-Biel, P. (2018). Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-stage Competition in Mathematics. *The Economic Journal*, pages 1–31.

Jurajda, Š. and Münich, D. (2011). Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities. *American Economic Review*, 101(3):514–518.

Lubienski, S. T., Robinson, J. P., Crane, C. C., and Ganley, C. M. (2013). Girls' and Boys' Mathematics Achievement, Affect, and Experiences: Findings from ECLS-K. *Journal for Research in Mathematics Education*, 44(4):634.

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

Niederle, M. and Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, 24(2):129–144.

Núñez-Peña, M. I., Suárez-Pellicioni, M., and Bono, R. (2016). Gender Differences in Test Anxiety and Their Impact on Higher Education Students' Academic Achievement. *Procedia - Social and Behavioral Sciences*, 228:154–160.

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

Paserman, M. D. (2007). Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players. *IZA Discussion Papers No. 2834*.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*, 115:94–110.

Putwain, D. W., Woods, K. A., and Symes, W. (2010). Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology*, 80(1):137–160.

Riener, G. and Wagner, V. (2017). Shying away from demanding tasks? Experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review*, 59:43–62.

Rodríguez-Planas, N. and Nollenberger, N. (2018). Let the girls learn! It is not only about math ... it's about gender social norms. *Economics of Education Review*, 62:230–253.

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

The Telegraph (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/.

The Times (2017). Oxford 'takeaway' exam to help women get firsts Retrieved on 9th June 2018 from https://www.thetimes.co.uk/edition/news/oxford-takeaway-exam-to-help-women-get-firsts-0v0056k8l.

von der Embse, N., Jester, D., Roy, D., and Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, 227:483–493.

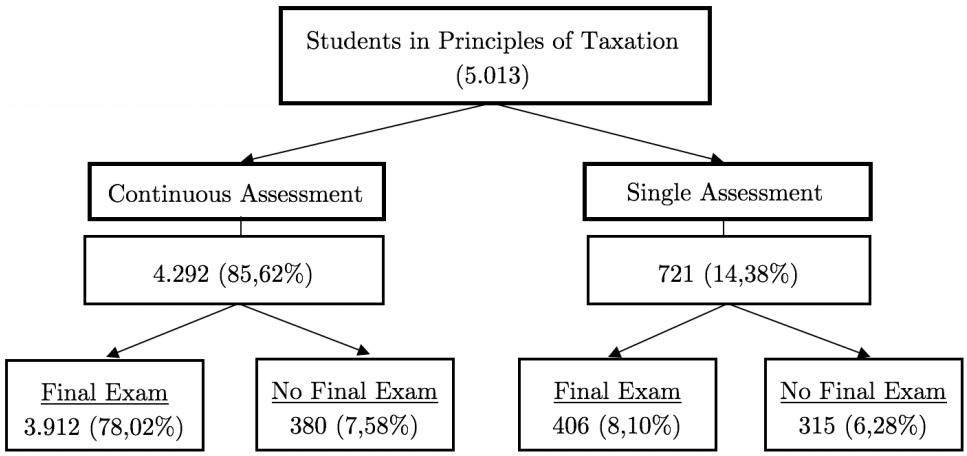
Wolfers, J. (2006). Diagnosing Discrimination: Stock Returns and Ceo Gender. *Journal of the European Economic Association*, 4(2-3):531–541.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*. South-Western, Cengage Learning.

Appendices

A Data

Figure A.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. 100%).

Table A.1: Descriptive statistics for the CA and final test grades - balanced sample

	CA grade			Final Test		
	Male	Female	Statistic	Male	Female	Statistic
1st Interval (N=1,449)						
N	684	765				
Percentage (%)	47.2	52.8				
$H_o^a : Pr(M_i) = 0.5$			2.128**			
$H_o^b : Pr(M_i) = Pr(F_i)$			3.001***			
Mean Test	5.34	5.44	-1.024	4.42	4.35	0.683
Median Test	5.44	5.5	0.201	4.38	4.38	0.0002
KS Test			0.051			0.023
10th percentile	2.82	3.04		1.88	1.88	
25th percentile	4.18	4.44		3	3	
50th percentile	5.44	5.5		4.38	4.38	
75th percentile	6.63	6.71		5.75	5.63	
90th percentile	7.58	7.63		7.13	6.88	
2nd Interval (N=2,463)						
N	1,337	1,126				
Percentage (%)	54.3	45.7				
$H_o^a : Pr(M_i) = 0.5$			-4.252***			
$H_o^b : Pr(M_i) = Pr(F_i)$			-6.013***			
Mean Test	4.74	5.13	-4.078***	5.22	5.15	1.009
Median Test	4.88	5.44	15.907***	5.25	5.13	1.788
KS Test			0.088***			0.034
10th percentile	1.32	1.57		2.88	2.9	
25th percentile	2.82	3.63		3.9	4	
50th percentile	4.88	5.44		5.25	5.13	
75th percentile	6.63	6.88		6.5	6.38	
90th percentile	7.82	8		7.6	7.4	

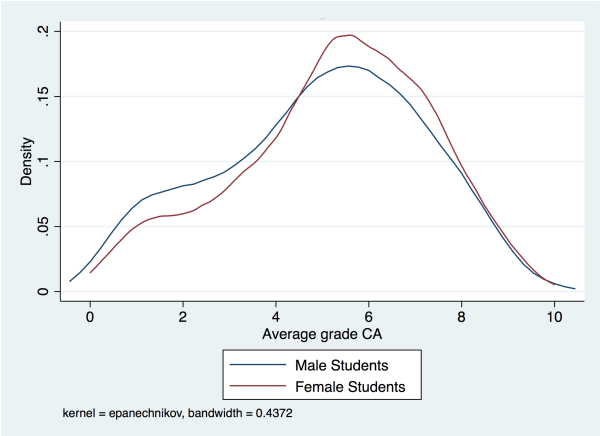
Notes: We define the first interval, 2009/10-2012/13, as being of high pressure and the second, 2013/14-2015/16, as being of low pressure. In line with the definition of a balanced sample, students opting for CA and sitting the final test are the same. Therefore, the number of males and females, their percentages and Tests *a* and *b* are the same for CA and the final test. The null hypothesis for Test *a* (H_o^a) is that the proportion of males (M_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 test 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 test, respectively) across gender, D-statistic. *** denotes significance at 1% level, ** the 5% level and * the 10% level.

Table A.2: Overview of all the variables

Variable	Description
1. Dependent variables	
CA grade	The final grade of the CA is standardised with mean 0 and standard deviation 1 at year level.
Final Test grade	The grade of the multiple choice part of the final exam, standardised with mean 0 and standard deviation 1 at year level.
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 their date of birth.
Nationality	Dummy variable which takes a value of 1 if Spanish national 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)*.
No of courses enrolled	The number of courses that the student enrolled on in the first term of the academic year in question, *Principles of Taxation* included.
Average grade from the courses	Average grade of all the courses that the student enrolled the first term of the academic year, *Principles of Taxation* included.
Repeat Student	Dummy variable which takes a value of 1 if the student has failed the course 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).
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 from 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* included.
Female teacher	Variable which ranges from 0 to 1. This is the proportion of female professors teaching that group, in ECTS terms.

Figure A.2: Kernel density estimations for the whole timespan - balanced sample

(a) CA grade by gender.



(b) Final Test by gender.

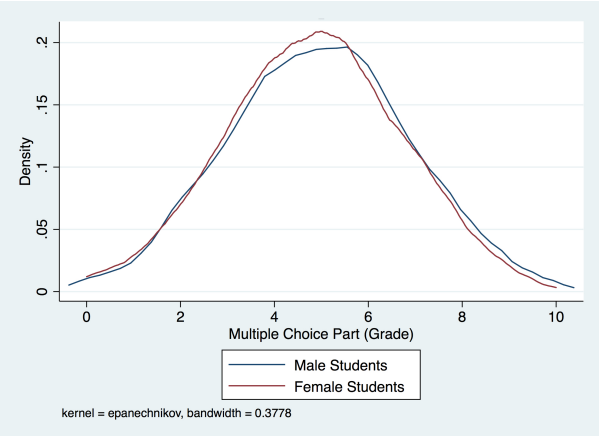
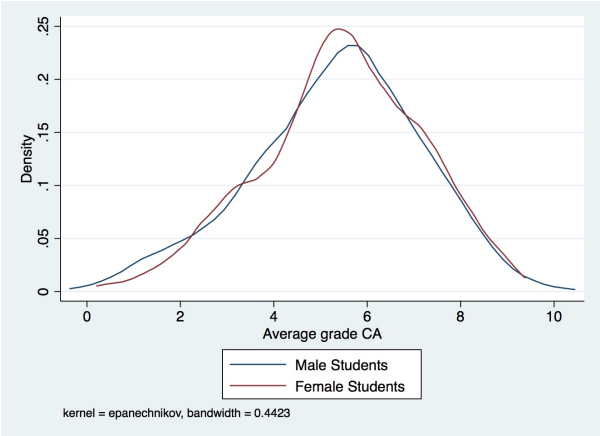
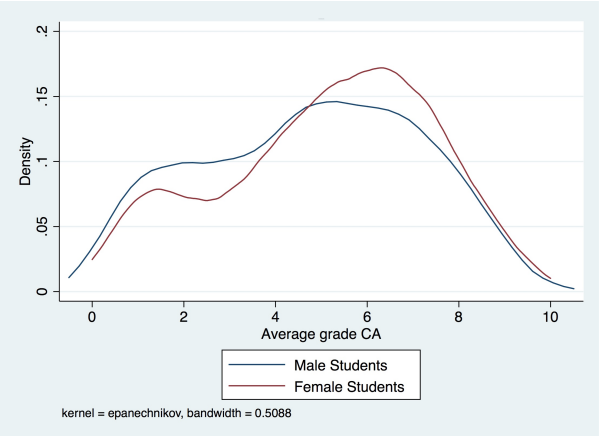


Figure A.3: Kernel density estimations for the whole timespan - balanced sample

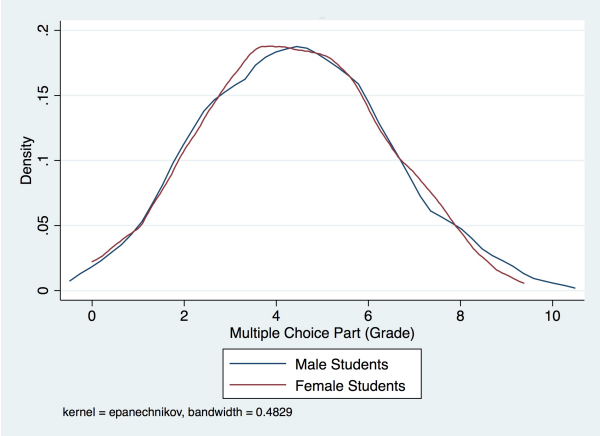
(a) CA grade by gender - 1st Interval.



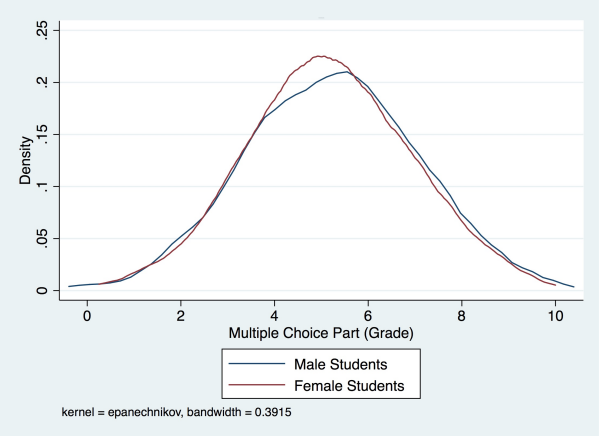
(b) CA grade by gender - 2nd Interval.



(c) Final Test by gender - 1st Interval.



(d) Final Test by gender - 2nd Interval.



B Additional Results

Table B.1: Gender gap in the first midterm (CA) and in the final test with the CA grade as an explanatory variable - balanced sample

	Panel A: 1st Midterm (CA)			Panel B: Final Test Grade		
	(1)	(2)	(3)	(4)	(5)	(6)
Ordinary Least Square						
Female	0.056*	0.057**	0.054*	-0.087***	-0.090***	-0.086***
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
Quantile Regressions						
$\tau = 0.10$	0.007	-0.009	-0.001	-0.014	-0.028	-0.029
	(0.057)	(0.057)	(0.059)	(0.056)	(0.052)	(0.057)
$\tau = 0.25$	0.092*	0.085*	0.091*	-0.032	-0.049	-0.0408
	(0.050)	(0.047)	(0.047)	(0.035)	(0.039)	(0.038)
$\tau = 0.50$	0.080**	0.086**	0.063	-0.144***	-0.125***	-0.128***
	(0.037)	(0.039)	(0.040)	(0.035)	(0.034)	(0.038)
$\tau = 0.75$	0.080**	0.087**	0.083**	-0.110***	-0.108***	-0.114***
	(0.038)	(0.039)	(0.041)	(0.040)	(0.041)	(0.042)
$\tau = 0.90$	0.031	0.045	0.013	-0.160***	-0.175***	-0.156***
	(0.038)	(0.039)	(0.044)	(0.043)	(0.043)	(0.045)
N	3,542	3,542	3,542	3,722	3,722	3,722
$Adj.\, R^2_{OLS}$	0.227	0.231	0.248	0.320	0.323	0.336
Individual var.	Yes	Yes	Yes	Yes	Yes	Yes
Group var.	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

Notes: The dependent variable measures the 1st midterm grade in columns (1) to (3) and the final test grade in 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 female student) for the OLS and QR. Standard errors, clustered at year-group level, for the OLS and bootstrapped standard errors with 1,000 replications for the QR. Standard errors are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. 190 observations were discarded because the students had no university access grade (having accessed the degree via another path) or no average grade for the first term (students who enroll but then fail to take any course). We do not dispose of the grade for the first midterm in the 2009/10 academic year. Individual variables: age, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and status as scholarship holder and repeat student. Estimations in columns (4) to (6) include the CA grade (not standardised) as an explanatory variable. 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.

Table B.2: Gender Gap - Main Setting (CA dummy variable)

	Pooled OLS	RE	FE
	(1)	(2)	(3)
$Female$	-0.0910***	-0.0910***	
	(0.0276)	(0.0276)	
CA	0.0189	0.0189	0.0189
	(0.0377)	(0.0377)	(0.0375)
$Female \cdot CA$	0.1337***	0.1337***	0.1337***
	(0.0374)	(0.0374)	(0.0372)
N	7,444	7,444	7,444
Number of Ind.	3,722	3,722	3,722
Individual var.	Yes	Yes	No
Individual FE	No	No	Yes
Period FE	Yes	Yes	Yes
Year-Group FE	Yes	Yes	No

Notes: The dependent variable measures student performance over the two periods: the CA grade in the first period and the final test 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. The CA *grade* takes a value of 1 if it refers to the first period (CA grade) and 0 to the second period (final test grade). Standard errors, clustered at year-group level are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Balanced sample.

Table B.3: Gender gap in the CA and final test grades across intervals - balanced sample

	CA Grade (low pressure)			Final Test Grade (high pressure)		
	(1)	(2)	(3)	(4)	(5)	(6)
1st Interval						
Female	-0.011	-0.012	-0.011	-0.117**	-0.115**	-0.115**
	(0.0227)	(0.0228)	(0.0227)	(0.0416)	(0.0418)	(0.0418)
$\tau = 0.10$	-0.0644	-0.0856	-0.0875	-0.0685	-0.0704	-0.0662
	(0.0875)	(0.0875)	(0.0801)	(0.0962)	(0.0947)	(0.0987)
$\tau = 0.25$	0.0131	-0.0119	-0.00369	-0.108*	-0.109	-0.138**
	(0.0572)	(0.0586)	(0.0593)	(0.0610)	(0.0664)	(0.0664)
$\tau = 0.50$	0.00547	-0.0131	-0.0113	-0.172***	-0.141**	-0.142**
	(0.0534)	(0.0516)	(0.0505)	(0.0592)	(0.0585)	(0.0623)
$\tau = 0.75$	0.00287	0.0104	0.0201	-0.0853	-0.0737	-0.0779
	(0.0504)	(0.0517)	(0.0534)	(0.0699)	(0.0676)	(0.0702)
$\tau = 0.90$	-0.0421	-0.0186	-0.0201	-0.204***	-0.172**	-0.186**
	(0.0608)	(0.0623)	(0.0645)	(0.0716)	(0.0687)	(0.0740)
N	1,305	1,305	1,305	1,305	1,305	1,305
$Adj.\, R^2_{OLS}$	0.400	0.406	.412	0.372	0.372	0.382
2nd Interval						
Female	0.0615	0.0640	0.0583	-0.0644*	-0.0695**	-0.0653*
	(0.0407)	(0.0409)	(0.0416)	(0.0320)	(0.0323)	(0.0323)
$\tau = 0.10$	0.0460	0.0102	0.0524	-0.00434	-0.0430	-0.0422
	(0.0565)	(0.0505)	(0.0575)	(0.0681)	(0.0689)	(0.0708)
$\tau = 0.25$	0.0850*	0.0846*	0.104**	-0.00245	-0.00759	0.0120
	(0.0502)	(0.0509)	(0.0511)	(0.0459)	(0.0486)	(0.0488)
$\tau = 0.50$	0.0973**	0.108**	0.0791	-0.112**	-0.0878*	-0.108**
	(0.0478)	(0.0452)	(0.0493)	(0.0476)	(0.0462)	(0.0498)
$\tau = 0.75$	0.0560	0.0667	0.0390	-0.0908*	-0.134***	-0.137***
	(0.0447)	(0.0429)	(0.0427)	(0.0523)	(0.0485)	(0.0515)
$\tau = 0.90$	-0.00717	-0.00876	-0.0362	-0.132***	-0.151***	-0.124**
	(0.0467)	(0.0469)	(0.0505)	(0.0499)	(0.0504)	(0.0539)
N	2,417	2,417	2,417	2,417	2,417	2,417
$Adj.\, R^2_{OLS}$	0.299	0.301	0.308	0.302	0.304	0.315
Individual var.	Yes	Yes	Yes	Yes	Yes	Yes
Group var.	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

Notes: The dependent variable measures the CA grade in columns (1) to (3) and the final test grade in 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 female student) for the OLS and QR. Standard errors, clustered at year-group level, for the OLS and bootstrapped standard errors with 1,000 replications for the QR. Standard errors are in parentheses and *** denotes significance at the 1% level, ** the 5% level and * the 10% level. Individual variables: age, nationality, university access grade, number of courses enrolled on in the first term, average grade obtained that first term, and status as scholarship holder and repeat student. Estimations in columns (4) to (6) include the CA grade (not standardised) as an explanatory variable. 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.

Table B.4: Gender gap - Main setting and heterogeneity analysis

	CA Grade (low pressure)		Final Test Grade (high pressure)	
	1st Int. (high)	2nd Int. (low)	1st Int. (high)	2nd Int. (low)
	(1)	(2)	(3)	(4)
Gender Gap (Pooled cross-section)	-0.0113	0.0583	-0.115	-0.0653
Gender Gap (Two-period panel data)	0.0017	0.0650	-0.1320	-0.0687

Notes: Coefficients from the pooled cross-section taken from Table B.3, column (3) and (6), OLS estimates. Coefficients for the two-period panel data taken from Table 9, the Gender Gap row.

2013

2013/1, Sánchez-Vidal, M.; González-Val, R.; Viladecans-Marsal, E.: "Sequential city growth in the US: does age matter?"
2013/2, Hortas Rico, M.: "Sprawl, blight and the role of urban containment policies. Evidence from US cities"
2013/3, Lampón, J.F.; Cabanelas-Lorenzo, P-; Lago-Peñas, S.: "Why firms relocate their production overseas? The answer lies inside: corporate, logistic and technological determinants"
2013/4, Montolio, D.; Planells, S.: "Does tourism boost criminal activity? Evidence from a top touristic country"
2013/5, Garcia-López, M.A.; Holl, A.; Viladecans-Marsal, E.: "Suburbanization and highways: when the Romans, the Bourbons and the first cars still shape Spanish cities"
2013/6, Bosch, N.; Espasa, M.; Montolio, D.: "Should large Spanish municipalities be financially compensated? Costs and benefits of being a capital/central municipality"
2013/7, Escardíbul, J.O.; Mora, T.: "Teacher gender and student performance in mathematics. Evidence from Catalonia"
2013/8, Arqué-Castells, P.; Viladecans-Marsal, E.: "Banking towards development: evidence from the Spanish banking expansion plan"
2013/9, Asensio, J.; Gómez-Lobo, A.; Matas, A.: "How effective are policies to reduce gasoline consumption? Evaluating a quasi-natural experiment in Spain"
2013/10, Jofre-Monseny, J.: "The effects of unemployment benefits on migration in lagging regions"
2013/11, Segarra, A.; García-Quevedo, J.; Teruel, M.: "Financial constraints and the failure of innovation projects"
2013/12, Jerrim, J.; Choi, A.: "The mathematics skills of school children: How does England compare to the high performing East Asian jurisdictions?"
2013/13, González-Val, R.; Tirado-Fabregat, D.A.; Viladecans-Marsal, E.: "Market potential and city growth: Spain 1860-1960"
2013/14, Lundqvist, H.: "Is it worth it? On the returns to holding political office"
2013/15, Ahlfeldt, G.M.; Maennig, W.: "Homevoters vs. leasevoters: a spatial analysis of airport effects"
2013/16, Lampón, J.F.; Lago-Peñas, S.: "Factors behind international relocation and changes in production geography in the European automobile components industry"
2013/17, Guío, J.M.; Choi, A.: "Evolution of the school failure risk during the 2000 decade in Spain: analysis of Pisa results with a two-level logistic mode"
2013/18, Dahlby, B.; Rodden, J.: "A political economy model of the vertical fiscal gap and vertical fiscal imbalances in a federation"
2013/19, Acacia, F.; Cubel, M.: "Strategic voting and happiness"
2013/20, Hellerstein, J.K.; Kutzbach, M.J.; Neumark, D.: "Do labor market networks have an important spatial dimension?"
2013/21, Pellegrino, G.; Savona, M.: "Is money all? Financing versus knowledge and demand constraints to innovation"
2013/22, Lin, J.: "Regional resilience"
2013/23, Costa-Campi, M.T.; Duch-Brown, N.; García-Quevedo, J.: "R&D drivers and obstacles to innovation in the energy industry"
2013/24, Huisman, R.; Stradnic, V.; Westgaard, S.: "Renewable energy and electricity prices: indirect empirical evidence from hydro power"
2013/25, Dargaud, E.; Mantovani, A.; Reggiani, C.: "The fight against cartels: a transatlantic perspective"
2013/26, Lambertini, L.; Mantovani, A.: "Feedback equilibria in a dynamic renewable resource oligopoly: pre-emption, voracity and exhaustion"
2013/27, Feld, L.P.; Kalb, A.; Moessinger, M.D.; Osterloh, S.: "Sovereign bond market reactions to fiscal rules and no-bailout clauses – the Swiss experience"
2013/28, Hilber, C.A.L.; Vermeulen, W.: "The impact of supply constraints on house prices in England"
2013/29, Revelli, F.: "Tax limits and local democracy"
2013/30, Wang, R.; Wang, W.: "Dress-up contest: a dark side of fiscal decentralization"
2013/31, Dargaud, E.; Mantovani, A.; Reggiani, C.: "The fight against cartels: a transatlantic perspective"
2013/32, Saarimaa, T.; Tukiainen, J.: "Local representation and strategic voting: evidence from electoral boundary reforms"
2013/33, Agasisti, T.; Murtinu, S.: "Are we wasting public money? No! The effects of grants on Italian university students' performances"
2013/34, Flacher, D.; Harari-Kermadec, H.; Moulin, L.: "Financing higher education: a contributory scheme"
2013/35, Carozzi, F.; Repetto, L.: "Sending the pork home: birth town bias in transfers to Italian municipalities"
2013/36, Coad, A.; Frankish, J.S.; Roberts, R.G.; Storey, D.J.: "New venture survival and growth: Does the fog lift?"
2013/37, Giulietti, M.; Grossi, L.; Waterson, M.: "Revenues from storage in a competitive electricity market: Empirical evidence from Great Britain"

2014

2014/1, Montolio, D.; Planells-Struse, S.: "When police patrols matter. The effect of police proximity on citizens' crime risk perception"
2014/2, Garcia-López, M.A.; Solé-Ollé, A.; Viladecans-Marsal, E.: "Do land use policies follow road construction?"
2014/3, Piolatto, A.; Rablen, M.D.: "Prospect theory and tax evasion: a reconsideration of the Yitzhaki puzzle"
2014/4, Cuberes, D.; González-Val, R.: "The effect of the Spanish Reconquest on Iberian Cities"
2014/5, Durán-Cabré, J.M.; Esteller-Moré, E.: "Tax professionals' view of the Spanish tax system: efficiency, equity and tax planning"
2014/6, Cubel, M.; Sanchez-Pages, S.: "Difference-form group contests"
2014/7, Del Rey, E.; Racionero, M.: "Choosing the type of income-contingent loan: risk-sharing versus risk-pooling"
2014/8, Torregrosa Hetland, S.: "A fiscal revolution? Progressivity in the Spanish tax system, 1960-1990"
2014/9, Piolatto, A.: "Itemised deductions: a device to reduce tax evasion"
2014/10, Costa, M.T.; García-Quevedo, J.; Segarra, A.: "Energy efficiency determinants: an empirical analysis of Spanish innovative firms"
2014/11, García-Quevedo, J.; Pellegrino, G.; Savona, M.: "Reviving demand-pull perspectives: the effect of demand uncertainty and stagnancy on R&D strategy"
2014/12, Calero, J.; Escardíbul, J.O.: "Barriers to non-formal professional training in Spain in periods of economic growth and crisis. An analysis with special attention to the effect of the previous human capital of workers"
2014/13, Cubel, M.; Sanchez-Pages, S.: "Gender differences and stereotypes in the beauty"
2014/14, Piolatto, A.; Schuett, F.: "Media competition and electoral politics"
2014/15, Montolio, D.; Trillas, F.; Trujillo-Baute, E.: "Regulatory environment and firm performance in EU telecommunications services"
2014/16, Lopez-Rodriguez, J.; Martinez, D.: "Beyond the R&D effects on innovation: the contribution of non-R&D activities to TFP growth in the EU"
2014/17, González-Val, R.: "Cross-sectional growth in US cities from 1990 to 2000"
2014/18, Vona, F.; Nicolli, F.: "Energy market liberalization and renewable energy policies in OECD countries"
2014/19, Curto-Grau, M.: "Voters' responsiveness to public employment policies"
2014/20, Duro, J.A.; Teixidó-Figueras, J.; Padilla, E.: "The causal factors of international inequality in co2 emissions per capita: a regression-based inequality decomposition analysis"
2014/21, Fleten, S.E.; Huisman, R.; Kilic, M.; Pennings, E.; Westgaard, S.: "Electricity futures prices: time varying sensitivity to fundamentals"
2014/22, Afcha, S.; García-Quevedo, J,: "The impact of R&D subsidies on R&D employment composition"
2014/23, Mir-Artigues, P.; del Río, P.: "Combining tariffs, investment subsidies and soft loans in a renewable electricity deployment policy"
2014/24, Romero-Jordán, D.; del Río, P.; Peñasco, C.: "Household electricity demand in Spanish regions. Public policy implications"
2014/25, Salinas, P.: "The effect of decentralization on educational outcomes: real autonomy matters!"
2014/26, Solé-Ollé, A.; Sorribas-Navarro, P.: "Does corruption erode trust in government? Evidence from a recent surge of local scandals in Spain"
2014/27, Costas-Pérez, E.: "Political corruption and voter turnout: mobilization or disaffection?"
2014/28, Cubel, M.; Nuevo-Chiquero, A.; Sanchez-Pages, S.; Vidal-Fernandez, M.: "Do personality traits affect productivity? Evidence from the LAB"
2014/29, Teresa Costa, M.T.; Trujillo-Baute, E.: "Retail price effects of feed-in tariff regulation"
2014/30, Kilic, M.; Trujillo-Baute, E.: "The stabilizing effect of hydro reservoir levels on intraday power prices under wind forecast errors"
2014/31, Costa-Campi, M.T.; Duch-Brown, N.: "The diffusion of patented oil and gas technology with environmental uses: a forward patent citation analysis"
2014/32, Ramos, R.; Sanromá, E.; Simón, H.: "Public-private sector wage differentials by type of contract: evidence from Spain"
2014/33, Backus, P.; Esteller-Moré, A.: "Is income redistribution a form of insurance, a public good or both?"
2014/34, Huisman, R.; Trujillo-Baute, E.: "Costs of power supply flexibility: the indirect impact of a Spanish policy change"
2014/35, Jerrim, J.; Choi, A.; Simancas Rodríguez, R.: "Two-sample two-stage least squares (TSTSLS) estimates of earnings mobility: how consistent are they?"
2014/36, Mantovani, A.; Tarola, O.; Vergari, C.: "Hedonic quality, social norms, and environmental campaigns"
2014/37, Ferraresi, M.; Galmarini, U.; Rizzo, L.: "Local infrastructures and externalities: Does the size matter?"
2014/38, Ferraresi, M.; Rizzo, L.; Zanardi, A.: "Policy outcomes of single and double-ballot elections"

2015

2015/1, Foremny, D.; Freier, R.; Moessinger, M-D.; Yeter, M.: "Overlapping political budget cycles in the legislative and the executive"
2015/2, Colombo, L.; Galmarini, U.: "Optimality and distortionary lobbying: regulating tobacco consumption"
2015/3, Pellegrino, G.: "Barriers to innovation: Can firm age help lower them?"
2015/4, Hémet, C.: "Diversity and employment prospects: neighbors matter!"
2015/5, Cubel, M.; Sanchez-Pages, S.: "An axiomatization of difference-form contest success functions"
2015/6, Choi, A.; Jerrim, J.: "The use (and misuse) of Pisa in guiding policy reform: the case of Spain"
2015/7, Durán-Cabré, J.M.; Esteller-Moré, A.; Salvadori, L.: "Empirical evidence on tax cooperation between sub-central administrations"
2015/8, Batalla-Bejerano, J.; Trujillo-Baute, E.: "Analysing the sensitivity of electricity system operational costs to deviations in supply and demand"
2015/9, Salvadori, L.: "Does tax enforcement counteract the negative effects of terrorism? A case study of the Basque Country"
2015/10, Montolio, D.; Planells-Struse, S.: "How time shapes crime: the temporal impacts of football matches on crime"
2015/11, Piolatto, A.: "Online booking and information: competition and welfare consequences of review aggregators"
2015/12, Boffa, F.; Pingali, V.; Sala, F.: "Strategic investment in merchant transmission: the impact of capacity utilization rules"
2015/13, Slemrod, J.: "Tax administration and tax systems"
2015/14, Arqué-Castells, P.; Cartaxo, R.M.; García-Quevedo, J.; Mira Godinho, M.: "How inventor royalty shares affect patenting and income in Portugal and Spain"
2015/15, Montolio, D.; Planells-Struse, S.: "Measuring the negative externalities of a private leisure activity: hooligans and pickpockets around the stadium"
2015/16, Batalla-Bejerano, J.; Costa-Campi, M.T.; Trujillo-Baute, E.: "Unexpected consequences of liberalisation: metering, losses, load profiles and cost settlement in Spain's electricity system"
2015/17, Batalla-Bejerano, J.; Trujillo-Baute, E.: "Impacts of intermittent renewable generation on electricity system costs"
2015/18, Costa-Campi, M.T.; Paniagua, J.; Trujillo-Baute, E.: "Are energy market integrations a green light for FDI?"
2015/19, Jofre-Monseny, J.; Sánchez-Vidal, M.; Viladecans-Marsal, E.: "Big plant closures and agglomeration economies"
2015/20, Garcia-López, M.A.; Hémet, C.; Viladecans-Marsal, E.: "How does transportation shape intrametropolitan growth? An answer from the regional express rail"
2015/21, Esteller-Moré, A.; Galmarini, U.; Rizzo, L.: "Fiscal equalization under political pressures"
2015/22, Escardíbul, J.O.; Afcha, S.: "Determinants of doctorate holders' job satisfaction. An analysis by employment sector and type of satisfaction in Spain"
2015/23, Aidt, T.; Asatryan, Z.; Badalyan, L.; Heinemann, F.: "Vote buying or (political) business (cycles) as usual?"
2015/24, Albæk, K.: "A test of the 'lose it or use it' hypothesis in labour markets around the world"
2015/25, Angelucci, C.; Russo, A.: "Petty corruption and citizen feedback"
2015/26, Moriconi, S.; Picard, P.M.; Zanaj, S.: "Commodity taxation and regulatory competition"
2015/27, Brekke, K.R.; Garcia Pires, A.J.; Schindler, D.; Schjelderup, G.: "Capital taxation and imperfect competition: ACE vs. CBIT"
2015/28, Redonda, A.: "Market structure, the functional form of demand and the sensitivity of the vertical reaction function"
2015/29, Ramos, R.; Sanromá, E.; Simón, H.: "An analysis of wage differentials between full-and part-time workers in Spain"
2015/30, Garcia-López, M.A.; Pasidis, I.; Viladecans-Marsal, E.: "Express delivery to the suburbs the effects of transportation in Europe's heterogeneous cities"
2015/31, Torregrosa, S.: "Bypassing progressive taxation: fraud and base erosion in the Spanish income tax (1970-2001)"
2015/32, Choi, H.; Choi, A.: "When one door closes: the impact of the hagwon curfew on the consumption of private tutoring in the republic of Korea"
2015/33, Escardíbul, J.O.; Helmy, N.: "Decentralisation and school autonomy impact on the quality of education: the case of two MENA countries"
2015/34, González-Val, R.; Marcén, M.: "Divorce and the business cycle: a cross-country analysis"

2015/35, Calero, J.; Choi, A.: "The distribution of skills among the European adult population and unemployment: a comparative approach"
2015/36, Mediavilla, M.; Zancajo, A.: "Is there real freedom of school choice? An analysis from Chile"
2015/37, Daniele, G.: "Strike one to educate one hundred: organized crime, political selection and politicians' ability"
2015/38, González-Val, R.; Marcén, M.: "Regional unemployment, marriage, and divorce"
2015/39, Foremny, D.; Jofre-Monseny, J.; Solé-Ollé, A.: "'Hold that ghost': using notches to identify manipulation of population-based grants"
2015/40, Mancebón, M.J.; Ximénez-de-Embún, D.P.; Mediavilla, M.; Gómez-Sancho, J.M.: "Does educational management model matter? New evidence for Spain by a quasiexperimental approach"
2015/41, Daniele, G.; Geys, B.: "Exposing politicians' ties to criminal organizations: the effects of local government dissolutions on electoral outcomes in Southern Italian municipalities"
2015/42, Ooghe, E.: "Wage policies, employment, and redistributive efficiency"

2016

2016/1, Galletta, S.: "Law enforcement, municipal budgets and spillover effects: evidence from a quasi-experiment in Italy"
2016/2, Flatley, L.; Giulietti, M.; Grossi, L.; Trujillo-Baute, E.; Waterson, M.: "Analysing the potential economic value of energy storage"
2016/3, Calero, J.; Murillo Huertas, I.P.; Raymond Bara, J.L.: "Education, age and skills: an analysis using the PIAAC survey"
2016/4, Costa-Campi, M.T.; Daví-Arderius, D.; Trujillo-Baute, E.: "The economic impact of electricity losses"
2016/5, Falck, O.; Heimisch, A.; Wiederhold, S.: "Returns to ICT skills"
2016/6, Halmenschlager, C.; Mantovani, A.: "On the private and social desirability of mixed bundling in complementary markets with cost savings"
2016/7, Choi, A.; Gil, M.; Mediavilla, M.; Valbuena, J.: "Double toil and trouble: grade retention and academic performance"
2016/8, González-Val, R.: "Historical urban growth in Europe (1300–1800)"
2016/9, Guio, J.; Choi, A.; Escardíbul, J.O.: "Labor markets, academic performance and the risk of school dropout: evidence for Spain"
2016/10, Bianchini, S.; Pellegrino, G.; Tamagni, F.: "Innovation strategies and firm growth"
2016/11, Jofre-Monseny, J.; Silva, J.I.; Vázquez-Grenno, J.: "Local labor market effects of public employment"
2016/12, Sanchez-Vidal, M.: "Small shops for sale! The effects of big-box openings on grocery stores"
2016/13, Costa-Campi, M.T.; García-Quevedo, J.; Martínez-Ros, E.: "What are the determinants of investment in environmental R&D?"
2016/14, **García-López, M.A; Hémet, C.; Viladecans-Marsal, E.:** "Next train to the polycentric city: The effect of railroads on subcenter formation"
2016/15, Matas, A.; Raymond, J.L.; Dominguez, A.: "Changes in fuel economy: An analysis of the Spanish car market"
2016/16, Leme, A.; Escardíbul, J.O.: "The effect of a specialized versus a general upper secondary school curriculum on students' performance and inequality. A difference-in-differences cross country comparison"
2016/17, **Scandurra, R.I.; Calero, J.**: "Modelling adult skills in OECD countries"
2016/18, **Fernández-Gutiérrez, M.; Calero, J.**: "Leisure and education: insights from a time-use analysis"
2016/19, Del Rio, P.; Mir-Artigues, P.; Trujillo-Baute, E.: "Analysing the impact of renewable energy regulation on retail electricity prices"
2016/20, Taltavull de la Paz, P.; Juárez, F.; Monllor, P.: "Fuel Poverty: Evidence from housing perspective"
2016/21, Ferraresi, M.; Galmarini, U.; Rizzo, L.; Zanardi, A.: "Switch towards tax centralization in Italy: A wake up for the local political budget cycle"
2016/22, Ferraresi, M.; Migali, G.; Nordi, F.; Rizzo, L.: "Spatial interaction in local expenditures among Italian municipalities: evidence from Italy 2001-2011"
2016/23, Daví-Arderius, D.; Sanin, M.E.; Trujillo-Baute, E.: "CO2 content of electricity losses"
2016/24, Arqué-Castells, P.; Viladecans-Marsal, E.: "Banking the unbanked: Evidence from the Spanish banking expansion plan"
2016/25 Choi, Á.; Gil, M.; Mediavilla, M.; Valbuena, J.: "The evolution of educational inequalities in Spain: Dynamic evidence from repeated cross-sections"
2016/26, Brutti, Z.: "Cities drifting apart: Heterogeneous outcomes of decentralizing public education"
2016/27, Backus, P.; Cubel, M.; Guid, M.; Sánchez-Pages, S.; Lopez Manas, E.: "Gender, competition and performance: evidence from real tournaments"
2016/28, Costa-Campi, M.T.; Duch-Brown, N.; García-Quevedo, J.: "Innovation strategies of energy firms"
2016/29, Daniele, G.; Dipoppa, G.: "Mafia, elections and violence against politicians"

2016/30, Di Cosmo, V.; Malaguzzi Valeri, L.: "Wind, storage, interconnection and the cost of electricity"

2017

2017/1, González Pampillón, N.; Jofre-Monseny, J.; Viladecans-Marsal, E.: "Can urban renewal policies reverse neighborhood ethnic dynamics?"
2017/2, Gómez San Román, T.: "Integration of DERs on power systems: challenges and opportunities"
2017/3, Bianchini, S.; Pellegrino, G.: "Innovation persistence and employment dynamics"
2017/4, Curto-Grau, M.; Solé-Ollé, A.; Sorribas-Navarro, P.: "Does electoral competition curb party favoritism?"
2017/5, Solé-Ollé, A.; Viladecans-Marsal, E.: "Housing booms and busts and local fiscal policy"
2017/6, Esteller, A.; Piolatto, A.; Rablen, M.D.: "Taxing high-income earners: Tax avoidance and mobility"
2017/7, Combes, P.P.; Duranton, G.; Gobillon, L.: "The production function for housing: Evidence from France"
2017/8, Nepal, R.; Cram, L.; Jamasb, T.; Sen, A.: "Small systems, big targets: power sector reforms and renewable energy development in small electricity systems"
2017/9, Carozzi, F.; Repetto, L.: "Distributive politics inside the city? The political economy of Spain's plan E"
2017/10, Neisser, C.: "The elasticity of taxable income: A meta-regression analysis"
2017/11, Baker, E.; Bosetti, V.; Salo, A.: "Finding common ground when experts disagree: robust portfolio decision analysis"
2017/12, Murillo, I.P; Raymond, J.L; Calero, J.: "Efficiency in the transformation of schooling into competences: A cross-country analysis using PIAAC data"
2017/13, **Ferrer-Esteban, G.; Mediavilla, M.:** "The more educated, the more engaged? An analysis of social capital and education"
2017/14, Sanchis-Guarner, R.: "Decomposing the impact of immigration on house prices"
2017/15, Schwab, T.; Todtenhaupt, M.: "Spillover from the haven: Cross-border externalities of patent box regimes within multinational firms"
2017/16, Chacón, M.; Jensen, J.: "The institutional determinants of Southern secession"
2017/17, Gancia, G.; Ponzetto, G.A.M.; Ventura, J.: "Globalization and political structure"
2017/18, González-Val, R.: "City size distribution and space"
2017/19, García-Quevedo, J.; Mas-Verdú, F.; Pellegrino, G.: "What firms don't know can hurt them: Overcoming a lack of information on technology"
2017/20, Costa-Campi, M.T.; García-Quevedo, J.: "Why do manufacturing industries invest in energy R&D?"
2017/21, Costa-Campi, M.T.; García-Quevedo, J.; Trujillo-Baute, E.: "Electricity regulation and economic growth"

2018

2018/1, Boadway, R.; Pestieau, P.: "The tenuous case for an annual wealth tax"
2018/2, Garcia-López, M.À.: "All roads lead to Rome ... and to sprawl? Evidence from European cities"
2018/3, Daniele, G.; Galletta, S.; Geys, B.: "Abandon ship? Party brands and politicians' responses to a political scandal"
2018/4, Cavalcanti, F.; Daniele, G.; Galletta, S.: "Popularity shocks and political selection"
2018/5, Naval, J.; Silva, J. I.; Vázquez-Grenno, J.: "Employment effects of on-the-job human capital acquisition"
2018/6, Agrawal, D. R.; Foremny, D.: "Relocation of the rich: migration in response to top tax rate changes from spanish reforms"
2018/7, García-Quevedo, J.; Kesidou, E.; Martínez-Ros, E.: "Inter-industry differences in organisational eco-innovation: a panel data study"
2018/8, Aastveit, K. A.; Anundsen, A. K.: "Asymmetric effects of monetary policy in regional housing markets"
2018/9, Curci, F.; Masera, F.: "Flight from urban blight: lead poisoning, crime and suburbanization"
2018/10, Grossi, L.; Nan, F.: "The influence of renewables on electricity price forecasting: a robust approach"
2018/11, Fleckinger, P.; Glachant, M.; Tamokoué Kamga, P.-H.: "Energy performance certificates and investments in building energy efficiency: a theoretical analysis"
2018/12, van den Bergh, J. C.J.M.; Angelsen, A.; Baranzini, A.; Botzen, W.J. W.; Carattini, S.; Drews, S.; Dunlop, T.; Galbraith, E.; Gsottbauer, E.; Howarth, R. B.; Padilla, E.; Roca, J.; Schmidt, R.: "Parallel tracks towards a global treaty on carbon pricing"
2018/13, Ayllón, S.; Nollenberger, N.: "The unequal opportunity for skills acquisition during the Great Recession in Europe"
2018/14, Firmino, J.: "Class composition effects and school welfare: evidence from Portugal using panel data"
2018/15, Durán-Cabré, J. M.; Esteller-Moré, A.; Mas-Montserrat, M.; Salvadori, L.: "La brecha fiscal: estudio y aplicación a los impuestos sobre la riqueza"

2018/16, Montolio, D.; Tur-Prats, A.: "Long-lasting social capital and its impact on economic development: the legacy of the commons"
2018/17, Garcia-López, M. À.; Moreno-Monroy, A. I.: "Income segregation in monocentric and polycentric cities: does urban form really matter?"
2018/18, Di Cosmo, V.; Trujillo-Baute, E.: "From forward to spot prices: producers, retailers and loss averse consumers in electricity markets"
2018/19, Brachowicz Quintanilla, N.; Vall Castelló, J.: "Is changing the minimum legal drinking age an effective policy tool?"
2018/20, Nerea Gómez-Fernández, Mauro Mediavilla: "Do information and communication technologies (ICT) improve educational outcomes? Evidence for Spain in PISA 2015"

IEB
Institut
d'Economia
de Barcelona

UNIVERSITAT DE
BARCELONA

ieb@ub.edu
www.ieb.edu

Public Policies

IEB Working Paper