Gender Gaps in different Grading Systems

Catarina Angelo[ab] Ana Balcão Reis[b]

Nova School of Business and Economics, Universidade Nova de Lisboa

November 2020

Abstract

This paper analyzes the impact of grading practices on the gender gap in student achievement. We examine the gender difference in the difference between teacher grading and scores on national exams to test whether there are gender differences associated with different grading systems. We use data on 21 subjects across humanities and sciences for students taking exams between the 4th and 12th grades from 2007 to 2018 in Portugal. Our results indicate that a grading system based on exams favors boys while one based on teacher assessment favors girls.

JEL Classification: I21, I24, J16.

Keywords: Student Achievement, Grading Practices, Gender Gap

[a]The author has benefited from the financial support of *FCT-Fundação para a Ciência e Tecnologia*, reference SFRH/BD/70215/2010, which is gratefully acknowledged.

[b]We acknowledge financial support from National Funds through Fundação para a Ciência e Tecnologia under the projects Ref.UID/ECO/00124/2013 and Ref.PTDC/IIM-ECO/6813/2014 and by POR Lisboa under the project Ref.LISBOA-01-0145-FEDER-007722.

1 Introduction

In recent years the gender gap in student achievement has been steadily closing in several developed countries, with girls catching up in almost every field and even surpassing boys in some fields.[1] Throughout the world girls generally outscore boys in fields that require more reading skills, whereas in mathematics the results are more mixed (Machin and Pekkarinen, 2008, Hyde et al., 2008, Guiso et al., 2008).

While the numbers are indisputable, the reasons behind the reversal of the gender gap in academic achievement are far from being settled. In this paper we look at the role of the evaluation system on the measured gender gap in student achievement. More specifically, we look at the differences between teacher assessment and evaluation through national exams.

The literature has discussed noncognitive skills as one potential source of the observed gender differences in achievement. In the words of Cornwell et al.(2013), who looked at primary schools in the US, "Boys who perform equally as well as girls on reading, math and science tests are graded less favorably by their teachers, but this less favorable treatment essentially vanishes when noncognitive skills are taken into account" (p.236). Different grading systems reward non-cognitive skills differently, implying that the choice of the grading system and its application at different levels of education may influence the gender gap in student achievement.

In general, educational systems rely heavily on teacher grading to determine student final scores. Teachers take into account several dimensions of student performance that are likely to be related to a set of non-cognitive skills (such as emotional maturity, empathy, and interpersonal skills)

[1] See for example the OECD (2012) PISA report for gender differences in mathematics, reading, and science.

that can influence students' behavior.[2] Girls may be more amenable to the learning process than boys, which makes teachers more prone to reward them with a higher score (Cornwell et al., 2013). But it may also be the case that higher scores obtained by girls is related to teachers' gender bias in evaluating students (Lavy (2008), Hinnerich et al. (2011), Hanna and Linden (2012), Carlana (2019)).

Compared with teacher grading, national exams are likely to be a more objective instrument to establish comparisons between students. Besides being equal for all students they are usually graded in a blind way, and according to grading criteria that are nationally defined. On the other hand, compared with written tests taken in the classroom exams are likely to have a higher associated stress component, especially when the stakes are high, which given the differential response of boys and girls to stressful situations and to low vs. high stakes testing may contribute to a relatively worse performance of girls (Falch and Naper (2013), Ors et al. (2013), Jurajda and Münich (2011), Cai et al. (2018), Azmat et al. (2016)).

This paper examines the gender difference in the difference between teacher grading and national exam scores to test whether there are gender differences associated with the different grading systems and whether they change across the academic path. We use teacher grading and national exam scores for several cohorts of Portuguese students, in the 4th, 6th, 9th, 11th and 12th grades.

Our work contributes to the existing literature on how grading practices may affect boys' and girls' scores differently, in two ways. The first is related to the subjects considered. Unlike most existing studies that focus on language and mathematics, we are able to examine scores for 21 subjects across humanities and sciences, using information about all students in the academic

[2]For a deeper understanding regarding the relevance of personality traits for both academic and professional success see, for example, Almlund et al. (2011).

track of secondary education who take at least one national exam. Second, we have data for 12 repeated cross sections for students from grade 9 to grade 12.

Our approach is to compare the overall impact of the two evaluation schemes on gender differences in student achievement. We identify a positive assessment gap in all the grades that is systematically larger for girls. The results obtained indicate a gap in assessment that is greater for girls in the large majority of the subjects under analysis, in both humanities and sciences, which suggests that a grading system based on exams favors boys, while one based on classroom evaluation favors girls.

The remainder of this paper is organized as follows. A review of the literature is conducted in the next section. Section 3 gives an overview of the Portuguese educational system. Section 4 presents the data used and a first analysis of the gender differences in the assessment gap, the latter defined as the difference between teacher grading scores and national exams scores. Section 5 presents the regression analysis and the main results, and the final section concludes.

2 Literature Review

This paper contributes to the literature on how different grading practices may affect boys' and girls' differently. Thus, it is related to the literature that examines the impact on the gender gap of different dimensions of the evaluation schemes as blind vs non-blind evaluations, higher vs lower stakes tests and teacher gender bias. We review this literature with more emphasis on those dimensions that may contribute to the overall difference between teacher grading and national exam scores in the context of the Portuguese educational system.

Some previous studies compare blind and non-blind scores of similar tests to examine teachers' gender bias in evaluating students' tests (Lavy, 2008, Hinnerich et al., 2011). Lavy (2008) uses data on blind and non-blind scores from matriculation exams of Israeli students for several cohorts of 12th graders in 9 subjects across humanities and sciences. Both scores result from single tests, taken one to three weeks apart from each other. Both tests follow exactly the same format and contain questions drawn from the same question bank. The main difference between the two scores comes from the fact that one is graded blindly by external examiners and the other is graded by school teachers. Under this setting Lavy (2008) finds evidence that strongly suggests a bias against boys in each subject. Hinnerich et al. (2011) also test for discrimination against boys, using data from Swedish high schools. They compare blind and non-blind scores of exactly the same Swedish Language test, for a random sample of about 1700 9th graders, and find no evidence of gender bias against boys in grading. It is worth emphasizing the fact that Hinnerich's study looks only at Swedish Language, and therefore does not allow ruling out discrimination in the remaining subjects.

Hanna and Linden (2012), conducting a field experiment in India, obtain no teachers' gender bias; however, their results indicate a statistical discrimination against lower-caste students. Carlana (2019), measuring teacher stereotypes by the Gender-Science Implicit Association Test applied in a sample of Italian schools, also obtains a positive impact of these stereotypes on the gender gap (in favor of boys) on math performance, but not on literature.[3] Overall, taking into account the different results presented in the literature coming from different educational systems, teachers' gender bias is certainly one channel that may contribute to the gender gap

[3] Going one step further, Carlana (2019) shows that this teacher gender bias has a negative impact on girls' subsequent achievement. Terrier (2020) and Lavy and Sand (2018) have also measured the impact of teachers' gender bias on students' subsequent achievement.

on the difference between teacher evaluation and exam scores. However, in most cases the two evaluations being compared are not exactly alike. So, other factors may contribute to explain the potential gender gap.

Other studies have looked at gender differences in response to stressful situations, as is the case of high-stakes exams. Falch and Naper (2013), using data on Norwegian student scores in the 10^{th} grade, study how boys and girls react to a change on test stakes. These authors use variation in exam stakes in different counties and also compare exam results to a low-stakes national test. On both cases they find evidence that female students have a relatively lower score at the lower stakes evaluation which goes against the hypothesis that girls obtain relatively worse results under stressful conditions. However, other studies have obtained different results. Ors et al. (2013) use data from students at HEC Paris, a top European business school, and compare male and female results in exams with different stakes. They obtain that males do relatively better at the high stakes entrance exam, but comparing the same group of students in other lower stakes exams done both earlier in time (the baccalaureat) and later in time (first year exams at HEC Paris), the advantage of male students is reversed or disappears. Jurajda and Münich (2011), for applicants to Czech universities with a varying level of competition, and Cai et al. (2018), comparing scores at the highly competitive China's National College Entrance Exam and at a low-stakes mock examination held two months earlier, also obtain that female students do relatively worse under more stressful conditions. Azmat et al. (2016) use data for high-school students from a private school in Barcelona, Spain, and are able to compare the gender gap in tests with low, medium and high stakes for several cohorts. They are also able to conduct an experiment which allows isolating the impact of stakes on the gender gap for a group of around 200 students. They obtain strong results showing that "overperformance by female students compared with male students in school is reduced or even negated as the stakes are increased substantially." Thus, it seems that most studies conclude in favor of the existence of gender differences in response to stressful situations with girls performing relatively worse than boys under higher stakes.

Contributing to this discussion but introducing a different perspective, Pekkarinen (2015) studied gender differences in the strategies adopted when answering multiple choice university entrance exams. He shows that boys and girls tend to adopt different strategies and it seems that strategies adopted by girls under more competitive pressure contribute to their lower performance relative to their results in less competitive evaluations. Thus, multiple choice exams may contribute to a gender gap in favor of boys. There is also some research on the impact of grading schemes on gender gaps. For example, Czibor et al. (2014) compare relative grading with absolute or criterion-referenced grading, noting that relative grading creates a more competitive environment. These authors implement a randomized controlled trial in a second year course of a Bachelor program in a Dutch university to compare grading on a curve (relative grading) to absolute grading and obtain that men tend to score higher under relative grading, thus decreasing the gender gap in exam performance.[4]

Another mechanism that may contribute to the gender gap in the difference between teacher grading and exam scores is the possibility that boys and girls react differently to continuous vs one-shot evaluations. Machin and McNally (2005) argue that increasing the weight of coursework elements in students' evaluation, as happened in the UK in 1988, improves girls'

[4]The impact of testing format on gender gaps has been studied by other authors (e.g. Riener and Wagner (2017)). We don't explore further this literature nor the literature on the impact of the type of grading as in the Portuguese educational system in class tests and national exams tend to have similar formats, including both multiple choice and open-ended questions, and both teacher and national exam scores are criterion–based graded (see next section). Thus, these mechanisms don't help to explain the differences in the gender gap between teacher scores and national exam scores observed in this paper.

relative performance.

Closer to our paper, some studies have compared the overall impact of different grading schemes. Marcenaro-Gutierrez and Vignoles (2015) compare teacher and test-based assessments in reading and mathematics for two cohorts of Spanish students aged 11 and 15. They find evidence that in mathematics the difference between teacher assessment and test-based results is significantly higher for girls than for boys, implying that teacher evaluation benefits girls. In reading they obtain no significant difference between genders. Falch and Naper (2013), using data on Norwegian student scores in the 10th grade, and Lindahl (2007), using Swedish student data for 9th graders, obtain that an evaluation system that relies heavily on teacher grading lowers boys' scores relative to girls'.

3 Institutional Setting

Portugal has 12 years of compulsory education.[5] Children attend school from the year they turn 6 to the year they turn 18 and private and public schools coexist in all levels of education.[6] At the end of the 9th grade students can choose between two different tracks of secondary education, the academic track and the professional track. The academic track targets students who want to pursue a university degree. In the 2011/2012 school year, students enrolled in this track represented roughly 57% of the secondary students. Within this track there are four major areas of study: Sciences and Technology; Economics; Languages and Humanities; Visual Arts (see

[5]In 2009, a law was approved extending compulsory education to 12 years or until the student is 18, with implications for students finishing secondary education from 2016 on. This may be one of the contributing factors for the steadily decrease of the percentage of students out of school at the age of 17, from 22.2% in 2007/08 to 3.8% in 2017/18. Also, the percentage of those aged 18-24 with at most lower secondary education (9th grade) and who were not in further education and training decreased between 2002 and 2019, from 50% to 10.6%.

[6]The percentage of students in private schools is around 13% at all cycles of basic education and of around 20% in secondary education.

Figure 1).

We use data from 2007 to 2018. However, regarding the 4th and 6th grade, national exams mandatory to all students started in 2013 and 2012 respectively, and were only in place untill 2015. For the 9th grade, all students were tested by means of national exams over the whole period. Finally, and also for the whole period, all students that follow the academic track of secondary education take national exams at the end of the 11th and 12th grade.

Portuguese and mathematics are the only subjects tested at the end of the 4th, 6th and the 9th grade, whereas at the end of secondary education there are national exams in 21 different subjects. The Portuguese Language national exam is the only one that is mandatory for all students at the end of secondary education. Each secondary student must take exams in 3 other subjects that are specific to their field of study at the end of the 11th or the 12th grade. Students are tested in the 11th grade on 2 of the biennial subjects attended during the 10th and 11th grade. In the 12th grade, besides the mandatory exam in Portuguese Language, students take the exam that is mandatory for their field of studies.

Exams happen in June, one to two weeks after the end of classes and after the publication of teacher grades. They are curriculum based, include both multiple choice and open-ended questions and take between 90 to 150 minutes depending on the subject. The evaluation of the exam copies is based on the answer key and grading scheme provided by an agency of the Ministry of Education. Exams are graded anonymously by teachers from a school different than the one attended by the student. They are solely identified by an id-number and the student name is not written in the exam.[7] Grades are given based on absolute or criterion-referenced

[7] The id-number that identifies the exam is not known to the student.

Basic Education – Mandatory*									Secondary education		
1st cycle - Primary				2nd cycle		3rd cycle			Professional track (several programs) / Academic track (4 areas of study)		
1st grade 6 y/o	2nd grade 7 y/o	3rd grade 8 y/o	4th grade 9 y/o	5th grade 10 y/o	6th grade 11 y/o	7th grade 12 y/o	8th grade 13 y/o	9th grade 14 y/o	10th grade 15 y/o	11th grade 16 y/o	12th grade 17 y/o
			National exams only in the period 2013-2015		National exams only in the period 2012-2015			National exams		National exams (only mandatory for academic track)	National exams (only mandatory for academic track)

Figure 1: Portuguese Education System.

* After 2016, all students in the Portuguese educational system face 12 years of compulsory education, therefore including secondary education.

evaluation. There is no "curving" or relative grading. Exams from previous years and their answer keys are available online to teachers and students and are frequently used as studying materials.

Teacher grades are based on several coursework elements, that include in-class tests but also homework, oral presentations, class participation and student behavior. The weight of the different elements varies across schools and subjects. In exam years it is known that teachers prepare their students for the exams and usually written tests follow the exams' model. Although there is no regulation on the number of written tests, the most common is to have two tests per trimester leading to around six written tests per year. In the Portuguese educational system curving is not used thus, like in national exams, teacher grades are based on absolute criteria.

In the 4^{th}, 6^{th} and 9^{th} grades, both teacher grades and exam scores are given in a 1 to 5 scale, where 3 to 5 are passing grades. In secondary education, teacher grades are given in a 0 to 20 scale but exam scores are published in a 0 to 200 scale. We converted the 0 to 200 scale to the 0 to 20 scale following the rules determined by the Ministry of Education to calculate the students' final GPA: the score is divided by 10 and rounded to the nearest whole number. The 0 to 20 scale was converted to the 1 to 5 scale using the criteria defined by the Ministry of Education: level 1 corresponds to the interval of scores [0,4), level 2 to [4,10), level 3 to [10,14), level 4 to [14,18), and level 5 to [18,20]. It is usual that teachers classify in-class tests using the 0-200 scale, but the final grade is always published in the 0 to 20 scale, the minimum passing grade being 10.

A student can take the national exam in a subject as an internal or as an external student. To be considered an internal student it must be the case that he/she attended classes in that school for the subject on which he/she takes the national exam. Moreover, secondary school students must obtain a final score from teacher grading of at least 10 (on a 1 to 20 scale) to be able to take the exam as internal students. External students are are all of the remaining. For students who take exams as internal students, 30% of the final score is determined by his/her exam score, the remaining 70% is determined by teacher grading.[8] Higher education access is determined by a weighted average of high-school GPA and scores on national exams, where the weight of the latter varies between 40% and 50% of their application score, depending on the university and degree they are applying to.

4 Data and descriptive statistics

The data sets used are available online (on the Ministry of Education's website) and contain yearly information from 2007 to 2018 on the results obtained by students in the 9th, 11th, and 12th grade exams.[9] Data is available from 2012 to 2015 for 6th graders, and from 2013 to 2015 for 4th graders. For each student we know age, gender, field of study, the school of enrollment and its geographical location. Unhappily, the data set does not allow us to link the different exams of the same student. Nonetheless, for each exam we know the exam score, the score obtained from teacher grading, if the student is enrolled as an internal student and the student characteristics refered above.

As we want to compare exam scores with teacher grading, we consider only internal students. Thus, the data set excludes the secondary students in the left tail of the scores distribution. This

[8] Students can cancel their enrollment in a subject and take the national exams as external students, in which case their final score is determined entirelly by the exam score. Unfortunately, the data does not allow to identify the students that cancel their enrollment.

[9] Although the database has existed since 1998, information on students' gender is available only since 2007.

may represent a problem if selection into the sample depends on gender when considering only internal students. Analysis of the data shows that from the exams taken by boys enrolled in the academic track of secondary education, 69% are for internal students. The corresponding value for girls is 71%. Although the average difference is of 2 p.p., the yearly difference has been steadily decreasing since 2012, reaching a 0.6 p.p. value in 2016. Although it does not seem to be the case that selection into the sample depends on gender when considering only internal students, we conducted a robustness analysis using a Heckman selection model that confirms that our results are not affected by selection issues, as explained in the following section.

In the following analysis the focus is the difference between genders in the gap between both assessments. Thus, we define the gap between assessments as the difference in scores between teacher grading and exam results.

Table 1 has the means and estimated standard errors for teacher grading and national exams, corrected for clustering at the school level, for the subjects in which students take exams in the 4th, 6th and the 9th grades. The scores examined are those attributed by teachers, the ones obtained in the national exam, and the difference between the two for both boys and girls. For all grades and subjects, teacher scores are on average higher than exam results. The gap between assessments for each gender is in columns (8) and (9) of Table 1. In all grades the assessment gap is positive and significant for boys and girls in Portuguese and mathematics (columns 8 and 9 in Table 1). In column (10) the difference tested is column (9) minus column (8), girls' assessment gap minus boys' assessment gap. The T test is for the difference in mean differences between girls' and boys' scores, and the T statistic in square brackets reflects standard errors in parentheses that are corrected for clustering at the school level. The difference between girls' and boys' assessment gap is positive and significant in both subjects (column 10 in Table 1), suggesting that teacher grading favors girls. The gender difference in the assessment gap decreases from the 4th to the 9th grade by half in Portuguese and by one fourth in mathematics.

Table 1: Means and estimated standard errors of teachers' and national exams' scores in grade 4 (2013-2015), grade 6 (2012-2015), and grade 9 (2007-2018) - Mean comparison tests

	Number of		Mean Teacher Score		Mean Exam Score		Mean Difference		T test[a)]
Subject	Observations	% Girls	Boys	Girls	Boys	Girls	Boys	Girls	[(9)-(8)]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Grade 4									
Portuguese	307012	49	3.64	3.85	2.99	3.13	0.65	0.72	0.06***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[6.06]
Mathematics	310947	49	3.67	3.64	3.03	2.92	0.64	0.72	0.08***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[7.01]
Grade 6									
Portuguese	428657	48	3.22	3.48	2.92	3.13	0.30	0.35	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[3.83]
Mathematics	434343	49	3.18	3.26	2.75	2.76	0.43	0.50	0.07***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[3.74]
Grade 9									
Portuguese	1076636	52	3.13	3.39	2.90	3.13	0.22	0.25	0.03*
			(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[2.52]
Mathematics	1082359	52	3.04	3.12	2.70	2.71	0.34	0.40	0.06***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[3.53]

Notes. The grading scale is 1 to 5. The mean difference in columns (8) and (9) is defined as mean teacher score minus mean exam score.
[a)] In column (10) the difference tested is column (9) minus column (8), girls' assessment gap minus boys' assessment gap. The T test is for the difference in mean differences between girls' and boys' scores and the T statistic in square brackets reflects standard errors in parentheses that are corrected for clustering at the school level. The significance of the assessment gap was estimated using Stata's clttest command. The number of clusters is 4932 for Portuguese and 4931 for mathematics in the 4th grade, 1284 in the 6th grade and 1476 in the 9th grade. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001

The results are in line with existing literature on student performance, according to which girls clearly score better than boys in reading while their advantage in mathematics is not as large.

Table 2 replicates the analysis from Table 1 for the subjects in which students take exams in secondary education. Similarly to what is observed in the earlier grades students obtain on average higher scores from teacher assessment than on national exams.

Table 2: Means and estimated standard errors of teachers' and national exams' scores in secondary education (2007-2018) - Mean comparison tests

Subject	Number of Observations	% Girls	Mean Teacher Score		Mean Exam Score		Mean Difference		T test[a)]
			Boys	Girls	Boys	Girls	Boys	Girls	[(9)-(8)]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Humanities									
English	432	77	14.3	14.1	12.3	11.4	1.9	2.7	0.8
			(0.42)	(0.25)	(0.54)	(0.42)	(1.05)	(0.61)	[0.67]
French	11994	71	12.8	13.4	11.5	12.0	1.2	1.3	0.2
			(0.07)	(0.07)	(0.11)	(0.10)	(0.22)	(0.13)	[0.61]
Geography	220022	63	13.1	13.3	11.4	10.7	1.6	2.4	0.8***
			(0.04)	(0.03)	(0.04)	(0.04)	(0.08)	(0.06)	[8.44]
German	9281	72	13.5	14.5	10.9	12.0	2.5	2.4	-0.1
			(0.09)	(0.09)	(0.18)	(0.13)	(0.33)	(0.21)	[-0.31]
History	168323	71	12.6	13.0	10.4	10.3	2.0	2.6	0.5***
			(0.03)	(0.03)	(0.05)	(0.05)	(0.10)	(0.06)	[4.46]
History B	8552	42	14.1	14.6	11.8	11.8	2.2	2.7	0.5
			(0.11)	(0.12)	(0.18)	(0.19)	(0.22)	(0.25)	[1.63]
History of Culture and Arts	34798	68	12.7	13.5	9.6	10.3	3.0	3.0	0.1
			(0.05)	(0.05)	(0.08)	(0.08)	(0.15)	(0.10)	[0.43]
Latin	1844	74	13.7	13.9	10.6	10.7	2.9	3.0	0.2
			(0.26)	(0.15)	(0.28)	(0.22)	(0.49)	(0.28)	[0.29]
Philosophy	72910	65	13.3	14.1	9.8	10.7	3.3	3.2	-0.1
			(0.04)	(0.04)	(0.09)	(0.07)	(0.12)	(0.09)	[-0.85]
Portuguese	755696	59	12.8	13.7	10.2	11.1	2.4	2.4	-0.1
			(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	[-1.07]
Portuguese Literature	21491	75	12.4	13.3	9.9	11.1	2.3	2.1	-0.2
			(0.06)	(0.06)	(0.11)	(0.10)	(0.19)	(0.11)	[-0.92]
Spanish	20868	70	14.4	15.5	12.8	13.7	1.5	1.7	0.2
			(0.09)	(0.09)	(0.11)	(0.10)	(0.19)	(0.13)	[0.69]
Sciences									
Biology and Geology	489443	57	13.6	14.0	9.8	9.9	3.6	3.9	0.3***
			(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	[3.81]
Descriptive Geometry	87543	49	14.7	14.3	11.4	9.9	3.2	4.2	1.0***
			(0.06)	(0.06)	(0.14)	(0.14)	(0.18)	(0.19)	[3.95]
Information and Technology	4246	42	14.9	14.7	11.7	10.5	3.2	4.1	0.9**
			(0.11)	(0.13)	(0.17)	(0.15)	(0.22)	(0.27)	[2.68]
Mathematics	519969	53	13.1	13.5	10.6	10.9	2.3	2.4	0.1
			(0.05)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)	[0.70]
Mathematics B	21320	64	12.7	13.3	10.2	10.8	2.3	2.3	0.0
			(0.05)	(0.06)	(0.14)	(0.13)	(0.20)	(0.16)	[0.12]
Mathematics SS	105073	71	12.8	13.4	10.1	10.6	2.5	2.7	0.2
			(0.03)	(0.03)	(0.07)	(0.07)	(0.12)	(0.08)	[1.41]
Physics and Chemistry	519607	53	13.1	13.5	9.2	9.3	3.8	4.1	0.3***
			(0.05)	(0.06)	(0.07)	(0.08)	(0.07)	(0.07)	[3.40]
Other									
Drawing	54551	67	14.8	15.5	12.4	12.8	2.3	2.6	0.3*
			(0.05)	(0.05)	(0.05)	(0.06)	(0.11)	(0.08)	[2.21]
Economics	84410	49	13.9	14.4	11.6	11.8	2.1	2.4	0.3*
			(0.05)	(0.05)	(0.08)	(0.07)	(0.11)	(0.10)	[2.03]

Note. The grading scale is 0 to 20. The mean difference in columns (8) and (9) is the gender assessment gap, defined as mean teacher score minus mean exam score. [a)] In column (10) the difference tested is column (9) minus column (8), girls' assessment gap minus boys' assessment gap. The T test is for the difference in mean differences between girls' and boys' scores and the T statistic in square brackets reflects standard errors in parentheses that are corrected for clustering at the school level. The significance of the assessment gap was estimated using Stata's clttest command. The number of clusters corresponds to the number of schools in Table A4 in the Apendix. * significant at $p < 0.05$, ** significant at $p < 0.01$, *** significant at $p < 0.001$

According to the data, girls achieve on average higher scores than boys when they are assessed by their teachers. This is evident in all but one of the humanities subjects and in almost every subject in sciences. Only in English, Descriptive Geometry and Information and Technology do boys obtain a higher score relative to girls when they are assessed by their teachers.

When we analyze exam scores boys perform on average better than girls in 5 of the 21 subjects under analysis. In Humanities courses boys obtain on average higher scores than girls in English, Geography and History. In Sciences courses boys obtain a higher score in Descriptive Geometry and Information and Technology.

While the assessment gap is significant and positive in all subjects, the difference between gender (defined as the girls' gap minus the boys' gap) is significant only for some of them, and for more subjects in Sciences than in Humanities. Regarding the subjects also analyzed in basic education, Portuguese Language and mathematics, the gender difference in the assessment gap we observe in the 12th grade is no longer significantly different from zero. However, for all the cases where the gender difference on the assessment gap is significant, it is positive. That is, the assessment gap is greater for girls. This suggests that teacher grading favors girls and exams favor boys. It is worth noticing that, when significant, the size of the assessment gap gender difference goes from 0.3 to 1, which in a grading scale from 0 to 20 is quite non-trivial.

5 Gender gap in different grading systems

5.1 Regression analysis

In the previous section we have shown that, on average, the difference between teacher score and national exams is greater for girls. In this section we extend the previous analysis running a regression to test if the difference in scores depends on gender, while controlling for student characteristics. We follow Marcenaro-Gutierrez and Vignoles (2015) and assume a linear specification for the gap equation to estimate the following model.

$$G_{ijts} = \alpha + \delta F_i + \beta X_{ijt} + \mu_t + \tau_s + \epsilon_{ijt} \tag{1}$$

where G_{ijts} is the gap between assessments for student i in subject j and time t and school s. As we cannot link the different exams of the same student, we run the regressions as if each student is observed only once.[10] G_{ijts} is assumed to be a function of the students' gender, F ($F = 1$ for female and $F = 0$ for male). The model considers a vector of co-variates X_{ijt}, which includes the student age, and a set of dummy variables controlling for whether the student is taking the exam with the purpose of applying for university, if he/she is trying to improve a previous score in the exam and if the school attended is private or public. The model also includes year fixed effects, μ_t, and school fixed effects, τ_s . The model is estimated using ordinary least squares pooling all observations from the available repeated cross sections.

As the assessment gap may be different for different ranges of the exam scores' distribution, as obtained by Lindahl (2016), it may bias the average gender gap in the assessment gap and produce misleading results. For this reason we also estimate the following model.

$$G_{ijts} = \alpha + \delta F_i + \beta X_{ijt} + \eta E_{ijtk} + \rho F_i E_{ijtk} + \mu_t + \tau_s + \epsilon_{ijt} \tag{2}$$

where k is the exam score level and E_{ijt} is an indicator variable equal to 1 if the score obtained by student i in the national exam for subject j in time t is equal to k. Significance of the

[10]Since we cannot identify the repeated exams of the same student within the sample, our unit of analysis is actually the exam. However, to make it clearer that we are considering the characteristics of the student that took that particular exam we decided to use the subscript i, for student.

coefficients associated to the interaction between E_{ijtk} and F_i, ρ, will indicate whether the "girl" effect on the assessment gap differs along the exam scores distribution. The remaining variables are the same as for Model 1. This approach allows us to estimate the marginal effect of being a female student conditional on the score obtained in the national exam.[11]

Finally, for secondary education students, and to account for the potential selection into the sample that may arise from considering only internal students, we will estimate a Heckman selection model where the outcome equation is the one for Model 2 and the selection equation is as follows.

$$Prob(Internal) = \alpha + \delta F_i + \beta X_{ijt} + \eta E_{ijtk} + \rho F_i E_{ijt} + \psi Enrolled_{it} + \mu_t + \tau_s + \epsilon_{ijt} \quad (3)$$

For identification purposes we included the indicator variable $Enrolled_{it}$. This variable takes the value 1 if student i took at least one national exam as internal student in year t.

5.2 Results

The following tables present the OLS estimation results of the model specifications described in Eq.(1) and Eq.(2) for the different grades and subjects.[12] Table 3 reports the results of Eq.(1) for Portuguese and mathematics for all grades. The table reports only the parameter of main interest, δ, measuring the impact of being a girl in the assessment gap.

[11]National exams are the same for every student in a given year and are graded anonymously. For this reason, scores from national exams are comparable across students.

[12]Although the dependent variable is discrete, which could suggest an ordered probit or logit model, the option for the linear regression is justifiable given the size of the sample and the fact that the focus is on the average marginal effect of gender. Nonetheless, we estimated an ordered probit for grades 4, 6 and 9 using the unstandardised assessment gap and all the qualitative results are maintained for both Portuguese Language and mathematics.

Table 3: Gender gap in assessments - Portuguese and mathematics.

Subject	Grade	Observations	Schools	(1)	(2)	(3)
Portuguese	4	307012	4932	0.081***	0.081***	0.082***
				(0.004)	(0.004)	(0.003)
	6	428535	1284	0.079***	0.076***	0.078***
				(0.004)	(0.003)	(0.003)
	9	1076636	1476	0.051***	0.048***	0.051***
				(0.003)	(0.003)	(0.002)
	12	755696	673	-0.024***	-0.024***	-0.024***
				(0.004)	(0.004)	(0.002)
Mathematics	4	310947	4931	0.113***	0.112***	0.108***
				(0.004)	(0.004)	(0.003)
	6	434222	1284	0.123***	0.119***	0.116***
				(0.004)	(0.003)	(0.003)
	9	1082359	1476	0.100***	0.095***	0.090***
				(0.003)	(0.003)	(0.002)
	12	519968	665	0.052***	0.051***	0.042***
				(0.005)	(0.005)	(0.002)
			Age	Yes	Yes	Yes
			Year	Yes	Yes	Yes
			Exam Call	Yes	Yes	Yes
			Private School	No	Yes	No
			[a]Exam Score Improvement	No	No	Yes
			[a]University Admission	No	No	Yes
			School	No	No	Yes

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. Standard errors in parentheses are robust in all models and corrected for clustering at the school level in models (1) and (2). Only the coefficient of interest, δ, is presented. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001. [a] Only applicable to 12[th] grade students

All the regressions reported in Table 3 were estimated for the sample of all internal students of each grade. The models in columns (1) to (3) differ only in the set of control variables considered, which are specified in the table for each of the models. The results are very consistent across the different specifications. For 4[th], 6[th], and 9[th] graders the results obtained confirm our initial analysis. The coefficient associated with gender is highly significant for both Portuguese and mathematics, meaning that teacher grading is more favorable for girls. For secondary school students, the results show a positive and significant coefficient for mathematics but a negative one for Portuguese. The estimated average effect for 12[th] graders in mathematics of 0.04 standard deviations is quite similar to the estimates obtained by Falch and Naper (2013) and Lavy (2008).

To better understand these results we looked at the percentage of girls and boys across the distribution of exam scores. It may be seen in Table A1 that within each subject the percentage of girls varies considerably along the scores distribution. If the assessment gap is different for different ranges of the exam scores' distribution, as obtained by Lindahl (2016), it may bias the average gender gap in the assessment gap and produce misleading results. That is the case in the Portuguese Language exam, where the percentage of girls in the upper tail of the exam scores' distribution is much higher than in the lower tail. Since the assessment gap is less for higher exam scores this leads to the misleading negative coefficient of Table 3. To address this problem we estimate the model from Eq.(2), where we control for the results obtained by students in national exams.

Table 4 reports the results of Eq.(2) estimation for Portuguese and mathematics for all grades, where we use specification (3) from Table 3. The table only presents the total impact of being a girl on the assessment gap for each level of exam scores, $\delta + \rho_k$, where k corresponds to the exam score level. The baseline score considered was the passing exam score, $k = 3$. As obtained by Lindahl (2016) we also obtain that the assessment gap is different for different ranges of the exam scores' distribution, as can be seen in Tables A2 and A3 in the Appendix, where we show the coefficient associated to each exam score level, from the estimation of Eq.(2), where we use specification (3) from Table 3. Moreover, the results in Table 4 indicate again that teacher grading is favorable for girls in all cases. For Portuguese, where when we do not take into account the level of exam scores the "girl" effect on the assessment gap is negative, we now obtain a positive effect at all exam score levels but level 5, where there is no significant difference between boys and girls. This result, together with the fact that in the 12$^{\text{th}}$ the share of girls at exam score 5 is 71% helps to explain the negative coefficient obtained previously.

We also look at how the impact of gender on the assessment gap changes as students progress through the grades. Although in Table 3 it seems clear that the impact of gender on the assessment gap decreases from grade 4 to grade 9, the estimates in Table 4 show a different evolution.

Table 4: Gender gap in assessment by exam score levels - Portuguese and mathematics.

Subject	Grade	Exam Score Level				
		1	2	3	4	5
Portuguese	4	0.06** (0.020) 36%	0.14*** (0.006) 44%	0.16*** (0.005) 48%	0.15*** (0.006) 53%	0.09*** (0.013) 58%
	6	0.16*** (0.030) 29%	0.17*** (0.004) 41%	0.18*** (0.004) 48%	0.19*** (0.006) 56%	0.12*** (0.015) 63%
	9	0.35*** (0.023) 30%	0.19*** (0.003) 42%	0.19*** (0.002) 52%	0.24*** (0.004) 61%	0.13*** (0.010) 68%
	12	0.13*** (0.012) 46%	0.12*** (0.003) 52%	0.15*** (0.003) 62%	0.10*** (0.005) 69%	0.01 (0.017) 71%
Mathematics	4	0.00 (0.014) 52%	0.04*** (0.005) 52%	0.07*** (0.005) 49%	0.02*** (0.006) 46%	-0.01 (0.010) 41%
	6	0.02** (0.006) 46%	0.05*** (0.004) 49%	0.14*** (0.005) 49%	0.14*** (0.006) 49%	0.08*** (0.009) 46%
	9	0.05*** (0.003) 52%	0.01*** (0.002) 52%	0.12*** (0.003) 50%	0.16*** (0.004) 52%	0.09*** (0.006) 54%
	12	-0.01 (0.005) 50%	0.05*** (0.003) 52%	0.11*** (0.003) 53%	0.10*** (0.004) 55%	0.05*** (0.005) 56%

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score level. The controls in the regression are the student age, call of the exam, year and school fixed effects. Standard errors in parentheses are robust.* significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table 5 presents the results from Eq.(2) estimation for secondary education, where we use specification (3) from Table 3.[13] In all the cases where the coefficient for female students is statistically significant, it is positive, meaning that the difference between teacher and exam score is greater for girls, and this happens in the large majority of the cases. This suggests that girls perform more poorly than boys in exams relative to teacher assessment and supports

[13] In Table A4 in the appendix we present the results for secondary education students for the different specifications considered in Table 3.

the hypothesis that teacher assessment favors girls while exam based assessments are more favorable to boys.

Table 5: Gender gap in assessment in secondary school by exam score level.

			Exam Score Level				
Subject Type/Subject	Grade	Observations	1	2	3	4	5
Humanities							
English	11	420	0.46*	-0.09	0.24	0.15	0.05
			(0.23)	(0.16)	(0.16)	(0.17)	(0.21)
French	11	11952	0.13	0.10***	0.16***	0.08**	0.10
			(0.13)	(0.03)	(0.02)	(0.03)	(0.08)
Geography	11	219994	-0.05	0.21***	0.24***	0.24***	0.17***
			(0.05)	(0.01)	(0.00)	(0.01)	(0.03)
German	11	9257	0.05	0.15***	0.14***	0.17***	-0.01
			(0.06)	(0.03)	(0.03)	(0.03)	(0.05)
History	12	168298	0.08***	0.15***	0.18***	0.17***	0.08**
			(0.02)	(0.01)	(0.01)	(0.01)	(0.03)
History B	11	8540	0.04	0.15***	0.21***	0.15***	0.14*
			(0.09)	(0.03)	(0.02)	(0.03)	(0.06)
History of Culture and Arts	11	34771	0.06	0.13***	0.14***	0.15***	0.02
			(0.03)	(0.01)	(0.01)	(0.02)	(0.07)
Latin	11	1831	0.41*	0.14*	0.12	0.12	-0.09
			(0.16)	(0.06)	(0.07)	(0.08)	(0.16)
Philosophy	11	72705	0.09***	0.09***	0.12***	0.10***	0.01
			(0.02)	(0.01)	(0.01)	(0.01)	(0.03)
Portuguese	12	755687	0.13***	0.12***	0.15***	0.10***	0.01
			(0.01)	(0.00)	(0.00)	(0.00)	(0.02)
Portuguese Literature	11	21471	0.09	0.09***	0.19***	0.09**	-0.11
			(0.04)	(0.02)	(0.02)	(0.03)	(0.09)
Spanish	11	20855	0.17	0.24***	0.25***	0.21***	0.09
			(0.19)	(0.03)	(0.02)	(0.02)	(0.05)
Sciences							
Biology and Geology	11	489438	0.04***	0.12***	0.15***	0.08***	0.03*
			(0.01)	(0.00)	(0.00)	(0.01)	(0.02)
Descriptive Geometry	11	87493	0.01	0.09***	0.10***	0.09***	0.04***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Information and Technology	11	4243	0.12	0.19***	0.13***	0.07	0.04
			(0.14)	(0.04)	(0.03)	(0.05)	(0.14)
Mathematics	12	519960	-0.01	0.05***	0.11***	0.10***	0.05***
			(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
Mathematics B	11	21301	-0.01	0.08***	0.09***	0.15***	0.11**
			(0.03)	(0.01)	(0.02)	(0.02)	(0.04)
Mathematics SS	11	105038	0.05***	0.12***	0.18***	0.17***	0.08***
			(0.02)	(0.01)	(0.01)	(0.01)	(0.02)
Physics and Chemistry	11	519598	0.05***	0.13***	0.18***	0.12***	0.04***
			(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
Other							
Drawing	12	54535	0.16	0.25***	0.21***	0.16***	0.09*
			(0.12)	(0.02)	(0.01)	(0.01)	(0.03)
Economics	11	84369	0.12***	0.18***	0.19***	0.15***	0.10***
			(0.04)	(0.01)	(0.01)	(0.01)	(0.02)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score level. Standard errors in parentheses are robust. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

The results hold when we consider the quintiles of exam score in place of the exam score itself

(Tables A5 and A6 in the Apendix), and also when we account for selection into the sample using the Heckman procedure (Tables A7 and A8 in the Apendix).[14]

The positive sign obtained for the coefficient of interest means that the existence of a national exam benefits boys to the detriment of girls. Or looking at the question from another perspective, girls have an advantage when assessed by their teachers. This implies that, if selection to university is based on national exams only, a higher percentage of boys will be selected than if selection is based on high-school GPA.

According to the results an educational system attributing a higher weight to teacher assessment will benefit girls while one attributing a higher weight to exams will benefit boys. Thus, the type of evaluation used to assess students has implications for the gender gap in achievement, and consequently for the gender composition of the higher education population and the labor force.

6 Concluding remarks

This paper looks at how grading practices affect boys' and girls' scores differently, looking at two different types of assessment: the one that is carried out by the students' teacher and the one done by means of national exams. While teacher assessment takes into account different aspects of student performance and relies on information collected throughout the year (written tests, student behavior in class, interest shown and homework delivered), national exams are a one shot assessment.

[14]Regarding the selection equation (Table A8 in the Apendix), and given the focus of the analysis, we chose tho present only the coefficient on gender. The effect of gender, although significant in most cases, shows no significant pattern.

Obtaining further evidence regarding the gender difference in achievement under different evaluation systems is especially valuable in light of the fact that female behavior seems to be more context-dependent than male behavior (Croson and Gneezy (2009)), and it is therefore important to obtain evidence for different countries. Also, not only we use data on Portuguese language and mathematics for the whole population of students taking exams at the end of the 4^{th}, 6^{th}, 9^{th}, 11^{th}, and 12^{th} grades, but also for 19 other secondary education subjects, across humanities and sciences, from 2007 to 2018, confirming and extending the results of the previous literature for language and mathematics.

Even though, on average, both boys and girls perform better when assessed by their teachers than in an exam, the results obtained indicate that a grading system relying on teacher assessment routinely rewards girls more than boys. The assessment gap, defined as the score obtained from their teacher minus the score obtained in the national exam, is greater for girls than for boys in all grades and subjects analyzed. This is true across the whole distribution of exam scores. For secondary school students, girls lower their score on the exam by around 0.15 of a standard deviation more than do boys in several subjects.

Taking into account the Portuguese educational system, our results could be due to the fact that the skills under evaluation in the classroom are different from the ones that determine a good score on national exams, with girls doing "better" than boys in the former. It may also suggest that teacher assessment relies on features of student performance for which the skills required develop later for boys. Our data does not allow to confirm the aforementioned mechanisms as source of the gender difference.

Other explanation discussed in the literature is that girls perform worse when stakes are high, which is unarguably the case for the 12th grade exams in Portugal. The literature has also discussed the differential response of boys and girls to novel situations and different testing formats. In Portugal students are usually "trained" to the exam, and the latter usually follows the same structure and types of questions, which allow us to exclude this mechanism as a source of the gender assessment gap we find in our data.

Our main conclusion is that, regardless of the mechanisms driving the results, girls do relatively better than boys under a grading system based on teacher assessment at all grades, and in the large majority of subjects.

Goldin (2014) has documented the converging roles of men and women in different dimensions of the labor market with particular emphasis on the earnings gender gap. The existence of evaluation systems that benefit girls may contribute to the increase in female enrollment rates at higher levels of education, with a potential subsequent impact on labor market characteristics. When used at younger ages, an evaluation system that benefits girls may motivate a greater percentage of boys to choose a vocational track, a decision that later on may hinder their access to higher education. Also, when universities select their students they consider teacher grading and exam scores, with different weights attributed to each of them, depending on the educational system. This implies that educational systems' choices in terms of the types of assessment considered and the relative weight given to each type have implications for the gender composition in higher education and also for the gender composition of the share of the labor force with higher educational attainment.

References

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Chapter 1 - Personality Psychology and Economics, In Handbook of The Economics of Education. Volume 4:pp.1–181.

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.

Cai, X., Lu, Y., Pan, J., and Zhong, S. (2018). Gender gap under pressure: Evidence from china's national college entrance examination. *The Review of Economics and Statistics*, 101.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

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

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Czibor, E., Onderstal, S., Sloof, R., and Praag, M. (2014). Does relative grading help male students? evidence from a field experiment in the classroom. tinbergen institute discussion paper 14-116/v. *SSRN Electronic Journal*.

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

Goldin, C. (2014). A Grand Gender Convergence: Its Last Chapter. *American Economic Review*, 104(4):pp.1–30.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, Gender and Math. *Science*, 320:pp.1164–1165.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Hinnerich, B. T., Höglin, E., and Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review*, 30(4):pp.682–690.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams, C. C. (2008). Gender Similarities Characterize Math Performance. *Science*, 321:pp.494–495.

Jurajda, S. and Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to czech universities. *American Economic Review*, 101(3):514–18.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10-11):pp.2083–2105.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279.

Lindahl, E. (2007). Comparing teachers assessments and national test results: Evidence from Sweden. *IFAU Working Paper*.

Lindahl, E. (2016). Are teacher assessments biased? Evidence from Sweden. *Education Economics*, 24(3):pp.224–238.

Machin, S. and McNally, S. (2005). Gender and student achievement in english schools. *Oxford Review of Economic Policy*, 21(3):357–372.

Machin, S. and Pekkarinen, T. (2008). Global Sex Differences in Test Score Variability. *Science*, 322:pp.1331–1332.

Marcenaro-Gutierrez, O. and Vignoles, A. (2015). A comparison of teacher and test-based assessment for Spanish primary and secondary students. *Educational Research*, 57(1):pp.1–21.

OECD (2012). PISA 2012 Report.

Ors, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics*, 31(3):443 – 499.

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

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.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77:101981.

Apendix

Table A1: Secondary education - percentage of girls by subject and exam score level (2007/18)

Subject Type	Subject	Exam Score Levels 1	2	3	4	5
Humanities	English	88	81	78	72	72
	French	71	69	70	75	80
	Geography	81	68	60	57	52
	German	55	67	74	79	77
	History	76	71	70	70	69
	History B	48	43	42	42	43
	History of Culture and Arts	62	65	70	76	80
	Latin	79	72	74	76	67
	Philosophy	55	61	68	71	70
	Portuguese	46	52	62	69	71
	Portuguese Literature	59	69	77	82	83
	Spanish	60	59	67	75	78
Sciences	Biology and Geology	59	57	58	59	57
	Descriptive Geometry	57	53	48	44	37
	Information and Technology	69	48	42	30	25
	Mathematics	50	52	53	55	56
	Mathematics B	60	61	64	67	70
	Mathematics SS	68	69	71	74	76
	Physics and Chemistry	55	52	54	54	52
Other	Drawing	49	62	66	70	70
	Economics	49	48	49	51	50

Table A2: Exam score level effect on the assessment gap - Portuguese and Mathematics.

Subject	Grade	Exam Score Level 1	2	4	5
Portuguese	4	1.28***	0.56***	-0.51***	-1.13***
		(0.013)	(0.005)	(0.005)	(0.011)
	6	1.77***	0.81***	-0.45***	-0.94***
		(0.016)	(0.004)	(0.005)	(0.012)
	9	1.87***	0.98***	-0.55***	-0.87***
		(0.013)	(0.002)	(0.003)	(0.009)
	12	2.38***	0.94***	-0.59***	-0.94***
		(0.009)	(0.003)	(0.004)	(0.015)
Mathematics	4	1.18***	0.46***	-0.37***	-1.01***
		(0.011)	(0.005)	(0.005)	(0.007)
	6	1.24***	0.46***	-0.31***	-0.73***
		(0.006)	(0.004)	(0.005)	(0.007)
	9	1.15***	0.54***	-0.41***	-0.69***
		(0.003)	(0.003)	(0.004)	(0.005)
	12	2.06***	0.91***	-0.50***	-0.69***
		(0.005)	(0.003)	(0.004)	(0.005)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the exam score level effect on the assessment gap. Standard errors in parentheses are robust.* significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A3: Exam score level effect on the assessment gap in secondary education.

Subject Type/Subject	Grade	Exam Score Level 1	2	4	5
Humanities					
English	11	1.73***	1.18***	-0.46*	-0.69***
		(0.23)	(0.21)	(0.21)	(0.20)
French	11	2.21***	0.90***	-0.41***	-0.58***
		(0.13)	(0.03)	(0.03)	(0.08)
Geography	11	2.58***	0.85***	-0.47***	-0.78***
		(0.05)	(0.01)	(0.01)	(0.03)
German	11	1.81***	0.74***	-0.56***	-0.61***
		(0.06)	(0.03)	(0.04)	(0.05)
History	12	2.10***	0.87***	-0.53***	-0.78***
		(0.02)	(0.01)	(0.01)	(0.02)
History B	11	1.86***	0.72***	-0.52***	-0.79***
		(0.07)	(0.02)	(0.02)	(0.04)
History of Culture and Arts	11	1.85***	0.73***	-0.49***	-0.72***
		(0.03)	(0.01)	(0.02)	(0.06)
Latin	11	1.48***	0.59***	-0.43***	-0.41**
		(0.16)	(0.08)	(0.09)	(0.13)
Philosophy	11	1.91***	0.88***	-0.56***	-0.88***
		(0.02)	(0.01)	(0.01)	(0.03)
Portuguese	12	2.38***	0.94***	-0.59***	-0.94***
		(0.01)	(0.00)	(0.00)	(0.01)
Portuguese Literature	11	2.04***	0.87***	-0.45***	-0.51***
		(0.04)	(0.02)	(0.03)	(0.08)
Spanish	11	2.46***	0.91***	-0.69***	-0.94***
		(0.15)	(0.03)	(0.02)	(0.05)
Sciences					
Biology and Geology	11	2.04***	0.81***	-0.57***	-1.10***
		(0.01)	(0.00)	(0.00)	(0.01)
Descriptive Geometry	11	1.42***	0.72***	-0.55***	-0.88***
		(0.01)	(0.01)	(0.01)	(0.01)
Information and Technology	11	1.76***	0.76***	-0.59***	-0.93***
		(0.13)	(0.03)	(0.03)	(0.09)
Mathematics	12	2.06***	0.91***	-0.50***	-0.69***
		(0.00)	(0.00)	(0.00)	(0.00)
Mathematics B	11	1.85***	0.79***	-0.53***	-0.65***
		(0.03)	(0.02)	(0.02)	(0.04)
Mathematics SS	11	1.97***	0.85***	-0.54***	-0.76***
		(0.02)	(0.01)	(0.01)	(0.02)
Physics and Chemistry	11	1.84***	0.83***	-0.53***	-0.95***
		(0.01)	(0.00)	(0.00)	(0.01)
Other					
Drawing	12	2.69***	0.85***	-0.71***	-1.26***
		(0.09)	(0.01)	(0.01)	(0.03)
Economics	11	1.97***	0.81***	-0.54***	-0.80***
		(0.03)	(0.01)	(0.01)	(0.01)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score quintile. Standard errors in parentheses are robust. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A4: Gender gap in assessment - Secondary Education.

Subject Type/Subject	Grade	Observations	Schools	(1)	(2)	(3)
Humanities						
English	11	432	42	0.28*	0.27*	0.32**
				(0.105)	(0.101)	(0.113)
French	11	11994	284	0.09***	0.09***	0.07***
				(0.023)	(0.023)	(0.017)
Geography	11	220022	639	0.34***	0.34***	0.33***
				(0.008)	(0.008)	(0.004)
German	11	9281	174	-0.02	-0.02	0.00
				(0.030)	(0.029)	(0.020)
History	12	168323	609	0.20***	0.20***	0.21***
				(0.007)	(0.007)	(0.005)
History B	11	8552	195	0.19***	0.19***	0.17***
				(0.031)	(0.032)	(0.019)
History of Culture and Arts	11	34798	310	0.03*	0.03*	0.07***
				(0.014)	(0.014)	(0.010)
Latin	11	1844	89	0.11	0.10	0.12*
				(0.064)	(0.065)	(0.048)
Philosophy	11	72910	642	-0.03**	-0.03**	-0.02*
				(0.011)	(0.011)	(0.007)
Portuguese	12	755696	673	-0.02***	-0.02***	-0.02***
				(0.004)	(0.004)	(0.002)
Portuguese Literature	11	21491	293	-0.05**	-0.06**	-0.02
				(0.018)	(0.018)	(0.014)
Spanish	11	20868	214	0.09***	0.09***	0.10***
				(0.021)	(0.021)	(0.013)
Sciences						
Biology and Geology	11	489443	668	0.12***	0.12***	0.11***
				(0.005)	(0.005)	(0.003)
Descriptive Geometry	11	87543	475	0.24***	0.23***	0.20***
				(0.013)	(0.013)	(0.006)
Information and Technology	11	4246	125	0.28***	0.28***	0.23***
				(0.048)	(0.048)	(0.028)
Mathematics	12	519968	665	0.05***	0.05***	0.04***
				(0.005)	(0.005)	(0.002)
Mathematics B	11	21320	274	0.02	0.02	0.06***
				(0.017)	(0.017)	(0.013)
Mathematics SS	11	105073	559	0.10***	0.10***	0.13***
				(0.009)	(0.009)	(0.006)
Physics and Chemistry	11	519607	669	0.13***	0.13***	0.12***
				(0.005)	(0.005)	(0.002)
Other						
Drawing	12	54551	314	0.12***	0.12***	0.15***
				(0.012)	(0.012)	(0.008)
Economics	11	84410	492	0.15***	0.15***	0.16***
				(0.010)	(0.009)	(0.006)
			Age	Yes	Yes	Yes
			Year	Yes	Yes	Yes
			Exam Call	Yes	Yes	Yes
			Private School	No	Yes	No
			Exam Score Improvement	No	No	Yes
			University Admission	No	No	Yes
			School	No	No	Yes

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. Standard errors in parentheses are robust and corrected for clustering at the school level. Only the coefficient of interest, δ, is presented. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A5: Gender gap in assessment by quintiles of exam score - Portuguese and mathematics.

Subject	Grade	Quintiles of Exam Score				
		1	2	3	4	5
Portuguese	4	0.07*** (0.007)	0.15*** (0.007)	0.14*** (0.008)	0.15*** (0.007)	0.09*** (0.007)
	6	0.16*** (0.005)	0.13*** (0.006)	0.17*** (0.006)	0.14*** (0.007)	0.13*** (0.007)
	9	0.17*** (0.004)	0.14*** (0.005)	0.16*** (0.004)	0.20*** (0.005)	0.17*** (0.006)
	12	0.11*** (0.005)	0.18*** (0.005)	0.18*** (0.005)	0.17*** (0.006)	0.10*** (0.006)
Mathematics	4	0.04*** (0.006)	0.09*** (0.007)	0.08*** (0.007)	0.08*** (0.007)	0.04*** (0.007)
	6	0.01 (0.006)	0.07*** (0.005)	0.11*** (0.006)	0.17*** (0.006)	0.14*** (0.006)
	9	0.01*** (0.004)	-0.02*** (0.004)	0.03*** (0.005)	0.09*** (0.005)	0.08*** (0.005)
	12	0.01** (0.005)	0.08*** (0.005)	0.14*** (0.005)	0.18*** (0.006)	0.12*** (0.005)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score quintile. Standard errors in parentheses are robust.* significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A6: Gender gap in assessment in secondary school by quintiles of exam score.

			Quintiles of Exam Score				
Subject Type/Subject	Grade	Observations	1	2	3	4	5
Humanities							
English	11	417	0.05	0.39	0.17	0.04	0.20
			(0.17)	(0.23)	(0.22)	(0.18)	(0.20)
French	11	5251	0.05	0.05	0.17***	0.16***	0.05
			(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
Geography	11	102451	0.14***	0.15***	0.19***	0.23***	0.24***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
German	11	4591	0.13**	0.19***	0.12*	0.26***	0.04
			(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
History	12	85531	0.10***	0.13***	0.17***	0.22***	0.22***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
History B	11	8540	0.14***	0.15***	0.21***	0.22***	0.09**
			(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
History of Culture and Arts	11	15406	0.10***	0.08***	0.11***	0.10***	0.07*
			(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Latin	11	329	0.09	0.25	0.25	0.20	-0.08
			(0.15)	(0.18)	(0.18)	(0.21)	(0.19)
Philosophy	11	72705	0.08***	0.13***	0.14***	0.12***	0.08***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Portuguese	12	329569	0.11***	0.18***	0.18***	0.17***	0.10***
			(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
Portuguese Literature	11	10271	0.06*	0.18***	0.18***	0.29***	0.14**
			(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Spanish	11	9066	0.20***	0.26***	0.21***	0.18***	0.11**
			(0.04)	(0.03)	(0.04)	(0.04)	(0.04)
Sciences							
Biology and Geology	11	206095	0.08***	0.12***	0.11***	0.11***	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Descriptive Geometry	11	34730	0.03*	0.08***	0.09***	0.07***	0.06***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Information and Technology	11	4243	0.22***	0.13**	0.13**	0.12**	0.06
			(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
Mathematics	12	240169	0.01**	0.08***	0.14***	0.18***	0.12***
			(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
Mathematics B	11	5349	0.05	0.02	0.05	0.11*	0.18***
			(0.04)	(0.03)	(0.04)	(0.04)	(0.05)
Mathematics SS	11	45291	0.08***	0.15***	0.22***	0.22***	0.18***
			(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Physics and Chemistry	11	213567	0.09***	0.14***	0.18***	0.17***	0.11***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Other							
Drawing	12	21099	0.27***	0.23***	0.19***	0.19***	0.12***
			(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Economics	11	42573	0.18***	0.20***	0.22***	0.17***	0.10***
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score quintile. Standard errors in parentheses are robust. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A7: Heckman Selection Model: Secondary Education Exams.

Subject Type/Subject	Grade	Observations	Lambda	Exam Score Level 1	2	3	4	5
Humanities								
English	11	27767	3.22	2.17	0.58	0.58	0.04	0.15
			(6.24)	(3.45)	(1.34)	(0.76)	(0.33)	(0.65)
French	11	15510	0.65***	0.21	0.10***	0.18***	0.12***	0.00
			(0.13)	(0.15)	(0.03)	(0.03)	(0.04)	(0.10)
Geography	11	260344	0.56***	-0.04	0.17***	0.21***	0.24***	0.18***
			(0.02)	(0.07)	(0.01)	(0.01)	(0.01)	(0.04)
German	11	10746	0.03	-0.08	0.13***	0.15***	0.17***	0.01
			(0.21)	(0.10)	(0.03)	(0.03)	(0.04)	(0.07)
History	12	210799	0.68***	0.10***	0.14***	0.21***	0.21***	0.12***
			(0.10)	(0.02)	(0.01)	(0.01)	(0.01)	(0.03)
History B	11	9923	0.74***	0.18	0.12***	0.16***	0.17***	-0.01
			(0.10)	(0.12)	(0.03)	(0.03)	(0.03)	(0.09)
History of Culture and Arts	11	49145	0.56***	0.00	0.08***	0.13***	0.19***	0.05
			(0.06)	(0.04)	(0.01)	(0.02)	(0.03)	(0.09)
Latin	11	2241	0.72*	0.59*	0.12	0.04	0.20	-0.08
			(0.31)	(0.30)	(0.07)	(0.08)	(0.12)	(0.27)
Philosophy	11	100274	0.43***	0.17***	0.12***	0.13***	0.10***	-0.03
			(0.04)	(0.02)	(0.01)	(0.01)	(0.01)	(0.04)
Portuguese	12	851420	1.37***	0.18***	0.15***	0.17***	0.11***	0.04
			(0.02)	(0.03)	(0.00)	(0.01)	(0.01)	(0.04)
Portuguese Literature	11	28323	0.46***	0.20**	0.12***	0.21***	0.09**	-0.01
			(0.06)	(0.07)	(0.02)	(0.02)	(0.03)	(0.10)
Spanish	11	30388	0.38**	0.44	0.25***	0.25***	0.19***	0.10
			(0.14)	(0.24)	(0.03)	(0.02)	(0.02)	(0.06)
Sciences								
Biology and Geology	11	737865	0.17***	0.03	0.11***	0.14***	0.08***	0.04
			(0.01)	(0.02)	(0.00)	(0.00)	(0.01)	(0.02)
Descriptive Geometry	11	131557	0.97***	-0.12***	-0.02	0.01	-0.03	-0.06***
			(0.04)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)
Information and Technology	11	5173	7.29	2.02	0.88	0.44	0.08	0.21
			(14.71)	(4.30)	(1.53)	(0.79)	(0.52)	(1.88)
Mathematics	12	718257	-0.21***	-0.01	0.04***	0.11***	0.11***	0.07***
			(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
Mathematics B	11	34749	0.68***	-0.06	0.08***	0.12***	0.18***	0.17***
			(0.10)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Mathematics SS	11	141218	0.75***	0.03	0.13***	0.20***	0.19***	0.08*
			(0.04)	(0.02)	(0.01)	(0.01)	(0.01)	(0.03)
Physics and Chemistry	11	783149	0.55***	0.06***	0.15***	0.20***	0.13***	0.06***
			(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
Other								
Drawing	12	63017	0.82***	0.18	0.24***	0.20***	0.17***	0.10*
			(0.05)	(0.10)	(0.02)	(0.01)	(0.01)	(0.04)
Economics	11	120393	0.05*	0.09	0.17***	0.19***	0.14***	0.11***
			(0.02)	(0.05)	(0.01)	(0.01)	(0.01)	(0.02)

Note. Dependent variable is the standardized assessment gap, $TeacherScore - ExamScore$. The coefficient presented represents the total effect of being a girl on the assessment gap at each exam score level. Standard errors in parentheses are robust. * significant at p <0.05, ** significant at p <0.01, *** significant at p <0.001.

Table A8: Heckman Selection Model: Secondary Education Exams - Selection Equation.

Subject Type/Subject	Grade	Observations	Exam Score Level 1	2	3	4	5
Humanities							
English	11	27767	0.60	0.24*	0.13	0.00	0.01
			(0.35)	(0.10)	(0.08)	(0.08)	(0.16)
French	11	15510	0.07	0.04	-0.04	0.08	0.09
			(0.17)	(0.04)	(0.05)	(0.06)	(0.15)
Geography	11	260344	-0.15	-0.15***	-0.03*	0.10***	0.25
			(0.10)	(0.01)	(0.01)	(0.03)	(0.15)
German	11	10746	-0.03	0.25***	0.22**	0.27**	0.18
			(0.14)	(0.06)	(0.08)	(0.09)	(0.11)
History	12	210799	0.08*	0.04**	0.20***	0.34***	0.38***
			(0.03)	(0.01)	(0.01)	(0.03)	(0.10)
History B	11	9923	0.34	-0.24***	-0.05	0.04	-0.27
			(0.20)	(0.06)	(0.07)	(0.10)	(0.32)
History of Culture and Arts	11	49145	-0.03	-0.12***	-0.07*	0.08	0.30
			(0.06)	(0.02)	(0.03)	(0.05)	(0.18)
Latin	11	2241	0.86**	0.02	-0.11	0.99**	4.75
			(0.28)	(0.10)	(0.19)	(0.36)	(877.55)
Philosophy	11	100274	0.34***	0.18***	0.03	0.03	0.00
			(0.03)	(0.01)	(0.02)	(0.03)	(0.08)
Portuguese	12	851420	0.17***	0.13***	0.10***	0.09***	0.31
			(0.03)	(0.01)	(0.01)	(0.03)	(0.17)
Portuguese Literature	11	28323	0.26**	0.25***	0.24***	0.11	-0.26
			(0.08)	(0.03)	(0.04)	(0.07)	(0.30)
Spanish	11	30388	-0.04	-0.03	0.06*	0.06	-0.09
			(0.19)	(0.04)	(0.03)	(0.03)	(0.09)
Sciences							
Biology and Geology	11	737865	-0.12***	-0.23***	-0.25***	-0.16***	-0.06
			(0.02)	(0.01)	(0.01)	(0.01)	(0.05)
Descriptive Geometry	11	131557	-0.22***	-0.21***	-0.20***	-0.27***	-0.29***
			(0.02)	(0.01)	(0.02)	(0.02)	(0.03)
Information and Technology	11	5173	0.62*	0.23***	0.15*	0.03	0.00
			(0.29)	(0.07)	(0.07)	(0.13)	(0.07)
Mathematics	12	718257	0.05***	0.10***	0.22***	0.26***	0.17***
			(0.01)	(0.01)	(0.01)	(0.02)	(0.03)
Mathematics B	11	34749	-0.07	0.03	0.16***	0.23***	0.33***
			(0.04)	(0.02)	(0.03)	(0.04)	(0.08)
Mathematics SS	11	141218	0.07**	0.08***	0.13***	0.13***	0.09
			(0.03)	(0.01)	(0.02)	(0.03)	(0.10)
Physics and Chemistry	11	783149	0.05***	0.12***	0.09***	0.04***	0.09***
			(0.01)	(0.00)	(0.01)	(0.01)	(0.03)
Other							
Drawing	12	63017	0.17	0.01	0.04	0.06	0.02
			(0.18)	(0.04)	(0.03)	(0.04)	(0.12)
Economics	11	120393	-0.11	-0.02	-0.03	0.01	0.03
			(0.08)	(0.02)	(0.02)	(0.02)	(0.06)

Note. Dependent variable is the probability of being an internal student. The coefficient presented represents the total effect of being a girl on the probability of being selected into the sample at each exam score level. Standard errors in parentheses are robust. * significant at $p < 0.05$, ** significant at $p < 0.01$, *** significant at $p < 0.001$.