Gender Gaps in different Grading Systems

Catarina Angelo[ab] Ana Balcão Reis[b]

Nova School of Business and Economics, Universidade Nova de Lisboa

April 2017

Abstract

This paper analyses 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 Portuguese data on 21 subjects across humanities and sciences for the whole population of students taking exams at the end of the 6th, 9th, 11th and 12th grades from 2007 to 2016. Results show that the difference in scores between teacher grading and exams is on average positive for boys and girls, but higher for the latter. This is verified across the whole distribution of exam scores. Thus, our results indicate that a grading system based on exams favors boys while one based on classroom evaluation 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.[1] Research shows that 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. One of the arguments recently put on the table hinges on non-cognitive skills as a potential source of the observed gender differences in achievement. In the words of Cornwell et al (2013), who looked at primary schools in US, "Even those boys who perform equally as well as girls on reading, math and science tests are nevertheless graded less favorably by their teachers, but this less favorable treatment essentially vanishes when non-cognitive skills are taken into account". Different grading systems reward non-cognitive skills differently. For this reason, the choice of the grading system used and its application at the different levels of education may influence the gender gap in student achievement.

Moreover, in many countries, students achievement is determinant in the choice between an academic or a vocational track and at the end of secondary schooling scores are used as a selection instrument for application to university. Different educational systems assign different weights to exam results and teacher grading. Thus, if at the end of secondary school we still

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

observe that boys are less favorably graded by their teachers, then the choice of the weight attributed to exam results and teacher grading will have an impact on the gender composition of the higher education student population and the labour market.

In general, educational systems rely heavily on teacher grading to determine a student final scores. Teacher grading is usually based on information regarding performance collected across the year, and so, it does not rely solely on scores obtained in the several tests taken in the classroom. When assigning a score the teacher also considers several other aspects, namely, student behavior in the classroom, if she/he keeps track of homework and class materials and if there is an improvement in performance throughout the year or not. Within this context, teachers can reward, either explicitly or not, a student's effort and commitment towards their course. These attitudes are likely to be related to a set of non-cognitive skills (like emotional maturity, empathy, interpersonal skills and verbal and non-verbal communication) that can influence a student's behavior.[2]

Although this allows teacher grading to reflect a more complete assessment of student achievement, it also leaves more room for discretionary judgement and assessment from teachers. Girls may be more amenable to the learning process than boys, which makes teachers more prone to reward girls with a higher score (Cornwell et al., 2013). Signals of perceived ability given by teachers when assigning an end of term score are likely to have an influence on a student's level of self-confidence and motivation, with the rewarding of elicited attitudes through a higher score being seen as a positive message from the teacher to the student that can influence her/his future achievement. Pavlova et al. (2010) provided evidence that not only does a positive message

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

enhance performance, whereas a negative message diminishes it, but that this effect is also more pronounced for girls, for whom a negative stereotype message has a stronger impact.

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 a written test taken in the classroom, exams are likely to have a higher stress component associated with them, though this may partly be mitigated by teachers usually investing in the preparation of students prior to exams.

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. We use teacher grading and national exam scores for several cohorts of Portuguese students, in the 6^{th}, 9^{th}, 11^{th} and 12^{th} grades. Portuguese Language and Mathematics are the only subjects that are subject to national exams in the 6^{th} and 9^{th} grades, while at the end of secondary education there are national exams for 21 subjects across humanities and sciences.

The fact that we have data from the 6^{th} to the 12^{th} grade makes it possible to test if gender differences associated with different grading systems change with age. Insights from psychology and psychiatry make it almost impossible not to acknowledge the differences between boys and girls, namely in what concerns personality traits. Cox (2005) argues that "the communication difficulties of boys are more noticeable than ever" and that society's demand for these skills to develop is faster than the pace at which they do develop in boys. If the personality traits that help determine a good performance in the classroom develop later in life for boys than for girls, we would expect the difference in scores from the two evaluation schemes analyzed in this work

to narrow at higher levels of education.

Some previous studies that compare blind and non-blind scores to measure a potential gender bias (Lavy, 2008, Hinnerich et al., 2011) use similar tests to conclude on discrimination. Lavy (2008) uses data on blind and non-blind scores from matriculation exams of Israeli students, for several cohorts of 12^{th} 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 both 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 9^{th} graders, and find no evidence of discrimination against boys in grading. It is worth emphasizing the fact that Hinnerich's study looks only at Swedish Language, and therefore does not allow to rule out discrimination in the remaining subjects.

Another set of studies, closer to our paper, compares different grading schemes. Falch and Naper (2013), using data on Norwegian student scores in the 10^{th} grade, and Lindahl (2016) using Swedish student data, obtain that an evaluation system that relies heavily on teacher grading lowers boys' scores relative to girls'. 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.

Our work contributes to the existing literature on how grading practices may affect boys' and girls' scores differently, in two ways. The first one 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 track of secondary education that take at least one national exam. Secondly, we have data for 10 different cohorts from grade 9 to grade 12. For the 6^{th} grade, since the national exam was only introduced more recently, we have only 4 cohorts. Investigating the gap between scores from teacher grading and national exams for different grades may give some insights on how grading practices will affect girls and boys differently over the life cycle.

The results obtained indicate a gap in assessment that is larger for girls in the majority of subjects under analysis, either in humanities or 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. The next section gives an overview of the Portuguese educational system. Section 3 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 4 presents a regression analysis and the main results and the final section concludes.

2 Institutional Setting

Portugal has 12 years of compulsory education, children attend school from the year they turn six to the year they turn 18. 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. Students enrolled in this track represented, in the 2011/2012 school year, 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. The professional track represents 33% of the students enrolled in secondary education. This track is tailored to students who want to obtain a professional qualification that allows them to enter the labour market. Regardless of the track chosen by students, they can always decide to pursue a university degree. Private and public schools coexist in all levels of education.

In the Portuguese Educational System, for the period under analysis, students are tested at the end of each cycle of studies by means of national exams. These exams are meant to provide a measure for the knowledge level of students in the core subjects of the curricula and are graded anonymously, by teachers from a different school to the one attended by the student.

Portuguese and Mathematics are the only subjects tested at the end of the 6th and the 9th grades, 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 common and mandatory for all students. Each secondary student is also tested in at least 3 other subjects, which are specific to their field of study, in the end of the eleventh or the twelfth grade. 30% of a student's final score in a subject for which he takes the national exam is determined by his/her exam score, the

remaining 70% are determined by teacher grading. Higher education access is determined by a weighted average of high-school GPA and scores on national exams.

3 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 2016, on the results of the students that take the national exams at the end of either the 9th grade or secondary education.[3] For 6th graders data is available only from 2012 to 2015. For each student, we know the school she/he was enrolled and its geographical location, their age, gender and field of study. Unhappily the data set does not allow to link the different exams of the same student. Nonetheless, for each exam we know the exam score, student characteristics and the score obtained from teacher grading, in case the student is enrolled as an internal student. As we want to compare exam scores with teacher grading, we only consider internal students.

Students enrolled in secondary education can only take the national exam in a subject as an internal student if the score assigned by their teacher in that subject is at least 8, a restriction that does not apply to 9th and 6th graders. In fact, for 11th and 12th graders, we have no observations in the data for students that take the national exams as internal students and simultaneously have a score from teacher grading below 10.[4] Thus, the data set excludes the secondary students in

[3] Although the database exists since 1998, information on students' gender is only available since 2007.

[4] Students can cancel their enrollment in a subject and still take the national exams, in which case their final score is fully determined by the exam score. Thus, when their teacher score is below 10, students prefer to cancel their enrollment, implying they are not internal students.

the left tail of the scores distribution.

For secondary education, the grading scale for teacher grading and the course final score is from 1 to 20 and for the national exam is 0 to 200. To unify the scale for both evaluations, the exam score was converted into a 1 to 20 scale. In the 6th and 9th grades, scores are all in a 1 to 5 scale.

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.

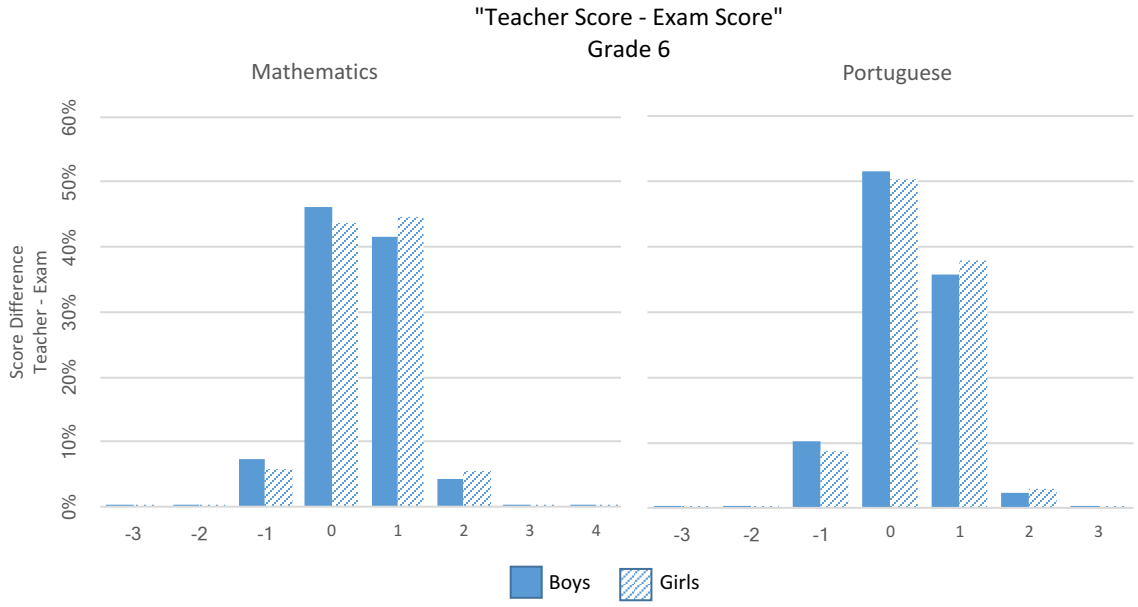

Figure 1: Distribution of the assessment gap in the 6th grade (2012-2015): Teacher Score-Exam Score

In figures 1 and 2 we can see the distribution of the assessment gap for Portuguese Language and Mathematics, the subjects that are tested in the 6th and the 9th grades. From the histograms we can conclude that on average, for both grades and subjects, teacher scores are higher than exam results. Also, in both grades and subjects there are relatively more boys than girls improving

their score in the exam relative to the score obtained from their teachers, an effect that is stronger in Mathematics. Another interesting aspect is that in Portuguese the level of similarity between the score attributed by teachers and the one obtained in the exam is higher than in Mathematics, with roughly 51% of boys and 50% of girls obtaining the same score, from their teachers and in the exam.

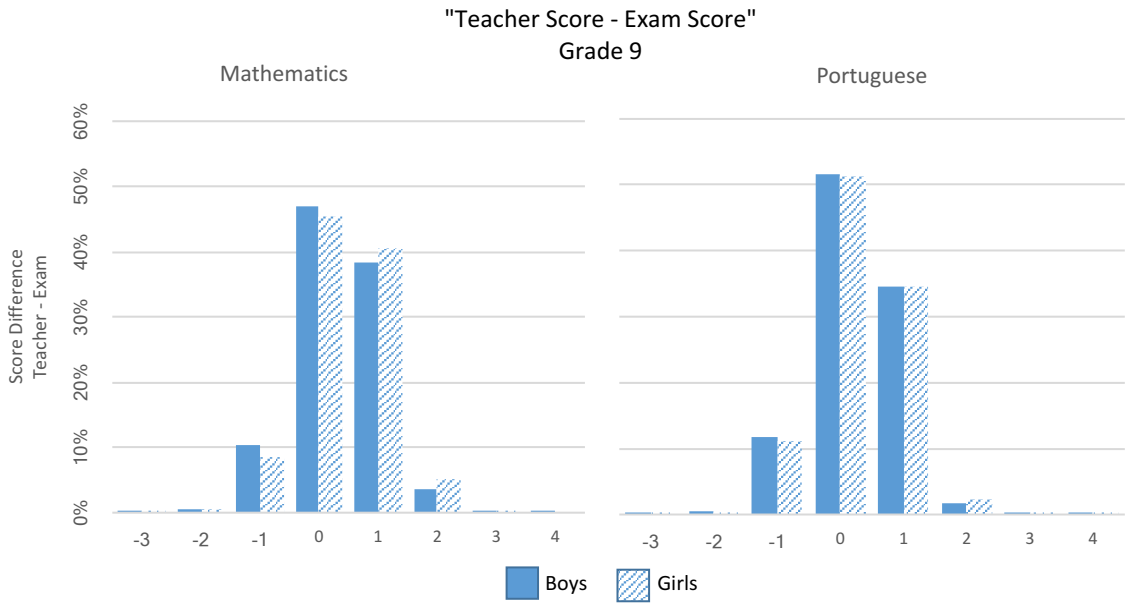

Figure 2: Distribution of the assessment gap in the 9th grade (2007-2016): Teacher Score-Exam Score

Figure 3 shows the kernel density estimates of the assessment gap distribution for the secondary education subjects under analysis.[5] We can observe that, as for the 6th and 9th grades, the average gap is always positive. For the subjects where the distributions do not coincide we observe a negative shift factor for the boys' distribution, indicating that on average, compared to teacher

[5]Using the Epanechnikov kernel function.

assessment, the boys' scores in the exams decrease less than the girls'.[6]

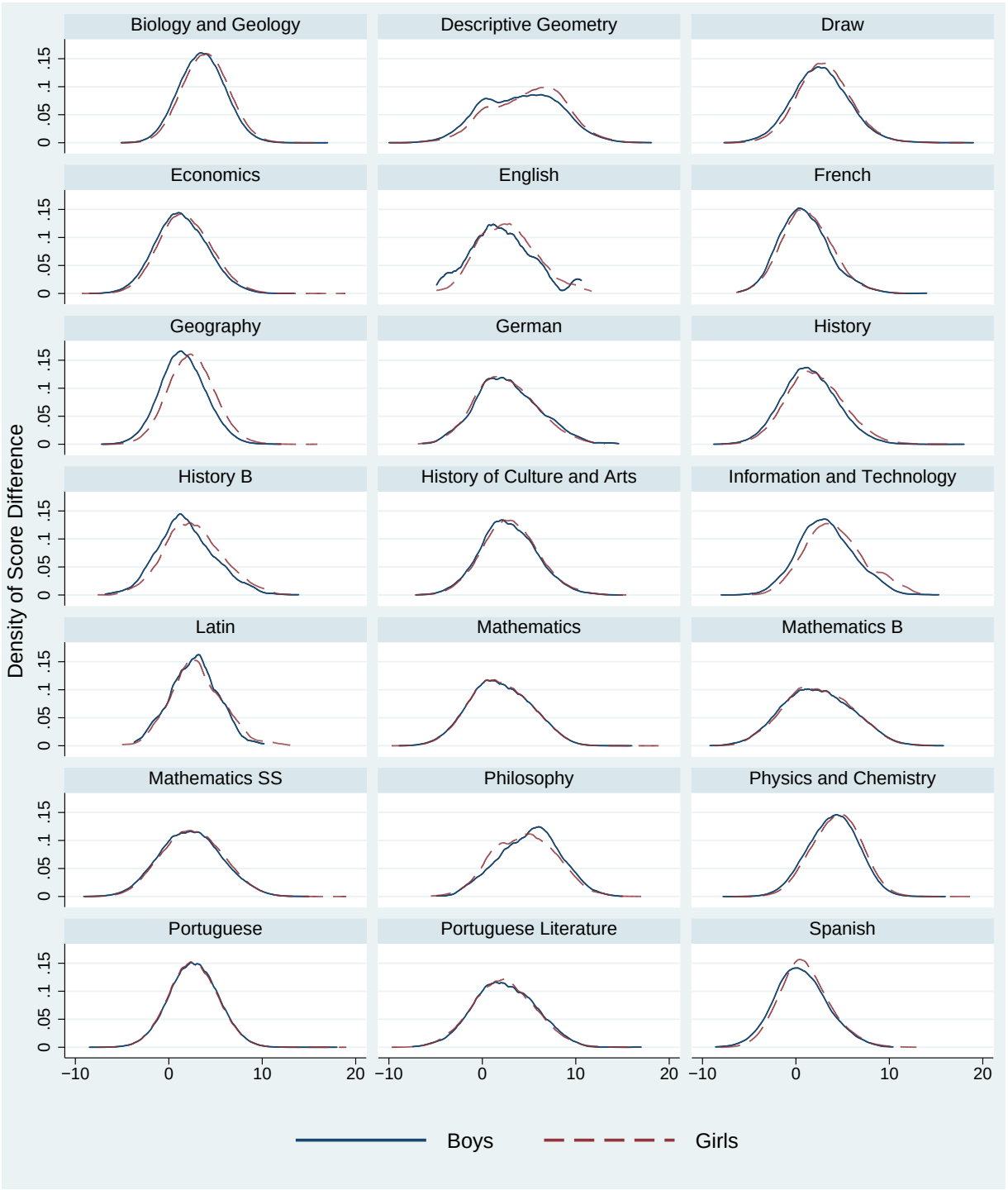

Figure 3: Kernel density estimates of the assessment gap in secondary education:Teacher Score-Exam Score

[6]The only exception is Philosophy, where the assessment gap is higher for boys.

Tables 1 and 2 present 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 6th and the 9th grades. The scores examined are the ones attributed by teachers, the ones obtained in the national exam and the difference between the two for both boys and girls. The gap between assessments is presented in columns (10) and (11) of tables 1 and 2. As already seen in figures 1 and 2, a common feature for both genders is that scores from teacher assessment are on average higher than the ones obtained in national exams. Remember that when assigning scores, teachers may also take into account other dimensions of student performance, like attention in class, interest shown towards the subject, punctuality and attendance, amongst others, which could also contribute to the assessment gap.

The results are clearly in line with existing literature on student performance, according to which, girls clearly score better in reading while their results in mathematics are not as good. In the 6^{th} grade girls perform better than boys in Portuguese and have almost the same results in Mathematics. The pattern is the same in the 9^{th} grade.

In both grades the assessment gap is positive and significant for boys and girls in Portuguese and Mathematics (columns 7 and 8 in tables 1 and 2). The difference between the girls' and boys' assessment gap is positive and significant in both subjects (column 9 in tables 1 and 2) suggesting that teacher grading favors girls. The gender difference in the assessment gap decreases from the 6^{th} to the 9^{th} grade for Portuguese and is unchanged in Mathematics.

Table 1: Means and standard deviations of teacher-score and national exams in grade 6 (2012-2015)

	Number of	Mean Teacher Score		Mean Exam Score		Mean Difference		T test for the difference in
Subject	Observations	Boys	Girls	Boys	Girls	Boys	Girls	Mean differences [(8)-(7)]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Portuguese	426681	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	432243	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)	[3.69]

Notes. The grading scale is 1 to 5. The mean difference in columns (7) and (8) is defined as mean teacher score minus mean exam score. In column (9) the difference tested is column (8) minus column (7). The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

In the appendix we present the assessment gap for boys and girls separating students by their exam scores. For both grades and for each exam score we obtain again that the assessment gap is larger for girls. The analysis conducted separately for each exam score could lead to different results because the distribution of scores is different for girls and boys.[7] The fact that the results are maintained reinforces our conclusions.

Tables 3*a* and 3*b* replicate the analysis from tables 1 and 2 for the subjects in which students take exams in secondary education. Similarly to what is observed in the 6th and 9th grades, and confirming what we saw in figure 3, students obtain on average higher scores from teacher assessment than in national exams.

[7]This can be verified in column 4 from tables A1 and A2, in the appendix.

Table 2: Means and standard deviations of teacher-score and national exams in grade 9 (2007-2016)

	Number of	Mean Teacher Score		Mean Exam Score		Mean Difference		T test for the difference in
Subject	Observations	Boys	Girls	Boys	Girls	Boys	Girls	Mean differences [(8)-(7)]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Portuguese	887541	3.12	3.37	2.86	3.10	0.26	0.28	0.02*
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[1.9]
Mathematics	891921	3.05	3.11	2.70	2.69	0.35	0.42	0.07***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	[4.33]

Notes. The grading scale is 1 to 5. The mean difference in columns (7) and (8) is defined as mean teacher score minus mean exam score. In column (9) the difference tested is column (8) minus column (7). The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

According to the results, girls achieve on average higher scores than boys when they are accessed by their teachers, this is evident in all of the humanities subjects and in almost every subject in sciences. Only in Descriptive Geometry and Information and Technology do boys obtain a higher score relative to girls when they are assessed by their teachers.
The pattern observed in scores obtained from teachers is partially reversed when we analyze exam scores. Under the latter type of assessment, boys perform on average better than girls in 6 of the 21 subjects under analysis and equally as well in 3 of them. In Humanities courses, boys obtain on average higher scores than girls in History and English. In Science courses girls outscore boys only in Mathematics and Biology.

While the gap in assessment is significant in all subjects, the difference between gender (defined as the girls' gap minus the boys' gap) is only significant when positive, that is, only in subjects where the gap is higher for girls, with the only exception being Philosophy. It is worth noticing that the assessment gap gender difference size is not only significant but also non trivial, ranging

Table 3*a*: Means and standard deviations of teacher-score and national exams in secondary education (2007-2016)

Subject Type	Subject	Grade	Number of Observations	% Girls	Mean Teacher Score		Mean Exam Score		Mean Difference		T test for the difference in Mean differences [(11)-(10)]
					Boys	Girls	Boys	Girls	Boys	Girls	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Humanities	English	11	390	0.79	13.9	13.9	11.9	11.0	2.0	2.9	0.9
					(0.42)	(0.24)	(0.83)	(0.55)	(1.08)	(0.61)	[0.68]
	French	11	11064	0.73	12.8	13.4	11.6	12.0	1.3	1.4	0.1
					(0.07)	(0.06)	(0.12)	(0.10)	(0.23)	(0.14)	[0.45]
	Geography	11	175901	0.64	13.1	13.3	11.4	10.7	1.7	2.6	0.8***
					(0.04)	(0.03)	(0.05)	(0.04)	(0.08)	(0.06)	[8.27]
	German	11	7929	0.74	13.4	14.4	10.6	11.8	2.9	2.7	-0.2
					(0.09)	(0.09)	(0.19)	(0.13)	(0.36)	(0.22)	[-0.53]
	History	12	129639	0.71	12.7	13.0	10.7	10.4	2.0	2.5	0.6***
					(0.03)	(0.03)	(0.06)	(0.05)	(0.10)	(0.06)	[4.73]
	History B	11	2818	0.43	14.1	14.5	11.9	11.7	2.2	2.8	0.6
					(0.11)	(0.12)	(0.19)	(0.23)	(0.25)	(0.28)	[1.54]
	History of Culture and Arts	11	28210	0.68	12.7	13.5	9.6	10.3	3.1	3.2	0.1
					(0.05)	(0.06)	(0.09)	(0.08)	(0.16)	(0.11)	[0.56]
	Latin	11	1259	0.74	13.7	13.9	10.7	10.7	3.1	3.2	0.2
					(0.26)	(0.16)	(0.26)	(0.23)	(0.52)	(0.29)	[0.39]
	Philosophy	11	29837	0.65	13.3	14.0	9.5	10.5	3.8	3.5	-0.3*
					(0.05)	(0.04)	(0.10)	(0.08)	(0.14)	(0.10)	[-1.65]
	Portuguese	12	614559	0.59	12.8	13.7	10.1	11.1	2.7	2.6	-0.1
					(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	[-0.81]
	Portuguese Literature	11	13020	0.75	12.4	13.3	10.0	11.1	2.3	2.2	-0.2
					(0.07)	(0.06)	(0.12)	(0.11)	(0.21)	(0.12)	[-0.61]
	Spanish	11	12254	0.70	14.4	15.5	12.6	13.6	1.8	1.9	0.2
					(0.09)	(0.09)	(0.13)	(0.11)	(0.21)	(0.14)	[0.61]

Note. The mean difference in columns (10) and (11) is the gender assessment gap, defined as mean teacher score minus mean exam score. In column (12), the difference tested is column (11) minus column (10), girls' assessment gap minus boys' assessment gap. The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

Table 3*b*: Means and standard deviations of teacher-score and national exams in secondary education (2007-2016)

Subject Type	Subject	Grade	Number of Observations	% Girls	Mean Teacher Score		Mean Exam Score		Mean Difference		T test for the difference in Mean differences [(11)-(10)]
					Boys	Girls	Boys	Girls	Boys	Girls	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Sciences	Biology and Geology	11	408908	0.57	13.6	13.9	9.7	9.8	3.8	4.1	0.3***
					(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	[3.56]
	Descriptive Geometry	11	72494	0.49	14.6	14.2	11.1	9.8	3.5	4.5	0.9***
					(0.06)	(0.06)	(0.14)	(0.14)	(0.18)	(0.19)	[3.51]
	Information and Technology	11	4132	0.41	15.0	14.7	11.7	10.5	3.2	4.2	0.9**
					(0.11)	(0.13)	(0.16)	(0.16)	(0.22)	(0.27)	[2.68]
	Mathematics	12	420417	0.61	13.0	13.4	10.6	10.9	2.5	2.5	0
					(0.05)	(0.06)	(0.08)	(0.08)	(0.09)	(0.08)	[0.23]
	Mathematics B	11	19685	0.63	12.7	13.3	10.2	10.7	2.5	2.6	0.1
					(0.05)	(0.06)	(0.14)	(0.13)	(0.20)	(0.15)	[0.45]
	Mathematics SS	11	84472	0.71	12.7	13.3	10.3	10.7	2.5	2.7	0.2
					(0.03)	(0.03)	(0.08)	(0.07)	(0.13)	(0.08)	[1.22]
	Physics and Chemistry	11	432694	0.54	13.0	13.3	9.0	9.0	4.0	4.3	0.3***
					(0.05)	(0.06)	(0.07)	(0.08)	(0.07)	(0.07)	[3.40]
Other	Draw	12	46214	0.66	14.8	15.5	12.3	12.7	2.5	2.8	0.3**
					(0.05)	(0.05)	(0.06)	(0.06)	(0.12)	(0.08)	[2.25]
	Economics	11	65718	0.49	13.9	14.3	11.8	11.8	2.2	2.5	0.3*
					(0.05)	(0.05)	(0.08)	(0.08)	(0.11)	(0.11)	[2.01]

The mean difference in columns (10) and (11) is the gender assessment gap, defined as mean teacher score minus mean exam score. In column (12), the difference tested is column (11) minus column (10), girls' assessment gap minus boys' assessment gap. The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

Table 4: Gender differences in the assessment gap for secondary education by exam score intervals (2007-2016)

Subject Type	Subject	[0,4.5)	[4.5,9.5)	[9.5,13.5)	[13.5,17.5)	[17.5,20)
Humanities	English	-.01	0.82	0.04	0.63	0.95
		[-.01]	[1.12]	[0.04]	[0.91]	[0.91]
	French	0.08	0.20	0.50**	0.30	0.13
		[0.23]	[0.92]	[2.88]	[1.36]	[0.48]
	Geography	0.11	0.38***	0.55***	0.59***	0.33**
		[0.82]	[5.26]	[7.07]	[6.88]	[3.04]
	German	-.24	0.58*	0.50*	0.48	0.16
		[-.68]	[1.98]	[1.67]	[1.46]	[0.55]
	History	0.34***	0.35***	0.44***	0.49***	0.22*
		[3.48]	[4.90]	[5.78]	[5.01]	[1.81]
	History B	0.04	0.43	0.44*	0.59*	0.29
		[0.11]	[1.61]	[1.73]	[2.26]	[1.01]
	History of Culture and Arts	0.21	0.40**	0.51***	0.62**	0.41
		[1.16]	[3.09]	[3.34]	[3.07]	[1.44]
	Latin	-.12	0.42	0.07	0.08	-.55
		[0.25]	[0.93]	[0.17]	[0.14]	[-.88]
	Philosophy	0.25*	0.29**	0.40***	0.33**	-.03
		[1.93]	[2.74]	[3.72]	[2.77]	[-.24]
	Portuguese	0.26***	0.38***	0.43***	0.32***	0.11
		[5.46]	[7.03]	[6.34]	[3.71]	[1.20]
	Portuguese Literature	0.45**	0.43**	0.63***	0.30	-.02
		[2.44]	[2.97]	[3.82]	[1.46]	[-.06]
	Spanish	0.88	0.78***	0.69***	0.78***	0.52*
		[1.65]	[3.32]	[3.42]	[3.68]	[1.97]
Sciences	Biology and Geology	0.08	0.28***	0.40***	0.21**	0.08
		[1.60]	[4.76]	[5.47]	[2.58]	[0.83]
	Descriptive Geometry	-.14	0.20	0.27*	0.29*	0.16
		[-1.13]	[1.56]	[1.83]	[1.93]	[1.26]
	Information and Technology	0.79	0.44	0.15	0.12	0.36
		[1.35]	[1.61]	[0.51]	[0.44]	[0.73]
	Mathematics	0.00	0.11*	0.24***	0.28**	0.13
		[0.07]	[2.01]	[3.19]	[3.04]	[1.53]
	Mathematics B	-.03	0.27*	0.42**	0.60**	0.65**
		[-.22]	[2.19]	[2.85]	[2.74]	[2.97]
	Mathematics SS	0.01	0.26***	0.53***	0.54***	0.34**
		[0.06]	[3.29]	[5.78]	[4.92]	[2.84]
	Physics and Chemistry	0.13**	0.31***	0.49***	0.35***	0.19**
		[3.08]	[5.10]	[5.55]	[3.52]	[2.35]
Other	Draw	0.27	0.68***	0.58***	0.50***	0.42**
		[0.84]	[4.71]	[4.45]	[3.94]	[2.55]
	Economics	0.34	0.33**	0.33***	0.34**	0.31**
		[2.08]	[3.06]	[3.17]	[2.85]	[2.69]

Note. The T statistic in square brackets reflects standard errors that are corrected for clustering at the school level.

from 0.3 to 0.9 (in a 0 to 20 scale).

When we divide students by their exam score to evaluate the assessment gap in the different ranges of the scores scale we obtain that in all the cases where the gender gap in assessment is significant it is positive. Even the exception previously found in Philosophy disappears. This is shown in table 4. In table A3 in the appendix we present the distribution of girls across the different brackets of exam scores.

4 Gender gap in different grading systems

4.1 Regression analysis

In the previous section we have shown that, on average, the difference between teacher score and national exams is higher 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 several student characteristics. We follow Marcenaro-Gutierrez and Vignoles (2015) and assume a linear specification for the gap equation to estimate the following model.

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

Where G_{ijt} is the gap between assessments for student i in subject j and time t. Each student is observed only at one point in time and G_{ijt} 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 students age, and a set of dummy variables controlling for whether the student is taking the exam with the purpose of applying for university, if she/he is trying to

improve a previous score in the exam and if she/he attends a private or a public school. The model also includes year fixed effects, μ_t, to account for the different cohorts under analysis. The model is estimated using ordinary least squares pooling all observations from the available repeated cross sections.

4.2 Results

The following tables present the OLS estimation results of the model specification described in Eq.(1) for the different grades and subjects.[8] Tables 5 and 6 report the results for Portuguese and Mathematics for all grades. Tables 7*a* and 7*b* report the results for all remaining subjects in secondary school. The tables report only the parameter of main interest, δ, measuring the impact of being a girl in the assessment gap. All the regressions presented in table 5 were estimated for the sample of all students. The models in columns (1) to (5) 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. Thus, in table 6 we focus on the specification of column 5 and present the results for regressions run separately for each exam score or each interval of exam scores.[9]

As can be observed in table 5, the results obtained for 6th and 9th graders taking Portuguese and Mathematics confirm our initial analysis. The coefficient associated to gender is highly significant for both subjects, meaning that teacher grading is more favorable for girls. For

[8]Although the dependent variable is discrete for 6th and 9th grade students, 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.

[9]For the 6th and 9th grades specifications 2 and 5 are equivalent.

secondary school students, the results in table 5 show a positive and significant coefficient for Mathematics but a negative one for Portuguese. However, the results of the regression run separately for each interval of exam scores shown in table 6, indicate again that teacher grading is favorable for girls in all cases.

Table 5: Gender gap in assessments - Portuguese and Mathematics.

Subject	Grade	Observations	Schools	(1)	(2)	(3)	(4)	(5)
Portuguese	6	426680	1261	0.054***	0.052***			
				(0.002)	(0.002)			
	9	887541	1440	0.029***	0.027***			
				(0.002)	(0.002)			
	12	614559	658	-0.034*	-0.034**	-0.012	-0.012	-0.026*
				(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Mathematics	6	432243	1261	0.087***	0.083***			
				(0.002)	(0.002)			
	9	891921	1440	0.081***	0.077***			
				(0.002)	(0.002)			
	12	420417	651	0.104***	0.103***	0.105***	0.104***	0.106***
				(0.016)	(0.016)	(0.016)	(0.015)	(0.015)
			Year	Yes	Yes	Yes	Yes	Yes
			Exam 1st Call	Yes	Yes	Yes	Yes	Yes
			Area Study	No	No	Yes	Yes	Yes
			Private School	No	Yes	No	Yes	Yes
			Exam Improval	No	No	No	No	Yes
			University Admission	No	No	No	No	Yes

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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.

In table A3 it can be seen that for the Portuguese Language exam, the percentage of girls in the higher exam scores bracket is much higher than in the lower ones. As the assessment gap is higher for lower exam scores this leads to the misleading negative coefficient of table 5. This shows the relevance of conducting the analysis separately by exam score.

We also look at the evolution of the impact of gender on the assessment gap as students progress

through the grades. Regarding the 6th and 9th grades the magnitude of the impact seems to be very similar. When the comparison is done separately for each exam score the impact is higher for Portuguese Language. To compare the magnitude of the impact of gender on the assessment gap for 6th and 9th graders with that for secondary students we need to take into account the change from a scale of 1 to 5 to a scale of 1 to 20. The results in table 6 suggest a decrease on the magnitude of the impact, at least for the higher exam scores. For Portuguese Language, in a scale of 1 to 20, the coefficient goes from 0.32 (0.08 in a 1 to 5 scale) for 9th graders to 0.15 for 12th graders, for students who had the highest scores in the exam, and from 0.6 to 0.37 for the previous bracket. For Mathematics, also in a 1 to 20 scale, the values go from 0.28 to 0.17 for top students, and from 0.48 to 0.28 for the previous bracket.

Table 6: Gender gap in assessment by exam score levels - Portuguese and Mathematics.

Subject	Grade	Exam Score				
		1	2	3	4	5
Portuguese	6	0.11***	0.12***	0.13***	0.13***	0.08***
		(0.02)	(0.00)	(0.00)	(0.01)	(0.01)
	9	0.25***	0.13***	0.13***	0.15***	0.08***
		(0.02)	(0.00)	(0.00)	(0.00)	(0.01)
	12[a]	0.40***	0.45***	0.48***	0.37***	0.15***
		(0.03)	(0.01)	(0.01)	(0.02)	(0.04)
Mathematics	6	0.01**	0.03***	0.10***	0.10***	0.05***
		(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
	9	0.04***	0.01*	0.09***	0.12***	0.07***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
	12[a]	0.01	0.10***	0.25***	0.28***	0.17***
		(0.02)	(0.01)	(0.02)	(0.02)	(0.02)

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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.

[a]For the 12th grade, scores 1 to 5 correspond to the brackets already presented in table 4.

This could be an indication that boys learn how to fulfill teachers' expectations regarding

their behavior in class, decreasing the gender difference in the assessment gap from basic to secondary education. Another explanation for this behavior could be self-selection, through which students choose between the academic and the professional track. In fact, supporting the latter hypothesis, the percentage of boys in the academic track of secondary education decreases with respect to the 9th grade, from 49% in the latter to almost 42% in the former.

Tables 7*a* and 7*b* present the results for secondary education when regressions are run separately for the different intervals of exam scores.[10] In all cases, the coefficient for female students is positive and significant, meaning that, the difference between teacher and exam score is higher for girls. This suggests that, on average, girls perform worse than boys in exams relative to teacher assessment and supports the hypothesis that teacher assessment favors girls while exam based assessments are more favorable for boys.

One possible explanation for this result could be that girls perform relatively worse than boys when stakes are high. However, in a recent work, Falch and Naper (2013) find evidence that does not support this hypothesis. Thus, even if national exams have a stronger high stakes component compared to teacher assessment, this factor may not justify the results obtained.

The 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.

[10]In tables A4*a* and A4*b* in the appendix we present the results for the sample of all students and for different specifications.

Table 7*a*: Gender gap in assessments - Humanities and other.

Subject Type	Subject	Grade	Observations	Schools	[0,4.5)	[4.5,9.5)	[9.5,13.5)	[13.5,17.5)	[17.5,20)
Other	Draw	12	46214	298	0.18	0.64***	0.53***	0.45***	0.34***
					(0.30)	(0.07)	(0.04)	(0.04)	(0.08)
	Economics	11	65718	459	0.37**	0.39***	0.46***	0.43***	0.37***
					(0.11)	(0.04)	(0.03)	(0.04)	(0.05)
Humanities	English	11	390	44	0.00	0.53	-.02	0.33	0.83
					(0.88	(0.36)	(0.51)	(0.48)	(0.66)
	French	11	11064	299	0.23	0.20*	0.47***	0.29**	0.06
					(0.22)	(0.09)	(0.07)	(0.09)	(0.16)
	German	11	7929	185	-0.12	0.50***	0.46***	0.45**	0.16
					(0.20)	(0.09)	(0.11)	(0.15)	(0.17)
	History	12	129639	583	0.25***	0.29***	0.41***	0.48***	0.23**
					(0.05)	(0.02)	(0.02)	(0.03)	(0.08)
	History B	11	6617	173	0.10	0.45***	0.41**	0.57***	0.31*
					(0.21)	(0.12)	(0.12)	(0.12)	(0.16)
	History of Culture and Arts	11	28210	294	0.15	0.32***	0.42***	0.57***	0.46**
					(0.09)	(0.04)	(0.05)	(0.10)	(0.18)
	Latin	11	1698	85	-0.10	0.42*	0.13	0.13	-0.48
					(0.31)	(0.19)	(0.22)	(0.28)	(0.30)
	Philosophy	11	45994	620	0.27***	0.29***	0.41***	0.37***	0.01
					(0.06)	(0.04)	(0.04)	(0.05)	(0.09)
	Portuguese	12	614559	658	0.40***	0.45***	0.48***	0.37***	0.15***
					(0.03)	(0.01)	(0.01)	(0.02)	(0.04)
	Portuguese Literature	11	17316	280	0.31**	0.31***	0.50***	0.23*	0.11
					(0.12)	(0.06)	(0.07)	(0.10)	(0.27)
	Spanish	11	17457	205	0.93*	0.67***	0.59***	0.58***	0.43***
					(0.42)	(0.10)	(0.06)	(0.06)	(0.10)

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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 7*b*: Gender gap in assessments - Sciences.

Subject Type	Subject	Grade	Observations	Schools	[0,4.5)	[4.5,9.5)	[9.5,13.5)	[13.5,17.5)	[17.5,20)
Sciences	Biology and Geology	11	408908	652	0.08***	0.24***	0.35***	0.20***	0.09***
					(0.02)	(0.01)	(0.01)	(0.02)	(0.03)
	Descriptive Geometry	11	72494	459	0.04	0.38***	0.46***	0.47***	0.33***
					(0.05)	(0.04)	(0.05)	(0.06)	(0.04)
	Geography	11	175901	613	0.17	0.41***	0.56***	0.63***	0.40***
					(0.09)	(0.02)	(0.02)	(0.03)	(0.07)
	Information and Technology	11	4132	123	0.99	0.46**	0.37*	0.27	0.67
					(0.66)	(0.17)	(0.14)	(0.15)	(0.34)
	Mathematics	12	420417	651	0.01	0.10***	0.25***	0.28***	0.17***
					(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
	Mathematics B	11	19685	269	-0.06	0.22***	0.32***	0.46***	0.52***
					(0.08)	(0.05)	(0.06)	(0.08)	(0.13)
	Mathematics SS	11	84472	531	-0.02	0.20***	0.48***	0.50***	0.30***
					(0.05)	(0.03)	(0.03)	(0.04)	(0.07)
	Physics and Chemistry	11	432694	653	0.14***	0.30***	0.48***	0.35***	0.19***
					(0.01)	(0.01)	(0.02)	(0.02)	(0.02)

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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.

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 so, also for the gender composition of the higher education population and the labour force.

5 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 other 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.

We contribute to the existing literature on grading systems and gender gaps by testing the gender difference in the difference between teacher grading and national exam scores. We use Portuguese data on 21 subjects across humanities and sciences for the whole population of students taking exams at the end of the 6^{th}, 9^{th}, 11^{th} and 12^{th} grades from 2007 to 2016.

Although, 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 systematically 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 higher 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 in the exam by around 0.5 more than boys in several subjects (in a 0-20 scale). We also obtain that for the subjects tested in the 6^{th}, 9^{th} and 12^{th} grades, the difference in the assessment gap between boys and girls seems to decrease as students progress from the 9^{th} to the 12^{th} grade.

Our results can be due to the fact that the skills under evaluation in the classroom are different from the ones that determine a good score in national exams, with girls doing "better" than boys in the former. It may also suggest that teacher assessment relies on aspects of student performance for which the skills required develop later for boys. To test this hypothesis further analysis is needed.

Goldin (2014) has documented the converging roles of men and women in different dimensions of the labour 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 potential impact on labour market characteristics. When used at younger ages, an evaluation system that benefits girls may motivate a larger percentage of boys to choose a vocational track, a decision that later on may restrain 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 of the labour force and particularly of the share of the labour 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.

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.

Cox, A. (2005). Boys of Few Words: Raising our Sons to Communicate and Connect. *The Guilford Press*.

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: It's 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.

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.

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.

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

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.

Pavlova, M. A., Wecker, M., Krombholz, K., and Sokolov, A. A. (2010). Perception of intentions and actions: gender stereotype susceptibility. *Brain Research*, 1311:pp.81–85.

Apendix

Table A1: Means and standard deviations of teacher grading by national exam score in grade 6 (2012-2015)

Subject	Number of Observations	Exam Score	% Girls	Mean Teacher Score		Mean Difference		T test for the difference in Mean differences [(8)-(7)]
				Boys	Girls	Boys	Girls	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Portuguese	3610	1	0.29	2.32	2.46	1.32	1.46	0.14***
				(0.01)	(0.02)	(0.01)	(0.02)	[5.59]
	114028	2	0.41	2.73	2.87	0.73	0.87	0.14***
				(0.00)	(0.00)	(0.01)	(0.01)	[13.90]
	190375	3	0.48	3.20	3.34	0.20	0.34	0.14***
				(0.01)	(0.01)	(0.01)	(0.01)	[12.74]
	106705	4	0.56	3.89	4.04	-.11	0.04	0.14***
				(0.01)	(0.01)	(0.01)	(0.01)	[8.74]
	11963	5	0.63	4.54	4.63	-.46	-.37	0.09***
				(0.01)	(0.00)	(0.02)	(0.01)	[4.45]
Mathematics	40319	1	0.46	2.16	2.20	1.16	1.20	0.04***
				(0.01)	(0.01)	(0.01)	(0.01)	[3.45]
	161831	2	0.49	2.65	2.70	0.65	0.70	0.05***
				(0.01)	(0.01)	(0.01)	(0.01)	[4.21]
	117409	3	0.50	3.31	3.42	0.31	0.42	0.11***
				(0.01)	(0.01)	(0.01)	(0.01)	[6.84]
	88845	4	0.49	4.10	4.14	0.05	0.14	0.10***
				(0.01)	(0.01)	(0.01)	(0.01)	[5.24]
	23839	5	0.46	4.70	4.74	-.30	-.26	0.05***
				(0.01)	(0.01)	(0.01)	(0.01)	[3.08]

Notes. The grading scale is 1 to 5. The mean difference in columns (7) and (8) is defined as mean teacher score minus mean exam score. In column (9) the difference tested is column (8) minus column (7). The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

Table A2: Means and standard deviations of teacher grading by national exam score in grade 9 (2007-2016)

Subject	Number of Observations	Exam Score	% Girls	Mean Teacher Score		Mean Difference		T test for the difference in Mean differences [(8)-(7)]
				Boys	Girls	Boys	Girls	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Portuguese	4863	1	0.30	2.35	2.60	1.35	1.60	0.25***
				(0.01)	(0.02)	(0.01)	(0.02)	[10.22]
	248042	2	0.43	2.78	2.91	0.78	0.91	0.13***
				(0.00)	(0.00)	(0.01)	(0.01)	[16.47]
	415394	3	0.52	3.09	3.23	0.09	0.23	0.14***
				(0.00)	(0.00)	(0.01)	(0.01)	[16.01]
	197194	4	0.62	3.73	3.89	-.27	-.11	0.16***
				(0.01)	(0.01)	(0.01)	(0.01)	[10.06]
	22048	5	0.69	4.51	4.58	-.49	-.42	0.08***
				(0.01)	(0.01)	(0.01)	(0.01)	[4.24]
Mathematics	103508	1	0.53	2.10	2.16	1.10	1.16	0.06***
				(0.01)	(0.01)	(0.01)	(0.01)	[6.22]
	340551	2	0.52	2.61	2.63	0.61	0.63	0.02*
				(0.00)	(0.00)	(0.01)	(0.01)	[1.98]
	225362	3	0.50	3.17	3.27	0.17	0.27	0.10***
				(0.01)	(0.01)	(0.01)	(0.01)	[9.12]
	166494	4	0.52	3.83	3.96	-.17	-.04	0.12***
				(0.01)	(0.01)	(0.01)	(0.01)	[7.57]
	56006	5	0.53	4.57	4.63	-.43	-.37	0.06***
				(0.01)	(0.01)	(0.01)	(0.01)	[4.14]

Notes. The grading scale is 1 to 5. The mean difference in columns (7) and (8) is defined as mean teacher score minus mean exam score. In column (9) the difference tested is column (8) minus column (7). The T statistic in square brackets reflects standard errors in parenthesis that are corrected for clustering at the school level.

v

Table A3: Secondary education - percentage of girls by subject and exam score intervals (2007-2016)

Subject Type	Subject	[0,4.5)	[4.5,9.5)	[9.5,13.5)	[13.5,17.5)	[17.5,20)
Humanities	English	0.76	0.83	0.83	0.73	0.68
	French	0.68	0.70	0.72	0.75	0.80
	Geography	0.81	0.71	0.62	0.58	0.51
	German	0.62	0.67	0.74	0.81	0.79
	History	0.77	0.72	0.70	0.70	0.69
	History B	0.50	0.45	0.41	0.43	0.40
	History of Culture and Arts	0.63	0.65	0.68	0.75	0.79
	Latin	0.81	0.73	0.75	0.74	0.77
	Philosophy	0.54	0.60	0.67	0.72	0.72
	Portuguese	0.46	0.51	0.60	0.68	0.72
	Portuguese Literature	0.62	0.70	0.76	0.82	0.83
	Spanish	0.56	0.59	0.67	0.74	0.77
Sciences	Biology and Geology	0.58	0.57	0.58	0.59	0.58
	Descriptive Geometry	0.57	0.53	0.49	0.45	0.39
	Information and Technology	0.66	0.48	0.43	0.31	0.26
	Mathematics	0.51	0.53	0.53	0.55	0.57
	Mathematics B	0.61	0.61	0.62	0.66	0.69
	Mathematics SS	0.68	0.70	0.71	0.74	0.76
	Physics and Chemistry	0.55	0.53	0.54	0.54	0.52
Other	Draw	0.52	0.62	0.65	0.69	0.70
	Economics	0.49	0.48	0.49	0.50	0.50

Note. The T statistic in square brackets reflects standard errors that are corrected for clustering at the school level.

Table A4*a*: Gender gap in assessments - Humanities and other.

Subject Type	Subject	Grade	Observations	Schools	(1)	(2)	(3)	(4)	(5)
Other	Draw	12	46214	298	0.346***	0.346***	0.346***	0.346***	0.345***
					(0.037)	(0.037)	(0.037)	(0.037)	(0.037)
	Economics	11	65718	459	0.417***	0.412***	0.419***	0.415***	0.414***
					(0.030)	(0.028)	(0.030)	(0.028)	(0.028)
Humanities	English	11	390	44	0.953*	0.938*	0.979*	0.957*	0.975*
					(0.378	(0.366)	(0.401)	(0.382)	(0.383)
	French	11	11064	299	0.211**	0.224**	0.212**	0.225**	0.227**
					(0.066)	(0.066)	(0.066)	(0.066)	(0.066)
	German	11	7929	185	-0.090	-0.082	-0.079	-0.071	-0.065
					(0.104)	(0.104)	(0.103)	(0.103)	(0.102)
	History	12	129639	583	0.632***	0.628***	0.632***	0.628***	0.629***
					(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
	History B	11	6617	173	0.573***	0.572***	0.573***	0.572***	0.577***
					(0.110)	(0.110)	(0.110)	(0.110)	(0.110)
	History of Culture and Arts	11	28210	294	0.119*	0.119*	0.125*	0.125*	0.121*
					(0.048)	(0.048)	(0.048)	(0.048)	(0.048)
	Latin	11	1698	85	0.324	0.309	0.329	0.313	0.305
					(0.189)	(0.193)	(0.189)	(0.193)	(0.191)
	Philosophy	11	45994	620	-0.253***	-0.258***	-0.236***	-0.237***	-0.230***
					(0.043)	(0.041)	(0.040)	(0.039)	(0.040)
	Portuguese	12	614559	658	-0.034*	-0.034**	-0.012	-0.012	-0.026*
					(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
	Portuguese Literature	11	17316	280	-0.120	-0.129*	-0.120	-0.130*	-0.141*
					(0.064)	(0.062)	(0.064)	(0.062)	(0.062)
	Spanish	11	17457	205	0.239***	0.243***	0.240***	0.244***	0.241***
					(0.052)	(0.052)	(0.052)	(0.052)	(0.052)
				Year	Yes	Yes	Yes	Yes	Yes
				Exam 1st Call	Yes	Yes	Yes	Yes	Yes
				Area Study	No	No	Yes	Yes	Yes
				Private School	No	Yes	No	Yes	Yes
				Exam Improval	No	No	No	No	Yes
				University Admission	No	No	No	No	Yes

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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 A4*b*: Gender gap in assessments - Sciences.

Subject Type	Subject	Grade	Observations	Schools	(1)	(2)	(3)	(4)	(5)
Sciences	Biology and Geology	11	408908	652	0.248***	0.247***	0.248***	0.247***	0.243***
					(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
	Descriptive Geometry	11	72494	459	0.918***	0.874***	0.404***	0.406***	0.409***
					(0.058)	(0.054)	(0.046)	(0.046)	(0.046)
	Geography	11	175901	613	0.847***	0.852***	0.814***	0.818***	0.816***
					(0.021)	(0.020)	(0.019)	(0.018)	(0.018)
	Information and Technology	11	4132	123	0.833***	0.831***	0.682***	0.680***	0.671***
					(0.147)	(0.148)	(0.126)	(0.127)	(0.127)
	Mathematics	12	420417	651	0.104***	0.103***	0.105***	0.104***	0.106***
					(0.016)	(0.016)	(0.016)	(0.015)	(0.015)
	Mathematics B	11	19685	269	0.125*	0.125*	0.124*	0.125*	0.131*
					(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
	Mathematics SS	11	84472	531	0.287***	0.284***	0.288***	0.286***	0.297***
					(0.031)	(0.031)	(0.031)	(0.031)	(0.031)
	Physics and Chemistry	11	432694	653	0.329***	0.327***	0.329***	0.327***	0.330***
					(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
				Year	Yes	Yes	Yes	Yes	Yes
				Exam 1st Call	Yes	Yes	Yes	Yes	Yes
				Area Study	No	No	Yes	Yes	Yes
				Private School	No	Yes	No	Yes	Yes
				Exam Improval	No	No	No	No	Yes
				University Admission	No	No	No	No	Yes

Note. Dependent variable is the assessment gap, $TeacherScore - ExamScore$. Standard errors in parenthesis 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.