UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Grading practices, gender bias and educational outcomes: evidence from Italy.

November 2, 2021

Abstract

We study if the Italian school system suffers from gender bias when judging students. To this aim, we use a differences-in-differences approach that compares the teachers' assessments and the standardized test scores that the students receive during the school year. We have census data for all Italian fifth and sixth graders in two different subjects, math and language that include a rich set of additional controls. Our evidence reveals that, since primary school, boys are graded less favourably than girls in both math and language. This result is also confirmed for middle school students (sixth graders), and it holds even when a) we separate the analysis between the most and least developed Italian regions, b) we control for possible gender specific-attitude towards cheating and teachers' manipulation, and c) we introduce class and school fixed effects in the models. Comparing the results obtained across different levels of schooling and subjects we cannot clearly identify the role of specific mechanisms in determining the gender bias. Overall the analysis suggests further study on the role of teachers' characteristics.

Keywords: Grading practices, Gender bias, Schooling outcomes, Italy.
J.E.L. Classification: I21, I24, O15.

1 Introduction

Is teachers' assessment influenced by students' gender? And, if so, is there any difference across different grades and subjects? These questions are of great interest as teachers' bias when assessing academic results may have long-lasting consequences on students' school performance and, through this, on their following labor market outcomes (Lavy and Sand, 2018; Terrier, 2020; Lavy and Megalokonomou, 2019).

International data show the presence of significant differences between boys and girls in educational attainments. In most countries, women with the same educational attainment as men are under-represented in many scientific and technical degrees, which typically lead to better-paid occupations (OECD, 2012; Eurydice, 2010). Moreover, when comparing students who perform equally well, boys are more likely to repeat school years. They are also the prevailing group among early school leavers, and data identify significant gender differences even in tertiary educational attainments in favor of girls (OECD, 2012, 2015).[1]

An extensive literature on education documents how these differences might be associated with teachers' gender bias in assessing students. In general, teachers' stereotypes and believes toward gender roles may affect students' attitudes, behaviors and through a self-fullfilling prophecies mechanism, their educational outcomes (Rosenthal, 1987; Stobart et al., 1992). There is a large literature describing how stereotypes might bias the teachers' interaction with pupils, and how in actual classroom practices teachers may reflect their biased views by giving different types of feedback to boys and girls (Ertl et al., 2017; Alan et al., 2018).

Further evidence also shows that boys are more likely than girls to be disruptive in class Matthews et al. (2009); Duckworth and Seligman (2006) and that teachers sanction these behaviours by giving boys lower marks. Again, this may produce a mechanism that leads to school failure and grade repetition since sanctions may further alienate boys from school (Jimerson et al., 2006; OECD, 2012, 2014).[2]

[1]The recent OECD averages tell that 51 % of 25-34 year-old women (34% in Italy) hold a tertiary qualification against only 39% of men (22% in Italy) (OECD, 2020)

[2]Data show that students who repeat a year are more likely to engage in high-risk behaviour and/or dropout. On

In this study we investigate whether there are systematic differences in teachers' grading between boys and girls in Italy. The Italian case represents an interesting context for two reasons. First, the Italian school system enables us to exploit rich administrative education data which includes a low-stake standardized test in addition to teacher assessed grades provided by the National Institute for the Evaluation of the Education System (henceforth INVALSI), an independent national public agency in charge for implementing the tests procedures. In Italy these tests are compulsory for all Italian students attending specific grades and provide the assessment of two different subjects, math and language. Our data cover almost all students attending two different grades of schooling during the schooling year 2010-11 and include a rich set of variables that enables us to analyse many important covariates at students, class and school level.[3] We investigate the presence of gender bias for both primary (fifth graders) and lower secondary school (sixth graders) students and look at how the bias evolves across school grades. In comparing the results, we consider that, despite there being only one year of difference, the two cohorts of students analysed belong to different levels of education, each with its own specificities (INDIRE, 2014) that may possibly influence the results. Second, the highly documented persistent duality between the developed North and the less developed South in Italy allows us to investigate the possible role of socio-economic and cultural factors. Womens' traditional role of wife and mother is still more persistent in the South of Italy, and this may be important for gender roles and expectations at school (Banca d'Italia, 2012; Albanese et al., 2016).

In order to test for the existence of gender bias in Italian schools, we follow the seminal contribution by Lavy (2008) and exploit the presence in our dataset of both teachers' assessment (our subjective or non-blind grades) and national standardized test results (our objective or blind grades) in two different subjects, math and language. Even if they cannot be considered fully blind tests, it is plausible to assume that the Italian standardized tests are "quasi-blind" objective assessments (Burgess and Greaves, 2013) since the INVALSI enforces a protocol for the administration to reduce discretion and the possibility of teachers' manipulation.[4] The presence of teachers' gender bias is

this, see also OECD (2014) and Lyche (2010).

[3]On the exclusion criteria and missing data see Section 1.3 in the Online Appendix.

[4]It is a written exam, nationally set and it is marked by an external teacher (for a sub-sample of students) or by another teacher of the same school. More on this in Section 3.

thus estimated using a differences-in-differences approach: we compare teacher-assigned grades and standardized test scores across genders, assuming that the latter are free from gender bias.

This methodology does not allow us to separate the effect of teachers' discrimination and stereotypes from the effect of a gender bias in teachers' grading practices due to the different design of the standardized tests and teachers' grades that measure different skills. There is ample literature that describes how different assessment techniques and the different testing environments affect outcomes in relation to gender, implying that girls and boys might perform differently to the two types of test, leading to a different gender gap in teachers' assessments and standardised tests (Murphy, 1982; Stobart et al., 1992; Burgess and Greaves, 2013).[5] However, we can still estimate an overall measure of gender bias, defined as the average difference between the non-blind and blind score for boys, minus the same gap for girls.

This issue is shared by the many papers that, as in our case, compare data on two similar but not identical measures of students' performance.[6] This literature has examined the relevance of different mechanisms to explain gender biases in grading. The role of the same-gender teacher (and other teachers' characteristics) has/have been found to be an important factor in Lavy (2008), Falch and Naper (2013) and Paredes (2014), but others studies find little or no effect (Holmlund and Sund, 2008; Breda and Ly, 2015; Borgonovi et al., 2018). More recently, Lavy and Megalokonomou (2019) use the value added approach to estimate an overall measure of teachers' effectiveness (Rivkin et al., 2005; Chetty et al., 2014) and find that gender bias is more prevalent among less effective teachers.

Additional research stresses the importance of the role of students' noncognitive skills measured using either students' self-reported motivations (Bonesrønning, 2008), or using other noncognitive skills indices reported by the teachers (Cornwell et al., 2013).[7] Finally, recent studies address the important issue of the effect of teachers' gender biases on students' progress and find that the latter leads to larger progress for the favoured gender several years later (Lavy and Sand, 2018; Lavy and Megalokonomou, 2019; Terrier, 2020).

Overall, most of these papers find that girls benefit from grading bias. Few exceptions to this rule

[5] See Section 2 for more details.

[6] An exception is Hanna and Linden (2012).

[7] Bonesrønning (2008) does not use the simple teachers' scores in the analysis. The focus is on the teachers' grading practices that are not directly observable, and are estimated using a model with the teachers' grade as dependent variable.

are present, such as Hinnerich et al. (2011) who find no gender bias, and Hanna and Linden (2012) that also represents the only study that focuses on a less developed country. Other analysis suggest more complex mechanisms, e.g., with discrimination that goes systematically against typical gender stereotypes (Breda and Ly, 2015) or depend on the students' socio-economic status (Cobb-Clark and Moschion, 2017).[8]

The evidence presented here complements this literature by showing the presence of teachers' grading bias against boys in Italy. To the best of our knowledge, we are the first to apply this methodology to examine this issue for the Italian case. This is also one of the few papers that focuses on census data for two different levels of schooling (primary and middle-school students) and two different subjects in a large industrialized country.[9] Indeed, compulsory testing represents an advantage in this setting since our data cover almost all Italian students and this reduces the possibility of selection problems. Additionally, the Italian case enables us to provide within-country evidence of the possible role of cultural factors and gender roles on the estimated teachers' gender bias.

This paper is also related to other studies that investigate the role of gender biased assessments using alternative methodologies to measure teachers' stereotypes against both a specific gender or ethnicity such as the use of the Implicit Association Tests (IAT) (Carlana, 2019; Alesina et al., 2018) or that exploit the comparison of pairs of siblings (Figlio, 2005).[10]

Finally, it also fits in the strand of literature that suggests that cultural factors and social conditioning affect gender differences in educational outcomes (Machin and Pekkarinen, 2008; Guiso et al., 2008; Else-Quest et al., 2010). These are cross-country studies that identify cultural differences as an important factor in determining gender imbalances in educational attainments and

[8] In this case the more "masculine" a subject is thought to be (such as math or physics), the more favored are the female candidates, while the opposite is true for the more "feminine" subjects (Breda and Ly, 2015). Cobb-Clark and Moschion (2017) find that girls in low and middle socio-economic status families have an advantage in reading, while boys in high socio-economic status families show an advantage in numeracy.

[9] See also Angelo and Reis (2021) with data on several cohorts of Portuguese students,Cobb-Clark and Moschion (2017) with data on Australian third graders and, even if the main focus of this study is on teachers' bias across ethnic minority and white pupils, Burgess and Greaves (2013) use five annual censuses of all state school pupils in England.

[10] Carlana (2019) conducted a survey to measure the teachers' gender stereotypes with IAT on a sample of 145 middle schools from 5 provinces in the North of Italy, and finds that stereotypes of literature teachers have no effect on reading performance of boys or of girls, while math teachers stereotypes have a negative influence on girls and no effect on boys.

stress the importance to pursue gender equality policies in order to narrow the gender gap in STEM studies.

Our results add to the growing body of evidence that shows that teachers' assessments act against male students and show that in Italy this is present in both math and language studies since primary school. The estimated bias against primary school boys is 0.19 of a standard deviation in language and 0.12 in math, while for middle school is 0.09 of a standard deviation in language and 0.22 in math. This main result of a negative and significant coefficient on the gender bias interaction term is robust to the inclusion of a rich set of covariates, of class and school fixed effects, and it holds in all our additional robustness checks.

We also replicate the analysis by firstly separating the sample between northern and developed regions and less developed southern ones and, then, by using a representative sample of monitored classrooms where external inspectors invigilate students during the standardized test. In the former case, our results are not consistent with the presence of cultural stereotypes against girls in southern Italian schools while, in the latter, they suggest the presence of teachers' manipulation during the standardized test procedures that might bias downward our gender bias estimates with census data. [11]

Overall, even if they are more suggestive rather than conclusive, our results add to the evidence that points to further investigate the role of the teachers' characteristics on grading gender bias, including their selection and training processes.

The rest of the paper is organized as follows. In Section 2 we briefly discuss the channels identified by the psychology and education literature between the presence of gender bias and school results. Section 3 describes the data. Section 4 discusses the empirical design, and Section 5 the main findings. Section 6 contains our robustness checks, and conclusions are in Section 7.

2 Interpreting the gender bias

In our setting, partial grading might be the result of three main factors, namely (i) the presence of teachers' stereotypes and discrimination, (ii) the use of different grading practices and (iii) the

[11] On the presence of manipulation and cheating behaviour during the standardized test procedures in the Italian school system, see Bertoni et al. (2013); Paccagnella and Sestito (2014).

students' discipline and behaviour.

The discussion below will provide a useful guide when we interpret our results.

2.1 Teachers' stereotypes and discrimination

There is a large literature describing how stereotypes might bias teachers' assessments and affect the interaction with pupils in class, i.e., they do not push a specific gender in class to excel or give more attention to boys/girls or praise and motivate them more or spend more time (Rosenthal, 1987; Ertl et al., 2017; Lavy and Sand, 2018). This behaviour might modify the students' expectations and self-image and finally influences their educational outcomes through a typical self-fulling prophecy mechanism (Rosenthal, 1987; Stobart et al., 1992).

Cultural stereotypes, e.g. about girls' science abilities ("girls are not good at math"), can also play an important role generating gender biased grading in specific subjects. Similar stereotypes may also hit boys in fields that are not stereotypically associated with them (Rosenthal, 1987; Ertl et al., 2017) and operate even without the intention to harm the stigmatized group (Bertrand et al., 2005). Moreover, gender bias could also arise when teachers try to *compensate* for a perceived discrimination of a specific gender. In this case, teachers would apply a policy of gender specific "easy grading" that affects the students' interest in a subject and/or the students' levels of effort thus influencing their performance (Hinnerich et al., 2011; Mechtenberg, 2009; Bonesrønning, 2008; Breda and Ly, 2015).

A number of papers have also identified more complex mechanisms linked to teachers' stereotypes. Dee (2005) stresses the importance of the teacher-student gender matching, and identifies a *passive* teacher effect that is not driven by an explicit teacher behaviour as described above.[12] Examples include the role-model effect, which occurs when the presence of the same-sex teacher raises a student's academic motivation, and the stereotype threat, arising when students expect stereotypes by a different-sex teachers and experience anxiety or uneasiness that negatively affect their achievement (Steele and Aronson, 1995; Dee, 2005).

Furthermore, another important characteristic in our analysis is the length of the student-

[12]Dee (2005) also identifies an active teacher effect described as biases in the teacher's "...prior expectations of and interactions with students who have different demographic traits", with mechanisms similar to those described above. See Dee (2005), p. 159.

teacher relationship. In fact, statistical discrimination results when economic agents have imperfect information about individuals they interact with, and we might expect that the more time teachers spend with their students, the less bias there is in their grades (Terrier, 2020). In our study, this point is important since a feature of this study is the comparison of two different cohort of students, the first that receives the scores by teachers who have known them for more than 4 years, the second from teachers who have known them for only for a few months.

2.2 Different grading practices

Partial grading might also results from different characteristics between the teachers' and the standardized test assessments. A number of studies stress that characteristics such as oral expression, self-confidence, anxiety or shyness are not gender neutral (Murphy, 1982; Stobart et al., 1992), and it also shows that the assessment techniques implemented, the type of the test (low or high stake) and even changes in the scale of the grading system, are possibly related to these characteristics and might thus affect outcomes in relation to gender.[13] This topic is related to the complex literature on how biology and experiences interact to produce skills and abilities that have both a genetic or an acquired character (Guiso et al., 2008; Cunha and Heckman, 2009; Stobart et al., 1992). We do not further investigate this issue here as an in-depth discussion of the nurture/nature debate goes beyond the scope of this paper.[14]

In our analysis we need to take into account that either the teachers' scores or the standardized test could be perceived by the students as a more pressured environment and different studies suggest that women are less effective than men under pressure or in competitive environments (Gneezy et al., 2003).[15] Moreover, the research suggests also the presence of significant differences in grading practices between math/science and language studies that might produce different gender biases (Prøitz, 2013; Stobart et al., 1992) and, even if the gap is narrowing, finds that boys do better than girls in multiple-choice items (Ben-Shakhar and Sinai, 1991).

[13] Hvidman and Sievertsen (2021) identify a different reaction between Danish boys and girls to a change in their high-stakes GPA caused by the implementation of a new grading system.

[14] See also Resh (2009); Prøitz (2013) for a survey of the empirical studies of teachers grading behavior and the nurture/nature debate to explain the gender differences in education.

[15] Carey et al. (2019) provide a literature review into the relationship between maths anxiety and performance and also report recent evidence on both Italian and UK students.

2.3 Students' discipline and behaviour

Finally, teachers use grades not only to assess the students' achievements on a subject but also to assess factors such as participation and behaviour in class and students' motivation. In Italy, the National Guidelines explicitly states for all schooling levels that *"Evaluation focuses on the learning process of pupils, their behaviour and their overall learning outcomes"* (INDIRE, 2014, p. 27).

Education scholars classify two main components in teachers' grading, a *universal* component, that is specifically related to the student's performance and knowledge, and a *differential* component that contains factors that are identified as gender non-neutral characteristics (Stobart et al., 1992; Mechtenberg, 2009). In fact, whether because of socialisation or innate differences, a number of studies find that boys are more likely than girls to be disruptive and physically active, and they are also less likely than girls to be able to delay gratification or set goals (Matthews et al., 2009; Duckworth and Seligman, 2006). Many teachers might sanction/reward these gender non-neutral skills by giving to students lower/higher marks and this would result in a difference between blind and non-blind scores in favour of girls.

Moreover, the differential component of grading is also more difficult to measure and more subjective, i.e. it is also more easily influenced by a stereotypical attitude towards boys or girls. This is why stricter guidelines for grading could, in principle, be effective in reducing the subjective part of the teacher's assessment and, thus, the gender bias. An example may be found in Norway that has recently changed the regulation for final grading. Norwegian teachers in all subjects are expected to assign final grades considering exclusively the students' performance and knowledge, the universal component of grading, without considering their motivation or attitude.[16]

3 Data and descriptives

We estimate the presence of gender bias on all Italian fifth and sixth graders. To this aim, we constructed two cross-sectional datasets with rich information on students, schools and areas characteristics with information referring to the 2010-11 school-year.

Our main data source is the census database provided by INVALSI. The INVALSI standardized

[16] See *The Knowledge Promotion reform* in 2006 and the revision of the Norwegian educational act, particularly the chapter on assessment and final grading amended in 2009 (Prøitz, 2013). On this, see also Bonesrønning (2008).

tests are compulsory for almost all Italian schools and students, both public and private, attending specific grades of schooling.[17] During the 2010-11 school year, the students with particular disabilities or specific learning disorders were exempted from the test, together with the educational institutions with fewer than ten students enrolled (INVALSI, 2011a).

The test is written, and the type of tasks that students have to complete includes multiple choice and closed-format short answer questions with a correction grid. In the latter case, teachers might have room for interpretation in assessing the answers and that can tell from the handwriting if the student is a boy or a girl, and this introduces the presence of a potential bias even in our blind test scores (Paredes, 2014). In order to reduce discretion and the possibility of teachers' manipulation, the INVALSI enforces a protocol for the administration of the standardized tests (INVALSI, 2011b). To this aim, the test is not administered by the class teachers but by other teachers of the school. Moreover, all school teachers are then simultaneously involved in the transcription process so that they cross-check each other.[18]

For both fifth and sixth graders the standardized tests are low-stake, since their scores are not taken into account for the students' final assessment but are mainly used for monitoring purposes. Therefore, when comparing the INVALSI with the teachers' scores, stress and anxiety should not play a significant role in our analysis as described in Section 2.2, since it is likely that the two types of assessments are equally stressful for Italian students.

Finally, the INVALSI identifies a representative sample of the population of Italian primary and lower secondary schools where the whole process is also administered and monitored by an external inspector. In our study, we use the data on all Italian students, but we also replicate the analysis on these two representative samples of inspected schools as a robustness check in order to control if the results are affected by possible students' cheating or teacher's manipulation.

This dataset includes also a measure of the score in both language and math assigned by the teachers during the first term. Unlike the standardized test grades that represent the result of a one-day exam, the teachers' assessment uses heterogenous grading practices that often refer to oral contributions and their scores are an overall assessment of a student during a four months period.

[17] More on this in the Online Appendix.

[18] For more on this see the Online Appendix and Lucifora and Tonello (2015).

Furthermore, as most studies in this field, the two assessments were performed at different times: the non-blind score is applied approximately four months after the teachers have begun instructing the class, and the time gap with the standardized test is about 3 months. Thus, in our analysis we assume that the educational environment remains almost unchanged throughout this period and that boys and girls have no different attitudes to studying during the school year. This assumption is confirmed by Lavy (2008), who documents that the gender grading gap is not influenced by the different grading timing.

As described in our Online Appendix, Section 1.3, our census datasets are not free from possible attrition problems due to missing observations. As detailed in Section 4, our identification is based on within-student variation in scores and the data include only students with both a teacher grade and a standardized score in a given subject. Hence, we lose all students observed only in one of the two scores. We first observe that, across all subjects and school levels, approximately 5% of the students miss the information on the teachers' scores. Second, in our 5th grade data, 5.6% of students in math and 4.4% in language has only the teacher's score, while for 6th graders, the percentage is approximately 4% in both math and language. The former case may be the results of the teachers' or students' mobility, while the absence of the standardized test score results may occur because, during the 2010-11 school year, the students with particular disabilities or specific learning disorders and the educational institutions with fewer than ten students enrolled were exempted from the test (INVALSI, 2011a). Overall, we consider these numbers reassuring and we do not expect significant selection problems.

Moreover, when we introduce control variables, we face a progressive loss of observations that reduce our sample by around 30%, which, as with other related papers in the field, can make some room for attrition and selection problems (Terrier, 2020). However, since student-level controls are not useful for the identification of gender bias, this potential issue will not bias our main estimates of interest (see on this Section 4).

Together with the students' assessment measures, during the schooling year 2010-11 the INVALSI has also collected a rich set of information through different questionnaires filled by the students, their families and by the school administrative staff.

Additional demographic information about the students includes: gender, citizenship (native,

first or second generation immigrant students), if she/he speaks a foreign language at home or an Italian dialect, her/his socio-economic background using the number of siblings, parental education and a dummy variable indicating whether her/his mother is a housewife.[19] The set of school characteristics lists the number of students per class and school, the proportion of female students per class, and a school-average socio-economics index. The latter is an index for student socio-economic background, analogous to the same one computed by OECD (the ESCS index) for the PISA tests.

In addition, the main dataset is also merged with additional variables from different sources that help to control for area characteristics, namely, the wealth level of the school catchment area (per capita value added), a measure of the level of criminality, a social capital indicator and macro-area dummies. In fact, previous studies find that geographical location is an important determinant of Italian students test scores, with students in the Northern area usually outperforming those living in the South and differences in both economic and cultural factors may play a role (Cipollone et al., 2010; Bratti et al., 2007; Di Liberto, 2008).

The full details of the variables can be found in Section A in the Appendix. See also the Online Appendix for more details on the Italian school system.

It is worth noting that there are differences between primary and middle schools that possibly influence our analysis.

These differences mainly relate to the teachers' characteristics that are not alike in the two levels of schooling. Unfortunately, the dataset does not contain any information on gender or other teacher characteristics and we only have the overall data on the percentage of women in primary and middle school.

A first notable difference concerns the percentage of female teachers. Italian data show that almost all primary school teachers are female (98%) and that this percentage is lower (even if still very high) in secondary schools (80%). Thus, if teachers' stereotypes are more likely with students who have different demographic traits, we may expect more gender bias against boys in our primary school evidence.

[19] First generation are students born abroad of foreign-born parents, while second generation students are native-born children of foreign-born parents.

Furthermore, even if the teachers' scores collected are both assigned at the end of the first semester, fifth graders' teachers got to know their students and have assessed them over a much longer period of time than their middle school colleagues. In details, primary school teachers usually follow students from grade 1 to grade 5, while middle school teachers from grade 6 to grade 8. Therefore, for year six students, the teachers' scores are assigned in the first semester of the first year of the middle school program, i.e. when teachers most likely do not yet know their students well. Conversely, fifth graders are assessed at the end of their five-year primary school program and, in most cases, they had the same teachers that know them well. In addition, with respect to primary school, in lower secondary schools there are also more teachers per class and each spends less time with each pupil. Therefore, according to the Section 2 discussion, this would imply the possibility of more gender bias due to statistical discrimination in our sixth graders cohort (Guryan and Charles, 2013; Terrier, 2020).

A final important difference between the teachers in the two levels of schooling refers to their training and qualifications: while lower secondary school teachers must hold a Master degree and they are specialists in one or few more subjects, primary school teachers are generalists and must hold a Master on primary education sciences.[20]

Tables 1a and 1b summarise the major characteristics of the variables used in regressions for, respectively, the fifth graders and the sixth graders. Comparing the two tables we find that the socio-economic characteristics of the two cohorts are, as expected, almost identical. With respect to class size, the rules between primary and lower secondary school are not alike and imply, as observed, smaller classes in primary school.

The main differences are observed for the scores (both teachers and standardized tests) in math and language that are higher and less dispersed for the fifth graders.

A number of studies indicate that middle school environments are more focused on grades and performance than are elementary schools and show that primary school teachers assign higher grades than do their middle school counterparts (e.g., Anderman et al. (1998)). This is also true for the standardized test scores: in fact, unlike primary school, the sixth grade tests introduce questions that are able to measure more accurately higher levels of the students' ability scale (INVALSI,

[20]Prior to 2000, they could also teach without a Master degree qualification. More on this in the Online Appendix.

2011b).

Together with numerous variables that identify her/his socio-economic status, the INVALSI also collects information on each students' beliefs and motivation in learning, such as students' *self-efficacy* and their *intrinsic motivation* in studying, characteristics that the existing psychology and education studies indicate as gender specific and consider being consistently related with academic achievements (Fan and Williams, 2018; Gilman and Anderman, 2006). The psychology and education literature suggests that students' competency beliefs and motivation in academic studies are not gender neutral, and they vary according to the subject being examined (Fan and Williams, 2018). Similar evidence is found in recent analysis that shows that women are less confident about their own ability in math and science than men, and this contributes to explain the observed differences in their academic performance and career choices (Bordalo et al., 2019; Carey et al., 2019).

Here we separately include some further descriptives on these measures of students' drive and motivation in studying (Tables 2a and 2b). These items were adapted from the OECD-PISA survey, and comprise a variety of psychological scales that specifically address the areas of intrinsic interest and self-efficacy. Two main areas of academic motivation were examined: questions A, B and C capture the students' self-efficacy in language and math, while questions D and E measure the intrinsic motivation of the students in learning language and math. Self-efficacy refers to the individuals' beliefs regarding their personal capabilities in learning a specific subject, while intrinsic motivations refer to the drive to perform an activity purely for the joy gained from the activity itself. As also found in OECD-PISA surveys, the numbers in Tables 2a and 2b show that the subject-specific propensity for learning and achieving is very different between boys and girls and that differences are present since primary schooling: Italian boys are more confident and enjoy more studying math, while girls are more confident in language studies.

Tables 3a and 3b check if there is a misalignment between the INVALSI standardized scores and the teachers' grades. Since the two scores are given in different grading scales, both test scores were standardized to a distribution with mean zero and standard deviation of one.[21] Tables 3a and 3b show that, for both primary and secondary school children, the standardized test scores in

[21]The INVALSI test results give grades on a scale from 0–100 while the teachers' assessment has a 0-10 range.

Italian language are always higher than the teachers' grades (or non-blind scores) for both for boys and girls. The opposite is true for math. Second, numbers show that in math boys outperform girls in the blind test scores but that, when assessed by teachers, girls obtain on average a higher score. For language the picture is different as girls are always better than boys in both types of tests. Moreover, the standard deviation values seem to confirm what the education experts call as the *spread* phenomenon, i.e. the scores on both teachers' assessment and standardized test are in most cases more spread out for boys than for girls (Sommers, 2013; Machin and Pekkarinen, 2008). The only exception is for the blind test in math primary school, where boys and girls show the same standard deviation values.

We finally compare the kernel-density distribution of the two types of z-scores by grade, gender and subject. Figure 1a shows that in primary school the direction of the difference between the distribution of the blind and non-blind scores is very similar between the two subjects. Indeed, for both subjects, the two distributions almost overlap for female students, i.e. the blind and non-blind assessments are consistent, while for males the teachers' scores distribution is always shifted to the left.

For sixth graders the picture is more heterogeneous: the teachers' and the blind scores distributions overlap for females in language and for males in math. But the final result in both subjects seem to be always detrimental for boys. In fact, as in primary school, in language the teacher-score distribution for boys is shifted to the left. Conversely, where the blind and non-blind score distributions are consistent for boys, we observe a rightward shift of the teacher-score distribution for the girls, i.e., there is evidence that teachers tend to inflate girls' scores in math (Fig. 1b).

4 Methodology

To measure the teachers' gender bias we use a double-difference strategy first introduced in a seminal paper by Lavy (2008). This approach assumes that the non-blind (NB) teachers' score y_{NB} in one subject for the student i is a linear function of a constant term, the students' ability, the teachers' potential prejudice of gender and an error term:

$$y_{iNB} = \alpha_{NB} + \beta Male_i + \gamma ability_i + \upsilon_{iNB} \tag{4.1}$$

where $Male$ is an indicator taking value 1 if student i is male and 0 otherwise. Assuming that ability is not correlated with gender, we could interpret β as a discrimination effect. If negative, male students are discriminated and, if positive, female students are discriminated. In other words, if discrimination does not occur, students producing the same quality of the test should get the same score, regardless of their gender. However, since ability is unobservable and can be correlated with gender, we need to solve the standard omitted-variable bias problem. We can take care of this endogeneity problem by considering an objective assessment, where the examiner has no information about the student's gender, which implies that for the blind-score case (B) the β coefficient is equal to zero in (1.1).

$$y_{iB} = \alpha_B + \gamma ability_i + \upsilon_{iB} \tag{4.2}$$

The difference between 4.1 and 4.2, assuming that both tests measure in the same way the same ability, gives us the standard difference-in-differences (DID henceforth) strategy where any individual characteristics, such as ability, are differenced away and β measures solely a discrimination effect:

$$\Delta y_i = \alpha + \beta Male_i + \upsilon_i \tag{4.3}$$

As described in Section 3, in our analysis, given the INVALSI testing protocol, we assume that the standardized test score is our objective score, while the subjective teacher's score may possibly reflect gender biased grading practices. However, the format of the two tests differs and thus might not measure the same abilities. If abilities are different between boys and girls, this strategy will yield biased estimates of gender discrimination. Finally, Section 3 also describes additional differences between the two scores, such as the test taking environment and the different dates at which they are filled.

Again, if boys and girls respond differently to stressful situations or competitive environments, the blind score cannot be seen as the counterfactual measure to the non-blind score to isolate teacher discrimination, but we can still estimate an overall measure of gender bias, defined as the average difference between the non-blind and blind score for boys, minus the same gap for girls. To this

aim, we exploit the data pooled over the two types of scores, and use the interaction formulation of the standard DID estimation strategy, separately for each subject, that leads to an algebraically equivalent estimation of β:[22]

$$y_{ij} = \alpha + \delta Male_i + \gamma T_{ij} + \beta(Male_i \times T_{ij}) + \upsilon_{ij} \tag{4.4}$$

where y_{ij} is an indicator of performance of student i in the test j where j is either the blind test or the teachers' grading procedure in language or math, $Male_i$ is the gender dummy (equal to one if male), and T (teacher) is the dummy identifying the teacher's score. Thus, the intercept is the average score obtained by female students on the standardized test scores, δ captures the score difference of male students in both types of tests, and γ measures the teachers' effect, that is, the average differences in scores due to the type of tests. The parameter of interest is the interaction term, β, that measures, given the double-differences strategy, the average difference between non-blind and blind scores for boys, minus this same gap for girls. As a result, we could interpret β as gender bias. If negative, the bias is toward male students and if positive, female students. Finally, note that the DID nature at the student level of the estimation of Eq. (4.4) implies that co-variates (or class and school fixed effects) are implicitly assumed away, as long as they have the same effect on the blind and non-blind score. The DID specification of Eq. (4.4) should, in fact, saturate all these effects, and lead to identical estimates of our main parameter of interest β.

5 Results

Tables 4 show the results in language studies for fifth (Tables 4a) and sixth graders (Tables 4b), while Tables 5a and 5b introduce the corresponding analysis for math. Standard errors are always clustered at class level. As said above, despite there being only one year of difference, the two

[22] That is, for both math and language, the data set is a stacked file including the teachers' and standardized tests' scores and the number of observations in our regressions will be twice the number of students.

cohorts of students analysed belong to different levels of education, each with its own specificities (INDIRE, 2014), implying that we do not necessarily expect similar results for the two levels of schooling. Looking at how the bias differs across school grades might thus provide interesting insights.

We set the scene with model 1 that includes the results of our most parsimonious specification.

In language studies, the coefficients on the male dummy for both levels of schooling (Tables 4a and 4b) show that, on average, female students in Italy always had higher achievements on the standadized test: our results show that girls have advantages of 0.06 (in primary school) and 0.209 (in middle school) of a standard deviation of the blind score distribution. A different picture emerges for math (Tables 5a and 5b) where the advantage is for male students: the coefficient on the gender dummy is positive and statistically significant in primary school (0.08 of a standard deviation) and middle school (0.125). Furthermore, the mean difference between the teachers' scores and the standardized test scores is always positive and significant for all levels of schooling and subjects. Finally, our main parameter of interest, the estimated coefficient of the interaction between the male students and the teachers score dummies, is always negative and significantly different from zero.

In sum, we observe a similar pattern in primary and middle school: for language studies, results indicate that teachers' scores widen an already existing female-male achievement difference, while, in math, the gender bias seems to overrule the male-female difference existing in the standardized test. All together this suggests that, starting from primary school, the teachers' grades act against boys in all subjects, confirming for Italy a piece of evidence already found for other countries.

However, the evidence from the two levels of schooling is similar but not identical. On the whole, sixth grade's gender bias coefficients are greater than those estimated with the fifth graders data. As discussed in Section 3, this evidence is consistent with statistical discrimination, since both language and math lower secondary schools teachers spent less time with each pupil than their primary school counterpart. Moreover, the estimated gender bias in primary school (0.19 of a standard deviation in language and 0.12 in math) is lower in math, while we find the opposite in middle school (0.09 of a standard deviation in language and 0.22 in math).

This inconsistency between the two levels of schooling in the two subjects seems to suggest that

the gender bias depends more on teachers than on students' characteristics. The two cohorts of students are, in fact, one year apart and, as seen in the descriptives, they are also very similar in terms of observed characteristics. Instead, as seen in the previous Section, the characteristics of the Italian primary and middle school teachers are not alike, especially in terms of their different selection and training and this might help explain our evidence. Nevertheless, more information on teachers' characteristics and future research is needed to further investigate these issues.

Models from 2 to 5 in Tables 4 and 5 (a and b) show the results when we introduce more variables in our models. To save on space, in this paper we only introduce the results of our main coefficients of interest, while the full set of results is reported in the Online Appendix. Model 2 includes a set of regional dummies

and it also introduces different variables that control for students' demographics. Model 3 further increases the specification with more family characteristics in order to take into account for the student's socio-economic background, while model 4 includes the school average socio-economic background (calculated by the ESCS index), the school size and the proportion of girls in each class. The latter variable enables us to analyse possible gender peer effects, and previous evidence shows that an increase in the proportion of girls improves both boys' and girls' cognitive outcomes (Lavy and Schlosser, 2011; Lee et al., 2014). Finally, model 5 contains additional area characteristics (standard proxies of regional economic performance and crime, and a social capital indicator) identified as important factors for explaining educational and cultural differences for Italian regions (Cipollone et al., 2010; Bratti et al., 2007; Di Liberto, 2008).

Overall, the results on these additional variables are consistent across grades and levels of schooling and their coefficients always show the expected sign. Moreover, as said in Section 4, the DID specification should saturate all these effects, and models from 2 to 5 should therefore lead to identical estimates of our parameter of interest β. As expected, we do not observe significant changes in our main coefficient of interest with respect to model 1 results.

6 Robustness checks

Since including new variables produces a progressive loss of observations in the sample due to missing information in the additional covariates, we firstly check if attrition might imply the presence

of selection problems in our estimates of Tables 4 and 5 (a and b). In principle, this might be the case since a number of variables have been collected by the school administrative staff and through a questionnaire filled by the students' parents. Therefore, it is likely that that the families that did not fill out the student questionnaire are the most socio-economically fragile, and that most missing observations are in less organized and possibly less performing schools and this might also affect our gender bias estimates. However, this is not the case: when we run the robustness check on the reduced samples, we find no significant changes in our main coefficients. The full set of results can be found in the Online Appendix.[23]

We then replicate the analysis separately for two Italian macro-areas, the richer North and the less developed South. As seen in Section 2, a large literature suggests that teachers use marks also to signal their judgment on behaviour in class and boys tend to show more disruptive classroom behaviour than girls. If this is more so particularly among boys of low socio-economic background (Borgonovi et al., 2018), then we might expect more bias against boys in both subjects in the poorer southern regions.

We might also expect differences in the two areas since gender roles might affect our coefficient of interests, and many indicators suggest that women's traditional role of wife and mother is still more persistent in the South of Italy than in the North. For example, labor market outcome indicators show that the employment rate among females aged 15 to 64 years in Italy ranges from about 60% in the northern regions to 30% in the south, while for males the gap is significantly lower (74% north vs 61% south).[24] An almost identical regional heterogeneity can be found using alternative gender equality indexes that also focus on different spheres of social life (Banca d'Italia, 2012).

All else equal, if cultural stereotypes against women play a role even inside schools, we might observe a change in the sign of our coefficient of interest, or less gender bias for boys, in the South. Alternatively, if teachers' stereotypes are also differentiated by subjects, such as "boys are good in math and science, while girls in humanities" we might observe less gender bias in math and possibly more gender bias in language against boys in the southern regions.

[23]For each level of schooling and subject, we use the same reduced sample of students with no missing values in the full set of controls. See Tables A5.1, A5.2, A6.1 and A6.2.

[24]ISTAT (Italian national institute of statistics), 2020.

Table 6a reports the results for Language in primary school, with Panel A and B showing the coefficients respectively of the northern and southern regions, Table 6b includes the same evidence for lower secondary school students, The same analysis is then replicated for math, and the full set of results is in Tables 7a (primary school) and 7b (middle school). This analysis confirms the presence in all subjects, grades and areas of a gender bias against boys: the coefficients on the interaction term are always negative and significant in all specifications. Moreover, when we compare the results between the two levels of schooling, this robustness check does not reveal a consistent pattern between the two areas of the country. We find that primary school results show larger gender bias against boys in the northern regions in all subjects (0.20 vs 0.18 of a standard deviation for language and 0.13 vs 0.12 for math), while the middle school results imply the opposite, i.e., again, a more favorable teachers' bias for girls in the South in all subjects (0.08 vs 0.10 of a standard deviation for language and 0.22 vs 0.24 for math).

This evidence does not identify a clear role of cultural stereotypes in schools. This analysis also seems to rule out a potential channel for the observed regional differences identified by Lavy and Megalokonomou (2019) who find that grading biases are larger for teachers of lower quality. In fact, if teachers were less effective in one area of the country, e.g. in the less developed South, we would expect to find larger biases in the same area of the country for both levels of schooling.

As a third step, we check if our results are affected by students' cheating and teachers' manipulation during the standardized test procedure. The dataset enables to identify a representative and random sample of monitored classrooms where external inspectors invigilate students during the test. This is an important feature of our dataset for two reasons. First, previous evidence shows that the attitude towards cheating may be differentiated by genders and Italian students in the non-monitored classrooms receive a more benevolent supervision, allowing student cheating behavior more easily (Lucifora and Tonello, 2015; Bertoni et al., 2013; Paccagnella and Sestito, 2014). Second, for this sample, we are also fully confident that the INVALSI test protocol has been thoroughly implemented and that teachers did not manipulate the scores and, eventually, discriminate by gender.[25]

[25]For this sample the inspectors verify all the steps of the procedure within the school, including the check of the tests results.

Cheating constitutes a violation of the academic norms, and evidence shows that men tend to perceive minor deviance and risk-taking to be part of the male gender role and, more important for us, that males are more likely to cheat than females (e.g., Baird Jr (1980)). If boys cheat more, we expect that estimates of their performance in the standardized test in the full dataset are biased upwards, therefore inflating the resulting gender bias estimates in Tables 4 and 5. In other words, if this mechanism affects our estimates, we expect to find smaller gender bias against boys in our inspected school sample.

Conversely, the presence of teachers' manipulation during the standardized tests would imply the opposite, i.e. a higher gender bias against boys in the inspected school sample. Moreover, even if the INVALSI are low-stake tests, the presence of external invigilators during the test might be perceived by the students as a more stressful situation (Montalbán and Sevilla, 2019). As described in Section 5, this may translate into a lower relative female performance (and a corresponding higher gender bias) in our sample of monitored classrooms.

Panel C results (in Tables 6 and 7 for language and math, respectively) show the results for the sample of inspected primary schools (Tables 6a and 7a) and middle schools (Tables 6b and 7b). The estimated coefficients on the interaction term confirm the main result of a gender bias against boys. Moreover, for all grades and subjects, we find that the new estimated coefficients are always higher than those found in the full sample. This suggests a greater role for the teachers' behaviour (and manipulation) that might bias downward our previous gender bias estimates, and evidence of the presence of teachers' manipulation during the standardized tests in Italy has been found by Pereda-Fernández (2019).[26] However, we cannot also rule out the alternative hypothesis that girls performance is negatively affected by the different testing environment.

We also replicate our analysis including fixed effects at class and school level. Class fixed effects should capture all unobservables which are constant within class that might possibly affect the gender bias, including teacher's characteristics such as, for instance, her/his severity or the student/teacher ratio, while school fixed effects control for unobserved factors that are shared by all individuals within the same school, e.g., the school principal management practices (on this, see

[26]This analysis uses data for the academic year 2012-13 and for different grades.

Di Liberto et al. (2015)).[27] In principle, this inclusion should not affect our gender bias estimates since, as described in Section 4, our identification strategy saturates class or school fixed effects as long as they have the same effect on the blind and non-blind score. Results are in the Online Appendix, Panel D and E (Table A7.1 and A7.2 for language and Table A8.1 and A8.2 for math) and confirm that when controlling for class and school fixed effects the gender gap coefficient is substantially invariant.

As a final robustness check, we further introduce in the set of additional controls discussed in Section 5 two measures of non-cognitive skills (model 6 in Tables 4a, 4b, 5a and 5b). Using the data on non-cognitive skills described in Section 3, we construct two dummy variables (for math and language) that should capture two different areas of self-reported non-cognitive abilities of the students, namely self-efficacy and intrinsic motivations. These are students' self-assessed measures collected at the same time as our score indicators, implying that these variables are potentially endogenous to student achievement. Hence, since we could not find a credible instrument for performing an IV analysis, this evidence must be seen as an exploratory exercise and the results interpreted accordingly.

Including these measures of motivation for studying did not change our results on teachers' gender bias. As expected, both Tables A2 and A3 in the Online Appendix show that our measures of self-efficacy and intrinsic motivation in studying a specific subject are positively related to the students' educational attainments in the same subject for both fifth and sixth graders. We have also introduced the cross-subject dummies (Bonesrønning, 2008), and we always find that more confidence and motivation in studying math/language is also positively related with the students' grades in language/math. The only exception is identified for sixth graders in math, where we find a negative on the cross-subject dummy. This might suggest that students who are motivated at studying humanities respond by reallocating their efforts from maths to their favorite subjects, but we do not further investigate this issue that represents an interesting suggestion for future research.[28]

[27] It is worth noting that, in our data, we observe only one class per school in 25% of classes in 5th grade and 22% in 6th grade, where class and school fixed effect would be overlapping.

[28] A similar result was found by Bonesrønning (2008), while evidence of cross-disciplinary effects for sixth grade Italian students has also been found in Meroni and Abbiati (2016).

We also replicate the analysis for two subsamples of students that share the same level of attitude for learning a specific subject: the first one selects the group of students who are very confident in studying and being proficient in a specific subject (self-efficacy), while the second includes those who strongly disagree with this statement. Results are reported in the Online Appendix in Tables A7.1 for fifth graders and A7.2 for six graders and language, and Tables A8.1 for fifth graders and A8.2 for six graders and math, with Panel A showing those for the students with high levels of the self-efficacy index (approximately 20% of Italian sixth graders), and Panel B including those for the group with a low attitude for studying (approximately 4-5 %).[29] The coefficient on the interaction term, always negative and significant, confirms that the teachers' bias is always against the boys. Interestingly, we find a different picture for the two subjects. For language (Tables A7.1 and A7.2), the estimated coefficients are in fact similar across the two subgroups of students in both primary and middle school, whereas for math we do not find a consistent pattern between primary and lower secondary school: for lower secondary school the results suggest that math teachers tend to punish more boys with higher self-esteem, while for primary school we find the opposite interpretation. Nevertheless, since endogeneity problems are likely to plague our estimates, this exercise must be seen as food for thought for future research on the important issue of the role of non-cognitive skills.

7 Conclusions

Gender and grading are key dimensions to consider to better target educational interventions since grades constitute the main feedback for students (and families) about their academic performance and may represent a crucial factor in their following schooling and labor market decisions. This study investigates if Italian teachers' evaluation practices are biased against a specific gender in two different levels of schooling: primary school (with data on fifth graders) and lower secondary school (sixth graders). To this aim, we apply a DID approach using information on teachers' (subjective) grades and the standardized test (objective) scores. Standardized tests are compulsory in the Italian school system, and we can exploit administrative data and their rich set of variables for the

[29] In detail, Panel A only includes students who strongly agree with the statement *I am proficient in math/language*. In panel B, the results are obtained using the sub-sample of students that strongly disagree with the same statement. For more on this, see Table 2a and 2b.

full set of Italian fifth and sixth graders. We thus compare two cohorts of students of similar age that attend two different levels of schooling, each with its specificities, and we can look at how the bias differs across school grades.

We find evidence that the same direction of gender bias exists in two very different subjects, math and language. In most studies, these subjects exhibit outcomes differentiated by gender in the standardized tests and Italy is no exception. While on average, boys outperform girls in math in the blind test scores, on the contrary, when assessed by the teachers, girls obtain the highest score. In the case of language studies, the picture is different: girls are always better than boys in both types of tests. Our analysis further reveals that teachers' grading is biased in favor of girls in both subjects, and we observe a similar pattern in primary and middle schools. For language studies, results indicate that teachers' scores widen an already existing female-male difference. In math, the gender bias overrules the male-female difference observed in the standardized test. This suggests that, starting from primary school, the teachers' grades act against boys in all subjects, confirming for Italy evidence already found for other countries. This main result is also confirmed by a series of robustness checks. First, it is robust to the inclusion of additional variables and class and school fixed effects. Second, the same evidence of a gender bias against boys holds even when we separate the analysis between the most and least developed Italian regions, and when we focus on a sample of inspected schools to control for possible gender-specific attitude towards students' cheating and teachers' manipulation.

When we compare the results obtained across the different levels of schooling and subjects and different robustness checks, we cannot identify the role of specific mechanisms that may be at the origin of gender bias, but our evidence provides some suggestive, even if not conclusive, hints. Overall, this paper contributes to the literature on gender bias grading, suggesting that boys difficulties at school in Italy might be, at least partly, driven by grading biases and that more investigations are needed on the role of the teachers' characteristics. In particular, the comparison of the results obtained across different levels of schooling, subjects and regions suggests further investigating the role of teachers selection and training processes on gender biased grading, an area where there is still little research. Up to now, the cross-sectional nature of the dataset and the lack of information on teachers' characteristics represent two important limits of this analysis that

prevented us from further investigate these issues, and we leave a more in-depth analysis to future research.

References

Alan, S., S. Ertac, and I. Mumcu (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics 100*(5), 876–890.

Albanese, G., G. De Blasio, and P. Sestito (2016). My parents taught me. evidence on the family transmission of values. *Journal of Population Economics 29*(2), 571–592.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2018). Revealing stereotypes: Evidence from immigrants in schools.

Anderman, E. M., T. Griesinger, and G. Westerfield (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology 90*(1), 84.

Angelo, C. and A. B. Reis (2021). Gender gaps in different grading systems. *Education Economics 29*(1), 105–119.

Baird Jr, J. S. (1980). Current trends in college cheating. *Psychology in the Schools 17*(4), 515–522.

Banca d'Italia (2012). Relazione annuale sul 2011. *Roma, Banca d'Italia*.

Ben-Shakhar, G. and Y. Sinai (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement 28*(1), 23–35.

Bertoni, M., G. Brunello, and L. Rocco (2013). When the cat is near, the mice won't play: The effect of external examiners in italian schools. *Journal of Public Economics 104*, 65–77.

Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review 95*(2), 94–98.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research 60*(3), 245–264.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

Borgonovi, F., A. Ferrara, and S. Maghnouj (2018). The gender gap in educational outcomes in norway.

Bratti, M., D. Checchi, and A. Filippin (2007). Geographical differences in italian students'mathematical competencies: Evidence from pisa 2003. *Giornale degli Economisti e Annali di Economia*, 299–333.

Breda, T. and S. T. Ly (2015). Professors in core science fields are not always biased against women: Evidence from france. *American Economic Journal: Applied Economics 7*(4), 53–75.

Burgess, S. and E. Greaves (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics 31*(3), 535–576.

Carey, E., A. Devine, F. Hill, A. Dowker, R. McLellan, and D. Szucs (2019). Understanding mathematics anxiety: investigating the experiences of uk primary and secondary school students.

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

Cartocci, R. (2007). Mappe del tesoro: atlante del capitale sociale in italia. *168*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–79.

Cipollone, P., P. Montanaro, and P. Sestito (2010). Value-added measures in italian high schools: problems and findings. *Giornale degli economisti e annali di economia*, 81–114.

Cobb-Clark, D. A. and J. Moschion (2017). Gender gaps in early educational achievement. *Journal of Population Economics 30*(4), 1093–1134.

Cornwell, C., D. B. Mustard, and J. Van Parys (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources 48*(1), 236–264.

Cunha, F. and J. J. Heckman (2009). Investing in our young people. *Investing in our Young People*, 387–417.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review 95*(2), 158–165.

Di Liberto, A. (2008). Education and italian regional development. *Economics of Education Review 27*(1), 94–107.

Di Liberto, A., F. Schivardi, and G. Sulis (2015). Managerial practices and student performance. *Economic Policy 30*(84), 683–728.

Duckworth, A. L. and M. E. Seligman (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of educational psychology 98*(1), 198.

Else-Quest, N. M., J. S. Hyde, and M. C. Linn (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin 136*(1), 103.

Ertl, B., S. Luttenberger, and M. Paechter (2017). The impact of gender stereotypes on the self-concept of female students in stem subjects with an under-representation of females. *Frontiers in psychology 8*, 703.

Eurydice, N. (2010). Gender differences in educational outcomes: Study on the measures taken and the current situation in europe. *Education, audiovisual, and culture executive agency*.

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

Fan, W. and C. Williams (2018). The mediating role of student motivation in the linking of perceived school climate and achievement in reading and mathematics. *3*, 50.

Fiaschi, D., L. Gianmoena, and A. Parenti (2011). The dynamics of labour productivity across italian provinces: convergence and polarization. *Rivista italiana degli economisti 16*(2), 209–240.

Figlio, D. N. (2005). Names, expectations and the black-white test score gap. *National Bureau of Economic Research Cambridge, Mass., USA*.

Gilman, R. and E. M. Anderman (2006). The relationship between relative levels of motivation and intrapersonal, interpersonal, and academic functioning among older adolescents. *Journal of School Psychology 44*(5), 375–391.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics 118*(3), 1049–1074.

Guiso, L., F. Monte, P. Sapienza, and L. Zingales (2008). Culture, gender, and math. *SCIENCE-NEW YORK THEN WASHINGTON- 320*(5880), 1164.

Guryan, J. and K. K. Charles (2013). Taste-based or statistical discrimination: the economics of discrimination returns to its roots. *The Economic Journal 123*(572), F417–F432.

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

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

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

Hvidman, U. and H. H. Sievertsen (2021). High-stakes grades and student behavior. *Journal of Human Resources 56*(3), 821–849.

INDIRE (2014). The italian education system. *I QUADERNI DI EURYDICE 30*.

INVALSI (2011a). Il servizio nazionale di valutazione, le rilevazioni degli apprendimenti a.s. 2010-11. *Technical Report*.

INVALSI (2011b). Rapporto tecnico sulle caratteristiche delle prove invalsi 2011. *Technical Report*.

Jimerson, S. R., S. M. Pletcher, K. Graydon, B. L. Schnurr, A. B. Nickerson, and D. K. Kundert (2006). Beyond grade retention and social promotion: Promoting the social and academic competence of students. *Psychology in the Schools 43*(1), 85–97.

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

Lavy, V. and R. Megalokonomou (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. *National Bureau of Economic Research*.

Lavy, V. and E. Sand (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.

Lavy, V. and A. Schlosser (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics 3*(2), 1–33.

Lee, S., L. J. Turner, S. Woo, and K. Kim (2014). All or nothing? the impact of school and classroom gender composition on effort and academic achievement.

Lucifora, C. and M. Tonello (2015). Cheating and social interactions. evidence from a randomized experiment in a national evaluation program. *Journal of Economic Behavior & Organization 115*, 45–66.

Lyche, C. S. (2010). Taking on the completion challenge: A literature review on policies to prevent dropout and early school leaving.

Machin, S. and T. Pekkarinen (2008). Global sex differences in test score variability. *Science*.

Matthews, J. S., C. C. Ponitz, and F. J. Morrison (2009). Early gender differences in self-regulation and academic achievement. *Journal of educational psychology 101*(3), 689.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The review of economic studies 76*(4), 1431–1459.

Meroni, E. C. and G. Abbiati (2016). How do students react to longer instruction time? evidence from italy. *Education Economics 24*(6), 592–611.

Montalbán, J. and A. Sevilla (2019). The gender gap in student performance: The role of the testing environment. *Mimeo*.

Murphy, R. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology 52*(2), 213–219.

OECD (2012). Equity and quality in education: Supporting disadvantaged students and schools. *OECD Publishing*.

OECD (2014). Ready to learn: Students' engagement, drive and self-beliefs. *OECD Publishing*.

OECD (2015). The abc of gender equality in education (pisa). *OECD Publishing*.

OECD (2020). Education at a glance. *OECD Publishing*.

Paccagnella, M. and P. Sestito (2014). School cheating and social capital. *Education Economics 22*(4), 367–388.

Paredes, V. (2014). A teacher like me or a student like me? role model versus teacher bias effect. *Economics of Education Review 39*, 38–49.

Pereda-Fernández, S. (2019). Teachers and cheaters: Just an anagram? *Journal of Human Capital 13*(4), 635–669.

Prøitz, T. S. (2013). Variations in grading practice–subjects matter. *Education Inquiry 4*(3), 22629.

Resh, N. (2009). Justice in grades allocation: Teachers' perspective. *Social Psychology of Education 12*(3), 315–325.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rosenthal, R. (1987). Pygmalion effects: Existence, magnitude, and social importance. *Educational Researcher 16*(9), 37–40.

Sommers, C. H. (2013). The war against boys: How misguided policies are harming our young men.

Steele, C. M. and J. Aronson (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology 69*(5), 797.

Stobart, G., J. Elwood, and M. Quinlan (1992). Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal 18*(3), 261–276.

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

A Data sources and description of variables

Dependent Variables:

Language-Standardized Test: INVALSI standardized test scores in language studies

Math-Standardized Test: INVALSI standardized test scores in math studies

Language-Teacher: Teachers' scores in language studies

Math-Teacher: Teachers' scores in math studies

Students characteristics:

Male: dummy=1 if male

N. siblings: number of siblings (4 indicates 4 or more)

Highly educated mother: dummy=1 if mother with a degree

Highly educated father: dummy=1 if father with a degree

Mother's secondary school attainment: dummy=1 if mother with a high school diploma

Father's secondary school attainment: dummy=1 if father with a high school diploma

Stay-at home mother: dummy=1 if mother housewife

Dialect: dummy=1 if language spoken at home is a dialect

Foreign language: dummy=1 if language spoken at home is not Italian

Immigrants (1st generation): dummy=1 if students are 1st generation immigrants

Immigrants (2nd generation): dummy=1 if students are 2st generation immigrants

Schools characteristics:

Class size: number of students per class

Ratio of females to males (class): ratio between females and males in the class

School size: number of students per school

ESCS: Average School Level ESCS Index. The INVALSI ESCS Index refers to the PISA index of Economic, Social and Cultural Status

Sample of inspected schools: dummy=1 if the school is selected for the external monitoring by INVALSI

Motivation for math: for details see notes in Tables 2a and 2b

Motivation for language: see notes in Tables 2a and 2b

All the variables listed above are from the 2010-11 datasets of the INVALSI

Area characteristics:

GDP per person employed: Total value added per capita, constant prices (base year 2000), 2001 data. Source: Fondazione Istituto Tagliacarne (2006). http://www.tagliacarne.it

Crime: Extortions (1999-2001): average rate of extortions over 10,000 inhabitants. Source: Fiaschi et al. (2011)

Social Capital: Social capital indicator. Source: Cartocci (2007)

Regional dummies:

North West: dummy=1 if regions are Liguria, Lombardia, Piemonte and Valle d'Aosta

North East: dummy=1 if regions are Emilia Romagna, Friuli Venezia Giulia, Trentino Alto-Adige and Veneto

Centre North: dummy=1 if regions are Lazio, Marche, Toscana and Umbria

Centre South: dummy=1 if regions are Abbruzzo, Campania, Molise and Puglia

Islands South: dummy=1 if regions are Basilicata, Calabria, Sicily and Sardinia

B Figures and Tables

Figure 1a: Kernel density in 5th grade

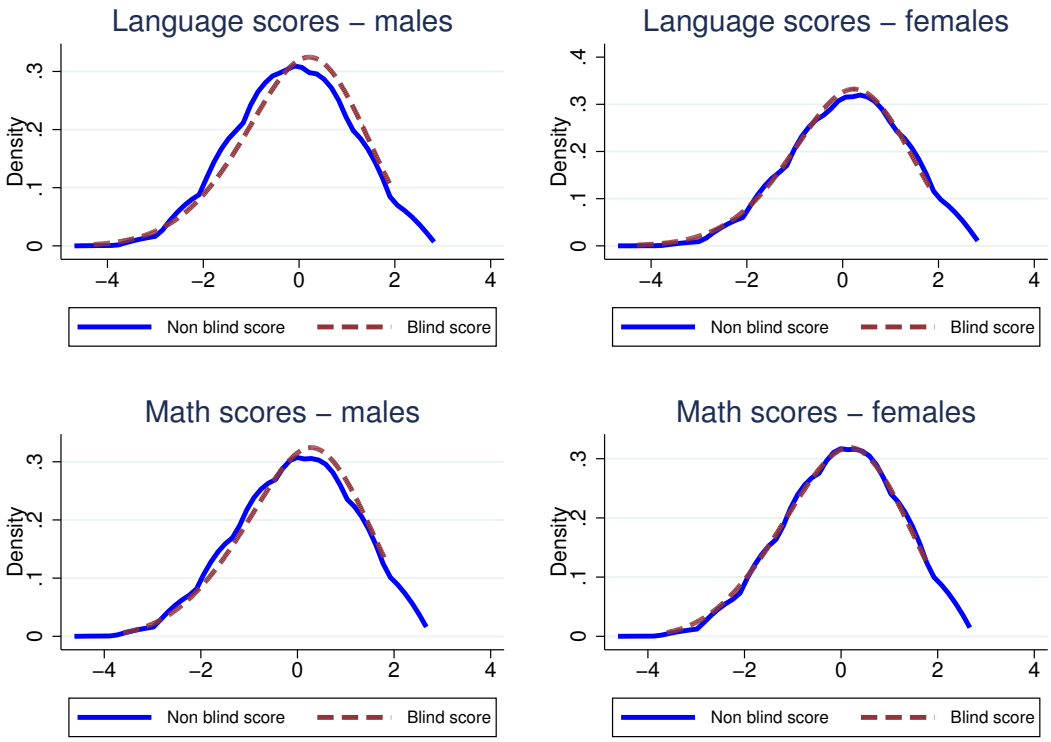

Figure 1b: Kernel density in 6th grade

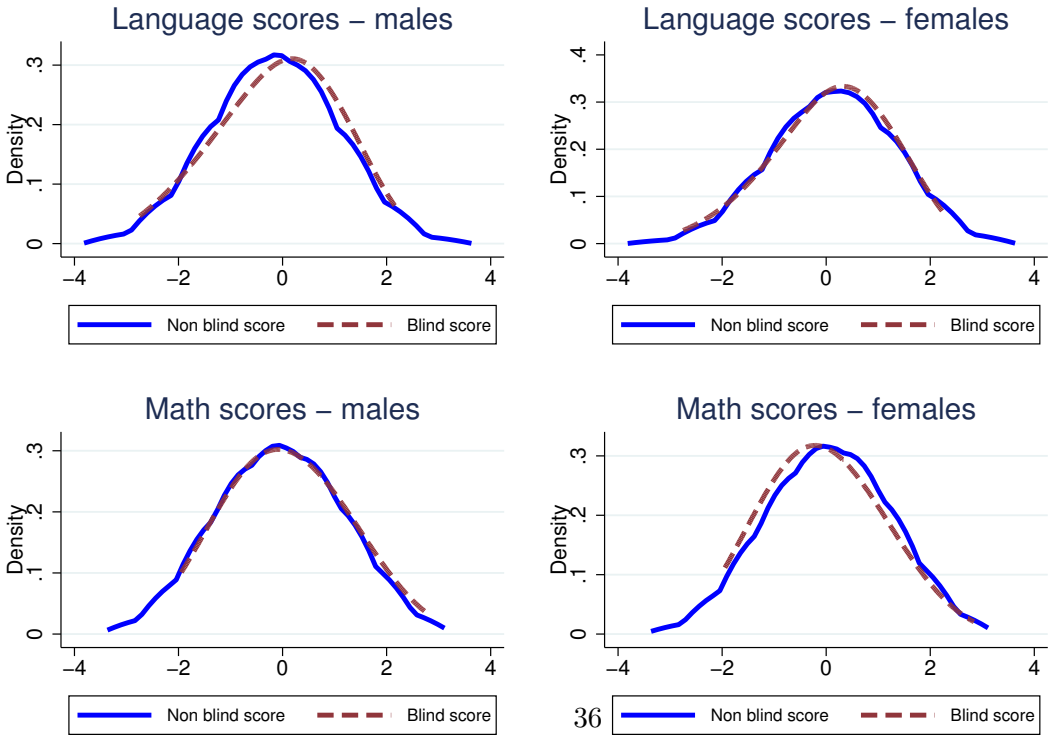

Table 1a: Fifth grade descriptive statistics: overall sample

Variable	Obs	Mean	Std. Dev.	Min	Max
Dependent variables					
Language-Standardized Test	474,366	74.22	14.05	0	100
Math-Standardized Test	474,366	69.86	16.78	0	100
Language-Teacher	474,366	72.99	12.77	0	100
Math-Teacher	474,366	73.97	13.24	0	100
Student and family characteristics					
Males	474,366	0.51	0.50	0	1
N. siblings	422,533	1.23	0.88	0	4
Highly educated mothers	411,848	0.13	0.33	0	1
Highly educated fathers	404,760	0.12	0.33	0	1
Mother's secondary school attainment	411,848	0.40	0.49	0	1
Father's secondary school attainment	404,760	0.35	0.48	0	1
Stay-at home mothers	414,845	0.40	0.49	0	1
Dialect	431,011	0.15	0.35	0	1
Foreign language	431,011	0.07	0.26	0	1
Immigrants (1st generation)	474,366	0.05	0.21	0	1
Immigrants (2nd generation)	474,366	0.04	0.20	0	1
School and Class characteristics					
Class size	474,366	19.276	4.35	1	35
Ratio of females to males (class)	474,366	0.47	0.11	0	1
School size	474,366	102.844	45.67	1	305
ESCS	467,177	0.004	0.46	-2.66	2.016
Area characteristics					
GDP per person employed	474,119	10.04	0.29	9.50	10.47
Crime	474,119	6.46	3.73	1.73	19.45
Social capital	474,119	-0.63	3.15	-6.43	5.47
North West	474,366	0.25	0.44	0	1
Centre North	474,366	0.18	0.38	0	1
Centre South	474,366	0.23	0.42	0	1
Islands South	474,366	0.16	0.37	0	1

Table 1b: 6th grade descriptives statistics: overall sample

Variable	Obs	Mean	Std. Dev.	Min	Max
Dependent variables					
Language-Standardized Test	498824	55.21	20.10	0	100
Math-Standardized Test	498824	40.65	20.89	0	100
Language-Teacher	498824	51.51	16.54	0	100
Math-Teacher	498824	52.50	19.70	0	100
Student and family characteristics					
Males	498824	0.51	0.50	0	1
N. siblings	462457	1.24	0.91	0	4
Highly educated mothers	418947	0.12	0.33	0	1
Highly educated fathers	412435	0.12	0.32	0	1
Mother's secondary school attainment	418947	0.38	0.49	0	1
Father's secondary school attainment	412435	0.33	0.47	0	1
Stay-at home mothers	424056	0.40	0.49	0	1
Dialect	467149	0.16	0.37	0	1
Foreign language	467149	0.07	0.26	0	1
Immigrants (1st generation)	498824	0.06	0.23	0	1
Immigrants (2nd generation)	498824	0.04	0.18	0	1
School and Class characteristics					
Class size	498824	21.74	3.86	1	34
Ratio of females to males (class)	498824	0.46	0.11	0	1
School size	498824	147.14	77.54	1	417
ESCS	486597	-0.01	0.47	-2.39	1.78
Area characteristics					
GDP per person employed	498824	10.04	0.29	9.50	10.47
Crime	498824	6.50	3.74	1.71	19.45
Social Capital	498824	-0.69	3.16	-6.43	5.47
North West	498824	0.25	0.43	0	1
Centre North	498824	0.18	0.38	0	1
Centre South	498824	0.23	0.42	0	1
Islands South	498824	0.16	0.37	0	1

Table 2a: 5th grade - Ability in Math and Language studies: boys vs. girls self-assessment

Please indicate if you agree with the following statements (mathematics) using the following scale: 1 - yes, 2 - No	**MALE**	**FEMALE**	**M vs. F**
Q1 A - I am good at maths - I am proficient in maths	83.67	75.93	7.74
Q1 B - Studying math is more difficult for me than for most od my classmates	20.87	28.21	-7.34
Q1 C - It is easy for me to learn maths	73.33	65.15	8.18
Q1 D - Studying mathematics is fun	72.26	65.65	6.61
Q1 E - I would like to study more math at school	49.11	39.73	9.38
Please indicate if you agree with the following statements (language) using the following scale: 1 - yes, 2 - No	**MALE**	**FEMALE**	**M vs. F**
Q2 A - I am good at language/Italian - I am procient in Language/Italian	74.50	85.40	-10.90
Q2 B - Studying Language is more difficult for me than for most od my classmates	29.78	16.11	13.67
Q2 C - It is easy for me to learn Italian/Language	66.71	79.96	-13.25
Q2 D - Studying Italian/Language is fun	58.24	73.50	-15.26
Q2 E - I would like to study more Italian at school	35.55	49.95	-14.40

Table 2b: 6th grade - Ability in Math and Language studies: boys vs. girls self-assessment

Please indicate how much you agree with the following statements (mathematics) using the following scale: 1-moderately disagree, 2-moderately disagree, 3-somewhat agree, 4 strongly agree	**MALE**	**FEMALE**	**M vs. F**
Q3 A - I am good at maths - I am proficient in maths	3.07	2.90	0.17
Q3 B - Studying math is more difficult for me than for most of my classmates	1.93	2.05	-0.12
Q3 C - It is easy for me to learn maths	3.05	2.87	0.18
Q3 D - Studying mathematics is fun	2.81	2.67	0.14
Q3 E - I would like to study more math at school	2.41	2.24	0.17
Please indicate how much you agree with the following statements (Language) using the following scale: 1-moderately disagree, 2-moderately disagree, 3-somewhat agree, 4 strongly agree	**MALE**	**FEMALE**	**M vs. F**
Q5 A - I am good at language/Italian - I am proficient in Language/Italian	2.90	3.08	-0.18
Q5 B - Studying Language is more difficult for me than for most of my classmates	2.03	1.79	0.24
Q5 C - It is easy for me to learn Italian/Language	2.96	3.20	-0.24
Q5 D - Studying Italian/Language is fun	2.57	2.92	-0.35
Q5 E - I would like to study more Italian at school	2.15	2.47	-0.33

Notes: These indexes are measured in a different ways for primary and lower secondary school students. During the survey, Italian students are asked to indicate how much they agree with five different statements about mathematics and language studies. For primary school the questions-answers type uses a simple agree-disagree scale. Because the adopted scale was 1 - yes, 2 - No, the variables on the self-assessed abilities are constructed as percentage of answer *YES*. For sixth graders the scale is: 1-moderately disagree, 2-moderately disagree, 3-somewhat agree, 4-strongly agree. We then report the average value.

Table 3a: 5th grade - Teachers' grades vs. Standardized test: average results by gender

	Gender	Obs	Mean	Std. Dev.	Min	Max
Language - Teachers	Male	240046	71.46	12.84	0	100
	Female	234320	74.57	12.50	0	100
Math - Teachers	Male	240,046	73.75	13.51	0	100
	Female	234,320	74.18	12.96	0	100
Language - Standardized test	Male	240,046	73.82	14.33	0	100
	Female	234,320	74.63	13.73	0	100
Math - Standardized test	Male	240,046	70.54	16.76	0	100
	Female	234,320	69.16	16.78	0	100

Table 3b: 6th grade - Teachers' grades vs. Standardized test: average results by gender

	Gender	Obs	Mean	Std. Dev.	Min	Max
Language - Teachers	Male	255032	49.12	16.56	0	100
	Female	243792	54.02	16.14	0	100
Math - Teachers	Male	255032	51.58	20.01	0	100
	Female	243792	53.45	19.32	0	100
Language - Standardized test	Male	255032	53.16	20.71	0.00	98.53
	Female	243792	57.36	19.21	0	100
Math - Standardized test	Male	255032	41.92	21.39	0	100
	Female	243792	39.31	20.28	0	100

Table 4a: Teachers' granding bias in Language - 5th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Males	-0.058***	-0.031***	-0.039***	-0.038***	-0.038***	-0.031***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Teacher score	0.094***	0.078***	0.082***	0.082***	0.082***	0.082***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Males X teacher score	-0.186***	-0.188***	-0.191***	-0.191***	-0.191***	-0.190***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Constant	0.029***	0.195***	0.020***	-0.031**	1.380***	0.626***
	(0.004)	(0.005)	(0.006)	(0.015)	(0.226)	(0.225)
Controls						
Student characteristics	No	Yes	Yes	Yes	Yes	Yes
Family characteristics	No	No	Yes	Yes	Yes	Yes
School and class characteristics	No	No	No	Yes	Yes	Yes
Area characteristics	No	No	No	No	Yes	Yes
Non-cognitive skills	No	No	No	No	No	Yes
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Observations	948,732	862,022	660,020	650,844	650,844	637,384
R-squared	0.008	0.051	0.119	0.119	0.120	0.140
N. classes	29244	29009	26467	26077	26077	26064

Table 4b: Teachers' granding bias in Language - 6th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Males	-0.209***	-0.166***	-0.174***	-0.171***	-0.171***	-0.178***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Teacher score	0.045***	0.040***	0.046***	0.047***	0.047***	0.046***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Males X teacher score	-0.088***	-0.090***	-0.091***	-0.092***	-0.092***	-0.092***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.107***	0.408***	0.200***	0.161***	0.794***	0.186
	(0.003)	(0.005)	(0.005)	(0.016)	(0.201)	(0.201)
Controls						
Student characteristics	No	Yes	Yes	Yes	Yes	Yes
Family characteristics	No	No	Yes	Yes	Yes	Yes
School and class characteristics	No	No	No	Yes	Yes	Yes
Area characteristics	No	No	No	No	Yes	Yes
Non-cognitive skills	No	No	No	No	No	Yes
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Observations	997,648	934,298	706,764	689,110	689,110	686,406
R-squared	0.016	0.093	0.184	0.185	0.187	0.217
N. classes	25819	25661	22928	22354	22354	22350

Notes: Dependent variable is the test results in Language (standardized test results and teachers' grading) in 5th grade (table 4a) and 6th grade (table 4b). Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Student characteristics include Dialect, Foreign language, Immigrants (1st and 2nd generation). Family characteristics are N. siblings, Highly educated mother, Highly educated father, Mother's secondary school attaintment, Father's secondary school attaintment, Stay-at home mother. School and class characteristics comprehend Class size, Ratio of females to males (class), School size, ESCS. Area characteristics include GDP per person employed, Crime and Social capital. Non-cognitive skills are Motivation for math and Motivation for language. Models from 2 to 6 includes the following regional dummies: North-West, Centre, South, South-Islands (North-East omitted). See Section A in Appendix for the full variables description. The number of observations is twice the number of students, since the dataset is stacked (for each student there are two observations, one for the teacher score and one for the INVALSI score).

Table 5a: Teachers' grading bias in Mathematics - 5th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Males	0.082***	0.108***	0.100***	0.104***	0.104***	0.081***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Teacher score	0.058***	0.045***	0.051***	0.050***	0.050***	0.048***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Males X teacher score	-0.115***	-0.116***	-0.117***	-0.118***	-0.118***	-0.117***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Constant	-0.042***	0.113***	-0.060***	-0.136***	1.279***	0.544**
	(0.004)	(0.006)	(0.007)	(0.016)	(0.240)	(0.236)
Controls						
Student characteristics	No	Yes	Yes	Yes	Yes	Yes
Family characteristics	No	No	Yes	Yes	Yes	Yes
School and class characteristics	No	No	No	Yes	Yes	Yes
Area characteristics	No	No	No	No	Yes	Yes
Non-cognitive skills	No	No	No	No	No	Yes
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Observations	948,732	862,022	660,020	650,844	650,844	637,384
R-squared	0.001	0.031	0.091	0.091	0.092	0.129
N. classes	29244	29009	26467	26077	26077	26064

Table 5b: Teachers' grading bias in Mathematics - 6th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Males	0.125***	0.163***	0.159***	0.163***	0.163***	0.086***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Teacher score	0.112***	0.113***	0.130***	0.131***	0.131***	0.130***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Males X teacher score	-0.219***	-0.221***	-0.223***	-0.224***	-0.224***	-0.224***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Constant	-0.064***	0.253***	0.032***	-0.014	0.314	-0.119
	(0.003)	(0.005)	(0.006)	(0.016)	(0.207)	(0.200)
Controls						
Student characteristics	No	Yes	Yes	Yes	Yes	Yes
Family characteristics	No	No	Yes	Yes	Yes	Yes
School and class characteristics	No	No	No	Yes	Yes	Yes
Area characteristics	No	No	No	No	Yes	Yes
Non-cognitive skills	No	No	No	No	No	Yes
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Observations	997,648	934,298	706,764	689,110	689,110	686,406
R-squared	0.003	0.070	0.151	0.152	0.154	0.234
N. classes	25819	25661	22928	22354	22354	22350

Notes: Dependent variable is the test results in Math (standardized test results and teachers' grading) in 5th grade (table 5a) and 6th grade (table 5b). Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. See Notes to Tables 4a and 4b for additional details regarding control variables. See Section A in Appendix for the full variables description. The number of observations is twice the number of students, since the dataset is stacked (for each student there are two observations, one for the teacher score and one for the INVALSI score).

Table 6a: Robustness checks, Language grades: North-South and Inspected school - 5th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Males	-0.058***	-0.040***	-0.041***	-0.044***	-0.044***	-0.035***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Teacher score	0.215***	0.199***	0.199***	0.198***	0.198***	0.200***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Males X teacher score	-0.201***	-0.201***	-0.202***	-0.202***	-0.202***	-0.201***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	415,188	382,650	296,014	291,972	291,972	285,356
Panel B: Southern regions						
Males	-0.067***	-0.024***	-0.042***	-0.037***	-0.037***	-0.032***
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)
Teacher score	-0.014*	-0.028***	-0.019**	-0.019**	-0.019**	-0.021**
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)
Males X teacher score	-0.178***	-0.181***	-0.183***	-0.184***	-0.184***	-0.183***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	362,676	326,880	254,050	250,584	250,584	246,130
Panel C: Sample of inspected schools						
Males	-0.035***	-0.012	-0.017	-0.021*	-0.021*	-0.014
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Teacher score	0.185***	0.184***	0.184***	0.184***	0.184***	0.185***
	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.016)
Males X teacher score	-0.220***	-0.220***	-0.217***	-0.217***	-0.217***	-0.216***
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Observations	59,332	57,872	47,322	47,322	47,322	46,332

Notes: Dependent variable is the test results in Language (standardized test results and teachers' grading) in 5th grade. Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. See Notes to Tables 4a and 4b for additional details regarding control variables. The northern regions include Emilia-Romagna, Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto Adige, Valle d'Aosta and Veneto, while the southern are Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia. Panel C includes students from a representative sample of selected schools for the external monitoring by INVALSI.

Table 6b: Robustness checks, Language grades: North-South and Inspected school - 6th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Males	-0.214***	-0.177***	-0.173***	-0.173***	-0.172***	-0.189***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Teacher score	0.141***	0.136***	0.134***	0.135***	0.135***	0.134***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Males X teacher score	-0.071***	-0.076***	-0.081***	-0.080***	-0.080***	-0.080***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	428,176	405,330	312,018	305,218	305,218	304,234
Panel B: Southern regions						
Males	-0.216***	-0.156***	-0.181***	-0.173***	-0.174***	-0.173***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Teacher score	-0.026***	-0.031***	-0.021***	-0.021***	-0.012***	-0.022***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Males X teacher score	-0.105***	-0.107***	-0.104***	-0.104***	-0.104***	-0.103***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Observations	392,322	363,858	275,138	267,764	267,764	266,522
Panel C: Sample of inspected schools						
Males	-0.208***	-0.170***	-0.179***	-0.175***	-0.176***	-0.184***
	(0.011)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
Teacher score	0.015	0.013	0.019	0.019	0.019	0.019
	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)
Males X teacher score	-0.108***	-0.112***	-0.112***	-0.112***	-0.112***	-0.112***
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	77,708	75,234	59,558	59,558	59,558	*59,464*

Notes: Dependent variable is the test results in Language (standardized test results and teachers' grading) in 6th grade. Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. See Notes to Tables 4a and 4b for additional details regarding control variables. The northern regions include Emilia-Romagna, Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto Adige, Valle d'Aosta and Veneto, while the southern are Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia. Panel C includes students from a representative sample of selected schools for the external monitoring by INVALSI.

Table 7a: Robustness checks, Math grades: North-South and Inspected school - 5th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Males	0.122***	0.141***	0.139***	0.141***	0.141***	0.107***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Teacher score	0.189***	0.182***	0.183***	0.182***	0.182***	0.181***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Males X teacher score	-0.129***	-0.131***	-0.130***	-0.130***	-0.130***	-0.129***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	415,188	382,650	296,014	291,972	291,972	285,356
Panel B: Southern regions						
Males	0.029***	0.067***	0.050***	0.057***	0.058***	0.046***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Teacher	-0.067***	-0.077***	-0.068***	-0.069***	-0.069***	-0.072***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Males X teacher score	-0.107***	-0.107***	-0.107***	-0.109***	-0.109***	-0.108***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	362,676	326,880	254,050	250,584	250,584	246,130
Panel C: Sample of inspected schools						
Males	0.106***	0.127***	0.117***	0.118***	0.118***	0.092***
	(0.012)	(0.012)	(0.013)	(0.012)	(0.012)	(0.012)
Teacher score	0.140***	0.141***	0.136***	0.136***	0.136***	0.135***
	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)	(0.017)
Males X teacher score	-0.134***	-0.135***	-0.120***	-0.120***	-0.120***	-0.119***
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Observations	59,332	57,872	47,322	47,322	47,322	46,332

Notes: Dependent variable is the test results in Math (standardized test results and teachers' grading) in 5th grade. Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. See Notes to Tables 4a and 4b for additional details regarding control variables. The northern regions include Emilia-Romagna, Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto Adige, Valle d'Aosta and Veneto, while the southern are Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia. Panel C includes students from a representative sample of selected schools for the external monitoring by INVALSI.

Table 7b: Robustness checks, Math grades: North-South and Inspected schools - 6th grade

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Males	0.132***	0.166]***	0.168***	0.173***	0.0173***	0.077***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Teacher score	0.171***	0.171***	0.180***	0.180***	0.180***	0.180***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Males X teacher score	-0.214***	-0.216***	-0.217***	-0.217***	-0.217***	-0.217***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	428,176	405,330	312,018	305,218	305,218	304,234
Panel B: Southern regions						
Males	0.105***	0.157***	0.142***	0.148***	0.147***	0.093***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)
Teacher score	0.075***	0.077***	0.095***	0.096***	0.096***	0.096***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Males X teacher score	-0.229***	-0.231***	-0.235***	-0.236***	-0.236***	-0.236***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	392,322	363,858	275,138	267,764	267,764	266,522
Panel C: Sample of inspected schools						
Males	0.142***	0.175***	0.165***	0.170***	0.169***	0.083***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Teacher score	0.122***	0.125***	0.143***	0.143***	0.143***	0.143***
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Males X teacher score	-0.244***	-0.245***	-0.249***	-0.249***	-0.249***	-0.249***
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	77,708	75,234	59,558	59,558	59,464	59,464

Notes: Dependent variable is the test results in Math (standardized test results and teachers' grading) in 6th grade. Standard errors in parenthesis, clustered at class level: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. See Notes to Tables 4a and 4b for additional details regarding control variables. The northern regions include Emilia-Romagna, Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto Adige, Valle d'Aosta and Veneto, while the southern are Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia. Panel C includes students from a representative sample of selected schools for the external monitoring by INVALSI.