British Journal of Sociology of Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cbse20

Do teacher and classroom characteristics affect the way in which girls and boys are graded?

A multilevel analysis of student–teacher matched data

Ilaria Lievore & Moris Triventi

To cite this article: Ilaria Lievore & Moris Triventi (2022): Do teacher and classroom characteristics affect the way in which girls and boys are graded?, British Journal of Sociology of Education, DOI: 10.1080/01425692.2022.2122942

To link to this article: https://doi.org/10.1080/01425692.2022.2122942

Published online: 17 Oct 2022.

Submit your article to this journal

Article views: 4205

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cbse20

BRITISH JOURNAL OF SOCIOLOGY OF EDUCATION
https://doi.org/10.1080/01425692.2022.2122942

Routledge
Taylor & Francis Group

Check for updates

Do teacher and classroom characteristics affect the way in which girls and boys are graded?

A multilevel analysis of student–teacher matched data

Ilaria Lievore and Moris Triventi

Department of Sociology and Social Research, University of Trento, Italy

ABSTRACT

Teachers' evaluations of students do not consider only academic competence, but are imbued with social considerations related to individual teacher and student characteristics, their interactions, and the surrounding context. The aim of this paper is understanding the extent to which teachers grade girls more generously than boys, and which characteristics of teachers and classrooms are likely to reduce this gender grading gap. We use Italian data from INVALSI-SNV, providing information on 10th-grade students linked with their teachers. The analysis relies on grade equation models in multilevel regression analysis, with students as first level, teachers/classrooms as second level, and schools as third level. Results show that, when comparing students who have identical subject-specific competence, teachers are more likely to give higher grades to girls. Furthermore, they demonstrate for the first time that this grading premium favouring girls is systemic, as teacher and classroom characteristics play a negligible role in reducing it.

ARTICLE HISTORY
Received 17 January 2022
Accepted 29 August 2022

KEYWORDS
Teachers' grades; gender; grading mismatch; academic performance; education; social inequalities

Introduction

During the last century, the existence of gender-related gaps in educational outcomes has been widely demonstrated almost worldwide (Goldin, Katz and Kuziemko 2006), and gender gaps in educational achievement are among most discussed topics in the field of educational studies. Typically, when comparing standardized test scores in most OECD countries, girls outperform boys in humanities, language, and reading skills, while boys show better scores in mathematics (PISA 2019; OECD 2019; IES 2009). Gender gaps in standardized test scores are even broader among students who demonstrate higher performance (OECD 2014).

At the same time, previous literature shows that, when skills are measured through grades, female students display higher educational attainment rates than male students in recent cohorts across all subjects (Downey and Yuan 2005). Female students outperform boys in terms of average school achievement, university enrolment (Becker, Hubbard, and Murphy 2010; Pekkarinen 2012), and tertiary degree attainment (Schofer and Meyer 2005).

CONTACT Ilaria Lievore ilaria.lievore@unitn.itilaria.lievore@gmail.com
Supplemental data for this article can be accessed online at https://doi.org/10.1080/01425692.2022.2122942
© 2022 Informa UK Limited, trading as Taylor & Francis Group

Therefore, even if men are still considered to be higher performing in scientific fields, and consequently women are underrepresented in highly financially rewarding professions (Card and Payne 2017), on average, girls have higher educational attainments. These findings have fostered the development of additional studies seeking to explain the sources of the disadvantages faced by boys in education, in particular by examining the roles of both the family and the school system (Mickelson 1989; Dumais 2002; Buchmann and DiPrete 2006).

Our work follows this stream of the literature, by aiming to improve the understanding of what role the school system, particularly teachers, plays in the production of gender differences in educational attainment. Some studies have shown that teachers assess girls' performances higher than boys' (Emanuelsson and Fischbein 1986); in addition, girls get significantly higher grades in both mathematics (Falch and Naper 2013) and language (Machin and Pekkarinen 2008; Falch and Naper 2013). However, from these studies it is not clear whether girls' higher grades stem from higher academic efforts and competence, or whether their marks reflect preferential treatment from their teachers.

This article investigates teacher grading—the way in which teachers assign marks to their students—as a potential source generating gender inequalities in education. Specifically, it examines whether teachers grade female students more generously compared to male students who have the same subject-specific competence, as measured via standardized test scores, and examines whether this putative female grading premium varies according to key teacher characteristics, features of the classroom, and elements of the school environment.

Investigating the way in which teachers assess their students is important for several reasons. First, grades might affect students' motivation and effort in education, and therefore influence their subsequent educational outcomes. Moreover, grades are an indicator of students' academic capacity sent by the school to parents, which thus might affect family decisions to invest more or less in a child's educational career. Teacher assessment is used and processed as an important source of information, not only by students and families, but also by teachers themselves, who base their educational recommendations on the prior indicators of student ability to which they have access: student grades. Grades are often determinant factors for students, in accessing the next level of education, finding a place among the best educational options, attaining rankings in the classroom, gaining a particular scholarship, or conquering admission to college (Blossfeld et al. 2016; Bonesrønning 2004; Bonner and Chen 2019). Moreover, teachers' assessments might have also long-term consequences, influencing occupational choices and thereby earnings, in students' adulthood (Lavy and Sand 2015; Borghans et al. 2016; Bonner and Chen 2019).

During the last few years, some researchers have attempted to assess whether teachers' grading involves systemic biases and differences related to students' gender in upper-secondary education. In Israel, Lavy (2008) found that girls obtained higher grades than boys in 'non-blind' tests in which the evaluator did know the student's identity, compared to 'blind' tests. Likewise, Lindahl (2007) observed that, when comparing students with the same level of competence, teachers in Sweden assessed girls more generously than boys; Angelo (2014) also reported similar findings for Portuguese high schools, as did Terrier (2015) for French high schools. Enzi (2015), analysing upper secondary education in Germany, found that gender plays a role in grading, a result that echoes that of Kiss (2013) in studying lower educational levels. Furthermore, among 15-year-old Czech students, there is a sizeable gender gap in teachers' grading which favours female students, in both language

and mathematics (Protivínský and Münich 2018). Although most studies have detected a grading premium favouring girls, one study of the Nordic countries did not find any gender gap in grading (Hinnerich, Höglin and Johannesson 2011).

Nonetheless, although this research stream provides important information on gender differences in education, it does not clarify whether biased grading is more likely to take place in some specific learning environments in comparison with others. This is particularly relevant from a sociological perspective (Reimer 2019), as teachers' grading behaviours not only take into account student competencies, but also reflect the social and the cultural contexts, the school environment (Farkas et al. 1990), the teachers' own beliefs (Chen and Bonner 2017), the relationship between students and their teacher (Costrell 1994), and teachers' sociodemographic characteristics and teaching approach (Bonner and Chen 2019).

This article aims to gain a more complete understanding of gender-based differences in teachers' grading by focusing on an understudied context: Italian upper-secondary schools, in which gender differences in academic performance, as measured via teachers' evaluations, are particularly pronounced. Moreover, in Italy the teaching profession is granted by the Constitutional right of 'autonomy' in duty delivery, being characterized by a certain degree of autonomy from other teachers or even the school administration regarding educational choices as grading (Bracci 2009). Additionally, the Italian educational system is characterized by vast territorial divides in school resources and socio-cultural environments (Checchi 2004; Montanaro 2008). For these reasons, Italy can be considered a 'best case' scenario, in which heterogeneity in grading practices related to student gender is likely to occur.

The empirical part of the current research relies on novel student–teacher matched data, which has only recently become available in Italy and permits in-depth investigation regarding the issue of our interest. We use a sample of about 39,000 students from two cohorts who were enrolled in 10th grade in the academic years 2015–2016 and 2016–2017. By relying on hierarchical linear regression models, we assess not only the overall grading gap between female and male students in two key subjects (language and mathematics), but also the extent to which this gap varies according to teacher characteristics, classroom composition, and school tracks.

The paper is organised as follows. The next section develops the theoretical framework and formulates potential explanations for the gender grading mismatch, and the third section provides relevant information about the Italian educational system. The fourth section includes information about the dataset, the variables, and the methodology. The fifth section presents the results of the empirical analysis, and finally, the last section discusses our findings and the conclusion.

The sources of the gender grading gap (GGG)

In this section, we develop a theoretical framework that may aid understanding of the mechanisms behind the gender grading gap (GGG). Teachers' grades result from a complex assessment process that incorporates multiple indicators of students' performance, reflecting not only students' levels of competence, effort and motivation, but also other important aspects, such as teachers' own expectations, considerations and beliefs (Chen and Bonner 2017), and the setting in which the relationships between students and their teacher develops (Wright, Horn and Sanders 1997).

In order to understand why teachers may be likely to give female students higher grades, teacher expectations theory is a useful tool. Accordingly, teachers have specific perceptions of, and therefore expectations about, their students. Such expectations involve an *a priori* evaluation of student competencies in specific domains (Cooper and Tom 1984). Several studies have attempted to assess the associational relationship between specific students' characteristics and teacher perception or expectation bias (as recent examples, see Hornstra et al. 2018; Zhu, Urhahne and Rubie-Davies 2018; and Riegle-Crumb and Humphries 2012). According to this prior literature, teacher expectations are strictly linked to stereotypes. Stereotypes, as representations of characteristics of specific groups (Bordalo et al. 2016), are means teachers use in order to process information about students in an easy and efficient manner. Together with students' ethnic background (see Strand 2012) and socioeconomic background (see Speybroeck et al. 2012), gender is one of the main sources generating teacher stereotypes. Regarding students' gender, internalised representations may result in biased judgements and in over or under-evaluation of specific students' performance. Teachers tend to perceive girls as more motivated, as more eager to learn (e.g. Ander, McElvany and Baumert 2010; Gentrup and Rjosk 2018), as more well-behaved (Glock and Kleen 2017), and as enacting less disruptive behaviour in the classroom (Terrier 2015). However, teachers may also have biased perceptions of gender-based relative talents for specific subjects, because, as research indicates, they are likely to believe maths is more difficult for girls than for identically performing boys (Riegle-Crumb and Humphries 2012). Therefore, teacher evaluation bias can result in a 'premium' or 'penalty' connected to dominant gender stereotypes. Thus, following previous research, we formulate our first hypothesis:

H1: Teachers are more likely to grade female students more generously than male students who have the same standardized test scores.

However, students' signals of their own cognitive abilities may be interpreted differently by teachers who have specific characteristics, or are operating in a given context. Some studies demonstrate how students benefit from having a same-gender teacher (Ammermueller and Dolton 2006). Accordingly, the 'stereotype threat' theory (Steel 1997) explains how the similarity between the demographic characteristics of students (such as gender) and those of their teachers improves communications and mutual understandings between teacher and student. This could lead teachers to unconsciously reward ways of behaving that are similar to their own. In this respect, it has been suggested that the increase in the share of female teachers may explain the gender gap in achievement that favours females, even if there is contrasting evidence on this topic (Neugebauer, Helbig and Landmann 2011).

Other teacher characteristics may influence the ways in which they grade their students according to gender. Carlana (2019) showed how the gender-science implicit association test—which measures gender stereotypes among teachers—correlates with some teachers' observable traits, such as their gender, field of study, or being the parents of daughters. It is reasonable to assume that other teacher characteristics may affect how they evaluate students, such as via age or seniority, as more experienced teachers could be less likely to be driven by stereotypes or expectations when evaluating students. Moreover, in general, teachers' working conditions may also influence their approach to grading students. For example, teachers' salaries, their level of stability in role, or their work opportunities (Basilio and Almeida 2018) may impact their investment in making high-quality assessments of

student abilities, through a 'disgruntled worker' effect (Vegas and de Laat 2003). Therefore, we expect that overall teacher characteristics contribute to moderate the GGG. More specifically, we develop two hypotheses:

H2a: Resemblance hypothesis: Male teachers are less likely to over-grade girls compared to female teachers.

H2b: Experience hypothesis: Older teachers, teachers with more years of in-school seniority, and those with a permanent contract display a lower GGG.

Very few studies have investigated how classroom or school features might affect the differences in teachers' grading of boys versus that of girls. However, insights from previous educational studies suggest possible moderation effects. For example, it has been demonstrated that in unfavourable contexts, such as classrooms that contain a large number of overall students; a high percentage of minority students; or a high proportion of students from a low socio-economic background, disciplinary problems are more frequently present (Rindermann 2007). In these problematic environments, teachers are more likely to be involved in conflict resolution, and less likely to interact with their students. This may reduce the time invested in assessing student achievement (Hochweber, Hosenfeld and Klieme 2013), and may make teachers more prone to rely on their expectations.

This reasoning may also apply to the type of school, because vocational and technical schools are more likely to be attended by students with lower school outcomes and poor educational expectations (Panichella and Triventi 2014). In such contexts, the gender gaps in grading might be larger, as teachers have fewer resources to accurately evaluate their students.

Moreover, the share of female students in the classroom may matter. Indeed, on one hand, boys may benefit from being in a classroom including a large share of girls, simply because the learning environment in such classrooms may be easier to manage for teachers. On the other hand, girls may be disadvantaged in such environments because individual personality traits may be inhibited from emerging over and above the gender-group related expectations. We can, then, hypothesise that teacher judgment accuracy might depend on classroom variables, because specific environments may lead teachers to process information about their students by relying on their own expectations. Thus, we expect that classroom composition and school track moderate the extent of GGG, specifically:

H3a: Structural hypothesis: Smaller classrooms put conditions in place that reduce the GGG.

H3b: Composition hypothesis: In high SES classrooms and in those with a higher share of female students, the GGG is smaller.

H3c: Tracking hypothesis: Academic tracks are associated to a smaller GGG.

Features of the Italian educational system

In the current Italian educational system, children enter schools at the age of 6, and school is mandatory until they are 16 years old. They attend eight years of comprehensive education, divided into five years in primary education, plus three years in lower secondary education. Primary school is commonly preceded by kindergarten, which lasts up to three years and is non-compulsory. In Italy's public schools, which in 2019 accounted for 92% of the total

number,[1] curricula are state-mandated and therefore similar across schools. After taking a national examination at the end of 8th grade, students are asked to make their first momentous choice among different educational programs, in order to attend upper-secondary schools.

There are several types of upper-secondary schools, but they can be broadly classified into three tracks: the academic track (*licei*), the technical track (*istituti tecnici*), and the vocational track (*istituti professionali*). The three main tracks have different curricula, subjects, educational purposes, and levels of heterogeneous prestige (Contini and Triventi 2016). In general, attending lyceums leads to university, whereas technical and vocational schools combine a general with a vocational education, and are usually considered to be less demanding.

There are significant differences by student gender in specific high school track enrolment. On average, more than 60% of students enrolled in lyceums are female, with around 70% of these enrolled in classical lyceums, but less than 50% enrolled in scientific/applied science lyceums, and up to 28% enrolled in scientific lyceum with sport curricula. Female students comprise the vast majority in linguistic lyceums (80%), humanities and social science lyceums (89%), and artistic lyceums (71%). In contrast, technical schools and vocational schools host larger proportions of boys; their respective student populations are approximately 70% and 57% male (Miur 2017).

After five years of upper-secondary schooling, the student population takes a national test (*esame di maturità*). Neither the final test score nor the type of chosen school is binding for continuing to university, even if several tertiary degree programs require ability-based entry tests and accept only a limited number of students.

Concerning teacher evaluations, the Italian Ministry of Education (Miur[2]) offers an evaluation regulation (*regolamento di valutazione*[3]) that includes precise guidelines for how teachers are expected to grade their students. These guidelines differ according to education level. In primary education, grades are expressed as a descriptive evaluation for each subject, which reflects four levels of learning: 'advanced', 'intermediate', 'basic', and 'in the process of first acquisition'. In secondary schools, grades are instead expressed as numbers from 1 to 10, where 6 is considered the passing mark and 10 is the highest score. During the academic years, students receive two report cards: one around February (*primo quadrimestre*) and one around June, at the end of the school year (*secondo quadrimestre*). Each subject has a distinctive grade: if at the end of the school year, students have not passed a given subject (meaning their subject grade is below 6), they need to take an exam on that subject before starting the new academic year. If the final report card contains three or more subjects below 6, the student needs to repeat their entire academic year. In the report card, grades are calculated as the average of results for several examinations and tests that students were required to take during the school year. Such examinations can vary according to the subject and to teacher preferences, but in most cases, they involve both oral and written testing.

Grading practices are also delineated in the formal educational agreement of co-responsibility (*patto educativo di corresponsabilità*[4]), also overseen by Miur. This document, which must be signed by both parents and students at the beginning of lower-secondary school, lists the principles and behaviours that schools, families, and students are expected to share and agree to respect. This emphasises the idea that grades are only one component of students' evaluation, that which also includes the context, students' progress throughout the

school year, analysis of their overall decline and progress patterns, and the school situation as a whole.

Despite the specificity of both regulatory documents (*regolamento di valutazione* and *patto educativo di corresponsabilità*), the extent to which these policies are implemented by schools and teachers remains unclear. Indeed, they are conceived as mere guidelines; therefore, each school can determine its own evaluation criteria, which are usually documented in school regulations. According to each school's specific regulations, teachers may have greater or lesser autonomy in deciding both the types of tests students must take, and their own grading practices.

Research design

Data

The empirical analysis is based on data collected by the Italian National Institute for the Evaluation of the Education System (INVALSI) within the National Evaluation System (SNV). The main mission of INVALSI is to perform periodic, systematic assessments of students' knowledge skills, of the quality of national educational institutions, and of the quality of vocational training. The INVALSI-SNV dataset provides a variety of information on the entire population of students enrolled in specific grades and academic years; data are gathered via administrative sources and student questionnaires. In particular, these data contain information for both teacher assessments of student abilities (teachers' grades) and student scores on standardized tests in Language and Mathematics (INVALSI test scores). An important advantage of this data source is that information on teachers' grades comes directly from the schools, and is not self-reported by students, which notably increases reliability.

In 2012, INVALSI handed out for the first time a CAWI[5] questionnaire, which was addressed to a random sample of Language and Mathematics teachers who worked in specific grade levels. The questionnaire gathered information on both teachers' socio-demographic characteristics and their teaching habits: their professional profiles and training, attitudes towards teaching, relationships with colleagues, and teaching practices. Thus, in the INVALSI-SNV dataset, it is possible to use student and class identifiers to link information about students with that of their teachers in the Language and Mathematics subjects. The final sample utilised in this study includes students in 10th grade matched with their teachers in the academic year 2015–16, pooled with students in 10th grade matched with their teachers in the academic year 2016–17. Our analytical sample includes 38,957 students with valid information for the variables included in the analysis.

Variables

Our dependent variables are teachers' grades in Mathematics and Language. Grade scoring ranges from 1 to 10, where 6 is the passing mark and 10 is the highest mark; grades are constructed as the average of oral and written performance marks. INVALSI collects information about grades from the midterm report card (around February), so information about teachers' grades at the end of the academic year is not available from their dataset. However, the INVALSI test is usually administered around March or April, making the two

different kinds of assessment relatively close in time. This limits the risk that differences between the two indicators, teachers' grades and INVALSI test scores, would be substantially affected by differential learning across genders.

Our main independent variable is student gender, recoded as 0 for a male student and as 1 for a female student. The main control variable is the student score on the INVALSI standardized test on Language and Mathematics. Through the INVALSI test score, we have information about the subject-specific competences of students, measured by applying Item Response Theory (IRT) to students' answers (see Lord 2012 for further information on IRT). This information provides a measure of student ability that is independent of teacher assessments, since it is blindly evaluated and therefore is considered to be 'unbiased' (Borghans et al. 2016). INVALSI test scores have a mean of 200 and a standard deviation of 40. Following Triventi (2020), we argue that INVALSI test scores are a good proxy of subject-specific competence compared to other indicators (such as the PISA test score), because they measure curriculum-related competences in the Italian educational system, and are therefore more directly comparable, in terms of knowledge, with what is expected to be assessed via school marks. Indeed, the main scope of INVALSI is assessing students' knowledge according to what national legislation expects for that specific grade. By definition, INVALSI tests do not aim to assess additional students' abilities such as communication skills, emotional and relational skills, participation and engagement. In fact, these dimensions are expected to be embedded in teacher assessed grades.[6] INVALSI test score results are also adjusted in order to reduce the risk of cheating during test administration (i.e. INVALSI 2018).

For other control variables at the student level, we use quarter of birth, migration background, geographical area, regularity in studies, attendance in kindergarten, an index of Economic and Social-Cultural Status (ESCS),[7] and a control for the academic year. These variables are included because previous studies suggest they are associated with teachers' grades.

As moderator variables at the second level, we introduced some teacher characteristics expected to affect the size of the GGG: gender, age,[8] within-school seniority,[9] and contract type. Concerning classroom composition, we consider the percentage of female students, the percentage of students with medium-high and higher ESCS,[10] and the classroom size.

As moderator variables at the school level, we control for the type of school attended by students (lyceum, technical school, and vocational school). Table 1 provides information about the recoding.

Analytical strategy

The objective of the analysis is twofold. The first goal is to test our first hypothesis, establishing to what extent female students are graded more generously than boys by their teachers. A common strategy used to assess whether teachers grade better one group compared to another with the same academic abilities and competences is the grade equation model (for examples, see Triventi 2020; Kiss 2013; and Hinnerich, Höglin and Johannesson 2011). The basic form of the grade equation model is a regression, in which a non-blind measure of student performance (in setting, teacher's grade) is expressed as a function of the variable identifying the group of interest (student gender)

Table 1. Description of the variables used in the analysis.

Variable	Description
Dependent variable	
Teacher grade in Mathematics Teacher grade in Language	Teacher grade in mid-term school report (pagella) in February. Average between grade in oral exam and grade in written exam. Scale 1–10 (where 6 is the passing mark)
Independent variable	
Student gender	Recoded as (0) if male; (1) if female
Student control variables	
INVALSI test score in numeracy INVALSI test score in literacy	Scores obtained by students in the INVALSI standardized tests. The scores have mean 200 and standard deviation 40. For the analysis, we standardized to have a mean of 0 and a standard deviation of 1. We use scores adjusted for potential cheating.
Quarter of birth	Recoded as (0) (January, February, March); (1) (April, May, June); (2) (July, August, September); (3) (October, November, December)
Migration background	Recoded as (0) if students are native born; (1) if students are II generation immigrants; (2) if students are I generation immigrants
Geographical area	Recoded as (0) North West; (1) North East; (2) Centre; (3) South; (5) Isles (as defined by ISTAT, national institute of statistics)
ESCS	Index provided from INVALSI that measures students' economic, social and cultural status. It is a synthesis of three indicators: 1. Parental occupational status (HISEI); 2. Parental level of education (PARED); 3. Possession of specific material assets (HOMEPOS), We recoded the variable in (0) Lower (1st quartile); (1) Medium-low (2nd quartile); (2) Medium-high (3rd quartile) (3) Higher (4rt quartile)
Regularity in studies	Recoded as (0) if students are regular; (1) if students are early starters (primina); (2) if students are late starters (including rejection)
Attendance in kindergarten	Recoded as (0) if students did not attend kindergarten; (1) if students attended kindergarten
Academic year	Recoded as (0) if students were in 10th grade in the A.Y. 2015/2016; (1) if students were in 10th grade in the A.Y. 2016/2017
Language and mathematics teacher variables	
Gender	Recoded as (0) if teachers are male; (1) if teachers are female
Age	Continuous variable, scale 28–77
Seniority in school	Recoded as (0) if teachers work in the school from 1 year or less; (1) from 2 to 3 years; (2) from 4 to 5 years; (3) from more than 5 years
Type of contract	Recoded as (0) if teachers have a fixed-term contract; (1) if teachers have a permanent contract
Classroom and school variables	
Class size	Number of students in the classroom, scale 5–30
% Female students	Percentage of female students in the classroom, scale 0–100
% High-background students	Percentage of students with medium-high (3rd quartile) or higher (4th quartile) ESCS in the classroom, starting from our recoding of ESCS at the student level, scale 0–100.
Type of school	Type of school attended by students, recoded as (0) Vocational schools; (1) Technical schools; (2) Lyceums

plus a blind measure of student ability (INVALSI test score) and a series of control variables.

The measure of teacher gender bias, or teacher grading premium, is therefore constructed by considering the average marks for boys and girls in Mathematics and Language in 'non-blind' classroom exams, and the respective mean scores of a 'blind' national exam (INVALSI test) that is marked anonymously.

Teacher grading premium or penalty for females is assessed relying on two different models for each subject (s), Language and Mathematics:

$$Grades_{i_s} = \alpha + \delta\left(female_{i_s}\right) + \beta\left(testscore_{i_s}\right) + \varepsilon_{i_s} \quad (1)$$

In this basic form of the grade equation model (Eq. (1)), δ represents the extent of teacher grading premium or penalty towards female students: if $\delta > 0$ then females are over-assessed, and if $\delta < 0$ then females are under-assessed compared to males. Test scores are considered in this framework as yardsticks against which teachers' marks can be compared.[11] The grade equation model allows us to flexibly handle the different measurement scales of the two competence measures (Dardanoni, Modica and Pennisi 2009; Hinnerich, Höglin and Johannesson 2011; Kiss 2013). In order to properly measure gender grading bias, we rely on two grading equation models embedded in hierarchical linear regression analysis, in which students comprise the first level, classrooms and teachers comprise the second level, and schools comprise the third level. The two main advantages of multilevel modelling over OLS regressions are: the possibility of considering the natural nesting of the data, thus solving the problem of dependency of observations; and the likelihood of providing correct standard errors, thus avoiding underestimating them, which would lead to incorrect inferences or interpretations.

In the first step of our analysis, we estimate a random intercept model for each of the two subjects (s), in which the average grade is allowed to vary by classroom, but the slope of the regression line is assumed to be fixed across schools and classrooms:

$$\begin{aligned} Grades_{ijk_s} &= \alpha + \delta\left(female_{ijk_s}\right) + \beta_1\left(testscore_{ijk_s}\right) + \beta_2\left(I_{ijk_s}\right) + \\ &\beta_3\left(T_{jk_s}\right) + \beta_4\left(C_{jk_s}\right) + \beta_5\left(Z_{k_s}\right) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s} \end{aligned} \tag{2}$$

These models follow the usual multilevel structure, where students (i) are nested in classrooms (j), which are nested in schools (k), and where:

- $Grades_{ijk_s}$ is the dependent variable that measures teacher assessment, with scores ranging from 1 to 10;
- δ represents the extent of teacher grading premium favouring females;
- $\beta_1\left(testscore_{ijk_s}\right)$ represents the main control variable at the individual level: the INVALSI test score, with its associated regression coefficient;
- $\beta_2\left(I_{ijk_s}\right)$ is a vector of control variables at the individual level, with its associated regression coefficients;
- $\beta_3\left(T_{jk_s}\right)$ is a vector representing teacher characteristics at the classroom level, with its associated regression coefficients;
- $\beta_4\left(C_{jk_s}\right)$ is a vector representing classroom composition variables, with its associated regression coefficients; and
- $\beta_5\left(Z_{k_s}\right)$ the type of school attended, with its associated regression coefficients.

The second step of the analysis is devoted to testing our second and third hypotheses, aimed at understanding which characteristics of teachers and classrooms are likely to enlarge or reduce the GGG. To do so, we estimate a series of random coefficient multilevel models, in which the GGG is allowed to vary across classrooms, and in which cross-level interactions are introduced to assess whether the GGG varies across: (i) teacher gender (resemblance hypothesis); (ii) teacher characteristics such as age, within-school seniority, and contract type (experience hypothesis); (iii) classroom size (structural hypothesis); (iv) percentage

of female students and percentage of students from higher social background (composition hypothesis); and (v) school track (tracking hypothesis).

The multilevel random slope regressions are estimated using the following model specifications, with cross-level interactions between gender and, respectively, teacher characteristics (Eq. (3)), classroom composition (Eq. (4)), and school track (Eq. (5)):

$$Grades_{ijk_s} = \alpha + \delta\left(female_{ijk_s}\right) + \beta_1\left(testscore_{ijk_s}\right) + \beta_2\left(I_{ijk_s}\right) + \beta_3\left(T_{jk_s}\right) + \beta_4\left(C_{jk_s}\right) + \beta_5\left(Z_{k_s}\right) + \delta\left(female_{ijk_s}\right)u_{1jk_s} + \beta_6\left(female_{ijk_s}\right)^*\left(T_{jk_s}\right) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s} \quad (3)$$

$$Grades_{ijk_s} = \alpha + \delta\left(female_{ijk_s}\right) + \beta_1\left(testscore_{ijk_s}\right) + \beta_2\left(I_{ijk_s}\right) + \beta_3\left(T_{jk_s}\right) + \beta_4\left(C_{jk_s}\right) + \beta_5\left(Z_{k_s}\right) + \delta\left(female_{ijk_s}\right)u_{1jk_s} + \beta_6\left(female_{ijk_s}\right)^*\left(C_{jk_s}\right) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s} \quad (4)$$

$$Grades_{ijk_s} = \alpha + \delta\left(female_{ijk_s}\right) + \beta_1\left(testscore_{ijk_s}\right) + \beta_2\left(I_{ijk_s}\right) + \beta_3\left(T_{jk_s}\right) + \beta_4\left(C_{jk_s}\right) + \beta_5\left(Z_{k_s}\right) + \delta\left(female_{ijk_s}\right)u_{1jk_s} + \beta_6\left(female_{ijk_s}\right)^*\left(Z_{k_s}\right) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s} \quad (5)$$

Results

Gender grading mismatch

As a first step, in Figure 1, we examine the average test scores and marks in Language and Mathematics among boys and girls. Analysing standardized INVALSI test scores makes evident that boys outperform girls in Mathematics, while girls perform significantly better in Language. This result is not surprising, compared with the gender gap in achievement across the OECD countries (OECD 2009, 2014; Fryer and Levitt 2010).

In both subjects, the difference in the average teacher grade between male and female students is approximately 0.4 points. The average grade in Language is around 6.2 for boys, while it is 6.6 for girls. In Mathematics, female students have on average a grade equal to 6.3, while boys are on average under the passing mark, performing at 5.9.

In Figure 2, we report the average grades in Language and Mathematics across boys and girls, along the distribution of the subject-specific competence level. The figure clearly shows a female grading premium: girls receive, on average, higher grades than boys at the same level of subject-specific competence. Additionally, it appears that such female advantage occurs across the entire distribution of students' subject-specific competences (INVALSI score). One might expect the gap to be larger in the middle of the competence distribution, because students around the mean, with average competences, may be graded more in accordance with their gender-related non cognitive skills and behaviours, rather than according to their actual competences. Indeed, students at the extremes of the distribution—those who perform very poorly at one end, or very well at the other end—are thought

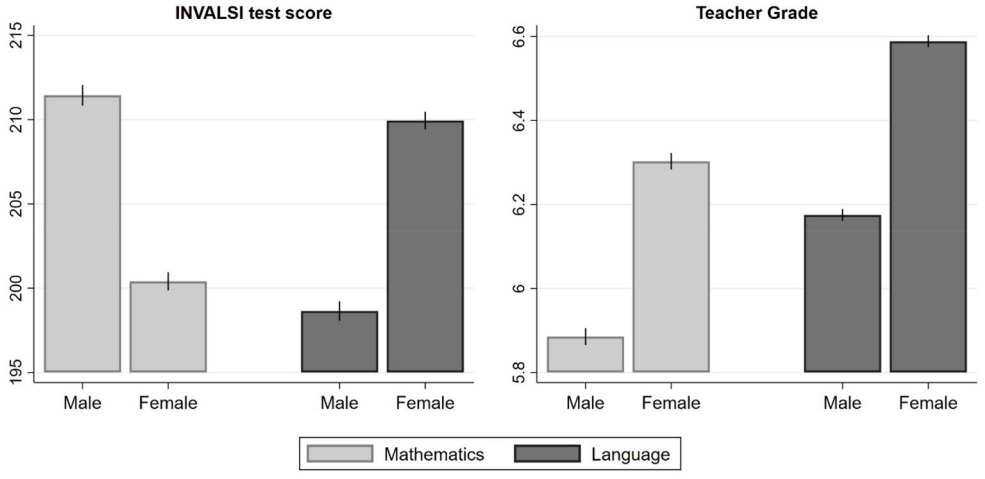


Figure 1. Average INVALSI test score in Numeracy and in Literacy for male and female students (left) and average teacher grades in Mathematics and Language for male and female students (right); 95% confidence intervals; N = 38,975.

to be evaluated in relation to their competences, as the stronger signal. However, Figure 2 shows clearly that the gender gap in teacher evaluation does not vary according to the level of competence. This gap is surprisingly constant in Language, in which girls are evaluated about 0.4 points higher than boys at each increment of the INVALSI score distribution. In Mathematics, the gap appears larger in the middle of the distribution, and slightly reduces for the top-performing students. Despite that, the difference in grades between boys and girls, even the top performing ones, is not negligible and statistically significant.

We next turn to the estimation of the GGG via regression models, as Figure 3 presents. The first three models are estimated via OLS linear regression. Model 1 represents the simple average gender gap in grades; model 2 adjusts for subject-specific competences, as measured by INVALSI test score; model 3 adjusts for INVALSI test score and individual level control variables; and model 4 is a three-level random intercept model with students nested in classrooms nested in schools, with the same independent variables as model 3.

Comparing the first two models, girls evidently obtain better grades than boys in both subjects; once we adjust for their subject-specific competence, the average differences are only slightly reduced. This means that only to a minor extent is the advantage of girls in grades due to their superior subject-specific competences, a result that is further confirmed in models 3 and 4. Comparing model 1 with model 4, the gap between boys and girls in teachers' grades appears to be slightly reduced in Language, even if it remains substantially relevant and statistically significant, whereas in Mathematics the GGG enlarges.

Teacher characteristics

Having established to what extent females are graded more generously than boys by their teachers, the next goal is understanding whether the sign and magnitude of this pattern is related to specific teacher characteristics: gender, age, within-school seniority, and contract type. This is achieved by relying on random slope models, in which student gender

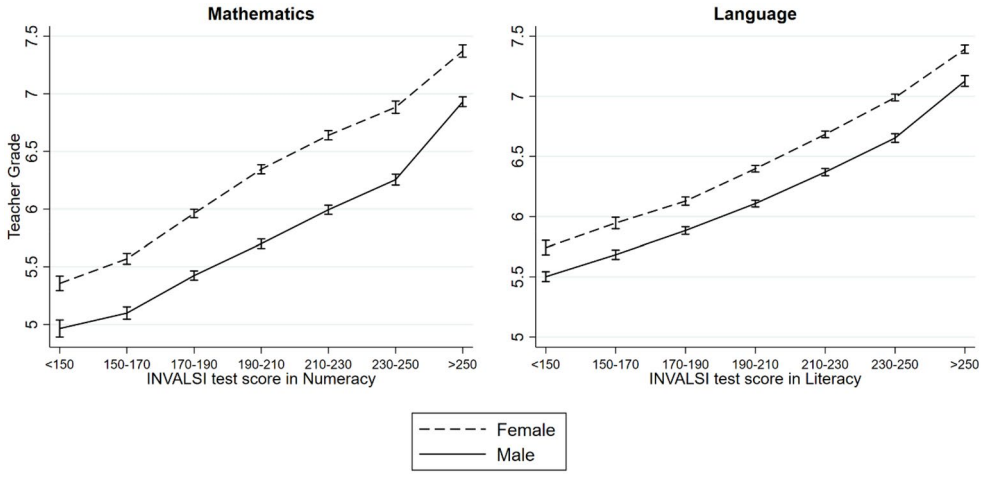


Figure 2. Average teacher grade for male and female students in Mathematics (left) and Language (right) for different intervals of INVALSI test score distribution; 95% confidence intervals; N = 38,975.

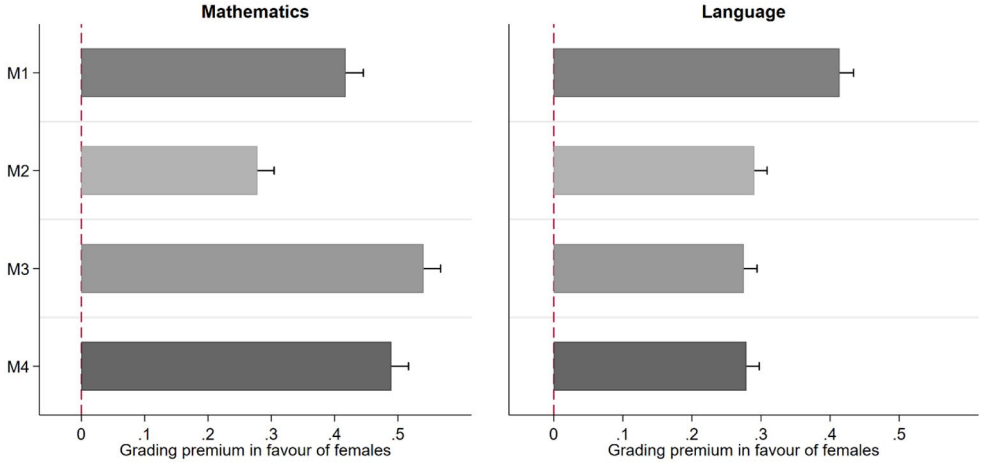


Figure 3. Teacher gap in grading in favor of female students in Mathematics (left) and in Language (right). Coefficients derived from four linear prediction models of teacher grade for female and male students (male = reference category); 95% confidence intervals; N = 38742. Note: model 1 represents the simple average gap in grade; model 2 controls for subject-specific competences (INVALSI test score); model 3 controls for INVALSI test score and individual level controls (student quarter of birth, migration background, geographical area, ESCS, regularity in studies, attendance to kindergarten, academic year); in model 4 linear predictions are derived from a hierarchical model with students nested in classrooms nested in schools, with the same controls as model 3.

is allowed to have a different effect on teachers' grades across classrooms, and which include a cross-level interaction between student gender and teacher characteristics. Overall, our results indicate that GGG in favour of females does not change according to teacher characteristics, either in Mathematics or in Language. In other words, the female grading premium is always present, irrespective of teachers' individual characteristics and practices. Figure 4 shows the predictive margins and average marginal effects of cross-level

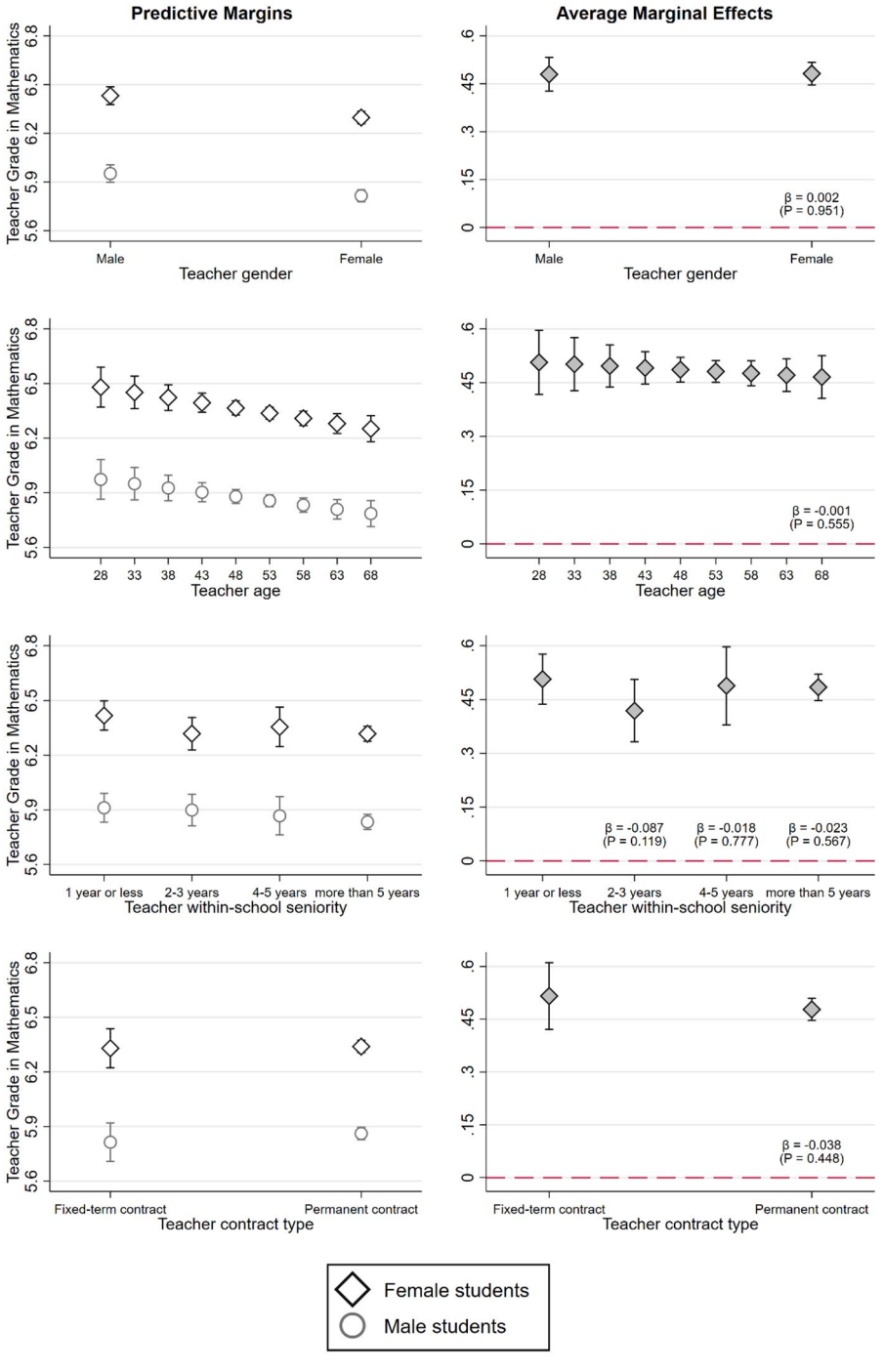


Figure 4. Predictive margins and average marginal effects of cross-level interactions between student gender and teacher characteristics in *Mathematics* (gender, age, within-school seniority, contract type); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.

interactions between student gender and teacher characteristics in Mathematics. The average marginal effects of cross-level interactions between student gender and teacher gender shows that both male and female maths teachers, on average, attribute higher grades to girls than to boys who have the same subject-specific competence. This suggests that male teachers are not (unconsciously) likely to better reward male students, and female teachers do not penalize more male students, as suggested by the same-sex resemblance hypothesis.

In Mathematics, while older teachers tend to give on average lower grades, having an older, and therefore more experienced, teacher does not affect the levels at which female and male students are graded. Examining within-school seniority yields similar results: maths teachers who have more than five years of seniority in school are more severe in grading, compared with teachers who have one year or less of seniority, but the difference in grading between girls and boys is similar irrespective of the teacher's length of experience within the school. Therefore, we cannot conclude that teachers with more seniority, who are likely to better know their students, are better able to provide more homogeneous grading for boys and girls who have the same mathematics competences. In addition, a teacher's type of contract does not affect either the extent of GGG, or the average grade of students.

Figure 5 shows a similar finding for the interaction between student gender and teacher characteristics in Language.

Classroom composition and school type

Regarding classroom composition, an increase in the size of the classroom corresponds to a slight decrease in the average grade. However, only for Mathematics the number of students per class significantly moderates the extent of the GGG. Indeed, the GGG is bigger when the classroom size is larger. As Figure 6 shows, in Mathematics classes, boys are more penalized than girls in terms of grading when the number of students per classroom increases. Interestingly, an increase of the percentage of female students in the classroom corresponds to an increase in the overall average grade for Mathematics, but also to a slight decrease in the overall average grade in Language. However, the share of girls in the classroom does not significantly moderate the GGG. Finally, a higher share of students with medium-high and higher ESCS is associated with a lower average grade only in Mathematics, but it does not moderate the extent of GGG either in mathematics or in Language. Overall, we can conclude that classroom composition features do not significantly contribute to enlarging or reducing the gender grading gap, with the exception of class size for Mathematics.

For the type of school attended,[12] results show that the average grade is lower in both lyceums and technical schools compared to vocational schools; this is true for both Language and Mathematics. Moreover, the type of school students attend seems to partially moderate the GGG, only in Mathematics; the gap is significantly larger in both technical schools and lyceums, compared with vocational schools. Therefore, in Mathematics, attending a technical school or lyceum rather than vocational school increases the gap in grading to the detriment of male students, who seem to be even more disadvantaged compared to their female counterparts. In contrast, regarding Language (Figure 7), attending a lyceum or a technical school rather than a vocational school does not affect the gap in grading between male and female students.

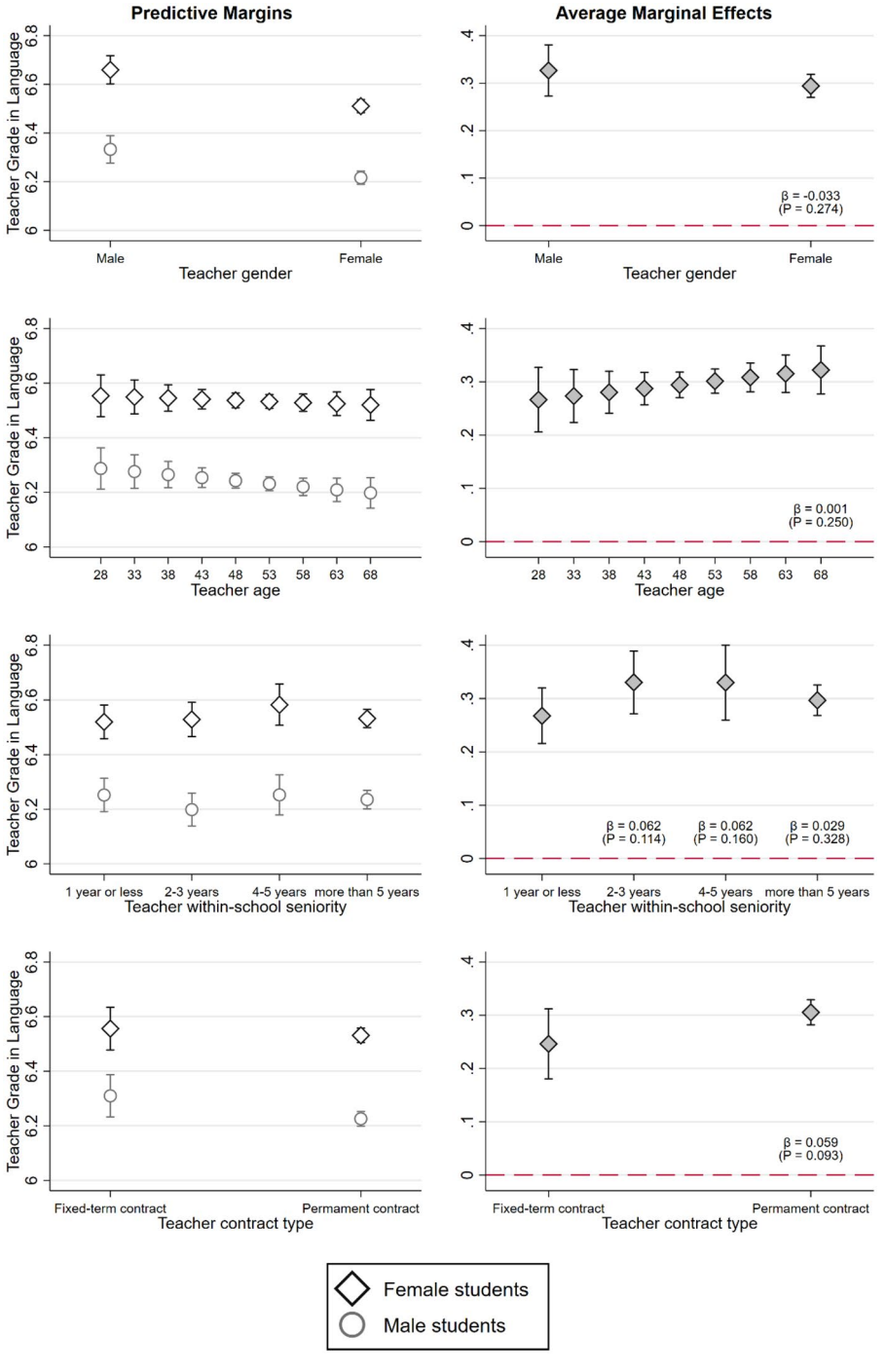


Figure 5. Predictive margins and average marginal effects of cross-level interactions between student gender and teacher characteristics in *Language* (gender, age, within-school seniority, contract type); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.

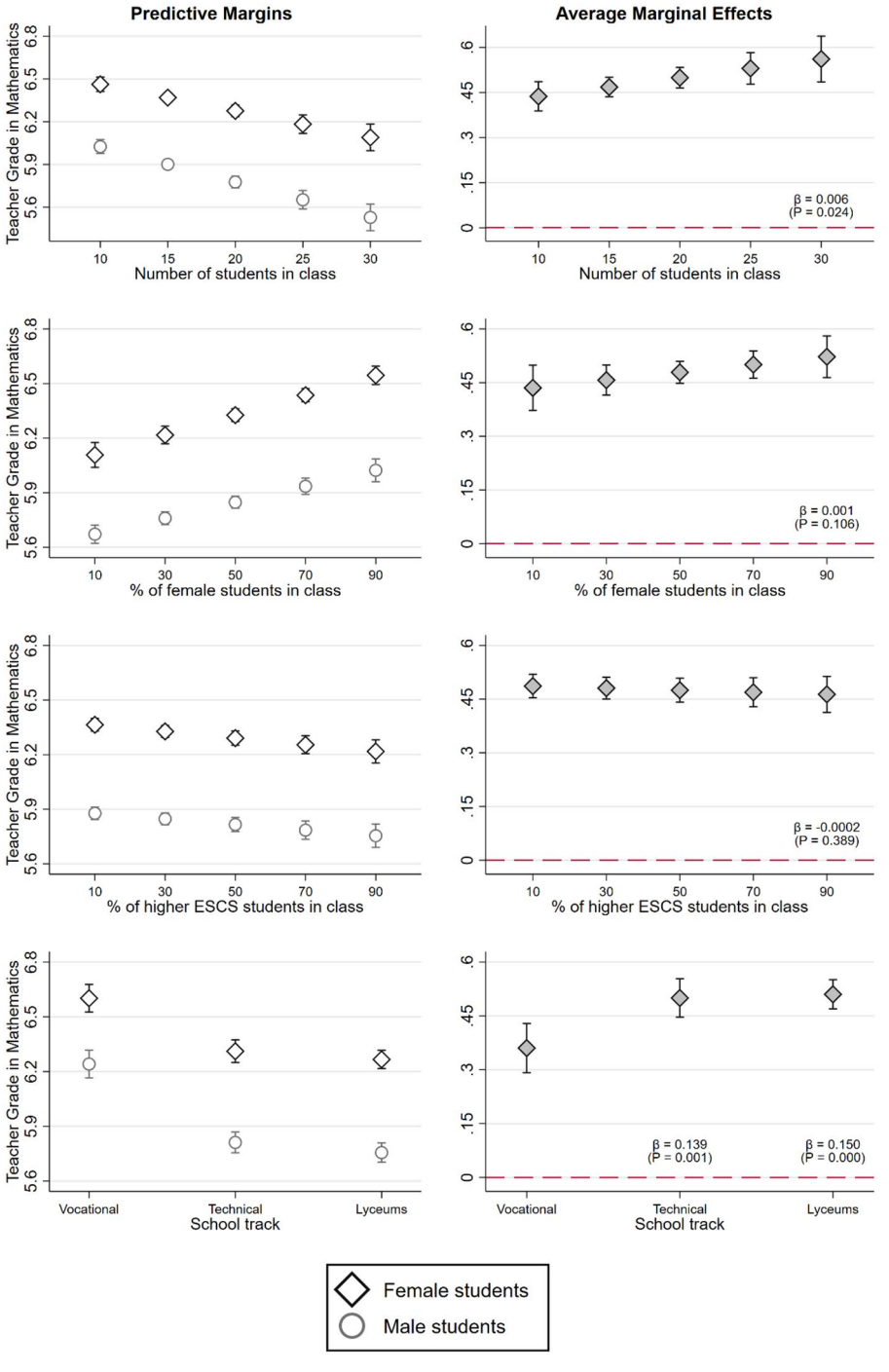


Figure 6. Predictive margins and average marginal effects of cross-level interactions between student gender and classroom and school characteristics in *Mathematics* (class size, % of female students, % of high-ESCS students, school track); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.

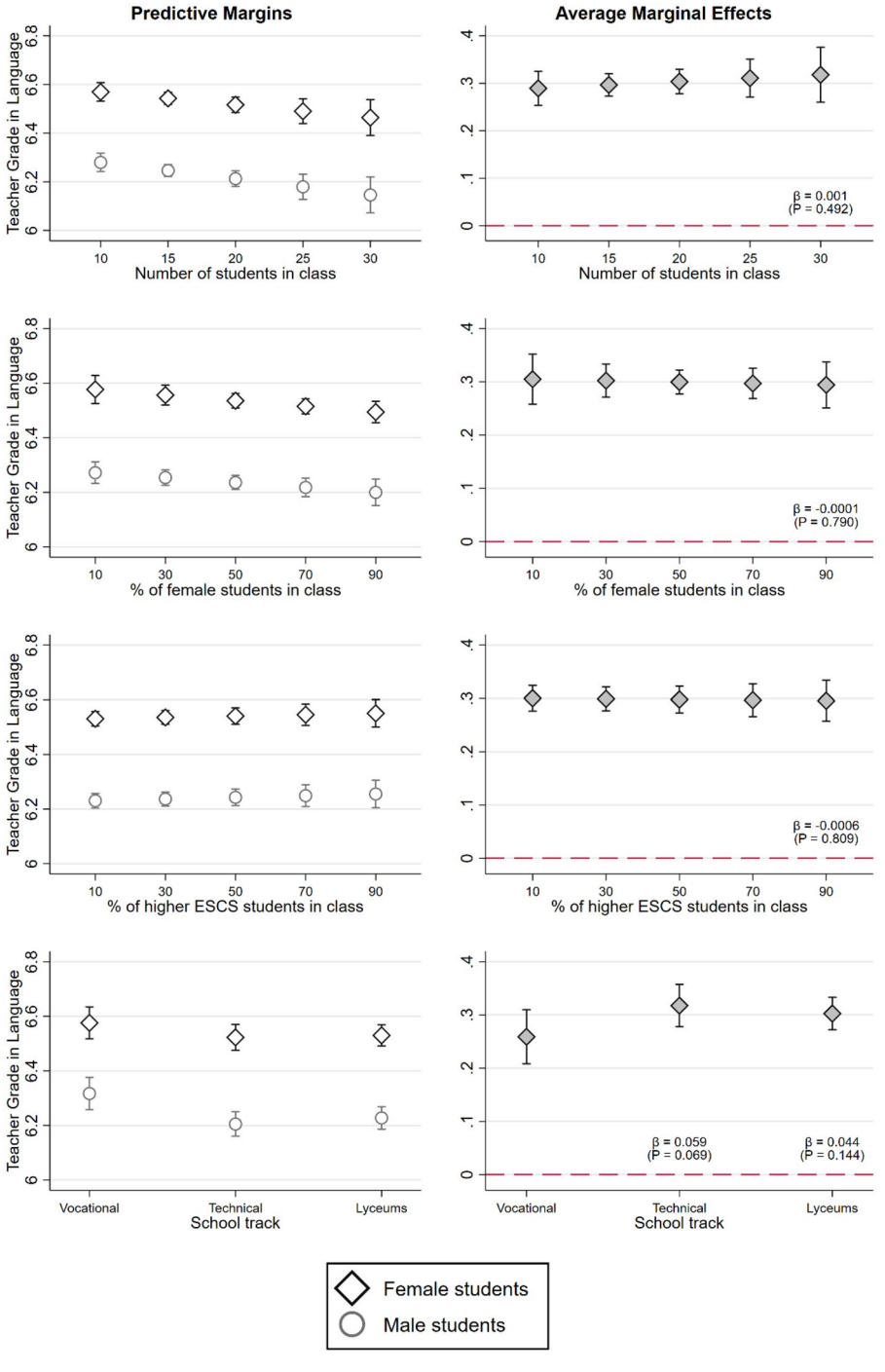


Figure 7. Predictive margins and average marginal effects of cross-level interactions between student gender and classroom and school characteristics in *Language* (class size, % of female students, % of high-ESCS students, school track); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.

Robustness checks

Three additional analyses are proposed in order to control for the robustness of the results (see online Appendices A5 for Mathematics and A6 for Language). The first analysis includes fixed-effect terms at the second level of analysis, in order to account for possible omitted-variable bias due to unobserved heterogeneity at the classroom level. Controlling for all possible confounders, including teacher characteristics, at the second and third levels, coefficients are almost identical, and they do not change the substantial interpretation of results. Therefore, our set of control variables at the classroom and school levels are good predictors of teacher grade in the two subjects.

The second additional analysis includes inverse probability weighting (IPW), in order to control for possible selection bias in our analytical sample. The final sample for Mathematics includes 42,707 students, while the final sample for Language includes 41,880 students. In order to compare coefficients across the two subject areas, we selected the smallest sample across both, which includes students who perform on both the literacy and numeracy tests. This implies a possible exclusion of students who decided to skip the exam for one of the two subjects, leading to biased estimates. To resolve this issue, we constructed different IPWs for Language and for Mathematics, based on the probability of the student being absent on the day of the test, using socio-demographic student variables as predictors, and including teacher grades. Results of the chosen models are robust.

The third additional analysis controls for measurement error in the INVALSI test score; in comparison with teacher grade, which is a repeated measure over the course of the academic year, the INVALSI test score may be subject to measurement error resulting from a variety of factors, such as a student's emotional or mental state while taking the test, or the conditions under which the test was administered. For this reason, GGG may be over- or under-assessed. Specifically, regarding Mathematics, as boys outperform girls in the INVALSI test score, the grading gap between the two groups may be under-assessed. In contrast, in Language, female students outperform male students in the INVALSI test score, resulting in a possibly over-assessed gap. To assess the magnitude of the bias in the estimates of subject-specific competences, following Bound, Brown and Mathiowetz (2011), we adopt an instrumental variable (IV) estimation approach. In this framework, the IV is constituted by a lagged student's subject-specific competence score. For each 10th-grade student in the analytical sample, the IV is represented by their individual test score from the 8th grade (in the academic years 2013–14 and 2015–16, respectively). This linkage is made possible by the INVALSI-SNV dataset structure, in which the SIDI code, a student unique identifier, allows the matching of students with their INVALSI test score results at previous grade levels. This approach enables correction of the extent of the teacher grading premium favouring female students. However, a considerable problem regarding this specification concerns the linkage with the SIDI codes, which implies a loss of cases and missing values due to school non-reporting, ID misclassifications, or students who did not pass the school exams and are thus not attending the originally planned academic year. Further, it implies a possible self-selection in the dataset of high-performing students which can affect the final estimates, and lead to a considerably smaller sample. Therefore, our preferred model remains the first. As expected, using the INVALSI test score from 8th grade as the instrument, the gap in Mathematics slightly increases, while in Language it slightly reduces; however, the overall findings are confirmed.

Conclusions and discussion

This paper has aimed to explore the influence of teachers on gender differences in academic outcomes in Italian upper-secondary schools. The goal was providing a measure of the gender-based grading mismatch—the difference in grade between male students and female students who have identical subject-specific competences. Previous research shown that girls are graded more generously than boys, even when they have the same level of competence, as measured via standardized tests (Lavy 2008; Lindahl 2007; Angelo 2014; Terrier 2015; Enzi 2015; Kiss 2013). The current paper addresses this gap, focusing on the Italian case, by demonstrating the existence of a gender-based grading mismatch that favours girls. Additionally, teacher premium in grading for girls versus boys appears to be different across two subjects, Language and Mathematics.

Focusing on grades in Language, for which subject females outperform males on average in the INVALSI test, it appears that teacher grading mismatch enlarges the gender differences in academic outcomes. In contrast, for Mathematics, in which boys outperform girls on the INVALSI test, the pattern is different. Teacher grading mismatch evens the gap, to the point that female students are always advantaged in classroom grades, despite their lower subject-specific competences. More specifically, girls are regularly assessed at 0.4 grade points higher than boys on average, in both Mathematics and Language. Whereas this estimate could be slightly overestimated for Language, the IV model specification suggests that in Mathematics the gap could be even larger. This may imply that teacher grading mismatch could be a significant discriminant in Mathematics, between scoring above or below the minimum passing grade. Therefore, we accept our hypothesis H1: teachers are more likely to grade female students more generously than male students who have the same subject-specific competences.

A further, substantive contribution of the current study is the assessment of whether specific teachers' characteristics, classroom composition, and type of school attended affect the ways in which females and males are evaluated by their teachers. These characteristics have been selected on the basis of theoretical considerations and the results of prior research. This study's results suggest that overall, the GGG is systemic: teacher characteristics, classroom composition, and school type do not appear to have any considerable effect on decreasing it. On the contrary, this study suggests that only classroom size and the type of school have a moderating role. For the Mathematics subject, some conditions such as classrooms with a large number of students, and classes within a technical or academic track rather than a vocational track, are associated with an increase in the gender grading mismatch. For Language instead, none of the considered features is significantly associated with the GGG.

Therefore, we can reject both the resemblance hypothesis and the experience hypothesis: teacher considered characteristics do not moderate the GGG. Regarding classroom characteristics, we can also reject the composition hypothesis, because the share of females and the share of higher-SES students do not affect the extent of the GGG.

However, we can partially accept the structural hypothesis, as a larger classroom size is significantly associated with an increase in the GGG, but only for Mathematics.

Finally, we can also partially accept the tracking hypothesis: the GGG is wider in both technical tracks and lyceums compared to vocational tracks, again only for Mathematics.

This (almost) systemic gender grading mismatch in favour of female students, which is surprisingly large especially in Mathematics, can be explained both methodologically and

theoretically. Technically, the implemented models with fixed effects at the classroom level enable controlling for all the characteristics of classrooms and schools that might aid in explaining the GGG. However, this does not account for students' specific educational signals[13] that work beyond competences, such as behaviour in the classroom, participation, engagement, perseverance, and effort. Indeed, students' attitudes and behaviours in the classroom are relevant criteria for grades attribution, and they partially enter in teacher's evaluation, but they are irrelevant criteria for results on the INVALSI test. One related theoretical stream interprets gender grading mismatch as also being a function of students' observed behaviours. School and classroom environments might indeed be adapted to traditionally female behaviours (Lavy 2008). Female students might thus adopt such *actual* behaviours during class, including precision, order, modesty, and quietness, which go beyond the individuals' academic performance, but which teachers may highly reward in terms of grades. Indeed, the idea that teachers may be prone to favour 'girly' attitudes in classroom is corroborated by other Italian findings in studies examining earlier school grades (Di Liberto et al. 2021). Conversely, teachers may be likely to associate such behaviours only with female students, because girls are traditionally thought of as possessing these traits. Consequently, teacher grading premium favouring females could also be related to teachers' *expectations* regarding their female students, rather than related to the actual behaviours of the latter during class.

Another theoretical explanation calls into play teacher overcompensation towards females. Girls are indeed often discussed in discrimination contexts, especially in speeches about gender differences in cognitive ability in dealing with the 'hard subjects'. A possible explanation for the reason teachers are more generous in grading female students could be that teachers wish to avoid possible discrimination against girls as an ability-stigmatized group. Therefore, teachers may over-assess girls in the same way they sometimes over-assess non-native students, to avoid negative stereotyping (Alesina 2018). From another perspective, it is also possible that the over-assessment of female students' competences in Mathematics partly represents a sort of 'push' to encourage the weaker students (Terrier 2015). As a consequence, this may translate into a positive discrimination which favours girls.

A final interpretation of our results could stem from a limitation of our data. Indeed, our measure of student grading is derived from the midterm report card, rather than the final report card of the academic year. Teachers may consciously adopt specific grading practices at the midterm to trigger students' effort differently according to their gender, to in turn obtain the best possible performance by the end of the academic year. Specifically, teachers could perceive that male students' effort is more easily triggered by lower grades, which could encourage them to study harder to achieve a better grade, while female students' effort to achieve higher grades may be more easily triggered via encouragement.

Disentangling such mechanisms is beyond the scope of this article, however; this contribution to the body of research focuses on providing an estimate of the total effect of students' gender on teachers' grades. The main limitation of this paper is the unavailability of indicators for students' behaviour and attitudes, which may be important mediators in explaining the teacher grading premium favouring females. Teacher assessment is indeed constituted by multiple different factors, also including students' non-cognitive indicators (Lipnevich et al. 2020); thus, it is not possible to entirely disentangle a pure gender discrimination bias from a broader grading gap related to students' attitudes and behaviours in class. Future studies should address the inclusion of students' gender-related behaviours

and attitudes towards the subject matter. Moreover, a measure of teacher gender stereotypes is absent from this study, and this would be determinant to conclude that GGG is driven by teachers' discrimination against boys—or in favour of girls. Overall, it would be necessary to take into account unobserved heterogeneity before speculating about the meaning and the implications of such a grading mismatch.

In conclusion, the magnitude of the bias against male students in not negligible, and may have negative consequences. This is especially true regarding Mathematics, where a teacher penalty may translate into a failing grade, since the average teacher grade for boys falls right on the passing mark. Indeed, it has been hypothesized that boys' struggles in the Italian system might be partially driven by grading biases (Di Liberto, Casula and Pau 2021).

The interpretation of results using teacher grades rather than scores from standardized tests may also have important implications in terms of public policy, as teachers' grades have previously been found to be strong predictors of a variety of important life outcomes (Borghans et al. 2016).

Notes

1. Calculation based on www.dati.Istat (Italian national statistical office), 2019 data.
2. Miur: *Ministero dell'Istruzione, dell' Università e della Ricerca*, divided in 2020 into two different ministries: *Ministero dell' Istruzione* (Ministry of Education) and *Ministero dell' Università e della Ricerca* (Ministry of University and Research).
3. https://www.miur.gov.it/valutazione
4. https://www.miur.gov.it/web/guest/patto-educativo-corresponsabilita
5. CAWI: Computer Assisted Web Interviewing
6. For more information about INVALSI test score: www.invalsiopen.it
7. Index provided from INVALSI that measures students' economic, social and cultural status. It is a synthesis of three indicators: (1) Parental occupational status; (2) Parental level of education; (3) Possession of specific material assets.
8. We tested the model fit introducing also age squared. Both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) suggest that age squared does not improve the model fit.
9. The expression "within-school seniority" refers to the length of teacher's experience within the school.
10. Medium-high and higher ESCS students correspond to the 3rd and 4th quartile of ESCS distribution
11. Another possibility would be to assess the teacher grading premium with a difference-in-difference design (Di Liberto, Casula and Pau 2021). Teacher grading premium can be understood as the difference between girls and boys, gathered via the difference between teacher grade ('non-blind' measure) and the standardized test score ('blind' and supposedly unbiased measure) (see Lavy 2008). Consequently, teacher grading mismatch would be defined as the average gap between for female students, minus the same gap for male students, as prior researchers have done to estimate gender discrimination (Falch and Naper 2013; Goldin and Rouse 2000). However, in the current context, this would also be feasible by standardizing the different scores in order to compare them.
12. It is important to note that these models also include controls for classroom composition. Therefore, our estimates regard GGG for female and male students in different types of schools that have similar classroom composition. Estimates without controls for classroom composition suggest similar patterns.
13. INVALSI-SNV data does not collect and provide such information.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This work was supported by Compagnia di San Paolo.

ORCID

Ilaria Lievore http://orcid.org/0000-0002-4652-004X
Moris Triventi http://orcid.org/0000-0002-2713-2957

References

Abbiati, G., G. Argentin, and T. Gerosa. 2017. "Different Teachers for Different Students? Evidence on Teachers-Student Matching and Its Consequences in the Italian Case." *Journal of Economic Policy* 33 (1): 13–58.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti. 2018. "Revealing Stereotypes: Evidence from Immigrants in Schools." NBER Working Paper: 25333.

Ammermueller, A., and P. Dolton. 2006. "Pupil-Teacher Interaction Effects on Scholastic Outcomes in England and the USA." ZEW Discussion Papers, 06–60.

Ander, Y., N. McElvany, and J. Baumert. 2010. "Die Einschätzung Lernrelevanter Schülermerkmale Zum Zeitpunkt Des Übergangs Von Der Grundschule Auf Die Weiterführende Schule: Wie Differenziert Urteilen Lehrkräfte? [Perception of Learning Relevant Student Characteristics at School Transition. How Differentiated Are Teacher Judgements?]." In *Der Übergang Von Der Grundschule in Die Weiterführende Schule: Leistungsgerechtigkeit Und Regionale, Soziale Und Ethnisch-Kulturelle Disparitäten*, edited by K. Maaz, J. Baumert, C. Gresch, and N. McElvany, 313–330. Bonn: Bundesministerium für Bildung und Forschung, Referat Bildungsforschung.

Angelo, C. S. 2014. "Is There a Bias Towards Girls in Non Anonymous Evaluation?" Working Paper.

Basilio, J. R., and A. M. F. Almeida. 2018. "Teacher Employment Contracts and Student Performance." *Revista Brasileira de Educação* 23: 1–23.

Becker, G. S., W. H. Hubbard, and K. M. Murphy. 2010. "Explaining the Worldwide Boom in Higher Education of Women." *Journal of Human Capital* 4 (3): 203–241. doi:10.1086/657914.

Bonesrønning, H. 2004. "Do the Teachers' Grading Practices Affect Student Achievement?" *Education Economics* 12 (2): 151–167. doi:10.1080/0964529042000239168.

Bonner, S. M., and P. P. Chen. 2019. "The Composition of Grades: Cognitive and Noncognitive Factors." In *What We Know about Grading: What Works, What Doesn't, and What's Next*, edited by T. R. Guskey and S. M. Brookhart, Chap. 3. Alexandria, VA: ASCD.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics* 131 (4): 1753–1794. doi:10.1093/qje/qjw029.

Borghans, L., B. H. H. Golsteyn, J. J. Heckman, and J. E. Humphries. 2016. "What Grades and Achievement Tests Measure." *Proceedings of the National Academy of Sciences of the United States of America* 113 (47): 13354–13359. doi:10.1073/pnas.1601135113.

Bound, J., C. Brown, and N. Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, edited by J. Heckman and E. Leamer, vol. 5, Chap. 59. Amsterdam. Elsevier Science.

Bracci, E. 2009. "Autonomy, Responsibility and Accountability in the Italian School System." *Critical Perspectives on Accounting* 20 (3): 293–312. doi:10.1016/j.cpa.2008.09.001.

Buchmann, C., and T. A. DiPrete. 2006. "The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement." *American Sociological Review* 71 (4): 515–541. doi:10.1177/000312240607100401.

Card, D., and A. A. Payne. 2017. "High School Choices and the Gender Gap in Stem." Technical Report, National Bureau of Economic Research.
Carlana, M. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *The Quarterly Journal of Economics* 134 (3): 1163–1224. doi:10.1093/qje/qjz008.
Checchi, D. 2004. "Da Dove Vengono le Competenze Scolastiche?" *Stato e Mercato* 72: 413–453.
Chen, P. P., and S. M. Bonner. 2017. "Teachers' Beliefs about Grading Practices and a Constructivist Approach to Teaching." *Educational Assessment* 22 (1): 18–34. doi:10.1080/10627197.2016.1271703.
Contini, D., and M. Triventi. 2016. "Between Formal Openness and Stratification in Secondary Education: Implications for Social Inequalities in Italy." In *Models of Secondary Education Inequality: An International Comparison*, edited by H. P. Blossfeld, S. Buchholz, J. Skopek, and M. Triventi, 305–322. Cheltenham, UK: Edward Elgar.
Cooper, H. M., and D. Y. H. Tom. 1984. "Teacher Expectation Research: A Review with Implications for Classroom Instruction." *The Elementary School Journal* 85 (1): 77–89. doi:10.1086/461393.
Costrell, R. M. 1994. "A Simple Model of Educational Standards." *The American Economic Review* 84 (4): 956–971.
Dardanoni, V., S. Modica, and A. Pennsi. 2009. "Grading across Schools." *The B.E. Journal of Economic Analysis & Policy* 9 (1) article 16. doi:10.2202/1935-1682.2065.
Di Liberto, A., L. Casula, and S. Pau. 2021. "Grading Practices, Gender Bias and Educational Outcomes: Evidence from Italy." *Education Economics*: 1–28. doi:10.1080/09645292.2021.2004999.
Downey, D., and A. S. V. Yuan. 2005. "Sex Differences in School Performance during High School: Puzzling Patterns and Possible Explanations." *The Sociological Quarterly* 46 (2): 299–321. doi:10.1111/j.1533-8525.2005.00014.x.
Dumais, S. A. 2002. "Cultural Capital, Gender, and School Success: The Role of Habitus." *Sociology of Education* 75 (1): 44–68. doi:10.2307/3090253.
Emanuelsson, I., and S. Fischbein. 1986. "Vive la Difference? A Study on Sex and Schooling." *Scandinavian Journal of Educational Research* 30 (2): 71–84. doi:10.1080/0031383860300202.
Enzi, B. 2015. "Gender Differentials in Test scores and Teacher Assessments: Evidence from Germany." Working Paper.
Falch, T., and L. R. Naper. 2013. "Educational Evaluation Schemes and Gender Gaps in Student Achievement." *Economics of Education Review* 36: 12–25. doi:10.1016/j.econedurev.2013.05.002.
Farkas, G., R. P. Grobe, D. Sheehan, and Y. Shuan. 1990. "Cultural Resources and School Success: Gender, Ethnicity, and Poverty Groups within an Urban School District." *American Sociological Review* 55 (1): 127–142. doi:10.2307/2095708.
Fryer, R. G., and S. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics* 2 (2): 210–240. doi:10.1257/app.2.2.210.
Gentrup, S., and C. Rjosk. 2018. "Pygmalion and the Gender Gap: Do Teacher Expectations Contribute to Differences in Achievement between Boys and Girls at the Beginning of Schooling?" *Educational Research and Evaluation* 24 (3–5): 295–323. doi:10.1080/13803611.2018.1550840.
Glock, S., and H. Kleen. 2017. "Gender and Student Misbehavior: Evidence from Implicit and Explicit Measures." *Teaching and Teacher Education* 67: 93–103. doi:10.1016/j.tate.2017.05.015.
Goldin, C., and C. Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–741. doi:10.1257/aer.90.4.715.
Goldin, C., L. F. Katz, and I. Kuziemko. 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap." *Journal of Economic Perspectives* 20 (4): 133–156. doi:10.1257/jep.20.4.133.
Hinnerich, B. T., E. Höglin, and M. Johannesson. 2011. "Are Boys Discriminated in Swedish High Schools?" *Economics of Education Review* 30 (4): 682–690. doi:10.1016/j.econedurev.2011.02.007.
Hochweber, J., I. Hosenfeld, and E. Klieme. 2013. "Classroom Composition, Classroom Management, and the Relationship between Student Attributes and Grades." *Journal of Educational Psychology* 106 (1): 289–300.
Hornstra, L., K. Stroet, E. van Eijden, J. Goudsblom, and C. Roskamp. 2018. "Teacher Expectation Effects on Need-Supportive Teaching, Student Motivation, and Engagement: A Self-Determination Perspective." *Educational Research and Evaluation* 24 (3–5): 324–345. doi:10.1080/13803611.2018.1550841.

IES. 2009. "NAEP Data Explorer." National Center for Education Statistics. http://nces.ed.
INVALSI. 2018. *Rilevazione nazionale degli apprendimenti 2017–2018: Rapporto risultati*. Roma: Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione. https://www.invalsi.it/invalsi/doc_evidenza/2018/Rapporto_prove_INVALSI_2018.pdf.
Jalava, N., J. S. Joensen, and E. Pellas. 2015. "Grades and Rank: Impacts of Non-Financial Incentives on Test Performance." *Journal of Economic Behavior & Organization* 115: 161–196. doi:10.1016/j.jebo.2014.12.004.
Kiss, D. 2013. "Are Immigrants and Girls Graded Worse? Results of a Matching Approach." *Education Economics* 21 (5): 447–463. doi:10.1080/09645292.2011.585019.
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. doi:10.1016/j.jpubeco.2008.02.009.
Lavy, V., and E. Sand. 2015. "On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases." NBER Working Paper Series.
Lindahl, E. 2007. "Comparing Teachers' Assessments and National Test Results: Evidence from Sweden." Working Paper 24, IFAU – Institute for Evaluation of Labour Market and Education Policy.
Lipnevich, A. A., T. R. Guskey, D. M. Murano, and J. K. Smith. 2020. "What Do Grades Mean? Variation in Grading Criteria in American College and University Courses." *Assessment in Education: Principles, Policy & Practice* 27 (5): 480–500. doi:10.1080/0969594X.2020.1799190.
Lord, F. M. 2012. *Applications of Item Response Theory to Practical Testing Problems*. New York: Routledge.
Machin, S., and T. Pekkarinen. 2008. "Global Sex Differences in Test Scores Variability." *Science (New York, N.Y.)* 322 (5906): 1331–1332. doi:10.1126/science.1162573.
Mickelson, R. A. 1989. "Why Does Jane Read and Write so Well? The Anomaly of Women's Achievement." *Sociology of Education* 62 (1): 47–63. doi:10.2307/2112823.
Miur. 2017. "Le iscrizioni al primo anno delle scuole primarie, secondarie di primo e secondo grado del sistema educativo di istruzione e formazione." Anno Scolastico 2017–2018 Giugno 2017, elaborazione su dati MIUR – Ufficio Statistica e Studi.
Montanaro, P. 2008. "I Divari Territoriali nella Preparazione degli Studenti Italiani: Evidenze dalle Indagini Nazionali e Internazionali." Questioni di Economia e Finanza. 14. Banca d'Italia.
OECD. 2009. *Equally Prepared for Life? How 15-year-old Boys and Girls Perform in School*. Paris: OECD (Organisation for Economic Co-Operation and Development).
OECD. 2014. *Are Boys and Girls Equally Prepared for Life?*. Paris: OECD (Organisation for Economic Co-Operation and Development).
OECD. 2019. PISA 2018: Insights and Interpretation.
Panichella, N., and M. Triventi. 2014. "Social Inequalities in the Choice of Secondary School." *European Societies* 16 (5): 666–693. doi:10.1080/14616696.2014.939685.
Pekkarinen, T. 2012. "Gender Differences in Education." *Nordic Economic Policy Review* 1 (1): 165–194.
PISA. 2019. *PISA 2018 Results*. Vol. II. Where All Students can Succeed; PISA, OECD Publishing. Paris.
Protivínský, T., and D. Münich. 2018. "Gender Bias in Teachers' Grading: What is in the Grade." *Studies in Educational Evaluation* 59: 141–149. doi:10.1016/j.stueduc.2018.07.006.
Reimer, D. 2019. "Teachers and Teacher Education: A Call for a Renewed Sociological Research Agenda." *International Studies in Sociology of Education* 28 (2): 90–109. doi:10.1080/09620214.2019.1601582.
Riegle-Crumb, C., and M. Humphries. 2012. "Exploring Bias in Math Teachers' Perceptions of Students' Ability by Gender and Race/Ethnicity." *Gender & Society* 26 (2): 290–322. doi:10.1177/0891243211434614.
Rindermann, H. 2007. "Die Bedeutung Der Mittleren Klassenfähigkeit Für Das Unterrichtsgeschehen Und Die Entwicklung Individueller Fähigkeiten [The Relevance of Class Ability for Teaching and Development of Individual Competences]." *Unterrichtswissenschaft* 35: 68–89.
Schofer, E., and J. W. Meyer. 2005. "The Worldwide Expansion of Higher Education in the Twentieth Century." *American Sociological Review* 70 (6): 898–920. doi:10.1177/000312240507000602.

Speybroeck, S., S. Kuppens, J. Van Damme, P. Van Petegem, C. Lamote, T. Boonen, and J. de Bilde. 2012. "The Role of Teachers' Expectations in the Association between Children's SES and Performance in Kindergarten: A Moderated Mediation Analysis." *PLoS One* 7 (4): e34502. doi:10.1371/journal.pone.0034502.

Steel, C. M. 1997. "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." *American Psychologist* 65 (5): 797–811.

Strand, S. 2012. "The White British–Black Caribbean Achievement Gap: Tests, Tiers and Teacher Expectations." *British Educational Research Journal* 38 (1): 75–101. doi:10.1080/01411926.2010.526702.

Terrier, C. 2015. "Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students' Achievement." CEP, Discussion Paper 1341.

Triventi, M. 2020. "Are Children of Immigrants Graded Less Generously by Their Teachers than Natives, and Why? Evidence from Student Population Data in Italy." *International Migration Review* 54 (3): 765–795. doi:10.1177/0197918319878104.

Vegas, E., and J. De Laat. 2003. "Do Differences in Teacher Contracts Affect Student Performance? Evidence from Togo." World Bank.

Wright, S. P., S. P. Horn, and W. L. Sanders. 1997. "Teacher and Classroom Context Effect on Student Achievement. Implications for Teacher Evaluation." *Journal of Personnel Evaluation in Education* 11 (1): 57–67. doi:10.1023/A:1007999204543.

Zhu, M., D. Urhahne, and C. M. Rubie-Davies. 2018. "The Longitudinal Effects of Teacher Judgement and Different Teacher Treatment on Students' Academic Outcomes." *Educational Psychology* 38 (5): 648–668. doi:10.1080/01443410.2017.1412399.