dECON
Facultad de Ciencias Sociales
UNIVERSIDAD DE LA REPÚBLICA

Documentos de Trabajo

Gender differences in teachers' assessments and blind test results – evidence from Uruguay

Marisa Bucheli - Florencia Amábile - Carmen Estrades

Documento No. 03/24
Abril 2024

ISSN 0797-7484

Gender differences in teachers' assessments and blind test results – evidence from Uruguay

Marisa Bucheli[*] - Florencia Amábile[*] - Carmen Estrades[*]

Abstract

This paper analyzes the existence of gender bias by public school teachers in Uruguay when grading students in the third and sixth years of primary level. The econometric strategy consists of estimating the effect of gender on the course score (non-blind outcome) when controlling by blind test scores and other relevant characteristics. We do not obtain evidence about a bias in the third year. However, we find an average bias in favor of girls in the sixth year, which responds to biases in the middle of the distribution of abilities (the extreme abilities are not gender-biased when assessed). The average results are robust to several checks. We rule out that sixth-year bias is mainly driven by statistical discrimination or explicit beliefs on talent gender stereotypes.

Keywords: gender differences, discrimination, stereotypes, teacher grading, blind-test, education.

JEL classification: I24, J16.

Resumen

Este trabajo analiza la presencia de sesgo de género por parte de los docentes de escuelas públicas de Uruguay al calificar a los estudiantes en tercer y sexto año de primaria. La estrategia empírica seguida consiste en estimar el efecto de género en la calificación del curso (resultado no ciego) con las obtenidas en pruebas estandarizadas corregidas por docentes que no observan las características de quienes las realizaron. No encontramos evidencia de la presencia de sesgo en tercer año. Sin embargo, encontramos la existencia de un sesgo promedio a favor de las niñas en sexto año, el cual responde a sesgos más

[*] Departamento de Economía, Facultad de Ciencias Sociales, Universidad de la República

Montevideo, Uruguay

Correo autor correspondencia: marisa.buecheli@cienciassociales.edu.uy

marcados en el centro de distribución de habilidades. Los resultados promedio son robustos a varias comprobaciones. Descartamos que el sesgo en sexto año sea principalmente impulsado por discriminación estadística o creencias explícitas sobre estereotipos de género en el talento de los estudiantes.

Palabras clave: diferencias de género, discriminación, estereotipos, calificación docente, pruebas ciegas, educación.
JEL classification: I24, J16.

1. Introduction

Numerous studies have analyzed whether there is a bias in teachers' grading toward several social categories, usually in developed countries, as documented in a review by Zanga and De Gioannis (2023). The teacher's bias refers to the difference in the grades they set between students with similar proficiency, which may be attributed to the social category of students, such as ethnicity or gender. Teachers' assessments are important because they give signals to students and parents considered in human capital decisions. Indeed, the final grade often serves as a pivotal information source for both students and their families, influencing subsequent study choices and potentially impacting educational continuity. Thus, grading bias may be crucial for students as long as teachers' assessments affect their efforts when investing in education, their decisions on dropout, and their choices of tracks and specialization fields (Bonesrønning, 2008; Terrier, 2020; Lavy and Sand, 2018). Besides, as long as the students may perceive bias associated with discrimination, there is evidence of potential adverse emotional effects reflected in an increased likelihood of depression, anger, and behavior disorders (Zanga and De Gioannis, 2023).

In this study, we use data from Uruguay to examine the gender differences in primary school students' grades set by teachers and the scores they obtained in a one-shot test administered by a third party. In Uruguay, grade retention and early dropout have been more likely among boys than girls for several decades (Bucheli and Casacuberta, 2000; Failache, Salas and Vigorito, 2018). There are various reasons behind this pattern, and the evidence has focused chiefly on the combined effect of low household resources and the gender gap in labor outcomes that favor boys. Thus, boys from disadvantaged backgrounds abandon school to enter the labor market. This paper looks at another issue related to the education system and grading practices that could cause gender differences in dropouts: a bias in teachers' assessments favoring girls over boys that demotivates school retention of boys and encourages retention of girls.

Most studies about gender gaps in teachers' grading in developed countries indicate higher assessments for girls than boys of similar proficiency, though the results are heterogeneous by subject and teacher characteristics (Lindahl, 2007; Lavy, 2008; Falch and Naper, 2013; Terrier, 2020; Protivinsky and Munich, 2018; Angelo and Reis, 2021). To our knowledge, the only evidence for a Latin American country is provided by Contreras (2023 who find a similar result for Chile.

Our study uses data from a study by a UNESCO project, TERCE (Tercer Estudio Regional Comparativo y Explicativo). The study administered tests covering several disciplines to a sample of children attending primary school's third and sixth years. We also have information on the teachers' assessments of the TERCE sample 2013 students provided by the National Educational Administration.

The standard methodology of the empirical literature relies on comparing, for specific disciplines, the scores set by teachers, who are aware of the gender and other attributes of the students, and the scores set by blind graders, which would indicate an objective measure of students' academic skills. The strategy consists of regressing the difference between scores on students and class characteristics or regressing the non-blind score on characteristics and the blind score. In this work, we have to adjust this approach because, in the Uruguayan grading system, the teacher does not give scores by discipline but sets a global performance score at the end of the school year. Thus, we regress this overall score set by teachers on the scores students obtained in the TERCE standardized tests of disciplines, plus other covariates usually used in the empirical literature.

This work contributes to the literature by providing information for a Latin American country in which, up to our knowledge, there is only evidence for Chile and Uruguay (Contreras, 2023; Bucheli and Contreras, 2018). Besides, it adapts the empirical strategy for a country where teachers grade the overall performance and not each discipline separately, with the risk of increasing discretion and bias by social categories. It also has the advantage of examining two scholastic years (and not just one).

The main finding is that, on average, teachers favor girls over boys when grading in the sixth grade, but there is no bias in the third grade. The main heterogeneity among the sixth-year students is that the extreme abilities are not gender biased assessed. Besides, in both scholastic years, teachers who favor boys and girls co-exist, and the bias against boys is higher for younger and the less experienced teachers, and for male than female teachers. Evidence for the presence of statistical discrimination is weak and gender unequal talent beliefs seem to deepen already existing grading biases.

The rest of the article is organized as follows. Section 2 reviews the related literature and Uruguay's educational background. Section 3 describes the data and methods. Section 4 shows the results and analyzes the findings, and Section 5 concludes.

2. Related literature and Uruguayan educational background

2.1. Literature review

A bias in grading refers to a systematic underestimation of a group of students' assessment based on a characteristic other than their actual performance. In a review of studies on grading bias, Zanga and De Gioannis (2023) report that the most analyzed students' biased characteristic is gender, followed by race or ethnicity and, less frequently, migration status, weight, and physical attractiveness, among others.

The main evidence comes from studies analyzing students' assessments in mathematics, language, and less frequently, science, history, and a foreign language in primary and secondary schools in developed countries. In the studies that focus on gender, the most common result is an average bias against boys, regardless of discipline and educational level, as found in studies of Sweden (Lindahl, 2007), Israel (Lavy, 2008), Norway (Falch and Naper, 2013), France (Terrier, 2020), Czech Republic (Protivinsky and Munich, 2018), Portugal (Angelo and Reis, 2021), and Chile (Contreras, 2023). Meanwhile, few cases report that, on average, there is bias against girls or no bias (Hinnerich, Höglin and Johansen, 2011; Doornkamp et al., 2022).

The method followed by empirical studies compares grades given by teachers who know students' characteristics with blind grades. Some evidence relies on experiments that provide a better-controlled frame and collect data from assessments of identical tests. Other studies use quasi-experimental data and compare teachers' grading and standardized tests graded by external examiners unaware of the students' social categories.

Most of the literature on grading bias is inclined to assign the found bias to teachers' behavior and to discuss discrimination issues. This interpretation requires ruling out explanations based only on the contents of the tests and the student's behavior. The content of the tests is one of the weaknesses of quasi-experimental data (Graetz and Karimin, 2022). Regarding students' behavior, the most quoted concern is that in the blind test, boys and girls put in different efforts or feel different anxiety levels, affecting the performance gap relative to the regular course achievements gap (Protivínsky and Münich, 2018). Despite these issues, most studies conclude that these explanations are not enough to explain the gaps.

One way to detect that the bias is due to teacher's behavior is to analyze if it varies among teachers' characteristics, such as gender and experience. Lavy (2008) points out that this finding suggests teachers's discrimination because there is no reason to expect students'

behavior to be consistent with the relation between the bias and teachers' characteristics. However, the relationship between bias and teachers' characteristics is not systematic across studies and even between disciplines within studies. For example, Lavy (2008) finds that while in mathematics, the gender bias is explained by the behavior of male teachers, mainly the oldest and the most experienced ones; in other disciplines, such as biology, chemistry, and physics, it is related to the behavior of the youngest and less experienced female teachers. Falch and Naper (2013) obtain same-sex punishment in Norwegian (native language) but not in mathematics and English (foreign language). Also consistent with same-sex punishment, Lindahl (2016) finds that female teachers are less generous with girls than boys when grading mathematics. Breda and Ly (2015), in a study of admission to a higher education institution, conclude that teachers overscore women in fields dominated by men (mathematics and philosophy) but favor men in fields dominated by women (literature and biology).

The empirical literature analyzes various potential explanations for grading bias linked to discrimination exerted by teachers. One is the statistical discrimination hypothesis, which states that under imperfect information, observed characteristics are signals of unobserved ones that correlate with academic performance. For example, if teachers suspect boys are more prone to cheat than girls, their assessments will be less reliable. Consequently, grading practices will be different for boys and girls. Studies by Lavy (2008) and Contreras (2023) rule out this channel at least as the unique main explanation of gender bias. On the contrary, Hanna and Linden (2012), who follow an experimental method that allows for analyzing the effect of the order in which participant teachers grade, find an order effect on the bias and conclude that this result is consistent with statistical discrimination.

Another hypothesis is that discrimination relies on gender stereotypes, such as teachers believing that boys and girls have different talent. Carlana (2019) and Doornkamp et al. (2022) analyze the role of gender stereotypes in explaining grading bias by eliciting teachers' explicit and implicit (non-conscious) attitudes and testing their effects. Carlana (2019) concludes that in maths, boys benefit from male teachers who believe that boys have innate advantages over girls, but she finds no effects of gender stereotypes about reading. Doornkamp et al. (2022) find that gender grading bias, which is null on average, varies with teacher's expectations: they gave higher scores to the gender of which they expect more talent and effort.

The studies of gender stereotpyes give support to the hypothesis of implicit discrimination. In this case, discrimination responds to inconscient attitudes and so, teachers would exert a bias unintentionally and without awerness.

Finally, students' behavior in class may bias teachers's scores. Cornwell, Mustard and Van Paris (2013) analyze the bias in reading, mathematics, and science grading in kindergarten and primary schools in the United States. Unlike previous studies, they account for non-cognitive characteristics measured by an attitudes-to-learning index. They find that the average grading bias favorable to girls responds to systematic gender differences in the index. Even in some disciplines, the bias of the sign changes when considering it. The authors explain that these changes rely on (consciously or subconsciously) teachers' behavior: teachers reward good attitudes toward learning, and better attitudes are more frequent among girls. Contreras (2023) finds a similar result for Chile: the average grading bias against boys vanishes when considering attendance rate and grade retention. Finally, in a study for Brazil, Ferman and Fontes (2022) explore the impact of the student's behavior, measured by an index based on teachers' qualitative assessments, on math scores, and conclude that teachers penalize bad behavior.

2.2. Institutional background

The Uruguayan education system comprises 14 compulsory schooling years classified into four levels: two years of early childhood education (the age of entrance is four), six years of primary education, three years of lower secondary education, and three of upper secondary education.

Children have almost universal primary education, and there are no gender differences in attendance. However, the repetition rates are lower for girls than boys, so the former enter secondary education at younger ages (Bernatzky and Cid, 2015; ANEP, 2023). Dropout starts in secondary education, and the fulfillment of compulsory education is weak. The official information reports that in recent years, just over 80% of a cohort finished lower secondary education, and only 40% completed upper secondary education (INEEd, 2022). Dropout is more frequent among boys than girls: the gender gap in completed secondary education has been around 15 percentage points in the last fifteen years (INEEd, 2022).

Public and private institutions offer education services, and all establishments offer the same curricula. In the last fifteen years, the public system has covered around 85% of enrollment at the primary level (ANEP, 2023).

Usually, the school year calendar begins in March and ends in December. To enter primary school, children must have age six before April 30. So, in the first grade, students are 6 or 7 years old, and those who do not repeat any grade finish school at ages 11 or 12.

At the primary level, teachers score students' academic performance using a 0 to 12 scale at the end of the school year. Students must attend 80% of classes and reach a minimum score of 6 in academic performance to pass. Teachers are the only ones responsible for their class's test preparation, assessments, and grading. The Uruguayan Administration of Education provides written guidelines to teachers covering different topics. In practice, in Uruguay, teachers have great discretion when setting grades.

The National Administration of Primary Education organizes the matching process of teachers to public schools. Each year, the Administration elaborates a list that starts with the so-called effective teachers; the effectivity status allows a permanent assignment in a school and is obtained through a selection process that includes merits and tests. The list continues with interim teachers and, finally, recent graduates. Within each group, the precedence order is based on performance. At the beginning of the year, the Administration posts the list of teachers and schools with vacancies. Teachers choose a school following the list's precedence order. The school's principal assigns teachers to classes: assignment guides state that the effective status should be prioritized for filling the first and sixth-year positions.

3. Data and Methods

3.1. Data and descriptives

In October 2013, *Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación* implemented tests in Uruguayan primary schools in the framework of a project led by UNESCO intending to set common anonymous tests in schools in 16 countries of Latin American and the Caribbean. This assessment, known as TERCE (*Tercer Estudio Regional Comparativo y Explicativo*), considered each country's curricula, aiming to be consistent with the expected performance outcomes. Using the plausible-values approach, it tested a sample of third-year and sixth-year classes and graded the tests by correctors (not the students' current teachers). Tests for third-year students covered mathematics, reading, and writing; tests for sixth-year students also included natural science. The details of the implementation and sample design are provided by UNESCO-OREALC (2016).

The mathematics, reading, and natural sciences tests contain over 90% multiple-choice questions and less than 10% short-answer questions. The score ranges from 250 to 1150. Meanwhile, the writing test assesses the production of long texts, and the score ranges from 1 to 4. The TERCE database provides information about the scores obtained in each discipline.

Besides, the TERCE project carried out surveys to be filled out by the children, a family adult, the school principal, and the teacher. Thus, the database contains information about several characteristics of the tested children and their families, schools, and teachers.

The National Administration of Primary Education provided additional information for the students attending public schools: the share of days that the students attended school in 2013 and their academic performance and classroom behavior scores set by teachers at the end of the schooling year. Thus, our database provides information on blind and non-blind scores only for the public school system: 1959 third-year and 2099 sixth-year students. [2]

Table 1 reports the average scores obtained in boys' and girls' academic performance in the schooling year and their grades in TERCE (only plausible value 1). All assessments are reported in z-scores, that is, the number of standard deviations below or above the mean value, which eases the interpretation of the results.

The grades set by teachers indicate that girls obtain higher scores than boys, though the academic performance gender gap is not statistically significant in the third year. Meanwhile, the TERCE tests indicate that girls perform better in writing in both academic years, whereas there are no gender differences in other subjects.

[2] In 2013, 82.5% of third and 84.2% of sixth-year students attended public schools. There is no gender selection bias in attending private/public schools.

Table 1. Mean scores set by teachers and obtained in TERCE (standard deviations into parenthesis). In z-scores.

Variables	Third-year			Sixth-year		
	Girls	Boys	Gender gap	Girls	Boys	Gender gap
Scores set by teachers						
Academic performance set by teachers	0,057	-0,053	-0,110	0,110	-0,121	-0,231**
	(1,0262)	(0,9728)		(0,9946)	(0,9908)	
Scores in TERCE						
Mathematics	0,071	-0,065	-0,135	-0,053	0,059	0,111
	(1,0078)	(0,9873)		(1,0283)	(0,9611)	
Natural Sciences				-0,003	0,004	0,007
				(0,9596)	(1,0446)	
Reading	0,070	-0,064	-0,134	0,039	-0,043	-0,081
	(1,0087)	(0,9866)		(0,9912)	(1,0075)	
Writing	0,173	-0,158	-0,331***	0,132	-0,146	-0,278***
	(0,9951)	(0,9769)		(0,9686)	(1,0144)	
Observations	981	978		1071	1028	

*** p<0.01, ** p<0.05, * p<0.1 for a test of means testing the null hypothesis that the variable is equal for girls and boys

Source: Own elaboration based on TERCE database

Table 2 shows descriptive statistics of relevant variables usually used to explain academic performance. Two individual characteristics of children are related to past education outputs: grade repetition before 2013 and preschool attendance, reported by parents. Repetition and preschool attendance are more frequent among boys than girls, though the gender gap is not statistically significant for the sixth-year sample. Besides, two variables refer to the behavior traits of students in 2013: the number of days that the child attended classes (provided by administrative records, as mentioned) and the parent's report of time spent studying at home measured by the number of days per week. The average values indicate that girls behave better than boys according to the two variables (though the difference in attended days for the sixth year is negligible).

Table 2. Average value of selected characteristics of students, their families, and schools (standard deviations into parenthesis).

Variables	Third-year			Sixth-year		
	Girls	Boys	Gender gap	Girls	Boys	Gender gap
Past educational outputs						
Grade repetition	0,159	0,268	0,109***	0,205	0,234	0,029
(Yes=1)	(0,375)	(0,432)		(0,399)	(0,429)	
Preschool	0,532	0,612	0,081**	0,491	0,556	0,065
(Yes=1)	(0,510)	(0,477)		(0,499)	(0,498)	
Behavioral traits						
Attendance	164,003	157,897	-6,106***	159,547	159,801	0,254
(days)	(17,135)	(25,220)		(28,464)	(24,249)	
Study at home	4,839	4,482	-0,357***	4,593	4,236	-0,357***
(days per week)	(1,3600)	(1,374)		(1,427)	(1,556)	
Family characteristics						
Father's education	0,420	0,432	0,012	0,343	0,388	0,044
(secondary/tertiary=1)	(0,505)	(0,484)		(0,468)	(0,495)	
Mother's education	0,464	0,527	0,063	0,481	0,485	0,003
(secondary/tertiary=1)	(0,511)	(0,488)		(0,492)	(0,508)	
Household income	4,181	3,790	-0,391	3,711	3,973	0,262
(range: 1 to 10)	(2,494)	(2,361)		(2,321)	(2,360)	
Homework supervision	2,657	2,527	-0,130*	2,520	2,578	0,058
(range: 0 to 3)	(0,678)	(0,686)		(0,856)	(0,821)	
Schools						
Rural areas	1,064	1,055	-0,008	1,054	1,038	-0,016
(Yes=1)	(0,250)	(0,224)		(0,222)	(0,194)	
Full-time school	0,557	0,571	0,014	0,501	0,512	0,012
(Yes=1)	(0,510)	(0,483)		(0,488)	(0,513)	
School inputs	0,596	0,686	0,090**	0,653	0,660	0,007
(yes=1)	(0,502)	(0,454)		(0,468)	(0,4831)	
School infraestructure	1,427	1,388	-0,039	1,503	1,419	-0,084***
(range: 0 to 5)	(0,979)	(0,937)		(0,886)	(0,927)	

*** p<0.01, ** p<0.05, * p<0.1 for a test of means testing the null hypothesis that the variable is equal for girls and boys

Average calculated on valid values. The number of missing values for third-year (sixth-year) are: grade repetition: 9 (12); preschool: 99 (123); father's and mother's education: 0(0); household income: 88(100); homework supervision:142 (170); rural areas: 0 (0); full-time school: 87(80); school inputs: (67 (52); attendance: 0 (0); study at home: 86 (106).

Source: Own elaboration based on TERCE database

In addition, Table 2 reports information based on the survey of children's families. The questionnaire inquires about household income based on income intervals, intending to reflect the income distribution's deciles. We used this report to build a variable ranging

from 1 (lowest income range) to 10 (highest). The survey also inquires about the educational level of the father and the mother. We built a dummy variable for each parent that distinguishes whether he/she has at least secondary education. The questionnaire also gathers information on whether parents supervise homework. Specifically, it asks how often parents a) make sure that children have done their homework, b) ask about what children did at school, and c) inquire about school grades. We assigned the value 1 to the answers "always" and 0 to the answers "never" or "sometimes" and we built a variable equal to the sum of these values -which ranges from 0 to 3-. As shown in Table 2, the only statistically significant gender difference is that third-grade girls are more supervised than their boys' classmates.

Finally, Table 2 informs about schools. We built a dummy variable to capture whether the school is located in a rural or urban area and another one to distinguish whether it is full-time. Besides, the questionnaire asks the principal if all schoolrooms have chalk or whiteboard markers, a teacher's table, a teacher's chair, tables for all students, and chairs for all students. We built a variable "school inputs" that takes value 0 when at least one of these inputs is lacking. Additionally, the principal informs if there is a computer room, an event room, a music/art room, a science lab, or a library in the school. We built a variable of school infrastructure, ranging from 0 to 5, equal to the sum of answers "yes". We find an average gender difference in two variables: availability of school inputs (in favor of boys in third-year) and infrastructure (in favor of sixth-year girls).

3.2. Empirical strategy

The gold standard in the literature is implementing a diff-in-diff strategy that compares the outcome of a blind and a non-blind test in a specific subject. However, we implement a different approach as teachers' assessment of academic performance is measured by one overall score. Thus, we state the following empirical representation:

$$y_{ij} = \alpha + \delta\, TERCE_{ij} + \theta X_{ij} + \beta\, boy_{ij} + \varphi_j + \varepsilon_{ij} \qquad (1)$$

where y_{ij} is the academic performance score (in z-scores) of child i in school j, TERCE is a vector that includes the scores in TERCE tests (three in the third year and four in the sixth year), X is a vector of individual, family and school characteristics, and φ are teacher fixed effects. The parameter of interest is β, the coefficient of a variable dummy (boy) that takes the value 1 when the student is a boy and 0 when she is a girl. We perform a weighted OLS with standard errors corrected by linearization.

We also build an aggregate measure of the TERCE tests (hereafter called the aggregate blind-test index), whose main challenge is calculating the weight of the disciplines. We estimate several specifications of equation (1) and interpret that the estimated coefficients δ measure the actual weight that teachers give to disciplines. The estimations, reported in Table A1 of the Annex, indicate that the coefficients do not have important variations between specifications within the scholastic year but vary between scholastic years. Thus, we estimate an aggregate index for the third and sixth years separately, with different weight structures. In the third year, all disciplines have the same weight. In the sixth year, the weights are 0.40 for mathematics, 0.21 for reading, 0.15 for writing, and 0.24 for sciences.

As our dependent variable refers to global academic performance, we have two sources of measurement error. First, the academic performance includes subjects not assessed in TERCE. We do not know their importance in the overall score and if there are systematic gender differences in these non-observed subjects' performance. Secondly, teachers assess performance during the academic year, whereas TERCE is a one-shot test. Systematic gender differences in tests, such as anxiety or attitudes toward challenges, may affect the interpretation of the estimated β.

Because of our awareness of omitted variables, we conducted robustness checks known as the AETO tests. Altonji, Elder and Taber (2005) propose a strategy that helps assess the estimated treatment effect (in our case, gender) when no instrumental variables are available. Oster (2019) extended and developed this proposal and provided recommendations that we follow to assess our estimated parameter of interest. We specifically report two calculations following complementary approaches (AETO tests).

One approach calculates the ratio of selection on unobservables to selection on observables required to attribute the entire obtained effect of gender to selection bias. If we name Rmax the R-squared of a regression that includes all observable and unobservable variables, we estimate the ratio δ for which the gender effect is zero. In other words, δ indicates how much correlated with treatment (gender) the unobservables would need to be to explain the entire association between treatment and the outcome of interest. Altonji, Elder and Taber (2005) suggest that $|\delta|=1$ would be an appropriate cut-off: note that when this equality holds, unobservable variables explain as much of the outcome as the actual controls. They argue that with rich datasets regarding explanatory variables, unobservables seem unlikely to be stronger than observables to explain the outcome.

The second approach consists of bounding the treatment effect. The procedure consists of running a baseline regression with only the gender variable as a regressor and another regression including the full set of available covariables. This procedure allows us to assess the magnitude and direction of the change in the gender parameter after including the observable characteristics. The bound of the effect of being a boy is:

$$\beta^{*} = \tilde{\beta} - \delta\left[\dot{\beta} - \tilde{\beta}\right]\left(\frac{R_{max} - \tilde{R}}{\dot{R}}\right) \quad (2)$$

where $\tilde{\beta}$ is the estimated coefficient of equation (1), $\tilde{R}$ is R-squared obtained from equation (1), and $\dot{\beta}$ and $\dot{R}$ are the coefficient and R-squared obtained with the estimation without controls. To estimate β*, we must make assumptions about δ and R_{max}. Because of the recommendations of Altonji, Elder and Taber (2005), we assume |δ|=1 for our calculus. As unobserved and observed controls fully explain the outcome, we may assume R_{max}=1. However, Oster (2019) argues that this value may lead to an over-adjustment and proposes to use an alternative R_{max} equal to 1.3 or 1.5 times $\tilde{R}$.

To analyze heterogeneity between students, we classify students in T groups and estimate the modified version of equation (1):

$$y_{ij} = \alpha + \delta\, TERCE_{ij} + \theta X_{ij} + \sum_{t=2}^{T} \emptyset^{t} q_{ij}^{t} + \sum_{t=1}^{T} q_{ij} * \beta^{t} boy + \varphi_{j} + \varepsilon_{ij} \quad (3)$$

The variables q^t are dummy variables that indicate if the student belongs to the group *t* (t=1,..,T) and the β^t set captures the heterogeneous bias between groups.

4. Results

4.1. Basic findings

In Table 3, we present the estimated gender gap (using equation 1) where a positive value indicates that, on average, teachers set a higher academic performance score for boys than girls.

In third grade, there is no raw gender difference in the score set by teachers. The controls of TERCE's scores and individual, family, and school characteristics do not change this result: the gender gap is not different from 0 at usual statistically significant levels, as reported in Columns (1) to (7).

Table 3. Estimated gender gap in grading (standard errors into parenthesis)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gender gap: Third-year	-0.110	0.029	0.108	0.094	0.079	0.069	0.063
	(0.117)	(0.078)	(0.093)	(0.087)	(0.079)	(0.077)	(0.079)
Gender gap: Sixth-year	-0.231**	-0.221***	-0.175***	-0.189***	-0.186***	-0.186***	-0.191***
	(0.097)	(0.065)	(0.062)	(0.062)	(0.061)	(0.056)	(0.054)
Controls:							
TERCE scores		Yes	Yes	Yes	Yes	Yes	Agg.index
Child's characteristics			Yes	Yes	Yes	Yes	Yes
Family background				Yes	Yes	Yes	Yes
School's characteristics					Yes	Yes	Yes
Teacher's fixed effect						Yes	Yes
R-squared: Third-year	0.003	0.367	0.414	0.447	0.494	0.647	0.645
R-squared: Sixth-year	0.013	0.409	0.496	0.513	0.536	0.663	0.662

*** p<0.01, ** p<0.05, * p<0.1

Note: The gender gap is the estimated coefficient of a dummy variable that takes value 1 for boys and 0 for girls. TERCE scores are plausible values 1. Number of observations: 1959 (third-year) and 2099 (sixth-year).

Source: Own elaboration based on TERCE database

On the contrary, in sixth year, the raw mean score is lower for boys than girls (Column 1) and slightly reduces when we control by TERCE's scores (Column 2). The gender gap declines again in the estimation of Column (3), that is, when we add children's characteristics. These variables reflect traits strongly related to performance: past grade repetition, preschool attendance, attendance days in the current year, and parents' reports of studying at home. These variables reflect attitudes toward learning and students' effort, and as reported in Table 2, girls outperform boys in most of these dimensions, as in with the studies for Chile (Contreras, 2023) and the USA (Cornwell, Mustard and Van Paris, 2013). The decline of the gender gap indicates that they are considered in teachers' grading but do not explain all the bias.

The rest of the columns indicate that the introduction of family background, school characteristics, and teacher's fixed effect does not change the main result about the gender gap. According to the full estimation reported in Column (6), sixth-year boys who score as well as girls in the TERCE tests, and after controlling other characteristics, receive teachers' grades that are, on average, 0.186 standard deviations lower, which lies within the 0.10-0.30 range most frequently found in the literature (Lavy, 2008; Falch and Naper, 2013; Terrier, 2020; Contreras, 2023; among others).

Finally, Column (7) reports the gender gap in a model using the aggregate blind-test index based on TERCE scores presented in Section 3.2. The result is similar to our preferred estimation in Column (6).

Thus, after controlling by relevant characteristics, we found no gender bias in teachers' grading in the third year and favoritism for girls over boys in the sixth year.

4.2. Robustness checks

The interpretation of the gap as a measure of teachers' bias requires ruling out the possibility that the relative performance of boys and girls is different in classroom and TERCE assessments. Indeed, the gap may result from boys' superior performance in TERCE compared to accomplishments in class relative to girls.

In a one-shot test, girls and boys could react differently because of gender differences in psychological attributes and preferences, such as anxiety, inclination to competition, or fear of feedback. For example, there is laboratory evidence that men are more competitive than women and put more effort into tournaments (Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007; Croson and Gneezy, 2009; Charness and Gneezy, 2012), though critical reviews of the behavioral economics literature on gender differences assess that both statistical and economic significances are not-robust or negligible (Sent and Van Staveren, 2018). As long as students interpret TERCE assessments as more competitive than class routines, boys may put more effort and perform better in the tests than in class, leading to the estimated gap. Teachers could exacerbate this phenomenon if they feel the student's performance reflects their own. Thus, in the interest of obtaining good student scores under the interpretation that they assess teachers' abilities, teachers could emphasize the importance of TERCE, provoking an outperformance of more competitive children (boys).

Another issue at stake is that TERCE tests do not form part of the formal school's assessment and do not have curricular consequences. Thus, we may expect that some children do not put all their effort into it even when teachers encourage them. If this response to TERCE is more frequent among girls, the obtained results could be explained by students' behavior and not by teachers' behavior.

The AETO tests, reported in Table 4, give insights into the relevance of the unobserved variables. We only did the estimates for the sixth year because the gender effect in the third grade is not statistically significant. As presented in Table 3, the gender raw gap is -0.231

and declines to -0.186 when observable controls are introduced. Thus, the change in the coefficient of interest indicates that the used controls (observable variables) are positively correlated with gender bias. If the unobservable and observable variables are positively correlated, introducing the former in the regression would risk vanishing the gender effect. So, we calculate the ratio of selection on unobservables to selection on observables required to attribute the entire obtained effect of gender to unobservables. As reported in Table 4, the estimated δ indicates that unobservables must be five times more important than observables for us to find an effect that is actually null, which exceeds the recommended cut-off.

Table 4 also reports the boundaries of the gender effect under different maximum R, assuming a positive correlation between observable and non-observable variables and δ equal to 1. The negative bounds indicate that the confidence interval does not include the null effect.

Table 4. Selection on unobservables in sixth year: estimates of δ for β=0 and Rmax=1, and adjusted β under various assumptions on Rmax and δ=1.

	(1)	(2)	(3)	(4)
Change in β	δ for β=0 and R_{max}=1	Adjusted β for δ=1 and:		
		R_{max}=1	R_{max}=1.5R	R_{max}=1.3R
0.0450	5.0636	-0.1588	-0.1594	-0.1705

Source: Own elaboration based on TERCE database

In addition, we did several robustness checks. First, we estimated the full model specification as stated in equation (1) but included the TERCE scores as polynomials of order 4. As shown in Table 5, the results are not sensitive to the functional form in which the TERCE scores are treated.

We also estimated the full model for several subsamples, dropping alternative cases: students with a missing value in any covariate, in classes with five students or less, in classes with ten students or less, and in classes of only boys or girls. The findings with these different subsamples are robust to the main results.

Finally, we interact teachers' fixed effects with the aggregate blind-tests index. This specification controls the eventual relationship between students' characteristics and teachers' grading practices (though not unexpectable and uneasy to explain if existing). As

reported in the last row of Table 5, the estimated gender gap is still not statistically significant in the third year and negative in the sixth year, although of a lower magnitude.

Table 5. Robustness checks: estimations of the gender gap in grading (standard errors in parenthesis and size sample in brackets)

	Third year	Sixth year
Polinomy on TERCE scores (order 4)	0.078	-0.182***
	(0.077)	(0.057)
	[1959]	[2099]
Subsample without missing values in all controls	0.014	-0.175**
	(0.087)	(0.069)
	[1376]	[1508]
Subsample of students in class sizes over 5	0.073	-0.179***
	(0.076)	(0.056)
	[1861]	[2016]
Subsample of students in class sizes over 10	0.076	-0.180***
	(0.076)	(0.058)
	[1669]	[1846]
Subsample of students in mixed classes	0.067	-0.185***
	(0.076)	(0.056)
	[1932]	[2084]
With teachers' fixed effects interacted with the aggregate	0.054	-0.153**
blind-tests index	(0.079)	(0.060)
	[1959]	[2099]

*** p<0.01, ** p<0.05, * p<0.1

Source: Own elaboration based on TERCE database

4.3. Heterogeneity of the grading bias among groups of students

Our basic estimation indicates that, on average, there is no bias in the third year and punishment for boys in the sixth year. This finding may be driven by score disparities in specific parts of the distribution of students' abilities. To explore this possibility, we classify students by the aggregate blind-test index quartiles and estimate equation (3) without including the TERCE vector as a dependent variable.[3] Table 6 reports the results.

In the third year, we do not find gender bias in any position. In the sixth year, the estimates are negative for all the quartiles. However, the coefficients are statistically significant only

[3] As mentioned in Section 1, previous works analyze the bias by discipline. Thus, when exploring the bias across the ability distribution, the percentiles of ability are the percentiles of the blind-score in the discipline (for example, Angelo and Reis, 2021; Gortazar, Martinez de la Fuente and Vega-Bayo, 2022).

for the second and third quartiles. They indicate that boys' scores are around 0.3 standard deviations lower, suggesting that the extreme abilities are not biasedly assessed.

Table 6. Estimated gender bias in grading by quartiles of the ability distribution measured by the aggregate blind-tests index (standard errors into parenthesis)

Groups of the ability distribution	Third-year	Sixth-year
First quartile	0.097	-0.160
	(0.121)	(0.137)
Second quartile	0.111	-0.341***
	(0.178)	(0.124)
Third quartile	-0.114	-0.289***
	(0.135)	(0.092)
Fourth quartile	0.060	-0.093
	(0.144)	(0.088)

Note: The estimated gaps are the estimated coefficients of the interaction of a gender dummy and dummies capturing the quartiles of the aggregate blind-tests index. The control variables are the child's characteristics, family background, school's characteristics, teacher's fixed effect, and dummies capturing the quartiles of the aggregate blind-tests index.

Source: Own elaboration based on TERCE database

We also analyze whether there is heterogeneity between the past outcomes (grade repetition and preschool attendance) and behavioral traits variables (current attendance and time spent studying at home). We present the results in Table 7. In the third year, we find three groups of students in which there is a bias favorable to boys: children who did not attend preschool and, with less precision, children having experienced grade repetition in the past and children with a current low attendance rate. In the sixth year, gender bias is negative for all subgroups of students, which are classified according to past outcomes and behavioral traits. Testing the null hypothesis that the gender gap between groups is null indicates no significant differences.

Table 7. Estimated gender gap in grading by subgroups based on students' behavior (standard errors in parenthesis)

	Third year		Sixth year	
Grade repetition				
No	0.050	(0.088)	-0.174***	(0.059)
Yes	0.225*	(0.131)	-0.230**	(0.096)
Preschool attendance				
No	0.207***	(0.069)	-0.164**	(0.078)
Yes	-0.019	(0.106)	-0.216**	(0.085)
Attendance: more than 170 days in the current year				
No	0.151*	(0.091)	-0.246***	(0.079)
Yes	-0.063	(0.072)	-0.078	(0.069)
Study at home: more than 4 days per week				
No	0.110	(0.100)	-0.191***	(0.068)
Yes	0.049	(0.099)	-0.163*	(0.084)

*** p<0.01, ** p<0.05, * p<0.1 testing the null hypothesis that the gender gap is null

Source: own elaboration based on TERCE database

Finally, we consider the income quintile distribution of the households where the children live, presented in Table 8. We do not find any in the quartiles of the third year, but there is a bias in sixth grade among children from deprived households and those with the highest income.

Table 8. Estimated gender gap in grading by income quartiles (standard errors in parenthesis)

Income quartile distribution	Third year		Sixth year	
First quartile	0.041	(0.116)	-0.402***	(0.091)
Second quartile	0.120	(0.080)	-0.101	(0.090)
Third quartile	0.132	(0.150)	-0.071	(0.099)
Fourth quartile	-0.047	(0.172)	-0.206***	(0.078)

*** p<0.01, ** p<0.05, * p<0.1 testing the null hypothesis that the gender gap is null

Source: own elaboration based on TERCE database

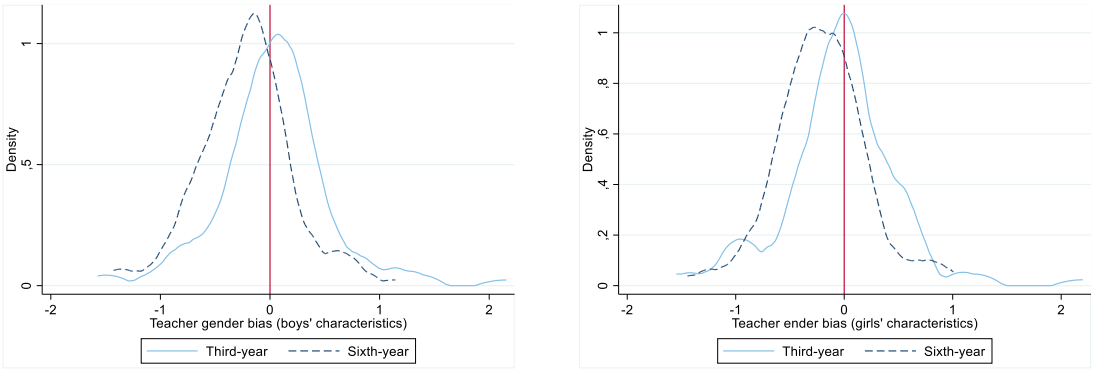
4.4. Heterogeneity of the grading bias among teachers

A bias that varies between teachers suggests that it responds to teachers' behavior, making the analysis of heterogeneity relevant. Thus, we estimate the grading bias distribution by scholastic year by pursuing the following strategy. We estimate equation (1), specifically the model reported in column 6 of Table 3, separately for boys and girls. Then, we predict

for each boy his grade and the grade he would have obtained using the parameters estimated for girls. The difference between these two scores is the gap attributable to the actual score and a counterfactual he would have obtained if treated as a girl. We calculate the bias of the teacher as the mean bias by class: it indicates that the teacher is biased in favor of boys when positive and against boys when negative. We analogously estimate the bias distribution based on girls' characteristics.

Figure 1 plots the teachers' average bias density functions based on the boys' characteristics (at left) and girls' characteristics (at right). The two graphs reflect the same patterns. Consistently with the estimates shown in Table 3, the function for the sixth year is at the left of the third year. The mode is negative for the sixth year and close to zero for the third year. The new features depicted in the pictures relate to the differences between teachers. In the sixth year, the function accumulates more observations below zero than above, indicating that the portion of teachers who favor boys is lower than the ones favoring girls. In the third year, where the average bias is statistically non-significant, the function indicates that teachers favoring boys and girls co-exist, and the share of the former is higher than the latter.

Figure 1. Density functions of the teachers' bias



Note: Estimations after regressing the academic score equation separately for boys and girls. At left, the bias is the difference between the predicted actual scores of boys and the predicted scores if treated as girls. At right, the bias is the difference between the predicted actual scores of girls if treated as boys and the predicted actual scores of girls.

Source: Own elaboration based on TERCE database

Heterogeneity between teachers raises the question of whether the bias is correlated with observable characteristics. We used the teachers' survey to select the variables (for

descriptives, see Table A2 of the Appendix). We classify students according this information and estimate equation (3). Table 9 shows the estimated gender gap by group and scholastic year.

Table 9. Estimated gender gap in in grading by subgroups based on teachers' characteristics (standard errors in parenthesis)

	Third-year		Sixth-year	
Age				
Up to 42	-0.018	(0.120)	-0.216***	(0.041)
43 or more	0.169**	(0.079)	-0.128	(0.127)
Years of experience				
Up to 15	0.042	(0.109)	-0.242***	(0.044)
16 or more	0.118	(0.091)	-0.032	(0.135)
Gender				
Male	-0.707***	(0.218)	-0.313***	(0.096)
Female	0.110	(0.068)	-0.166***	(0.062)
Permanent assignment				
No	0.117	(0.154)	-0.301***	(0.046)
Yes	0.044	(0.079)	-0.112	(0.081)
Official documents available				
No	0.090	(0.080)	-0.181***	(0.053)
Yes	-0.006	(0.200)	-0.195*	(0.115)
Teacher's performance assessment				
No	0.176**	(0.071)	-0.177***	(0.066)
Yes	0.044	(0.131)	-0.175*	(0.093)

*** p<0.01, ** p<0.05, * p<0.1 testing the null hypothesis that the gender gap is null

Source: own elaboration based on TERCE database

The three first classifications refer to demographic variables that, according to previous works, are potentially related to performance and grading practices. The patterns are similar for both scholastic years: bias against boys is higher for younger teachers (age below the median), the less experienced (below the experience years median), and for male than female teachers. However, the precision of the difference between groups is weak, and the only statistically significant difference is the teachers' gender effect in the third year.

We also analyze three variables related to school policies, which we expect to reduce biases. One is the permanent assignment status in the school, with an incidence of 65%, higher in the sixth than in the third year. A permanent assignment may capture better skills, but not having it may encourage additional effort to obtain it. Thus, the expected sign is

ambiguous. The second variable is the teacher's report about the access to official documents about teaching practices, including assessments, which could alert them about eventual grading bias (with an average incidence of 37%). The third classification is based on whether there is or not a teacher's assessment in the school (44% of teachers inform there is). We expect that assessments lead to less biased grading practices.

The overall findings about school policies does not satisfy our expectations because, in general, there is no bias in the third year and a bias against boys in the sixth year in all groups. However, the bias gap in the sixth year is less pronounced when teachers have a permanent assignment to the school than when do not have. Bsides, in the third year, there is a bais agains girls when teachser are regularly assessed at school whereas there is no bias when they are.

4.5. Statistical discrimination and gender stereotypes

An explanation of grading bias explored in the literature is the statistical discrimination hypothesis based on gender differences in abilities. If teachers expect girls to overperform boys, they rationally opt for a gender-different behavior in their grading decisions. There are reasons for teachers to rely on expectations, such as not putting enough effort into assessing quality or having incomplete information. Another quoted channel is the lack of information coming from the limited confidence in testing instruments and their capacity to capture skills, mainly if there are gender differences in cheating (Lavy, 2008; Hanna and Linden, 2012).

To test the appropriateness of this hypothesis, we follow the proposal by Lavy (2008), followed by other studies on grading bias, such as Gortazar, Martinez de la Fuente and Vega-Bayo (2022). The basic idea is that the school's average relative performance of boys and girls in blind exams accurately measures the teacher's expected relative cognitive skills. Thus, within schools, the gender with the best blind test grade will have the best non-blind course grade. We classify the students into three groups (students in classes where girls outperform boys in all tests, in classes where boys overperform girls, and the rest of the students) and estimate equation (3). We report the results in Table 10.

In the third and sixth years, the bias is non-significant for students in classes with no average overperformance of one gender in the blind tests. Besides, in none of the years does the bias sign depend on whether girls are better than boys or vice versa: it is positive in the third year and negative in the sixth year. The only result consistent with statistical

discrimination is that in the third year: favoritism for boys is stronger in classes where they show more skills than girls. Balancing all the aspects of this picture, we interpret that statistical discrimination is not the primary explanation for the observed gender disparities between teachers' grading and blind tests in sixth grade.

Table 10. Estimated gender gap in grading by subgroups based on relative gender performance in TERCE tests (standard errors in parenthesis)

Variable	Third year	Sixth year
Class type: average performance in all TERCE tests is higher for girls than boys	0.197*	-0.238***
	(0.103)	(0.0501)
Class type: average performance in all TERCE tests is higher for boys than girls	0.417***	-0.524***
	(0.121)	(0.0767)
Class type: the rest of the classes	-0.0127	-0.106
	(0.0963)	(0.0796)

*** p<0.01, ** p<0.05, * p<0.1 testing the null hypothesis that the gender gap is null

Source: own elaboration based on TERCE database

We turn to explore the role of gender stereotypes. Previous studies point out that teachers with gender-unequal expectations about talent are more likely to display grading biases based on gender. So, we use information about teachers' beliefs about the abilities of boys and girls reported in the teacher's survey. Specifically, the questionnaire asks teachers who has the greatest ease in learning the language. The teacher must select one of three answers: girls, boys, or both have the same ease. Analogous questions inquire about mathematics and, in the case of sixth-year teachers, sciences. Thus, we have opinions signaling teachers' stereotypes based on explicit attitudes.

In our sample, 86% of teachers support the idea that boys and girls have the same ease in language, 90% in maths and 97% in sciences. When reporting a gender difference, almost everyone believes girls are better at language and boys at mathematics. A disadvantage of self-reported attitudes collection is that people are, in general, reluctant to endorse gender stereotypes as the result of a social desirability bias (Carlana, 2019). Thus, the actual incidence of beliefs about gender differences may be higher than the observed. We may speculate that we capture the most radical believers in talent gender stereotypes.

Table 11 reports the estimated gap in grading by teachers' beliefs based on equation (2). In the third year, teachers who believe that talent differs between genders seem to favor boys, but the precision is weak except in the specific case of endorsing that boys have greater

ease in mathematics than girls. In the sixth year, the estimates indicate that teachers of all groups, on average, favor girls. As in the third year, beliefs on unequal-gender talent deepen the grading bias, though the precision is weak.

Table 11. Estimated gender gap in grading by subgroups based on beliefs about talent and gender (standard errors in parenthesis)

	Third year		Sixth year	
Girls are more talented in language than boys				
Yes	0.241	(0.355)	-0.325***	(0.0973)
No	0.0611	(0.0774)	-0.162***	(0.0617)
Boys are more talented in maths than girls				
Yes	0.855	(0.277)***	-0.266*	(0.137)
No	0.0526	(0.0764)	-0.176***	(0.0609)
Gender-unequal talent in language or maths				
Yes	0.234	(0.347)	-0.262***	(0.0842)
No	0.0612	(0.0775)	-0.166**	(0.0663)

*** p<0.01, ** p<0.05, * p<0.1 testing the null hypothesis that the gender gap is null

Source: own elaboration based on TERCE database

The grading system in Uruguay does not allow studying the relationship between beliefs and grading bias by subject. This is a disadvantage because beliefs of differentiated talent related to a subject may only affect this particular subject and have a weak impact on the overall assessment made by the teacher. So, it is particularly interesting to find that gender stereotypes (girls better than boys in language and boys better than girls in maths) seem to deepen already existing grading biases whose origin would stem from other reasons.

5. Final comments

This work studies whether gender bias in grading exists among students in Uruguay's third and sixth grades of primary schools. The empirical strategy relies on estimating the effect of gender on teachers' academic grading, which is a non-blind score when controlling for blind test scores (TERCE) and other characteristics of the child, the family background, the school and teacher's fixed effects.

The results show that, on average, there is not a gender bias in grading in the third year, but there is a bias against boys (or in favor of girls) in the sixth year. This result holds after several robustness checks. The bias against boys in the sixth year is concentrated among children with regular (and not extreme) abilities, the poorest and the richest.

The existence of a bias in the sixth year is especially relevant since this is the time prior to entering high school. Thus, this bias may affect boys' enthusiasm or confidence in their abilities at the beginning of a new education cycle. In addition, it could affect parents' decisions regarding the institution where their children will attend high school.

Note that we obtain this result when controlling students' behavior. This control is important because, as in studies for other countries (Cornwell, Mustard and Van Paris, 2013; Contreras, 2023; Ferman and Fontes, 2022), gender differences in students' behavior affect grading. In our study, girls perform better in past outcomes (preschool attendance and grade retention) and behavioral traits (current regular attendance and at-home study), generating higher scores for females than males.

The main limitation of our empirical procedure comes from the educational grading policy: teachers set a global academic performance score, which makes it impossible to compare scores by discipline. Besides, the TERCE tests do not include all the disciplines taught at school, so gender differences in the performance of these disciplines could potentially explain that gender gaps in teachers' grades persist after controlling by blind scores. Another limitation, common to other studies on grade bias, is that eventually, there is a systematic difference in behavior between boys and girls in blind tests and the classroom. The hypothesis of this type of gender difference, built on some pieces of evidence, could explain gender differences in classroom and blind test performance.

We performed some procedures to assess the influence of these limitations on our results. Our overall findings suggest that the gender gap in the sixth year is due to teachers' behavior and not to different contents tested by teachers and TERCE or to gender behavioral differences in class and blind tests. We support this idea mainly through the results of two procedures. First, we discard that the result comes from unobservable variables following the proposal of AETO tests. Second, we find heterogeneity among teachers, and no explanations justify attributing the relationship between teachers' characteristics and bias to a gender behavior difference among students.

To analyze the channels of teachers' behavior, we performed a usual procedure to detect whether the result responds to the hypothesis of statistical discrimination, whose main explanation is that teachers exercise a bias because gender provides information about proficiency that the usual evaluations do not detect. Our results do not support this hypothesis. We also explored the role of teachers' reported beliefs about gender differences in talents. The results suggest that gender stereotypes, informed as explicit attitudes in

talent, may underlie the gender bias magnitude but not explain all the bias. Unfortunately, we do not have information to assess the extent to which there may be other gender stereotyping beliefs behind this.

Our evidence does not allow us to explain the finding of gender bias in the sixth and not in the third year. However, we can speculate about the reasons for behavioral differences between teachers of different years. As mentioned, sixth grade is the last year of elementary school. So, teachers may feel a greater need to send signals in sixth grade than in other years, more or less permeated by prejudices or feelings. For example, they may wish to favor girls on the grounds that they need positive signals to increase self-confidence to succeed in secondary school. Or they might penalize boys, understanding that this is a way to get them to make a greater effort in their future educational steps.

Acknowledgments

The authors would like to thank the National Administration of Primary Education (Administración Nacional de Educación Pública – ANEP) for providing information on students attending public schools. We would also like to thank all the comments received at the seminar in the Department of Economics. All remaining errors are ours.

6. References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113(1), 151–184.

ANEP (2023). Observatorio: https://observatorio.anep.edu.uy/documentos/tendencias-educativas

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

Bernatzky, M., & Cid, A. (2015). Brecha de género en la educación secundaria: singularidades de la mujer y el varón en las estrategias educativas. *Páginas de Educación* 8(1), 99-122.

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

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

Bucheli, M., & Casacuberta, C. (2000). Asistencia escolar y participación en el mercado de trabajo de los adolescentes en Uruguay. *El Trimestre Económico*, 67(267), 395-420

Bucheli, M., & Contreras, C. (2018). *Discriminación de género en las calificaciones de las escuelas públicas uruguayas* (No. 2018008). Banco Central del Uruguay.

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

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of economic behavior & organization*, *83*(1), 50-58.

Contreras, D. (2023). Gender differences in grading: teacher bias or student behaviour?. *Education Economics*, 1-24.

Cornwell, C., Mustard, D. B., & Van Parys, J. (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.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, *47*(2), 448-474.

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

Doornkamp, L., Van der Pol, L. D., Groeneveld, S., Mesman, J., Endendijk, J. J., & Groeneveld, M. G. (2022). Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs. *Teaching and Teacher Education*, *118*, 103826.

Failache, E., Salas, G., & Vigorito, A. (2018). Desarrollo en la infancia y trayectorias educativas de los adolescentes. Un estudio con base en datos de panel para Uruguay. *El trimestre económico*, *85*(337), 81-113.

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

Ferman, B., & Fontes, L. F. (2022). Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. *Journal of Public Economics*, *216*, 104773.

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, *94*(2), 377-381.

Gortazar, L., de Lafuente, D. M., & Vega-Bayo, A. (2022). Comparing teacher and external assessments: Are boys, immigrants, and poorer students undergraded?. *Teaching and Teacher Education*, *115*, 103725.

Graetz, G., & Karimi, A. (2022). Gender gap variation across assessment types: Explanations and implications. *Economics of Education Review*, *91*, 102313.

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, *4*(4), 146-168.

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

INEEd (2022). "Informe sobre el estado de la educación en Uruguay 2019-2020. Tomo 1 (edición revisada)": https://www.ineed.edu.uy/images/ieeuy/2019-2020/Informe-estado-educacion-Uruguay-2019-2020-Tomo1.pdf

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

Lindahl, E. (2007). *Comparing teachers' assessments and national test results-evidence from Sweden* (No. 2007: 24). IFAU-Institute for Evaluation of Labour Market and Education Policy.

Lindahl, E. (2016). "Are teacher assessments biased?–evidence from Sweden." *Education economics* 24(2), 224-238.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*, 37(2), 187–204.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much?. *The quarterly journal of economics*, *122*(3), 1067-1101.

Protivínský, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, *59*, 141-149.

Sent, E. M., & van Staveren, I. (2019). A feminist review of behavioral economic research on gender differences. *Feminist Economics*, *25*(2), 1-35.

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

UNESCO-OREALC. (2016). Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE. Santiago, Chile: https://unesdoc.unesco.org/ark:/48223/pf0000247123

Zanga, G., & De Gioannis, E. (2023). Discrimination in grading: A scoping review of studies on teachers' discrimination in school. *Studies in Educational Evaluation*, *78*, 101284.

Annex

Table A1. Estimated coefficients of TERCE variables

	(1)	(2)	(3)	(4)	(5)
Third-year					
Mathematics	0.294***	0.237***	0.219***	0.244***	0.292***
	(0.050)	(0.050)	(0.050)	(0.045)	(0.039)
Reading	0.236***	0.203***	0.188***	0.196***	0.215***
	(0.041)	(0.033)	(0.035)	(0.030)	(0.037)
Writing	0.205***	0.211***	0.214***	0.227***	0.283***
	(0.062)	(0.050)	(0.045)	(0.039)	(0.029)
Sixth-year					
Mathematics	0.315***	0.250***	0.245***	0.260***	0.284***
	(0.047)	(0.039)	(0.037)	(0.037)	(0.032)
Reading	0.148***	0.129***	0.142***	0.156***	0.160***
	(0.045)	(0.042)	(0.037)	(0.035)	(0.034)
Writing	0.127***	0.101**	0.084**	0.087**	0.119***
	(0.046)	(0.043)	(0.041)	(0.042)	(0.039)
Sciences	0.206***	0.172***	0.162***	0.149***	0.143***
	(0.060)	(0.053)	(0.052)	(0.049)	(0.039)
Controls:					
TERCE scores	X	X	X	X	X
Child's characteristics		X	X	X	X
Family background			X	X	X
School's characteristics				X	
Teacher's fixed effect					X
R-squared: Third-year	0.367	0.414	0.447	0.494	0.647
R-squared: Sixth-year	0.409	0.496	0.513	0.536	0.663

*** p<0.01, ** p<0.05, * p<0.1 when testing the null hypothesis that the difference between the third and sixth year is equal to zero

Source: own elaboration based on TERCE database

Table A2. Teachers' characteristics by scholastic year

Variables	All	Third-year	Sixth-year
Mean age	41.1	40.1	42.1
Median age	42	40	43
Mean of years of experience	15.7	14.3	17.0**
Median of years of experience	15	13	16
Proportions:			
Females	0.892	0.941	0.844**
Permanent assignment	0.651	0.639	0.664
Official documents available	0.365	0.294	0.434**
Teacher's performance assessment	0.443	0.447	0.440
Agree with closed questions and multiple choice tests	0.550	0.561	0.539
Agree that all students have to do the same test	0.278	0.256	0.301
Prefer equal tests to different tests	0.108	0.084	0.131
Observations	241	119	122

*** p<0.01, ** p<0.05, * p<0.1 when testing the null hypothesis that the difference between the third and sixth year is equal to zero

Source: own elaboration based on TERCE database