Economics of Education Review 77 (2020) 101981

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Boys lag behind: How teachers' gender biases affect student achievement☆

Camille Terrier*

University of Lausanne., Switzerland

ARTICLE INFO

JEL classification:
I21
I24
J16
Teachers
Gender biases
Progress
Achievement inequalities

ABSTRACT

I use a combination of blind and non-blind test scores to show that middle school teachers favor girls in their evaluations. This favoritism, estimated as individual teacher effects, has long-term consequences: as measured by their national evaluations three years later, male students make less progress than their female counterparts. On the other hand, girls who benefit from gender bias in math are more likely to select a science track in high school. Without teachers' bias in favor of girls, the gender gap in choosing a science track would be 12.5% larger in favor of boys.

Boys are increasingly falling behind girls at school.[1] This disadvantage has important consequences: boys who fall behind are at risk of dropping out of school, not attending college or university, and/or being unemployed. In OECD countries, 66% of women entered a university program in 2009, versus 52% of men, and this gap is increasing (OECD, 2012). In Europe, 43% of women aged 30–34 completed tertiary education in 2015, compared to 34% of men in the same age range. Because this gap has increased by 4.4 percentage points in the last ten years, there is a growing interest in identifying its roots.[2] Some recent studies have highlighted the role of school-related inputs, such as school quality (Autor, Figlio, Karbownik, Roth, & Wasserman, 2016), peer socio-economic status (Legewie & DiPrete, 2012), teacher gender (Dee, 2005), and teaching focus on literacy or numeracy (Machin & McNally, 2005).

This article complements this literature by demonstrating how teachers' gender biases affect their pupils' progress and schooling decisions. A number of papers have shown that stereotyping can bias teachers' assessments and grades, but the impact of such biases has rarely been studied.[3] Prior research on this topic is limited, and it has focused on specific mechanisms through which gender bias could affect progress. Research shows that teachers' biases generate self-fulfilling prophecies (Jussim & Eccles, 1992), produce stereotype threats[4] (Hoff & Pandey, 2006; Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995), affect students' interest in a subject (Bonesrønning, 2008; Marsh & Craven, 1997; Trautwein, Ludtke, Marsh, Koller, & Baumert, 2006), and affect students' effort.[5] This paper provides empirical evidence on how teachers' gender biases affect pupils' progress and schooling decisions.

☆ This paper benefited from discussions with and helpful comments from Joshua Angrist, David Autor, Esteban Aucejo, Elizabeth Beasley, Thomas Breda, Ricardo Estrada, Marc Gurgand, Victor Lavy, Eric Maurin, Stephen Machin, Sandra McNally, Steve Pischke, Jonah Rockoff and anonymous referees and participants at various seminars and conferences. I am especially grateful to Francesco Avvisati, Marc Gurgand, Nina Guyon, and Eric Maurin for sharing their dataset, as well as to the Direction de l'Evaluation, de la Prospective et de la Performance (DEPP) of the French Ministry of Education for giving me access to complementary data used in this paper. A previous version of this paper circulated as a CEP Discussion Paper No 1341 (March 2015).

* Corresponding author.
E-mail addresses: Cterrie1@unil.ch, c.terrier@lse.ac.uk.

[1] In OECD countries, "15-year-old boys are more likely than girls, on average, to fail to attain a baseline level of proficiency in reading, mathematics and science" (OECD, 2015).

[2] In France, 49.6% of women aged 30–34 had completed tertiary education in 2015, compared to only 40.3% of their male counterparts.

[3] See for instance Bar and Zussman (2012), Burgess and Greaves (2013), Hanna and Linden (2012) on teachers' gender bias, and Tiedemann (2000) and Fennema, Peterson, Carpenter, and Lubinski (1990) for the existence of a gender bias in mathematics. Several papers have exploited blind and non-blind scores (teachers' grades) to test for such biases in teachers' grades, a methodology introduced in a seminal paper by Lavy (2008). Some papers find that girls benefit from grade discrimination (Lindahl (2007), Lavy (2008), Robinson and Lubienski (2011), Falch and Naper (2013), Cornwell, Mustard, and Parys (2013)), while others find no gender bias (Hinnerich, Höglin, & Johannesson, 2011). Ouazad and Page (2013) and Dee (2007) observed that gender biases depend on the teacher's gender. Breda and Ly (2015) found that discrimination depends on the degree to which the subject is "male-connoted".

[4] "Stereotype threats" arise when girls or minority groups perform poorly because they fear confirming the stereotype that their group performs poorly.

[5] Mechtenberg (2009) models school results as a combination of talent and effort, and posits that gender biases may affect effort.

https://doi.org/10.1016/j.econedurev.2020.101981
Received 22 December 2018; Received in revised form 1 March 2020; Accepted 12 March 2020
0272-7757/ © 2020 Elsevier Ltd. All rights reserved.

I use a rich student-level dataset produced by Avvisati, Gurgand, Guyon, and Maurin (2014) that follows 4490 pupils from grade 6 until grade 11. To quantify teachers' gender biases in math and French, I exploit an essential feature of the data: it contains both blind and non-blind scores. An external grader without knowledge of student's characteristics provides schools with blind scores. These scores are presumably free of teachers' biases. Teachers provide non-blind scores for in-class exams. Both scores are designed to measure the same skills—an assumption that I discuss and test in Section 4.5. In addition, the dataset contains blind scores up to grade 9, the type of high school each student attends in grade 10 (general, professional, or technical), and students' course choices in grade 11 (scientific, literature, or social sciences). This data allows me to study the effect of teachers' gender biases on pupils' progress, schools attended, and course choices.

To identify the impact of teachers' gender biases on pupils' progress, I use a novel identification strategy that relies on both the variation in teachers' gender biases and the quasi-random assignment of students to biased teachers. The identification therefore stems from a comparison of the progress of girls (relative to boys) in classes with more biased teachers to the progress of girls in classes with less biased teachers. To measure teachers' gender biases, I follow in the footsteps of many previous studies and use a double-difference (DiD) methodology (Blank, 1991; Breda & Ly, 2015; Falch & Naper, 2013; Goldin & Rouse, 2000; Lavy, 2008). Gender bias is defined as the average gap between non-blind and blind scores for girls, minus this same gap for boys.

My identification strategy requires that boys and girls are not differentially assigned to teachers with different degrees of bias. To test this assumption, I follow Pei, Pischke, and Schwandt (2019) by conducting both a right-hand-side (RHS) and a left-hand side balancing test. The two tests confirm that girls who scored higher on the baseline blind score are not more likely to be assigned a biased teacher than boys. Selection is also similar for boys and girls across age, social background (high and low), grade repetion, and number of boys and girls in the class.

The main finding is that teachers' gender biases have a large and significant effect on boys' progress relative to girls in both math and French.[6] For two classes where the achievement gap between boys and girls would be identical in 6th grade, quasi-randomly assigning a teacher who is 1 SD more biased against boys to one of the classes decreases boys' progress in that class relative to girls by 0.123 SD in math and by 0.106 SD in French. Over the four years of middle school, teachers' gender bias against boys accounts for 6% of boys falling behind girls in math. Analyzing the effect separately for boys and girls, I find that having a math teacher who is 1 SD more biased against boys does not impact boys' progress, but significantly increases girls' progress. Conversely, a biased French teacher does not impact girls' progress, but significantly reduces boys' progress.

Moving to outcomes in high school (four years after students may have been exposed to a biased teacher), I find that having a math teacher who is 1 SD more biased in favor of girls increases girls' probability of selecting a scientific track in high school by 3.6 percentage points compared to boys. Interestingly, without teachers' bias in favor of girls, the gender gap in choosing a science track—a predictor of careers in STEM fields—would be 12.5% larger in favor of boys. On the other hand, teachers' gender biases do not impact boys' relative probability of attending a general high school (rather than a professional or technical one) or of repeating a grade. I also rule out some potential mechanisms. Teachers' biases do not have a cumulative effect: being reassigned to the same biased teacher for a second year does not further impede boys' relative progress. Similarly, teachers' gender biases have no spillover effect: a bias in one subject does not impact boys' relative progress in the other subject.

My results confirm the findings of two contemporaneous studies. Lavy and Sand (2018) analyze the effect of teachers' gender biases on boys' and girls' respective progress. Using the same identification strategy, they find that teachers' biases in favor of one gender lead to a larger progress for that gender several years later, and that teachers' biases in math encourage the favored gender to enroll in science or math courses. Carlana (2018) uses the Gender-Science Implicit Association Test to measure gender stereotypes of teachers. She finds that teachers with stronger gender stereotypes have a large negative effect on girls' progress and induce them to self-select into less demanding tracks. It is particularly interesting that we find very similar results in three different institutional contexts: primary schools in Israel in Lavy and Sand (2018), middle schools in Italy in Carlana (2018) and middle schools in France in this paper.[7]

Taken together, these results build upon an important literature that suggests teachers' grades are biased. My findings confirm the existence of such biases, but more importantly, they highlight the fact that teachers' gender biases can have long-lasting effects on boys and girls' human capital accumulation, and therefore on the evolution of gender inequalities at school and in the labor market. These findings could open the door to new policies. If policy-makers want to reduce achievement gaps—whether between boys and girls or students from different ethnicities or social backgrounds—teachers' evaluation methods and behavior could be an instrument to achieve that goal.

1. Data

1.1. Datasets

I use a French dataset that covers 35 middle schools, 191 classes, and 4490 pupils. Fig. A.1 presents a timeline of the data. All students are first observed during grade 6 (11 years old), the first year of middle school. Blind and non-blind test scores are available for each student. Students obtain the blind score when they complete a standardized test at the beginning and end of grade 6. The French Education Ministry created this test, taken annually by all French pupils, to assess students' cognitive skills. Identical across all schools, it tests knowledge on French (reading and writing) and mathematics. Importantly, this test is externally graded, and graders do not know the names, genders, social backgrounds, or behavior of the pupils they evaluate. We can therefore assume that these scores are free of any bias caused by stereotypes from an external examiner. Each student also receives grades from teachers on in-class exams. A pupil has a different teacher in each subject, and each teacher reports their pupils' average grades on end-of-term report cards. In this study, I use information on the average grade given by teachers in math and French during the first and last terms of grade 6. Because teachers have permanent contact with the pupils they teach, these average grades could be biased by teachers' stereotypes.

The standardized test and class exams are designed to measure the same abilities. Appendix A.1 describes these tests and their content. Both tests are taken under the same conditions: pupils fill in both tests in their usual classrooms, and their teachers give instructions. Blind and non-blind tests include questions with different degrees of difficulty. The national

[6] My analysis identifies three effects that I cannot completely disentangle: (1) teachers' gender bias in grades, (2) teachers' potentially biased evaluation methods (for instance, some teachers might use more homework as an evaluation tool, and boys and girls might perform differently on homework), and (3) teachers' behavior in class, which might favor girls or boys. I try to disentangle this last effect by measuring students' progress over a period when they do not interact with a biased teacher.

[7] This article also contributes to the recent and growing literature on the impact of teachers' discretion in grading on students' success (Apperson, Bueno, & Sass, 2016; Dee, Dobbie, Jacob, & Rockoff, 2016; Diamond & Persson, 2016). Papageorge, Gershenson, and Kyungmin (2016) also develop a structural econometric model of biases in teacher expectation and estimate the impacts of those biases on students' attainment.

evaluation relies heavily on written questions: in French, only 18% of questions are multiple choice, with the remaining 82% requiring written answers. The percentage is even higher in math, where 95% of the questions require written answers. The reliance on written questions makes the national evaluation format similar to in-class exams, where multiple-choice questions are quite rare.[8] This similarity is partly due to grade 6 teachers: 49% of French teachers and 47% of math teachers report using the standardized evaluation provided by the ministry as a benchmark to create their own class exams (French Ministry of Education, 2005). However, despite featuring similar types of questions, the formats of the two tests might differ. The standardized test consists of two sessions of 45 minutes over two days, while teachers' assessments rely primarily on in-class exams and possibly some home work. The stakes also differ between tests. The standardized tests are not high-stakes for the students.[9] They are an administrative tool aimed at reporting average achievement by school. Unlike in-class exams, a pupil's result on the standardized test does not factor into his/her end-of-term average score or have a bearing on the grade repeat decision at the end of the year. This dataset also contains a rich set of measures of grade 6 pupils' disruptive behavior. Records include official "disciplinary warnings", definitive exclusions from school, temporary exclusions from school or class, and detentions. Temporary exclusions signal violent behavior or repeated transgressions of the rules and are decided by the school head.

Blind scores and schooling decisions are available several years after grade 6, which enables me to estimate the effect of gender bias on pupils' progress, school choices, and course choices. Pupils receive blind scores at the beginning of grade 6, at the end of grade 6, and at end of grade 9. The test completed at the end of grade 6 is extremely similar to the one pupils take when they enter grade 6. Both the beginning and end-of-year exams test similar knowledge and are created by the French Education Ministry, are identical across schools, and are graded externally. At the end of grade 9 (which is the end of middle school), all pupils take a national exam to obtain the Diplome national du brevet. This externally graded score constitutes the final blind measure of pupils' ability in middle school.[10] The dataset includes information about pupils' choice of high school and course. After students complete middle (and compulsory) school in grade 9, they must choose between general, vocational, or technical training. Pupils who decide to follow general training have to specialize when they enter grade 11 by choosing sciences, humanities, or economics and social sciences. I use this information to estimate the effect of teachers' gender biases on four outcomes: pupils' probability of undergoing general training, likelihood of choosing a scientific track, likelihood of choosing a literature track, and likelihood of repeating a grade. I conduct a detailed analysis of attrition in Section Appendix E.

Finally, the dataset contains information on teachers' genders, birth dates, and years of experience, as well as administrative information on children: gender, parents' professions, grade retention, and birth date. The schools included in this dataset are mostly located in deprived areas. Therefore, they do not perfectly represent all French pupils, an issue I discuss in a Appendix E.

1.2. Descriptive statistics

The first column of Table 1 presents descriptive statistics for all students; subsequent columns compare the characteristics of boys and girls. 48.1% of the pupils are girls, and 68.6% of them have low SES parents, which is consistent with most schools located in the deprived administrative area of Creteil. In grade 6, 50% of math teachers and 85% of French teachers are female. 45% of students in the dataset attended a general high school in grade 10, but this percentage is higher for girls (50.9%) than for boys (40.3%). Around 16% of the sample attended the scientific track of a general high school in grade 11. In the analysis, all test scores are standardized—the mean is zero and the variance is one. Scores are standardized within evaluation (blind or non-blind), subject, and term.

Table 1
Descriptive statistics for boys and girls.

	All Mean (1)	Boys Mean (2)	Girls Mean (3)	Difference (4) = (3)-(2)	*p*-value
Pupils' test scores in grade 6					
Blind - French	0.000	−0.211	0.223	0.434***	(0.000)
Blind - Math	0.000	0.072	−0.075	−0.147***	(0.000)
Non-Blind - French	0.000	−0.224	0.236	0.460***	(0.000)
Non-Blind - Math	0.000	−0.083	0.087	0.170***	(0.000)
Pupils' characteristics in grade 6					
% Grade repetition	0.062	0.074	0.049	−0.026***	(0.000)
% Disciplinary warning	0.062	0.097	0.025	−0.072***	(0.000)
% Excluded from class	0.056	0.086	0.023	−0.064***	(0.000)
% Temporary exclusion from school	0.036	0.062	0.008	−0.054***	(0.000)
Parents' characteristics in grade 6					
% High SES	0.178	0.185	0.170	−0.015***	(0.000)
% Low SES	0.686	0.672	0.701	0.028***	(0.000)
% Unemployed	0.117	0.120	0.114	−0.006***	(0.000)
Teachers' characteristics in grade 6					
% Female teachers - Math	0.499	0.504	0.494	−0.011***	(0.000)
% Female teachers - French	0.846	0.846	0.845	−0.001***	(0.000)
Teachers' age - Math	34.378	34.354	34.403	0.049	(0.599)
Teachers' age - French	37.942	37.894	37.993	0.098	(0.423)
Schools and courses attended after grade 10					
% General high school (grade 10)	0.457	0.403	0.509	0.106***	(0.000)
% Scientific track (grade 11)	0.165	0.162	0.167	0.005***	(0.000)
% Literature track (grade 11)	0.063	0.030	0.095	0.065***	(0.000)
Number of observations	4490	2332	2158		

†Notes: This table presents differences between boys' and girls' characteristics. Column 4 reports the coefficients of the regression of various dependent variables on a dummy indicating that the pupil is a girl. All scores are standardized. For standard errors, we use the White estimator of variance.
Parents' professions: Parents belong to the *high SES* category if they belong to the French administrative category "corporate manager" or "executive." Parents are classified as *low SES* if they belong to the categories "worker" or "white-collar worker." For both variables, the dummy takes the value 1 if at least one of the parents belongs to the category.
Stars correspond to the following *p*-values: $*p<.05$; $**p<.01$; $***p<.001$.

Figs. 1 and 2 display the distributions of blind and non-blind French scores at the beginning of grade 6. Girls strongly outperform boys in French, and this premium is not affected by the nature of the evaluation (blind or non-blind). As reported in Table 1, girls' average score is 0.434 points higher than boys when the score is blind and 0.460 when it is non-blind. However, the story is different in mathematics. Figs. 3 and 4 show that boys outperform girls when grades are blind, but the opposite is true when teachers assess their own pupils: girls' average score at the beginning of grade 6 is 0.147 points lower than boys when the score is blind, but it is 0.170 points higher when the score is non-blind. Graphically, girls' score distribution clearly shifts to the right of boys' distribution when comparing blind and non-blind math scores. These distributions reflect the difference-in-difference (DiD) methodology that is widely used to measure gender bias in teachers' grades: boys and girls might perform differently, but if the achievement gap is systematically stronger in favor of girls when the grades are non-blind, this higher achievement gap indicates a gender bias in teachers' grades in favor of girls (or equivalently, a bias against boys).

[8] Machin and McNally (2005) suggests that the mode of assessment could affect the gender achievement gap.

[9] For teachers, their evaluations or salaries do not depend on their pupils' results on standardized tests, so they have no incentive to "teach to the test".

[10] Unlike the grade 6 blind scores, the grade 9 blind scores are high-stakes for the pupils.

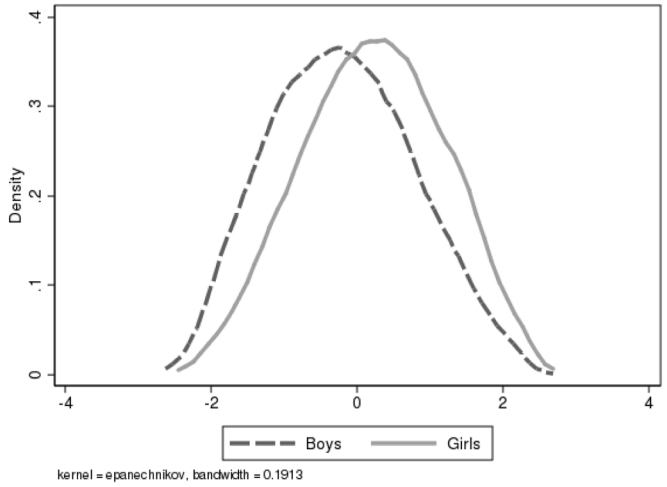


Fig. 1. Distribution of Blind scores (Grade 6) French. Test scores are standardizedthe mean equals zero and the variance equals one. Scores are standardized within test (blind or non-blind) and subject.

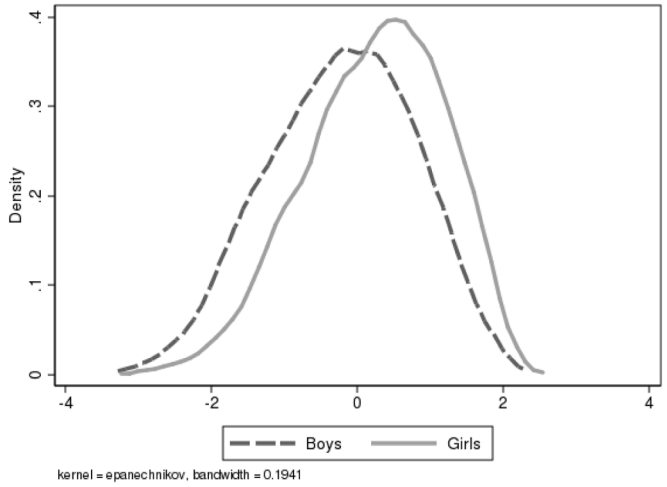


Fig. 2. Distribution of Non-blind scores (Grade 6) French. Test scores are standardizedthe mean equals zero and the variance equals one. Scores are standardized within test (blind or non-blind) and subject.

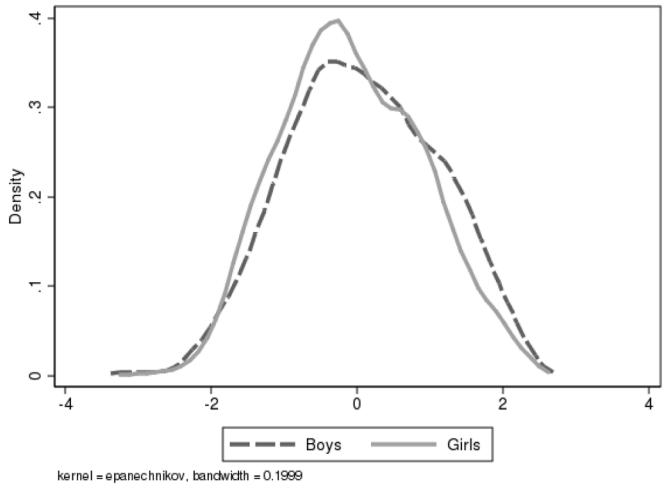


Fig. 3. Distribution of Blind scores (Grade 6) Math. Test scores are standardizedthe mean equals zero and the variance equals one. Scores are standardized within test (blind or non-blind) and subject.

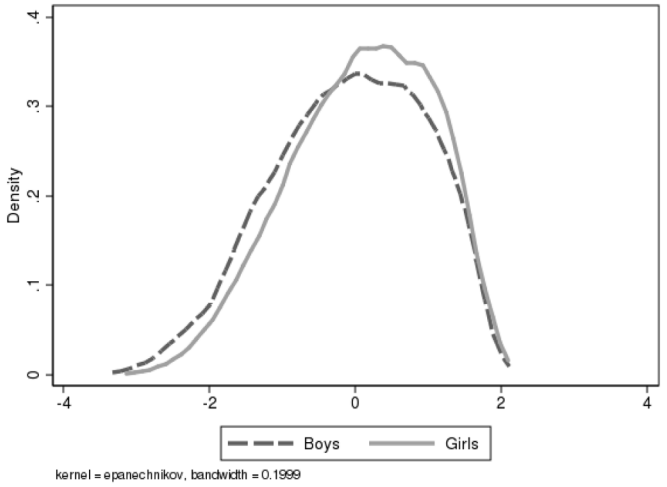


Fig. 4. Distribution of Non-blind scores (Grade 6) Math. Test scores are standardizedthe mean equals zero and the variance equals one. Scores are standardized within test (blind or non-blind) and subject.

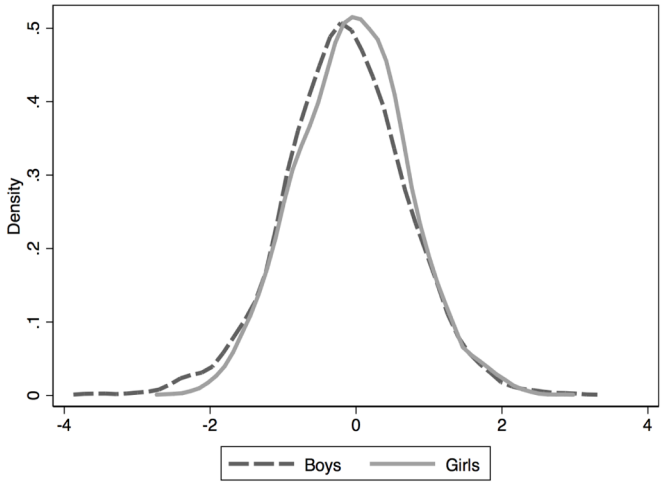


Fig. 5. Boys and girls' progress over middle school. French. Notes: Figs. 5 and 6 plot the distribution of boys' and girls' progress over middle school, that is, between the beginning of grade 6 and the end of grade 9. The solid lines represent girls and the dotted lines represent boys. I define progress as the difference between the blind score at the end of grade 9 and the blind score at the beginning of grade 6. Because both scores are standardized, progress corresponds to a higher ranking over time in the score distribution.

Figs. 5 and 6 plot the distribution of boys' and girls' progress during middle school—between the beginning of grade 6 and the end of grade 9. I define progress as the difference between the blind score at the end of grade 9 and the blind score at the beginning of grade 6. Because both scores are standardized, progress corresponds to a higher ranking over time in the score distribution. There is clear evidence that boys progress less than girls in mathematics, whereas progress in French is similar.[11] Girls start from a lower baseline but catch up with and even overtake boys in math. Throughout middle school, girls maintain their lead in French. I show in this paper that teachers' biased behavior against boys can explain part of this differential progress in math and the observed inequalities in choice of STEM courses in high school.

[11] At the beginning of grade 6, girls' average math score is 0.075 points below the mean. It is only 0.021 points below the mean at the end of the 6th grade, and 0.029 points above the mean by the end of grade 9, hence a total increase of 0.104 points of the SD.

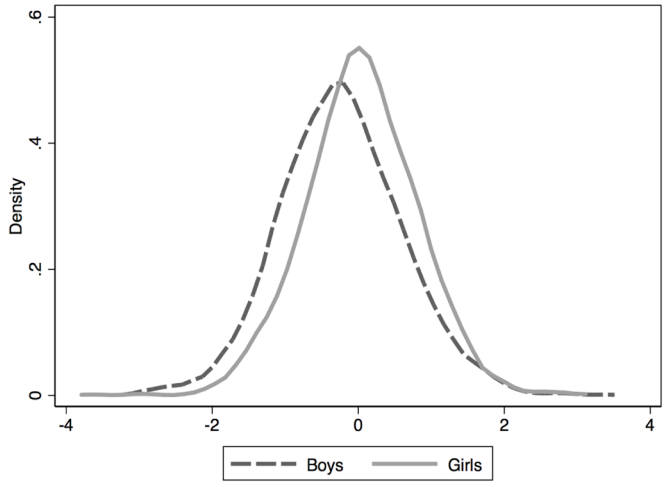


Fig. 6. Boys and girls' progress over middle school. Math.

2. Model of pupil's progress

I define a simple model aimed at isolating the effect of teachers' gender biases on pupils' progress. Eq.(1) describes a blind score B_{1i} given at the beginning of a period. This score is a noisy measure of a student's ability θ_{1i}. The term ϵ_{B1i} captures measurement error. Eq. (2) describes a blind score given at the end of the period. All variables and parameters referring to the end of the period are indexed by (2). A biased grade is the difference between a student's ability θ_i and the non-blind grade NB_i given by the teacher.[12]

$$B_{1i} = \theta_{1i} + \epsilon_{B1i} \tag{1}$$

$$B_{2i} = \theta_{2i} + \epsilon_{B2i} \tag{2}$$

$$Bias_i = NB_i - \theta_i \tag{3}$$

I model the evolution of a pupil's ability between the beginning and end of the period as:

$$\theta_{2i} - \theta_{1i} = \beta Bias_{1i} + \eta G_i + \mu_i T_i + \gamma\theta_{1i} + \omega_i. \tag{4}$$

T_i is a teacher effect that captures a teacher's quality (also referred to as their value added). θ_{1i} is a pupil's ability. Including this term allows students starting from different baselines to progress at different rates. Because low achievers have more room for improvement, they might make more progress than their high-achieving counterparts. G_i is a dummy variable for girls. Girls' unobserved characteristics might be correlated with both teachers' gender biases and girls' progress.[13]

A pupil's progress is measured by the evolution of his/her blind score over time:

$$\begin{aligned} B_{2i} - B_{1i} &= \theta_{2i} + \epsilon_{B2i} - \theta_{1i} - \epsilon_{B1i} \\ &= \beta Bias_{1i} + \eta G_i + \mu_i T_i + \gamma\theta_{1i} + \omega_i + \epsilon_{B2i} - \epsilon_{B1i} \\ &= \beta(NB_{1i} - B_{1i}) + \eta G_i + \mu_i T_i + \gamma B_{1i} + \epsilon_{B2i} + (\beta - 1 - \gamma)\epsilon_{B1i} \\ &\quad + \omega_i \end{aligned} \tag{5}$$

By aggregating this equation at the gender-by-class level, I define a first-difference specification in which the dependent variable becomes the gap between girls' and boys' progress in class c:

$$\begin{aligned} ((B_{2G} - B_{2B}) - (B_{1G} - B_{1B}))_c &= \eta + \beta[(NB_{1G} - B_{1G}) - (NB_{1B} - B_{1B})]_c \\ &\quad + \gamma(B_{1G} - B_{1B})_c + (\omega_G - \omega_B)_c \end{aligned} \tag{6}$$

I use Eq. (6) to estimate the effect of having a gender-biased teacher on girls' relative progress (captured by the coefficient β). Thanks to the first-difference specification, the simple difference $(NB_{1i} - B_{1i})$ becomes a double difference $(NB_{1G} - B_{1G}) - (NB_{1B} - B_{1B})$, a frequent measure of gender biases in teachers' grades (Breda & Ly, 2015; Falch & Naper, 2013; Goldin & Rouse, 2000; Lavy, 2008). The gender bias is the difference between the gender gap in the blind and non-blind test scores. Using this double difference formula to measure each teacher's gender bias is equivalent to estimating α_3 in the following regression (run separately for each teacher):

$$S_{in} = \alpha_0 + \alpha_1 G_i + \alpha_2 NB_{in} + \alpha_3(G_i \cdot NB_{in}) + \alpha_4 X_i + \epsilon_{in}. \tag{7}$$

S_{in} is the grade a pupil receives ($n = 1$ for non-blind and 0 for blind). G_i is a dummy variable for girls. NB_{in} is a dummy variable equal to 1 if the score was given non-anonymously by the pupil's teacher. X_i is a set of potential control variables.[14]

When estimating a gender bias at the teacher level, I account for estimation error arising from sampling variation by constructing empirical Bayes estimates of teacher gender bias, as detailed in Appendix C (Chetty, Friedman, & Rockoff, 2014; Jacob & Lefgren, 2005; Kane & Staiger, 2008). With small samples, a few students can have a large impact on test scores. In my sample, the average number of sents per teacher is 36.3 in math and 31.8 in French.[15] With sampling error, the variance of the estimated teacher biases has two components: the true variance of the teacher bias and the average sampling variance. Without accounting for sampling bias, the estimation error would suffer from attenuation bias when I use the teacher bias measure as an explanatory variable for students' progress in Eq. (6).

The aggregation and first-difference specification allows us to rule out two endogeneity concerns that would have arisen in the individual-level specification. First, because Eq. (6) is specified as a difference between boys' and girls' average scores at the class level, teacher effects disappear so long as they similarly affect boys and girls within a class.[16] The first-difference specification therefore ensures that the effect of the gender bias I estimate is not explained by a correlation between teachers' value added and their potentially biased behavior against boys. This is important as Lavy and Megalokonomou (2019) show that gender biases are more prevalent among low value added teachers than among more effective teachers. Averaging scores at the class level also significantly reduces the measurement error affecting individual-level blind score as well as concerns of reversed causality at the individual level. Of course, measurement error and reversed causality might still exist at the class level. The next section addresses some of these concerns.

3. Identification strategy

For identification, I take advantage of the variation of the gender bias across teachers and the quasi-random assignment of pupils to teachers with different degrees of bias.

[12] At this stage of the model, a biased grade does not refer to a gender bias. It might correspond to a teacher's tough or lenient grading practice that applies to both genders.

[13] The coefficient β can capture several channels through which grade biases can affect a pupil's progress (motivation, discouragement, effort, self-confidence...). I will not be able to distinguish between different channels, which are all captured by the coefficient β.

[14] In practice, I estimate the gender bias by running a regression of the difference between the non-blind and the blind score on a dummy for girls and control variables for pupils' blind score, grade repetition, and social background. That regression gives the same gender bias estimate as Eq. ((7)).

[15] At the school level, Kane and Staiger (2002) found that among the smallest schools, more than half (56%) of the variance in mean gain scores is due to sampling variation and other non-persistent factors.

[16] Teachers are only observed once in grade 6, which rules out estimating teacher fixed effects. Fox (2016) estimates teacher's value-added separately for boys and girls and finds that teachers are approximately equally effective across sex.

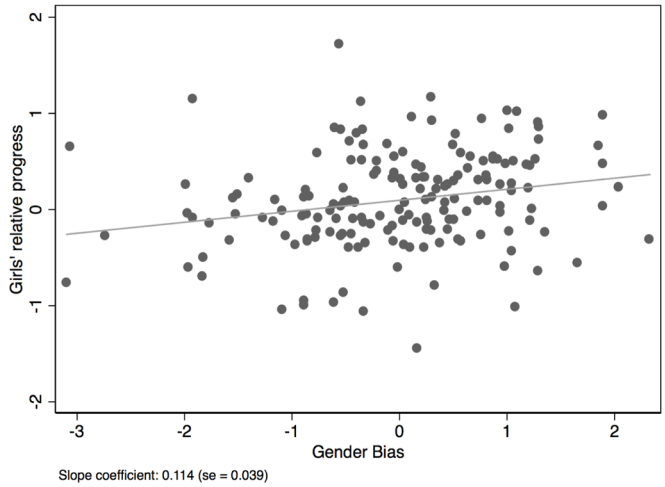


Fig. 7. Correlation between teachers' gender biases and girls' relative progress over middle school. French. Notes: Figs. 7 and 8 display correlation between the gender bias measure (on the horizontal axis) and girls' progress relative to boys during middle school (on the vertical axis). Gender bias is defined as the class average difference between the non-blind and the blind scores for girls, minus this same difference for boys. On the vertical axis, girls' progress relative to boys is measured as the difference between their blind score at the end of grade 9 and this blind score at the beginning of grade 6, minus the same difference for boys.

3.1. Variation in teachers' gender bias

Figs. 7 and 8 display the gender bias on the horizontal axis and girls' progress relative to boys (during middle school) on the vertical axis for each class in the sample. Girls' progress relative to boys is measured as the difference between their blind score at the end of grade 9 and their blind score at the beginning of grade 6, minus the same difference for boys. We see high variation in the degree of teachers' gender biases. Further, the degree of gender bias in favor of girls is positively correlated with girls' progress relative to boys'. The slope coefficient is 0.122 in math (SE = 0.030) and 0.114 in French (SE = 0.039).

I also check if the bias distribution changes over time by loking at blind and non-blind scores at the beginning of the year and at the end. If teachers' biases are mainly driven by statistical discrimination, we might expect end-of-year grades to be less biased (and the variance to be smaller) because teachers acquire information about students during the year. On the other hand, if teachers' biases are mainly taste based, bias should not change over time. In that case, end-of-year in-class grades should produce similar bias variance than first-semester grades. The mean and variance of the bias are very similar at the beginning of the year and at the end, suggesting that gender favoritism is mainly taste based.

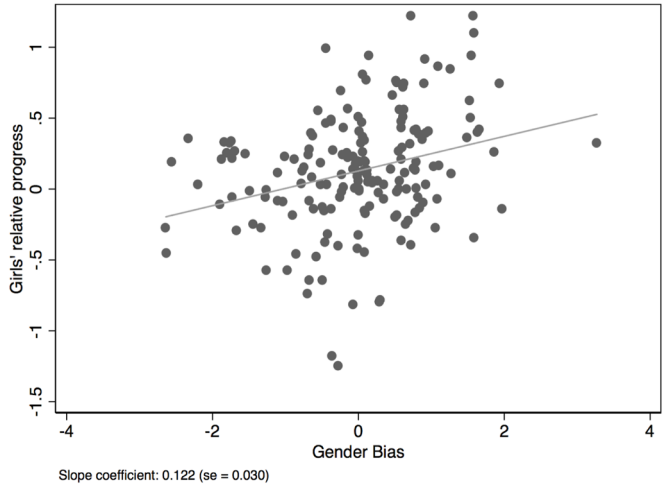


Fig. 8. Correlation between teachers' gender biases and girls' relative progress over middle school. Math.

Finally, note that the variation in gender bias is not driven by the ability of blind score graders to guess pupils' gender based on their handwriting. Some graders might be able to guess a student's gender. If they suffer from the same biases as teachers, the difference between the blind and the non-blind exam would be attenuated. However, this attenuation should be the same in all classes, so that the variation of the gender bias between teachers would not change.

3.2. Quasi-random assignment of students to biased teachers

In Eq. (6), the coefficient β identifies the effect of being assigned a teacher who is 1 SD more biased against boys on boys' average progress relative to girls' after controlling for the initial achievement gap between boys and girls. Due to that first difference specification, the identifying assumption requires no differential selection of boys and girls to biased teachers. The distinction between *selection* and *differential selection* of boys and girls is important. Assignment of students to biased teachers does not need to be quasi-random—for instance disadvantaged students might be more likely to be assigned a biased teacher. However, disadvantaged girls must not be more likely to be assigned a biased teacher than disadvantaged boys.

Institutional features make differential selection of boys and girls to biased teachers unlikely. Pupils considered in this study are in grade 6, which is the first year of middle school in France. All of them were enrolled in a different school the year before. Hence, when deciding the composition of classes, school heads had very little information on these new pupils. It is unlikely that they can predict students' progress, and therefore influence their assigned class and teacher.

I test the non-diffential selection of boys and girls to biased teachers by following Pei et al. (2019). I start by conducting a right-hand-side (RHS) balancing test, whereby I regress the gender bias on the class-level difference between boys' and girls' characteristics. A significant coefficient indicates that boys and girls with a given characteristic (high social background for instance) are not equally likely to be assigned a biased teacher, which would violate the identifying assumption. For each regression, I control for class level differences between girls' and boys' blind scores because that variable is systematically controlled for in the analysis.[17]

The results in columns 1 and 7 of Table 2 show that girls who scored higher on the baseline blind score are not more likely to be assigned a biased teacher than boys who scored similarly. This rules out reversion to the mean, which would make the bias in favour of girls mechanically larger in classes where they perform relatively worse on the blind score.[18] Table 2 also shows no differential selection of boys and girls on age, social background (high and low), grade repetion, or number of

[17] When correlating bias and initial achievement, the OLS estimate of the correlation is biased with finite class size because a high ability student affects the initial test score and the gender bias measure. To eliminate bias due to the own-observation problem, I omit the own blind and non-blind test scores from the measure of teacher bias for student *i*. Hence, I proxy for teacher bias using a leave-out mean peer exposure to bias (Chetty et al., 2011). I use this jackknife estimate for both the right-hand-side and left-hand-side balancing tests.

[18] Note that the coefficients of the gender gap in blind score is never significant but is often large in magnitude. This confirms the importance of controlling for the gender gap in blind score to ensure that my estimates do not capture a mechanical reversion-to-the-mean effect, whereby the gender bias effect would capture both the true effect of teacher gender bias and the fact that the bias in favour of girls is mechanically larger in classes where they perform relatively worse on the blind score. In these classes, girls have a high chance of facing a less negative shock on their end-of-year blind score, which would mechanically increase their progress, for reasons unrelated to the gender bias of their teacher. Controlling for the blind score gender gap at the beginning of the year mitigates this potential bias.

Table 2
Right hand side balancing test.

	Gender bias in Math						Gender bias in French					
RHS variable: $(\bar{X}_B - \bar{X}_G)_c$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Blind score	0.258	0.219	0.203	0.167	0.160	0.141	0.348	0.330	0.375	0.369	0.378	0.389
	(0.178)	(0.202)	(0.190)	(0.183)	(0.178)	(0.178)	(0.198)	(0.194)	(0.205)	(0.209)	(0.210)	(0.222)
Age		−0.231	−0.195	−0.221	−0.129	−0.098		−0.147	−0.220	−0.223	−0.151	−0.150
		(0.303)	(0.290)	(0.286)	(0.311)	(0.280)		(0.397)	(0.375)	(0.372)	(0.391)	(0.389)
High socio-eco status			0.546	0.423	0.382	0.412			−1.112*	−1.150*	−1.178*	−1.181*
			(0.428)	(0.389)	(0.392)	(0.396)			(0.495)	(0.505)	(0.521)	(0.527)
Low socio-eco status				−0.385	−0.352	−0.318				−0.120	−0.099	−0.093
				(0.469)	(0.471)	(0.461)				(0.591)	(0.592)	(0.618)
Grade repetition					−1.062	−1.090					−0.677	−0.662
					(0.810)	(0.823)					(1.021)	(1.043)
Number of girls						−0.033						−0.005
						(0.071)						(0.066)
Number of boys						−0.068						0.011
						(0.073)						(0.050)
Constant	0.043	0.012	0.019	0.015	−0.005	1.206	−0.147	−0.153	−0.190	−0.186	−0.201	−0.276
	(0.029)	(0.057)	(0.054)	(0.052)	(0.053)	(1.642)	(0.082)	(0.089)	(0.099)	(0.103)	(0.111)	(1.291)
R^2	0.412	0.415	0.421	0.425	0.434	0.441	0.321	0.323	0.350	0.350	0.354	0.354
N	177	177	177	177	177	177	172	172	172	172	172	172
$P - val$		0.180	0.190	0.272	0.070	0.111		0.236	0.151	0.231	0.333	0.540

†Notes: This table reports coefficients from regressions of the class-level gender bias on class-level differences between boys' and girls' characteristics. For instance, the second row reports the coefficient of the age difference between boys and girls. Each regression includes school fixed effects and controls for differences between girls' and boys' blind scores (because that variable is systematically controlled for in the regression of girls' relative progress on gender bias). The gender bias corresponds to the difference between the gender gap in the non-blind test score and the gender gap in the blind test score. I estimate it using a leave-out mean peer exposure to bias (Chetty et al., 2011). I use this jackknife estimate for all correlations reported in this table. The last row of the table reports the p-value of the joint test that all right-hand-side variables are jointly equal to zero. Standard errors are clustered at the school-level. Stars correspond to the following p-values: $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

boys and girls in the class. Only high socioeconomic status appears unbalanced, but in French only and the coefficient is only significant at 10%. The fact that the gender bias is uncorrelated with the number of boys and girls in a class indicates that my estimate of gender bias is not driven by sampling error in classes with fewer girls or boys. The last row of the table reports the p-value of the joint test that all variables are jointly insignificant. Both in math and French, the test does not reject the hypothesis that all controls are balanced, with a p-value of 0.54 in French and 0.11 in math for the full specifications.

As a robustness check, I conducted a left-hand-side (LHS) balancing test in which I regress each potential confounding variable (now placed on the left-hand side) on the gender bias (Pei et al., 2019). To match the first-difference specification, I use the class-level difference between boys and girls' characteristics as the left-hand side variable, and I control for the baseline achievement gap. Table 3 reports coefficients for the LHS balancing test and confirms that teachers' gender bias is uncorrelated with class-level differences between boys and girls in all regressions but one. The last row reports the LHS joint balancing test, which is an F-test for the joint significance of the gender bias coefficients in the six regressions. This test accepts the hypothesis that all six variables are balanced with a p-value of 0.231 in French and 0.058 in math. The p-value in math is relatively low, but Pei et al. (2019) note that both the LHS and RHS balancing tests with robust standard errors have a size distortion under the null hypothesis and reject too often. The p-value of 0.058 in math is therefore a conservative lower-bound.

As a third balancing test, I check if the gender bias of math teachers is correlated to the gender bias of French teachers. As class size is relatively small, even a random assignment of students to classes might hide differences in students' unobserved characteristics due to small sample size.[19] However, if students' unobserved characteristics (such as motivation, stress, or other non-cognitive skills) equally affect the bias of math and French teachers who teach the same students, we would expect their bias to be correlated. Yet, I find no correlation (the coefficient is 0.009 with SE = 0.094), which indicates that the gender bias is driven by differences between teachers in their level of bias, rather than by differences in students' characteristics across classes.

3.3. Interpreting the gender bias

The fact that the distribution of students' characteristics across teachers is balanced does not rule out the possibility that two teachers who face the same students might put different weights on their cognitive and non-cognitive skills when they evaluate them. The blind test is a standardized test created by the French Education Ministry. Its content does not vary across teachers. Non-blind evaluations are designed to measure the same competencies, but their format might differ. Some teachers might rely more on homework than others, or load more on students' non-cognitive skills (such as perseverance, conscientiousness, grit, motivation). If boys' and girls' non-cognitive skills differ, or if their diligence for homework differs, the estimated teacher gender bias might capture these differences in teachers' evaluation methods.[20] These potential differences between teachers in the skills they measure affect how we interpret the gender bias, but not its effect on students' progress, so long as there is no differential selection of boys and girls with higher homework diligence, perseverance, or motivation to teachers who load more on homework or non-cognitive skills.

The gender bias might also capture teachers' biased behavior. Teachers who tend to be biased against boys in their evaluations might

[19] The risk of a non-random assignment of students to teachers would be attenuated if we were observing several classes per teacher, as we would be able to use leave-out type estimators. Unfortunately, my data only contains one class per teacher.

[20] Borghans, Golsteyn, Heckman, and Humphries (2016) find that teacher set grades load on non-cognitive skills with more weight than achievement tests. Cornwell et al. (2013), using data from the 1998–1999 ECLS-K cohort of primary school pupils, took into account pupils' non cognitive skills, and found that controlling for how well a pupil is "engaged in the classroom" significantly reduces or completely removes the bias in teachers' grades.

Table 3
Left hand side balancing test.

	Age	High SES	Low SES	Grade repet	Nb boys	Nb girls	Age	High SES	Low SES	Grade repet	Nb boys	Nb girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Bias Math	−0.022	0.021	−0.022	−0.018	−0.191	0.030						
	(0.029)	(0.018)	(0.023)	(0.009)	(0.198)	(0.237)						
Achievement gap	−0.162**	0.035	−0.091*	−0.021	−0.417	0.198						
	(0.050)	(0.036)	(0.038)	(0.019)	(0.274)	(0.315)						
Bias French							−0.012	−0.035*	0.007	−0.008	0.053	−0.065
							(0.032)	(0.014)	(0.024)	(0.012)	(0.130)	(0.189)
Achievement gap							−0.113	0.060*	−0.063	−0.001	−0.776*	0.407
							(0.065)	(0.028)	(0.034)	(0.018)	(0.330)	(0.379)
Constant	−0.133***	−0.005	0.000	−0.030***	12.242***	11.405***	−0.044	−0.035**	0.040**	−0.026**	12.635***	11.215***
	(0.008)	(0.006)	(0.006)	(0.003)	(0.045)	(0.052)	(0.027)	(0.012)	(0.014)	(0.007)	(0.137)	(0.157)
R^2	0.238	0.236	0.251	0.216	0.403	0.478	0.185	0.253	0.218	0.202	0.415	0.485
N	177	177	177	177	177	177	172	172	172	172	172	172
$P-val$				.058						.231		

†Notes: This table reports coefficients from regressions of the class-level differences between boys' and girls' characteristics on the gender bias. For instance, the first row reports the coefficient of a regression of the class-level differences between boys' and girls' age on the gender bias. Each regression includes school fixed effects and controls for differences between girls' and boys' blind scores (because that variable is systematically controlled for in the analysis). I measure teacher bias using a leave-out mean peer exposure to bias (Chetty et al., 2011). I use this jackknife estimate for all correlations reported in this table. The last row of the table reports the LHS joint balancing test, which is an F-test for the joint significance of the gender bias coefficients in the six regressions. Standard errors are clustered at the school-level. Stars correspond to the following p-values: *$p < .05$; **$p < .01$; ***$p < .001$.

also engage in other unobserved classroom practices that make boys less likely to succeed. They might be less encouraging, less friendly, focus less attention on boys, or be more critical. The confounding effect of teachers' behavior is a concern when measuring the gender bias at the very beginning of grade 6 and measuring students' progress during grade 6 (between September and June) because pupils experience the gender bias in evaluations at the beginning of the year and then potentially experience the biased behavior of their teacher throughout the entire year. To partially disentangle these two effects, I use the bias measured at the end of the grade 6—instead of the beginning—and pupils' progress between the beginning of grade 7 and the end of grade 9. This ensures that the progress is measured over a period when pupils are less affected by the biased behavior of their teacher.

4. Empirical results

4.1. Average gender bias across all teachers

I start by estimating the average gender bias across all teachers using Eq. (7). The first column of Table 4 presents the results without control variables. In math, the coefficient of the interaction term Girl × Non-Blind is high and significant—0.259 points of the SD—indicating a strong bias against boys in math. Conditional on blind scores, boys' non-blind scores are on average 5.2% lower than girls in math. On the other hand, the results do not show any gender bias in French. I present results using blind and non-blind evaluation at the beginning of the year in Appendix A and results decomposed by teachers' characteristics in Appendix B.[21]

These results partially confirm what Lavy (2008) observes in his analysis: despite the commonly held belief that girls are discriminated against, teacher biases favor girls. Similarly, Robinson and Lubienski (2011) found that teachers in elementary and middle schools consistently rate females higher than males in both math and reading, even when cognitive assessments suggest that males have an advantage. Contrary to both previous studies, I find a bias only in math. Breda and Ly (2015) also found that discrimination goes in favor of females in more "male-connoted" subjects (e.g., math).

4.2. Effect of teachers' gender biases on progress

I now turn to results on the effect of being assigned a biased teacher on students' outcomes. Estimates are based on Eq. (6) and are reported in Table 5. The dependent variable is girls' relative progress between the end of grade 6 and the end of grade 9.[22] The variable of interest is the gender bias of the grade 6 teacher—measured at the end of the year—and all regressions control for the gender achievement gap measured at the beginning of grade 6. For inference, I use a two-step bootstrapping method because the bias variable is a generated regressor. I correct for the sampling error that affects the standard errors of the coefficient β by using the same method as in Ashraf and Galor (2013). Appendix D contains a detailed presentation of the method.

Results reported in column 1 suggest that teachers' gender biases have a high and significant effect on girls' progress relative to boys in both math and French. For two classes where the achievement gap between boys and girls would be identical in grade 6, randomly assigning a teacher who is one standard deviation more biased against boys to one of the classes would decrease boys' relative progress in that class by 0.123 SD in math and by 0.106 SD in French.

As shown in Table 1, during the four years of middle school girls catch up with—and even overtake—boys in math and French. At the beginning of grade 6, boys' blind score is 0.147 SD higher than girls in math. By the end grade 9, the achievement gap is in favor of girls, whose math score is 0.058 SD higher than boys. This represents a relative falling behind of boys compared to girls of 0.205 SD over the four years of middle school. Trying to understand how much of this is due to the gender bias, I find that 6% of boys' falling behind girls in math can

[21] All results presented here are based on blind and non-blind scores given at the end of grade 6. To test if the gender bias differs at the beginning of the year, I use the blind and non-blind grades given during the first term and replicate the analysis. The gender bias is slightly larger during the first term than during the last term.

[22] Note that, unlike the grade 6 blind scores, the grade 9 blind scores are high-stakes for the pupils. If girls tend to be relatively less effective than boys when stakes are higher, the measure I use might show lower progress for girls than for boys (Azmat, Calsamiglia, & Iriberri, 2016). However, this would not explain my results so long as stakes-sensitive boys and girls are not differentially selected to biased teachers, which is what the balancing tests suggest.

Table 4
Estimation of the gender bias - third term.

	(1)	(2)	(3)	(4)	(5)
Girl × Non-Blind	0.259***	0.251***	0.220***	0.251***	0.213***
	(0.035)	(0.041)	(0.042)	(0.036)	(0.044)
Girls	−0.045	−0.039	−0.106	−0.053	−0.108
	(0.037)	(0.051)	(0.053)	(0.036)	(0.053)
Non-Blind Score	−0.120	−0.182	−0.139	−0.111	−0.132
	(0.069)	(0.089)	(0.089)	(0.067)	(0.088)
Controls for punishment			.		.
Punishment			−0.700***		−0.683***
			(0.050)		(0.052)
Punishment × Non-Blind			−0.168**		−0.163**
			(0.045)		(0.043)
Punishment × Non-Blind × Girl			0.036		0.031
			(0.098)		(0.095)
Controls for grade repetition				.	.
Grade repetition				−0.372***	−0.297**
				(0.074)	(0.084)
Repetition × Non-Blind				−0.119	−0.114
				(0.144)	(0.193)
Repetition × Non-Blind × Girl				0.132	0.126
				(0.132)	(0.166)
Constant	−1.398***	2.220***	1.672***	−1.384***	1.737***
	(0.066)	(0.125)	(0.125)	(0.065)	(0.123)
Class FE	Yes	Yes	Yes	Yes	Yes
R^2	0.122	0.126	0.181	0.131	0.188
Number of observations	7714	4460	4460	7714	4460

Notes: This table reports DiD estimates of teachers' gender biases. The dependent variable is the score (both blind and non-blind) obtained by a pupil in math during the last term of grade 6. Each pupil has two observations: one for the blind score and one for the non-blind score. The punishment variable takes a value of 1 if a pupil has received a disciplinary warning from the class council during the third term of grade 6 or if he/she was temporarily excluded from the school. Column 2 presents results of the standard DiD regression implemented on the sample of students for whom punishment information is available. The sample used in columns 2, 3, and 5 does not include pupils for whom a punishment variable is missing. All regressions include a class fixed effect. Standard errors are clustered at the school-level. Stars correspond to the following p-values: *$p < .05$; **$p < .01$; ***$p < .001$.

Table 5
Effect of teachers' gender biases.

	Progress		General HS		Science		Literature		Repetition	
	Bias in		Bias in		Bias in		Bias in		Bias in	
	Math (1)	French (2)	Math (3)	French (4)	Math (5)	French (6)	Math (7)	French (8)	Math (9)	French (10)
Gender bias	0.123***	0.106*	0.026	0.029	0.036*	0.021	−0.008	0.005	−0.030	−0.020
	(0.037)	(0.045)	(0.020)	(0.023)	(0.015)	(0.016)	(0.011)	(0.011)	(0.018)	(0.021)
Achievement gap	−0.323***	−0.604***	0.209***	0.184***	0.154***	0.099***	0.003	0.047**	−0.069**	−0.061
	(0.058)	(0.056)	(0.028)	(0.034)	(0.023)	(0.025)	(0.015)	(0.019)	(0.024)	(0.036)
Constant	0.108***	0.333***	0.102***	0.020	0.005	−0.042*	0.068***	0.049***	−0.099***	−0.071***
	(0.025)	(0.033)	(0.014)	(0.019)	(0.010)	(0.017)	(0.007)	(0.011)	(0.014)	(0.019)
N	177	172	177	172	177	172	177	172	177	172
R^2	0.242	0.398	0.203	0.150	0.227	0.084	0.031	0.004	0.052	0.030

†Notes: This table reports estimates of the effect of teachers' gender biases. In columns 1 and 2, the dependent variable is the gap between girls' and boys' progress between the beginning of grade 7 and the end of grade 9. In columns 3 and 4, the dependent variable is the gap between girls' and boys' probability of choosing a general high school in grade 10. In columns 5 and 6, the dependent variable is the gap between girls' and boys' probability of choosing a scientific track in grade 11. In columns 7 and 8, the dependent variable is the gap between girls' and boys' probability of choosing a literature track in grade 11. In columns 9 and 10, the dependent variable is the gap between girls' and boys' probability of repeating a grade. For all regressions, I use the empirical Bayes estimate of teacher bias. The unit of observation is a class. Standard errors are estimated with a two-step bootstrapping method. Stars correspond to the following p-values: *$p < .05$; **$p < .01$; ***$p < .001$.

be ascribed to teachers' gender bias against them. Moving from no gender bias to the average estimate of teachers' bias (0.259) makes boys progress 0.032 points less than girls. These results are very much in line with Carlana (2018), who finds that the gender gap in math progress triples in classes assigned to a math teacher who implicitly associates boys more than girls with mathematics.

When interpreting the previous coefficients, we should keep in mind that the effect is relative: saying that teachers' gender biases reduces boys' relative progress is equivalent to saying that it increases girls' relative progress. For consistency, I will use the first construction. As the outcome corresponds to the difference between girls' and boys' progress, the positive coefficient I find could correspond to higher

progress for girls than for boys, or a blind score that remains constant for girls over time but decreases for boys (due to their feeling of being negatively discriminated against compared to girls, for instance). To check which effect dominates, I estimate Eq. (6) by using successively girls' and boys' progress as outcomes.[23] The results suggest that having a math teacher who is one SD more biased against boys does not impact boys' progress but significantly increases girls' progress (coef = 0.103, SE = 0.037). Again, these results confirm those from Carlana (2018), which finds that teachers' stereotypes have no effect on boys, while they lower girls' math performance. Interestingly, I find the opposite effect in French where having a biased teacher significantly reduces boys' progress (coef = −0.061, SE = 0.037) but positively impacts girls' progress (coef = 0.047, SE = 0.040), although the coefficients are not significant.[24]

4.3. Effect of teachers' gender biases on course choice

Grade 9 is the last grade of middle (and compulsory) school. After this grade, pupils choose between a vocational, technical, or general high school. Then, for the pupils who decide to attend a general high school, everyone attends the same courses during grade 10, but pupils have to specialize when they enter grade 11. Three track options are available: sciences, humanities, or economics and social sciences. In our sample, 50.9% of girls chose a general high school, as did 40.3% of boys. This highly unbalanced statistic raises a first question: do teachers' gender biases impact the type of high school boys choose compared to girls? In this sample, among girls in general high school, 32.8% chose the scientific track, compared with 40.2% of boys. This reversal of the gender probability is striking, as the scientific path is the most prestigious. It is also the path that leads to higher education in science, technology, engineering, and math (STEM) fields. These fields of study are highly gender-unbalanced in most countries, which raises a second question: do teachers' gender biases impact the relative probability that girls enroll in a scientific track?

Using Eq. (6), I analyze the effect of teachers' gender biases on four additional outcomes: boys' relative probability to attend a general high school, to choose the scientific track, to choose the literature track, or to repeat a grade. I present my results in columns 3 to 10 of Table 5. All regressions are run on all pupils to avoid any selection effect. For instance, the regression of the probability to choose a scientific course in grade 11 is not conditional on attending a general high school.

I find that being assigned a teacher who is one SD more biased against boys in grade 6 decreases boys' relative probability of attending a general high school (rather than a professional or technical one) by 2.6 percentage points, although that coefficient is not statistically significant. Yet, a back-of-the-envelope calculation confirms the sign and the size of this effect. I find that having a teacher who is 1 SD more biased against boys in math decreases their relative progress by 0.123 SD, and a simple regression shows that a one-SD drop in boys' relative achievement at the end of middle school reduces their relative probability of attending a general high school by 20.7 percentage points. By combining these two effects, I get an upper-bound effect of a biased teacher on boys' relative probability of attending a general high school of 0.025. This is very much in line with the coefficient I obtain (0.026).[25]

The results reported in column 5 suggest that teachers' biases in math positively affect girls' relative probability of choosing the scientific track in grade 11. More precisely, having a teacher who is one SD more biased in favor of girls increases girls' probability of selecting a scientific track by 3.6 percentage points compared to boys. This would reduce the gap between boys' and girls' probabilites of choosing a scientific track (initially 7.4 percentage points) by 48%. This observation is important, as the scientific path is the most prestigious one, and the one that leads to higher education in STEM fields. This result is in line with Lavy and Sand (2018), who found that "the estimated effect of math teachers' stereotypical attitude [in favor of boys] on enrollment in advance studies in math is positive and significant for boys (0.093, SE = 0.049) and negative and significant for girls (-0.073, SE = 0.044)."

I also calculate what share of the observed gender gap in the scientific track choice is due to the average value of teachers' gender biases in math (estimated around 0.259 points of a SD). Teachers' average biases in math reduce the gender gap in scientific track enrollment by 12.5%.

Finally, the biases of French teachers have no impact on girls' relative probability of selecting the scientific track in grade 11. Teachers' biases against boys in math and French seem to increase boys' relative probability of repeating a grade, although the coefficients are not statistically significant.

4.4. Exploiting between-subjects variation in gender bias

So far, the analysis has exploited the gender bias variation between teachers within a given subject. Another interesting approach is to exploit the gender bias variation *between subjects* within students. I investigate the effect of an increase in the gender bias of the math teacher (relative to the gender bias of the French teacher) on student progress in math (compared to French). This alternative approach has two main advantages. First it allows to run student-level (instead of class-level) regressions. It also gives me the opportunity to investigate whether teacher gender bias has a different effect for boys and girls.

I use the following regression:

$$B_{2is} - B_{1is} = a + b.\ GenderBias_{is} + c.\ T_{is} + \gamma_i + \epsilon_i \quad (9)$$

The outcome of interest is student progress between the beginning of grade 7 and the end of grade 9. Note that the above equation is equivalent to Eq. (5) with two exceptions. First, the unit of observation is a student-by-subject instead of a student. Second, the variable of interest is a teacher gender bias (now indexed by subject). γ_i is a student fixed effect. T_{is} is a proxy for teacher value-added, which I measure using the following steps: I start by running a regression of student progress (between the beginning and the end of the year) on student initial blind score, gender, and social background. Then, I aggregate the residual of that regression at the teacher level and use that measure as a proxy for teacher value-added.

The results, reported in Table 6, show that increasing the gender bias in favor of girls in math (compared to French) by 1 SD increases students' progress in math (compared to French) by 0.052 SD. I test a specification with student fixed effects and a specification without fixed effects. The latter controls for gender, social background, and baseline blind score. The sign and magnitude of the results are similar in both specifications, but the results are only significant in the specification that does not include student fixed effects.

The key question is whether boys and girls react differently to their

[23] As explained in Section 2, the first-difference specification ensures that the estimates do not capture the effect of teachers' quality, which might be correlated to a teacher' gender bias. In that sense, the first-difference specification might provide a better identification than a specification that uses boys' and girls' outcomes separately.

[24] The differences observed between subjects and genders are consistent with a simple model that would take into account two parameters: (1) the importance attached to grades (assumed to be higher for girls than for boys) and (2) the lack of self-confidence (assumed to be higher in French for boys and in math for girls). If students are more impacted by encouragement in a subject where they lack self-confidence, we would intuitively expect a bias in math and French to impact boys and girls differently.

[25] This is an upper bound due to the high endogeneity in the second regression of boys' relative probability to attend a general high school on their relative achievement at the end of middle school.

Table 6
Effect of teachers' gender bias on student progress - between-subject analysis.

	All students		Girls		Boys	
	(1)	(2)	(3)	(4)	(5)	(6)
Gender Bias	0.039*	0.052	0.096***	0.080	−0.021	0.023
	(0.016)	(0.034)	(0.021)	(0.046)	(0.023)	(0.049)
Teacher VA	−0.002	0.021	−0.006	0.013	0.001	0.029
	(0.017)	(0.034)	(0.021)	(0.043)	(0.026)	(0.053)
Blind score	−0.042**		−0.022		−0.066**	
	(0.016)		(0.023)		(0.024)	
Girl	0.111***					
	(0.031)					
High SES	0.027		0.093		−0.030	
	(0.042)		(0.060)		(0.059)	
Low SES	0.020		0.015		0.031	
	(0.038)		(0.054)		(0.054)	
Constant	−0.139***	−0.068***	−0.040	−0.015	−0.133**	−0.124***
	(0.038)	(0.015)	(0.050)	(0.021)	(0.050)	(0.022)
Student FE	No	Yes	No	Yes	No	Yes
Observations	6011	6011	3064	3064	2947	2947
R-Square	0.008	0.623	0.015	0.615	0.007	0.627

Notes: This table reports coefficients from regressions of student progress (between the beginning of grade 7 and the end of grade 9) on teachers' gender biases. The unit of observation is a student-by-subject, so that each student has two observations, one in math and one in French. The regressions in columns (2), (4), and (6) include student fixed effects and control for teacher value-added. The regressions in columns (1), (3), and (5) do not include student fixed effects, but instead control for student gender, social background, and blind score at the beginning of grade 6, in addition to teacher value-added. For all regressions, I use the empirical Bayes estimate of teacher bias. Standard errors are estimated with a two-step bootstrapping method. Stars correspond to the following *p*-values: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

teachers' relative gender bias in favor of girls. I find a larger effect for girls than for boys. For girls, increasing the gender bias in math (compared to French) by 1 SD increases their progress in math by 0.08 SD compared to their progress in French. For boys, it only increases relative progress by an insignificant 0.023 SD. That interesting difference provides additional evidence on boys' and girls' different reactions to teacher biases. It is in line with my previous results which show that having a math teacher who is more biased against boys does not impact boys' progress, but significantly increases girls' progress.

4.5. Discussion of potential mechanisms

4.5.1. Bias in grades or bias in skills measured?

As mentioned in Section 3.3, the impact of the gender bias is likely to capture the effect of teachers' bias in grades, in evaluation methods and potentially in their behavior. I show that the blind and non-blind evaluations are measuring very similar skills, thus it is unlikely that the effect of teachers' bias is driven by differences in how teachers weight students' cognitive and non-cognitive skills. To show that grades given by teachers and standardized evaluations are measuring similar skills, I use a simple model that describes what blind and non-blind scores measure. I model blind scores B_i as a noisy measure of a pupil's ability θ_{iB} :

$$B_i = \theta_{iB} + \epsilon_{iB}, \tag{10}$$

where ϵ_{iB} is measurement error on the blind score. Non-blind scores NB_i are measuring a different ability θ_{iNB}. They can also be affected by a student' gender G_i:

$$NB_i = \alpha_0 + \theta_{iNB} + \alpha_2 G_i + \epsilon_{iNB}. \tag{11}$$

Abilities measured by blind and non-blind scores might differ. I model the relationship between the two scores as

$$\theta_{iNB} = \rho\theta_{iB} + \nu_i, \tag{12}$$

where ν_i captures any specific ability measured by class exams but not by standardized tests. The skills measured by blind scores (θ_{iB}) might include pupils' long-term memory and their ability to synthesize knowledge acquired in the last few months, while ability measured by non-blind scores (θ_{iNB}) might also integrate homework or non-cognitive skills such as motivation or perseverance.[26] By replacing θ_{iNB} by its formula in Eq. (11), and by replacing θ_{iB} by $(B_i - \epsilon_{iB})$, we obtain

$$NB_i = \alpha_0 + \rho B_i + \alpha_2 G_i + (\epsilon_{iNB} + \nu_i - \rho\epsilon_{iB}). \tag{13}$$

By estimating Eq. (13), I can test if ρ is significantly different from 1. If not, it is safe to assume that both tests measure similar skills. Due to the measurement error affecting the blind score, I use two instrumental variable approaches for this estimation. In the first approach, I instrument the third-term blind score with a dummy variable indicating whether the student is born at the end of the year (between July and December). In the second approach, I instrument the third-term blind score with the first-term blind score. The methods and results are fully detailed in Appendix F, which also discusses the independence and exclusion restriction assumptions. As reported in Table F.2, the IV coefficient of the blind score is equal to 0.941 (with blind score instrument) and 1.009 (with birth date instrument) in French. In math, it ranges from 0.864 (with blind score instrument) to 1.050 (with birth date instrument). In three regressions out of four, I cannot reject the hypothesis that $\rho = 1$. The effect of the gender bias that I estimate is therefore unlikely to be driven by differences between teachers in how much they weight cognitive and non-cognitive skills when they evaluate their students.[27]

4.5.2. Spillovers of teachers' Ggender biases

I test the existence of between-subjects effects to understand if the biases of math teachers can impact the progress of students in French,

[26] This model of blind and non-blind scores is highly simplified and relies on two important hypotheses. I suppose a linear relation between non-blind scores, ability, and gender, and I assume that non-blind scores do not depend on blind scores. This hypothesis is likely to be satisfied in our context because blind tests were not graded by teachers but by independent correctors.

[27] In addition, the IV estimate of the gender bias measure (α_2) ranges from 0.178 to 0.206 in math and from 0.072 to 0.087 in French, values that are very similar to the coefficients obtained from DiD.

Table 7
Effect of teachers' biases with spillovers.

	Progress over 3 years in				Science course	
	Math		French			
	(1)	(2)	(3)	(4)	(5)	(6)
Gender Bias Math	0.123***	0.118**		0.069	0.036*	0.035*
	(0.037)	(0.039)		(0.044)	(0.015)	(0.016)
Gender Bias French		0.047	0.106*	0.103**		0.004
		(0.044)	(0.045)	(0.037)		(0.029)
Achievement Gap	−0.323***	−0.332***	−0.604***	−0.603***	0.154***	0.155***
	(0.058)	(0.061)	(0.056)	(0.066)	(0.023)	(0.025)
Constant	0.108***	0.099***	0.333***	0.335***	0.005	0.008
	(0.025)	(0.025)	(0.033)	(0.036)	(0.010)	(0.011)
Observations	177	170	172	170	177	170
R-Square	0.242	0.265	0.398	0.417	0.227	0.227

†Notes: This table reports estimates of the effect of teachers' gender biases on girls' progress relative to boys and on girls' relative course choice. The unit of observation is a class. In columns 1 to 4, the dependent variable is the gap between girls' and boys' progress between the beginning of grade 7 and the end of grade 9. In columns 5 and 6, the dependent variable is the gap between girls' and boys' probability of selecting a science course in grade 11. For all regressions, I use the empirical Bayes estimate of teacher bias. Standard errors are estimated with a two-step bootstrapping method. Stars correspond to the following *p*-values: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

and vice-versa. To this end, I estimate the effect of the gender bias in math and French simultaneously on boys' relative outcomes. Including both biases in a regression is also a good means to test and confirm that the gender bias of French and math teachers are independent. Including the bias in French in a regression should not change the effect of the gender bias in math.

I present results from the standard specification (without spillovers) in columns 1, 3, and 5 of Table 7. In columns 2 and 4, I regress girls' relative progress in a given subject on both the bias in this subject and the bias in the second subject. The results show a complete absence of spillovers: boys' relative progress in math over middle school is affected by their teachers' biases in math, but not by their teachers' biases in French. The reverse is true in French: boys' relative progress in French is not affected by their teachers' gender biases in math. The last column reports the result for boys' relative probability to select a scientific track, and again, there is no spillover. In addition, it is important to notice that, between columns 1 and 2, the coefficient of the bias in math does not change when the bias in French is included in the regression, confirming the two variables are independent.

4.5.3. Cumulative effect over time of the gender bias

Teachers' biases affect boys' relative progress over middle school. In this section, I test if this effect corresponds to a cumulative effect of being assigned a biased teacher for several consecutive years. Pupils assigned to more biased teachers might have a higher probability to be re-assigned the same teacher in later grades.[28] If this is the case, and if the effect of teachers' gender biases is cumulative over time, the effect I observe will be additive. To test this, I have information on the teacher a pupil is assigned to during grades 6 and 7. I check if the probability that a pupil is assigned the same teacher during grade 7 is correlated to his/her teachers' gender biases.[29] The results suggest that being assigned a grade 6 teacher with a one-SD higher gender bias increases a pupil's probability of being reassigned to the same teacher in grade 7 by 4.1 percentage points in math (SD = 0.004), but decreases a pupil's probability by 2.5 percentage points in French (SD = 0.05). Both coefficients are statistically significant, and the estimates are very similar for boys and girls. Then I check if the effect of the bias is cumulative—in other words, if being reassigned a biased teacher further impedes boys' relative progress. For each class, I calculate the percentage of pupils in the class that are reassigned to the same teacher in grade 7. I add this variable, and its interaction with the gender bias, to the previous specification. The results presented in Table 8 clearly indicate that the effect of teachers biases' is not cumulative over time: the interaction term added is close to 0 in math and French. Being re-assigned the same biased teacher does not further reduce boys' relative progress. This result is unsurprising if we think that students might become aware of the gender biases of their teachers, so that the effect fades out. If effort and achievement are substitutes, boys could even increase their effort once they realize that they perform relatively poorly compared to girls. My findings differ from Alan, Ertac, and Mumcu (2018) which find that the effect of being exposed to teachers with traditional gender views is amplified with longer exposure to the same teacher.

4.5.4. Contrast effects, stereotype threat, and mistrust

Finally, prior research can help interpret my results. Some papers highlight a "contrast effect" according to which a student's academic self-concept is positively influenced by his or her individual achievement, but negatively affected by other peers' average achievement—usually composed of peers in the classroom—after controlling for individual achievement (Marsh & Craven, 1997; Murphy & Weinhardt, 2013; Trautwein et al., 2006). This helps explain why a gender bias in a given subject—which is a bonus for girls compared to their male peers—could increase girls' progress in this subject, but reduce boys'. Positively rewarding girls, relative to boys, could also reduce the stereotype threat effect. In situations where stereotypes are perceived as important, some girls perform poorly for the sole reason that they fear confirming the stereotypes (Spencer et al., 1999). If girls perceive math as highly affected by teachers' stereotypes, over-grading girls in this subject would reduce their anxiety of being judged as poor performers, and therefore favor their progress in math. Conversely, if boys become aware of the gender biases of their teachers, they might develop behavior that confirms that bias. Finally, if teachers' gender biases are too obvious during grade 6, boys and girls might increasingly mistrust their grades. Mechtenberg (2009) suggests that girls are

[28] Pupils cannot have the same teachers in earlier grades since grade 6 is the first grade of middle school. All pupils were in a different school the year before.

[29] More specifically, I regress a dummy indicating if a pupil has the same teacher during grades 6 and 7 on the gender bias of the grade 6 teacher. This regression is run on a sample of 3761 pupils for which I have information on their grade 7 teacher.

Table 8
Cumulative effect of teachers' biases.

	Math		French	
	(1)	(2)	(3)	(4)
Gender Bias	0.123***	0.130***	0.106*	0.114*
	(0.037)	(0.044)	(0.045)	(0.057)
Achievement Gap	−0.323***	−0.323***	−0.604***	−0.608***
	(0.058)	(0.059)	(0.056)	(0.057)
Gender Bias*Pct Same Teacher		−0.083		−0.085
		(0.251)		(0.458)
Observations	177	177	172	172
R-Square	0.242	0.248	0.398	0.400

†Notes: This table reports estimates of the effect of teachers' gender biases on girls' progress relative to boys. The unit of observation is a class. In columns 1 to 4, the dependent variable is the gap between girls' and boys' progress between the beginning of grade 7 and the end of grade 9. For all regressions, I use the empirical Bayes estimate of teacher bias. Standard errors are estimated with a two-step bootstrapping method. Stars correspond to the following *p*-values: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

reluctant to internalize good grades in math because they believe their grades are biased.

5. Robustness checks

First, I run a placebo test where teachers are randomly assigned to different classes. Running the standard regression with boys' relative progress as a dependent variable in both math and French gives insignificant coefficients in both subjects (in math: $\beta = -0.032$, SD = 0.027, while in French: $\beta = -0.019$ and SD = 0.036).

Second, the blind and the non-blind tests are not taken at exactly the same date, which can affect my estimates. Pupils take the standardized blind test during one of the last days of the school year, while teachers' assessments are an average of several grades given between April and mid-June (last term of the academic year). Hence, teachers' scores measure a pupil's average ability about one and a half months before the end of the school year. This time lag between the dates of the blind and non-blind scores might be worrisome if girls tend to progress more than boys during this period, especially if girls' better progress is higher in classes where teachers are more biased. Yet, because the blind score is measured after the non-blind score, the double-difference coefficient, which captures the gender bias, would be a lower bound for the true gender bias if girls tend to progress more. Most importantly, the higher the gender bias of a teacher, the larger the downward bias, so that the time lag would tend to shrink the variance of the gender bias. As a result, my estimates would tend to underestimate the impact of teachers' gender biases on students' progress and subsequent outcomes.

6. Conclusion

A number of papers have shown that teachers' stereotypes can bias their assessments and grades, yet none of these papers has gone one step further by studying the impact of teachers' gender biases on students' subsequent progress and schooling trajectories. This paper takes that next step. I use a new identification strategy based on the variation of gender biases between teachers and the quasi-random assignment of students to these different teachers to study longer-term outcomes. To measure gender biases, I use a standard double-difference methodology that exploits the availability of both blind and non-blind scores for each student.

The key finding is that teachers' gender biases have a high and significant effect on girls' progress relative to boys' in both math and French. Over middle school, teachers' gender bias against boys explains 6% of boys falling behind girls in math. Moving to other outcomes, I find that having a teacher who is one SD more biased in math increases girls' probability of selecting a scientific track in high school by 3.6 percentage points compared to boys'. Teachers' average bias in math reduces the gender gap in choosing scientific courses by 12.5%.

I use a dataset that has been collected from schools in a relatively deprived educational district in France. Teachers assigned to deprived areas are on average younger than teachers in more advantaged schools, and we have seen that inexperienced teachers are more biased. Similarly, pupils in these areas might face more constraints (financial or self-censorship) regarding their schooling decisions. This should be kept in mind when interpreting the results.

An interesting follow-up would look at the channels through which gender bias affects boys' relative achievement. Rewarding pupils could provide motivation, increase effort and self-confidence, and reduce the effects of stereotype threat. On the other hand, if pupils consider effort and abilities as substitutes, a higher grade might be an incentive to reduce effort and work. Unfortunately, I am not able to disentangle these effects, which might compensate for or reinforce each other. This is an interesting question for future research. Another follow-up would look at long-term effect of teacher gender bias. Lavy and Megalokonomou (2019) started exciting work in that direction. Finally, replicating this analysis based on students' ethnicity would also be an interesting direction for future work.

This analysis provides policy-relevant results. Teachers' gender biases can have a strong impact on the achievement gap between boys and girls. This provides a new explanation for boys increasingly falling behind girls at school, and for girls choosing relatively fewer scientific courses in high school. These findings open the door for new policies. If the main objective of policy-makers is to reduce achievement gaps—-whether between boys and girls or students from different ethnicities or social backgrounds—teachers' evaluation methods and behavior could be considered an instrument to achieve this goal.

Appendix A. Estimation of gender biases at the beginning of grade 6

Table A.1
Skills measured by standardized tests and class exams.

Standardized tests	Class exams
Math	
Space and geometry Recognize and draw two-dimensional figures Properties of alignment, perpendicular, parallel, and symmetry Recognize a cube shape and parallelepiped rectangle	**Geometry** Two-dimensional figures Symmetry of a straight line Parallelepiped rectangle
How to exploit numerical data Solve a problem using proportionality Solve pbs with addition/substraction/multiplication/division Read and interpret a table, diagram, and graphic	**Organize and understand data** Proportionality Read information in tables Read information on axis, diagrams/graphics
Size and measurement Knowledge and use of measurement units (length, mass, volums, and duration)	**Size and measurement** Length, mass, and duration Angles Area: measure, comparison, and calculus Volumes
Knowledge of natural whole numbers Knowledge of integers Use and writing of fractions Use decimal numbers	**Numbers and calculus** Integer numbers and decimals Fractions
Calculus Knowledge of the four operations	Operations
French	
Knowledge and recognition of words Understand the formation of words Exploit time-space indications Knowledge of verb tenses	**Grammar** Classes of words (noun, pronoun, verb) Conjugation Tenses (present/past/future) **Spelling** Grammatical spelling Lexical spelling
Understanding of words Decipher rare words Understand the meaning of a word with its context Classify and link information	**Vocabulary** **Reading**
Production of a text Add punctuation to a text Produce a coherent text Transform a text Use of usual words	**Writing** Use of punctuation Production of a text (one page max)
	Oral expression (reading aloud, recitation) **Initiation to art history**

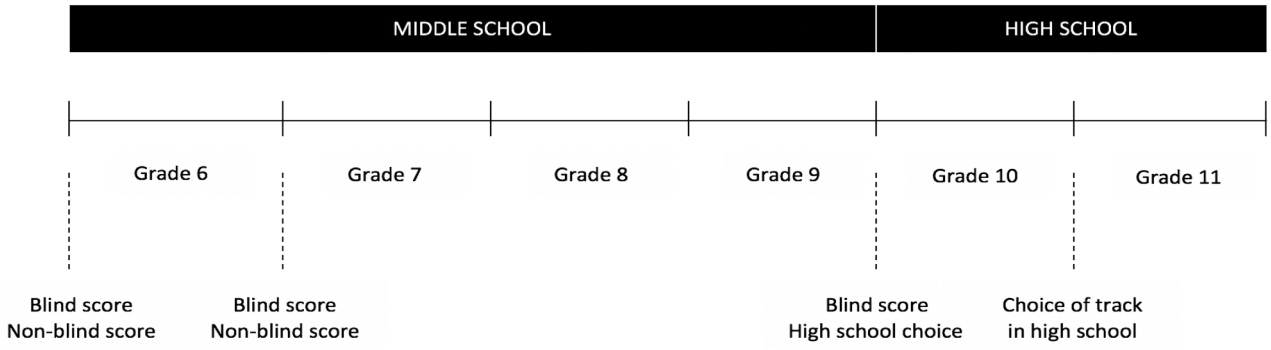


Fig. A.1. Timeline.

Table A.2
Estimation of the gender bias - first term.

	(1)	(2)	(3)	(4)	(5)
Girl × Non-Blind	0.318***	0.327***	0.317***	0.313***	0.318***
	(0.027)	(0.031)	(0.029)	(0.027)	(0.031)
Girl	−0.152***	−0.146**	−0.211***	−0.160***	−0.221***
	(0.028)	(0.040)	(0.039)	(0.028)	(0.039)
Non-Blind Score	−0.156**	−0.170*	−0.147*	−0.150**	−0.145
	(0.052)	(0.064)	(0.067)	(0.051)	(0.071)
Controls for punishment			.		.
Punishment			−0.566***		−0.546***
			(0.076)		(0.075)
Punishment × Non-Blind			−0.153*		−0.152*
			(0.071)		(0.070)
Punishment × Non-Blind × Girl			−0.301		−0.296
			(0.157)		(0.170)
Controls for grade repetition				.	.
Grade repetition				−0.352***	−0.383**
				(0.090)	(0.125)
Repetition × Non-Blind				−0.076	−0.007
				(0.133)	(0.096)
Repetition × Non-Blind × Girl				0.077	−0.040
				(0.112)	(0.165)
Constant	2.361***	4.717***	5.034***	2.492***	5.170***
	(0.133)	(0.062)	(0.067)	(0.127)	(0.072)
Class FE	Yes	Yes	Yes	Yes	Yes
R^2	0.118	0.105	0.136	0.125	0.143
N	8329	4413	4413	8329	4413

Notes: This table reports DiD estimates of teacher gender bias. The dependent variable is the score (both blind and non-blind) obtained by a pupil in math during the first term of grade 6. Each pupil has two observations: one for the blind score and one for the non-blind score. The punishment variable takes a value of 1 if a pupil has received a disciplinary warning from the class council during the first term of grade 6 or if he/she was temporarily excluded from the school. Column (2) presents results of the standard DiD regression implemented on the sample of students for which the punishment information is available. The sample used in columns (2), (3), and (5) does not include pupils for which a punishment variable is missing. All regressions include a class fixed effect. Standard errors are clustered at the school-level. Stars correspond to the following p-values: *$p < .05$; **$p < .01$; ***$p < .001$.

Appendix B. Gender bias and teachers' characteristics

Contrary to prior research that finds that girls tend to benefit from discrimination in all subjects (Cornwell et al., 2013; Falch & Naper, 2013; Lavy, 2008; Lindahl, 2007; Robinson & Lubienski, 2011), my results suggest that girls are favored in math only. To explain this difference, I focus on teachers' characteristics that could influence their grading practices, specifically characteristics that would be different for math and French teachers (such as their gender or experience). As displayed in Table 1, the share of male and female math teachers is the same, but the pattern is very different in French, where 85% of the teachers are female. On average, math teachers are 3.5 years younger than French teachers.

Several studies show that the interplay between student and teacher gender plays a role in teachers' assessments (Dee, 2005; Falch & Naper, 2013; Lavy, 2008; Lindahl, 2007; Ouazad & Page, 2013). To test if teachers' genders explain their biased behavior, I run the previous DiD regressions separately on the sub-sample of male and female teachers (Fig. B.1). I find that the gender bias does not differ by teachers' gender in French, and only marginally in math. In math, female teachers are less biased in favor of girls than male teachers: the average gender bias is 0.294 for female teachers and 0.343 for male teachers, but this difference is not significant.[30]

Second, I test if teachers' experience affects gender bias. I decompose the sample into three groups of teachers based on their years of experience: first year, two to five years, and more than five years. 58.1% of math teachers and 45% of French teachers have five or fewer years of experience. I run the DiD regression on each of the three samples. The results suggest that, in mathematics, teachers in their first year of teaching are more biased than more experienced teachers: the average gender bias represents 0.571 points of a SD for new math teachers, versus 0.295 for teachers with more than five years of experience. In French, teachers' experience has no effect on their gender bias.

[30] My findings are in line with those of Falch and Naper (2013) who found a limited or zero effect of teachers' gender on the gender bias in grades. They do not confirm Lavy (2008), whose results suggest that the gender bias in math is driven by male teachers.

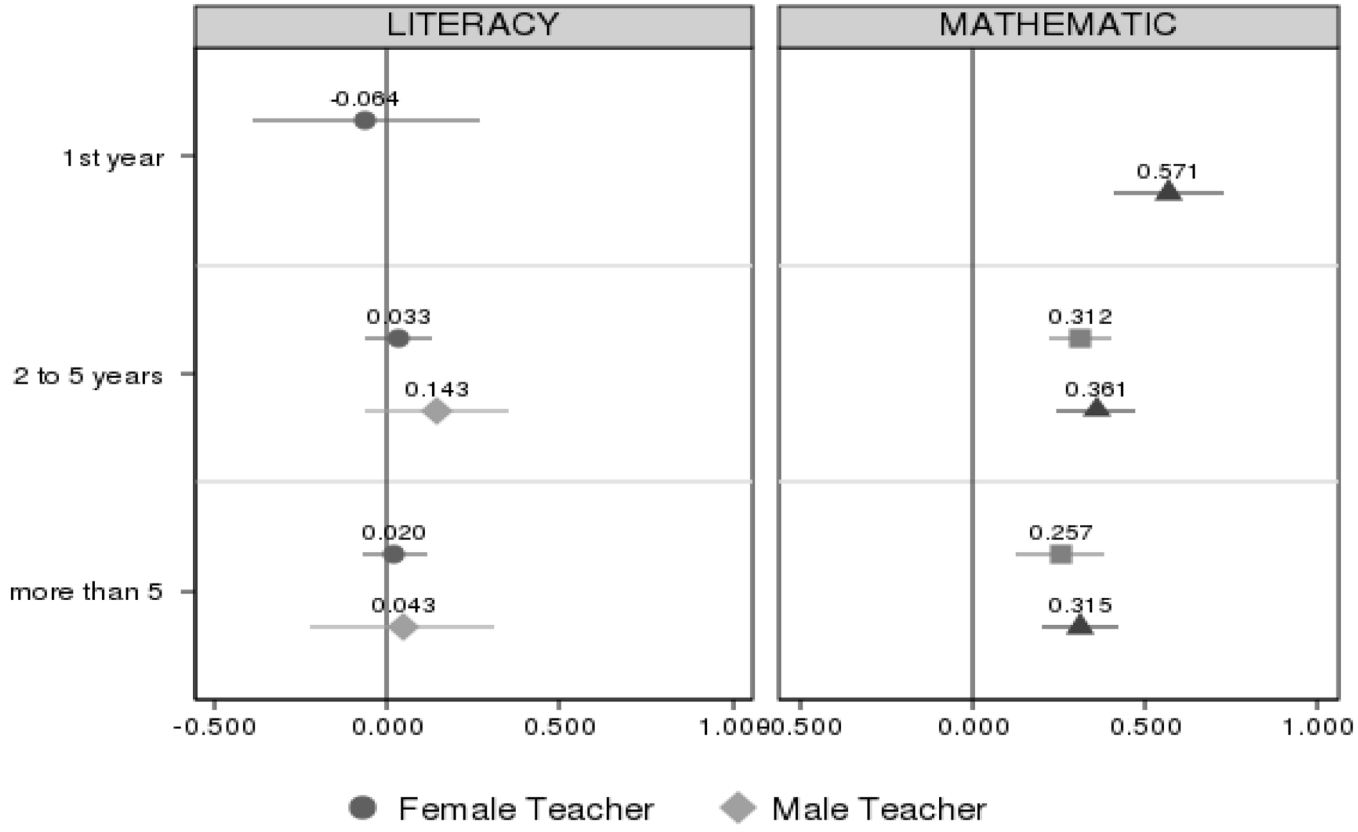


Fig. B.1. Gender bias measure by teachers' gender and years of experience.

Appendix C. Empirical Bayes estimates of teacher bias

A concern when estimating measures of teachers' gender biases involves estimation error arising from sampling variation. With small samples, a few students can have a large impact on test scores. In my sample, the average number of students per teacher is 36.3 in math and 31.8 in French. At the school level, Kane and Staiger (2002) found that among the smallest schools, more than half (56%) of the variance in score gain is due to sampling variation and other non-persistent factors. With sampling error, the estimated teacher bias t_j is the sum of the true teacher bias θ_j plus some error ϵ_j, where ϵ_j is uncorrelated with t_j. The variance of the estimated teacher biases has two components: the true variance of the teacher bias and the average sampling variance. Without accounting for it, the estimation error would lead to attenuation bias when I use the teacher bias measure in regressions as an explanatory variable for students' progress. To address this problem of sampling error, I construct empirical Bayes estimates of teacher gender bias. This approach was suggested by Kane and Staiger (2002) for measures of schools' accountability measures. The basic idea of the empirical Bayes approach is to multiply a noisy estimate of each teacher bias by an estimate of its reliability. Thus, less reliable estimates are shrunk back toward the mean (0, since the teacher estimates are normalized to be mean 0). Several recent applications have used this methodology to estimate teacher value added (Chetty et al., 2014; Jacob & Lefgren, 2005; Kane & Staiger, 2008). For each teacher, the reliability ratio of the noisy estimate of the gender bias is the ratio of signal variance to signal plus noise variance, where the noise corresponds to the squared standard-error of the bias estimate. It is relatively simple to estimate this ratio by using the observed estimation error from each teacher bias estimation. The first necessary step is to estimate the gender bias using Eq. (7). Following that regression, I save a gender bias coefficient for each teacher as well as the standard error of that coefficient. I compute a measure of the true variance $V(\theta)$ by subtracting the mean error variance (the average of the squared standard errors of the estimated teacher bias) from the variance of the observed bias: $V(\theta) = V(t) - E[V(\epsilon_j)]$.

$$RR_j = \frac{V(\theta)}{V(\theta) + V(\epsilon_j)} = \frac{V(t) - E[V(\epsilon_j)]}{V(t) - E[V(\epsilon_j)] + V(\epsilon_j)} \tag{14}$$

Finally, I construct an empirical Bayes estimator of each teacher's bias by multiplying the initial bias estimate by an estimate of its reliability: $t_j^{EB} = t_j \times RR_j$. After adjusting for estimation error, the standard deviation of teacher bias is 0.047 in math and 0.112 in French. Before the shrinkage, these SDs were 0.25 and 0.37, which shows that most of the variation between teachers in the degree of the estimated bias is due to sampling noise. I use the adjusted estimators for teachers' gender biases in all forthcoming regressions of students' progress on teachers' biases. Jacob and Lefgren (2005) showed that using the empirical Bayes estimates as an explanatory variable in a regression yields point estimates that are unaffected by the attenuation bias that would result from using standard OLS estimates.

Appendix D. Bootstrap method

I use a two-step bootstrapping method to compute the standard errors in all regressions that use the estimated gender bias. Two-step estimation methods yield inconsistent estimates of the standard errors in the second-stage regression because they fail to account for the presence of a generated regressor (Murphy & Topel., 1985; Pagan, 1984). This causes Naäve statistical inferences to be biased in favor of rejecting the null hypothesis. To deal with this concern, I use a two-step bootstrapping method to compute the standard errors (Ashraf & Galor, 2013). The bootstrap estimates of the standard errors are constructed in the following manner. First, for each teacher, I draw a random sample of pupils with replacement. The first stage regression is estimated on this random sample, and the corresponding OLS coefficient on teachers' gender bias are obtained. The second-stage regression—based on Eq. (6)—is then estimated on a random sample of classes with replacement, and the OLS coefficients are stored. This process of two-step bootstrap sampling and least-squares estimation is repeated 1000 times. The standard deviations in the sample of 1000 observations of coefficient estimates from the second-stage regression are thus the bootstrap standard errors of the point estimates of these coefficients.

Appendix E. Balance Check of Attrition

Three different outcomes are used to estimate the causal effect of teachers' gender biases on students: the blind score at the end of grade 9, the school attended during grade 10, and pupils' subject choices during grade 11. Two types of attrition exist: attrition at the class level, when scores are missing for all pupils in a class, and attrition at the individual level, when scores are missing for some pupils within a class. There is no attrition at the class level in my sample: all classes for which the bias is estimated at the end of grade 6 are observed in grades 10 and 11. The second type of attrition exists, but it would only be problematic if student attrition is correlated with teacher bias. To test for this, I check if the percentage of girls or boys missing in a class is correlated with the degree of bias of their teacher. I regress the percentage of girls missing (per class) on the gender bias. For each gender, I run six different regressions (corresponding to the six columns of Table E.1), where each of the potentially missing variables takes a turn as the dependent variable: blind score in French and math at the end of grade 9, information on school choice during grade 10, and information on course choice during grade 11. None of the coefficients are significant.

Table E.1
Balance check of attrition at the individual level.

	Grade 9		Grade 10		Grade 11	
	Math (1)	French (2)	Math (3)	French (4)	Math (5)	French (6)
Dep var: % girls missing	0.005	−0.004	−0.002	−0.001	−0.002	−0.001
	(0.011)	(0.009)	(0.011)	(0.009)	(0.011)	(0.009)
Dep var: % boys missing	0.001	−0.003	0.001	−0.008	0.001	−0.008
	(0.010)	(0.009)	(0.009)	(0.010)	(0.009)	(0.010)
Number of observations	177	172	177	172	177	172

†Notes: This table reports estimates from a class-level regression of the percentage of girls (resp boys) with a missing score on the gender bias. This is done for both boys and girls. In columns 1 and 2, the dependent variable is the percentage of girls (respectively boys) for whom the blind score is missing at the end of grade 9 (blind score missing in math in column 1 and French in column 2). In columns 3 and 4, the dependent variable is the percentage of girls (respectively boys) for whom the high school attended from grade 10 is missing. In columns 5 and 6, the dependent variable is the percentage of girls (respectively boys) for whom the course choice in grade 11 is missing. For standard errors, we use the White estimator of variance. Stars correspond to the following p-values: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Appendix F. Correlation between the skills measured by the blind and the non-blind scores

In mathematical terms, the assumption that both tests measure the same ability is equivalent to $\rho = 1$ and $v_i = 0$ in Eq. (12) from Section 4.5: $\theta_{iNB} = \rho\theta_{iB} + v_i$. I can test the validity of the hypothesis by directly estimating the reduced-form equation below and checking if the coefficient ρ is significantly different from 1. If not, both tests can be assumed to measure very similar abilities.

$NB_i = \alpha_0 + \rho B_i + \alpha_2 G_i + (\epsilon_{iNB} + v_i - \rho\epsilon_{iB})$

Since B_i is a noisy measure of ability θ_{1i}, it is correlated to the measurement error ϵ_{iB}. I solve this endogeneity issue by using two instrumental variables approaches. The first approach uses a dummy for being born in the second part of the year as an instrument (Angrist & Krueger, 1991; Bedard & Dhuey, 2006; Crawford, Dearden, & Meghir, 2007). The second approach uses the first-term blind scores as an instrument for the third-term blind scores.

When using pupils' month of birth as an instrument, I start by testing the correlation between last-term blind scores and pupils' months of birth by running a regression of blind scores in French and math on a set of 11 dummies for each month of birth. January is the reference month. Fig. F.1 presents the correlation coefficients.

There is clear evidence that pupils born at the end of the year have lower results than those born at the beginning of the year. To avoid including too many instrumental variables in the equation, I create a dummy variable for pupils born after September. Results of the first-stage regression are displayed in columns (1) and (2) of Table F.1. After controlling for covariates, being born at the end of the year has an important negative effect on blind scores—0.185 points of the SD in math and 0.151 in French. Columns (3) and (4) show that first-term blind score is also very correlated to third-term blind score, and provides large F statistics both in French and in math.

Being born at the end of the year will only be a valid instrument if the exclusion restriction holds. In other words, the only reason why being born at the end of the year affects teachers' grades should be because being born at the end of the year impacts a student's ability—measured by the blind

score—after controlling for other covariates. This restriction seems valid, provided that I control for pupils' behavior, parents' professions, and grade retention, three variables that might be correlated with being born at the end of the year.[31]

The second instrument relies on a different exclusion restriction, namely that the first term blind score has no effect on the grades given by teachers during the third term other than through its effect on the third term blind score. Although the plausibility of this assumption might be questioned, finding similar results with both instruments is reassuring.

I estimate Eq. (13) using both OLS and 2SLS. Results are presented in Table F.2 and discussed in the body of the paper.

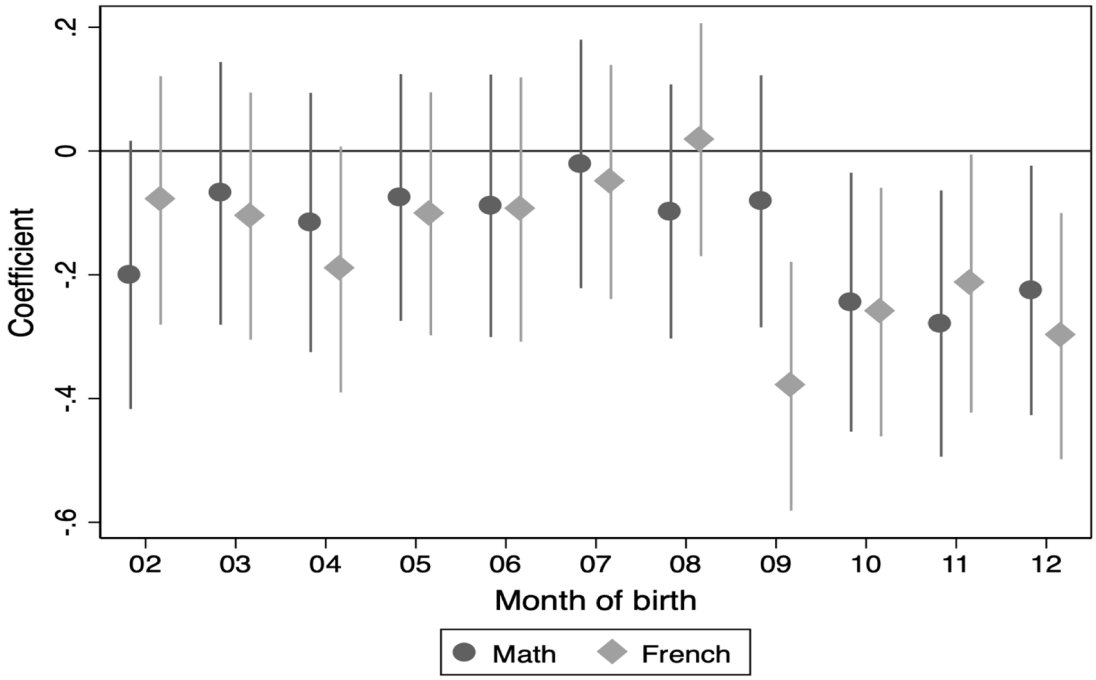


Fig. F.1. Correlation between pupils' month of birth and blind score.

Table F.1
First stage.

	Instrument: Born end of year		Instrument: First term blind score	
	Math (1)	French (2)	Math (3)	French (4)
Born End of Year	−0.185***	−0.151***		
	(0.045)	(0.044)		
First Term Blind Score			0.814***	0.650***
			(0.013)	(0.018)
Girl	−0.098*	0.291***	0.118***	0.144***
	(0.042)	(0.041)	(0.027)	(0.036)
Punishment	−0.672***	−0.584***	−0.221***	−0.069
	(0.069)	(0.069)	(0.061)	(0.071)
Grade repetition	−0.196*	−0.381***	−0.152	−0.257**
	(0.082)	(0.084)	(0.090)	(0.092)
High SES	0.424***	0.393***	0.050	0.116**
	(0.054)	(0.051)	(0.034)	(0.043)
Constant	0.197***	0.017	−0.027	−0.089**
	(0.037)	(0.037)	(0.022)	(0.030)
R2	0.086	0.113	0.660	0.454
Observations	2101	2133	1835	1803
F stat	15.05	8.38	6190.84	2748.50

Notes: The first two columns of this table report first stage estimates of the effect of being born at the end of the year (between September and December) on students' standardized blind scores. Columns (3) and (4) report first stage estimates of the effect of the first term blind score on last term blind score. The dependent variable is the blind score obtained by a pupil during the last term of grade 6. Standard errors are in parentheses and have been clustered at the school-level. Stars correspond to the following p-values: *$p < .05$; **$p < .01$; ***$p < .001$. All tests scores are standardized.

[31] Buckles and Hungerman (2012) showed that family background characteristics have strong relations with both season of birth and later educational outcomes. Similarly, Elder (2010) showed that children born in the latter part of the cohort are more likely to be diagnosed with ADHD. If these students are also more likely to cause disruptions in the classroom, controlling for behavior is important for the exclusion restriction to hold.

Table F.2
OLS and 2SLS estimates.

	Math			French		
	OLS	IV Birth date	IV Blind score	OLS	IV Birth date	IV Blind score
	(1)	(2)	(3)	(4)	(5)	(6)
Blind Score	0.703***	1.050***	0.864***	0.537***	1.009***	0.941***
	(0.033)	(0.076)	(0.029)	(0.026)	(0.104)	(0.044)
Girl	0.169***	0.206***	0.178***	0.204***	0.072	0.087*
	(0.036)	(0.041)	(0.036)	(0.035)	(0.054)	(0.037)
Constant	−1.062	−1.469	−1.218	−0.031	−0.221	−0.163
	(0.966)	(0.877)	(0.918)	(0.701)	(0.848)	(0.832)
R2	0.555	0.441	0.537	0.428	0.224	0.279
Observations	2101	2101	2033	2133	2133	2063
P-val (Blind = 1)		0.51	0.00		0.93	0.19

Notes: This table reports OLS and 2SLS estimates of the correlation between the non-blind and blind scores. The dependent variable is the third-term non-blind score. The unit of observation is a pupil. In columns (2) and (5), the instrument is a dummy variable for pupils born between September and December. In columns (3) and (6), the instrument is the first-term blind score. Control variables include grade retention, punishment, and high socio-economic status, and class fixed effects. Standard errors are in parentheses and have been clustered at the school-level. Stars correspond to the following p-values : $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.econedurev.2020.101981 .

References

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

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics, 106*(4), 979–1014. https://doi.org/10.2307/2937954.

Apperson, J., Bueno, C., & Sass, T. (2016). *Do the cheated ever prosper? The long-run effects of test-score manipulation by teachers on student outcomes* Working Paper No. 155

Ashraf, Q., & Galor, O. (2013). The 'out of Africa' hypothesis, human genetic diversity, and comparative economic development. *American Economic Review, 103*(1), 1–46.

Autor, D., Figlio, D., Karbownik, K., Roth, J., & Wasserman, M. (2016). *School quality and the gender gap in educational achievement. SEII Working Paper*.

Avvisati, F., Gurgand, M., Guyon, N., & Maurin, E. (2014). Getting parents involved : A field experiment in deprived schools. *Review of Economic Studies, 81*(1), 57–83.

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association, 14*(6), 1372–1400.

Bar, T., & Zussman, A. (2012). Partisan grading. *American Economic Journal: Applied Economics, 4*(1), 30–48.

Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics, 121*(4), 1437–1472.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American economic review. *The American Economic Review, 81*(5), 1041–1067.

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

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences, 113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113.

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. https://doi.org/10.1257/app.20140022.

Buckles, K. S., & Hungerman, D. M. (2012). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics, 95*(3), 711–724. https://doi.org/10.1162/REST_a_00314.

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

Carlana, M. (2018). *Implicit stereotypes: Evidence from teachers' gender bias* Harvard Kennedy School Working Paper No. 18-034

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project star. *The Quarterly Journal of Economics, 126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Cornwell, C., Mustard, D. B., & Parys, J. V. (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. https://doi.org/10.3368/jhr.48.1.236.

Crawford, C., Dearden, L., & Meghir, C. (2007). *When you are born matters: The impact of date of birth on child cognitive outcomes in England* CEE Discusion Paper

Dee, T., Dobbie, W., Jacob, B., & Rockoff, J. (2016). *The causes and consequences of test score manipulation: Evidence from the New York regents examinations* NBER Working Paper No. 22165

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

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *The Journal of Human Resources, 42*(3), 528–554. https://doi.org/10.2307/40057317.

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes tests* NBER Working Paper No. 22207

Elder, T. E. (2010). The importance of relative standards in ADHD diagnoses: Evidence based on exact birth dates. *Journal of Health Economics, 29*(5), 641–656.

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

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21*(1), 55–69.

Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy, 11*(1), 70–96.

French Ministry of Education (2005). *Les pratiques d'évaluation des enseignants en collège* Dossier 160

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90*(4), 715–741. https://doi.org/10.1257/aer.90.4.715.

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. https://doi.org/10.1016/j.econedurev.2011.02.007.

Hoff, K., & Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *The American Economic Review, 96*(2), 206–211. https://doi.org/10.2307/30034643.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* NBER Working Paper No. 11463

Jussim, L., & Eccles, J. (1992). Teachers expectations II: Contruction and reflection of student achievement. *Journal of Personality and Social Psychology, 63*, 947–961.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91–114.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* NBER Working Paper No. 14607

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009.

Lavy, V., & Megalokonomou, R. (2019). *Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study* NBER Working Paper No. 26021

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.

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review, 77*(3).

Lindahl, E. (2007). *Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden.* Uppsala University Working paper

Machin, S., & McNally, S. (2005). Gender and student achievement in english schools. *Oxford Review of Economic Policy, 21*(3), 357–372. https://doi.org/10.1093/oxrep/gri021.
Marsh, H. W., & Craven, R. G. (1997). Academic self-concept: Beyond the dustbowl. In G. D. Phye (Ed.). [*Handbook of classroom assessment*]. San Diego, CA: Academic Press.
Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *Review of Economic Studies, 76*(4), 1431–1459.
Murphy, K. M., & Topel., R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics, 3*(4), 370–379.
Murphy, R., & Weinhardt, F. (2013). *The importance of rank position* CEP Discussion Paper No. 1241
OECD (2012). *Education indicators in focus.* OECD Publishing.
OECD (2015). *The abc of gender equality in education: Aptitude, behaviour, confidence.* OECD Publishing.
Ouazad, A., & Page, L. (2013). Students' perceptions of teacher biases: Experimental economics in schools. *Journal of Public Economics, 105*, 116–130. https://doi.org/10.1016/j.jpubeco.2013.05.002.
Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *Inter Economic Review, 25*(1), 221–247.
Papageorge, N. W., Gershenson, S., & Kyungmin, K. (2016). *Teacher expectations matter* IZA Discussion Paper No. 10165
Pei, Z., Pischke, J. S., & Schwandt, H. (2019). Poorly Measured Confounders are More Useful on the Left than on the Right. *Journal of Business and Economic Statistics, 37*(2), 205–216.
Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302. https://doi.org/10.3102/0002831210372249.
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797.
Tiedemann, J. (2000). Gender related beliefs of teachers in elementary school mathematics. *Educational Studies in Mathematics, 41*, 191–207.
Trautwein, U., Ludtke, O., Marsh, H. W., Koller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98*(4), 788–806.