NBER WORKING PAPER SERIES

PERSISTENCY IN TEACHERS' GRADING BIAS AND EFFECTS ON LONGER-TERM OUTCOMES: UNIVERSITY ADMISSIONS EXAMS AND CHOICE OF FIELD OF STUDY

Victor Lavy
Rigissa Megalokonomou

Working Paper 26021
http://www.nber.org/papers/w26021

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

Victor Lavy acknowledges financial support from the European Research Council through ERC Advance Grant 323439 and from CAGE at the Department of Economics at the University of Warwick. Rigissa Megalokonomou acknowledges financial support from The University of Queensland BEL Early Career Grant No: UQECR1833757. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Persistency in Teachers’ Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study
Victor Lavy and Rigissa Megalokonomou
NBER Working Paper No. 26021
June 2019
JEL No. I24,J16,J24

ABSTRACT

Recent research has focused on what shapes gender differences in academic achievement and students’ choice of university field of study. In this paper we examine how teachers’ gender role attitudes and stereotypes influence the gender gap by affecting the school environment. We explore the extent to which teachers’ gender bias in high school influences students’ school attendance and academic performance in high-stakes university admission exams and students’ choice of university field of study. We use data from a large number of high schools in Greece, where the performance in these high-stakes exams determines university admission. We measure teachers’ bias as the difference between a high school student’s school exam score and national exam score. We then define a teacher bias measure at the class level by the difference between boys’ and girls’ average gap between the school score and the national score. We link teachers over time to obtain a persistent teacher bias measure based on multiple classes, and to estimate the effect for later cohorts’ performance. We find a very high correlation of within-teacher gender biases measured in different classes, which reveals high persistency in teachers’ gender favoritism behavior. We then find substantial effects of these teacher biases on students’ school attendance and performance in university admission exams, quality of enrolled degree and the given field of study at the university. We also find that gender biases are more prevalent among low value added teachers, while the more effective teachers have an approximately neutral gender attitude. This suggests that less effective teachers can harm their students twice, by being a bad teacher and by discriminating against one of the genders.

Victor Lavy
Department of Economics
University of Warwick
Coventry, CV4 7AL
United Kingdom
and Hebrew University of Jerusalem
and also NBER
v.lavy@warwick.ac.uk

Rigissa Megalokonomou
Department of Economics
University of Qeensland
r.megalokonomou@uq.edu.au

1 Introduction

A robust stylized fact established in recent years in many countries is that girls out-perform boys in school achievements in primary and secondary school. The gap is larger for school tests that are graded by school teachers and smaller for external exams that are graded 'blindly'. The gaps are smaller in STEM subjects, often still showing boys' advantage followed universally by higher college enrollment rates of men in these fields of study. For example, the National Center for Educational Statistics (NCES) 2015 report shows that 57 percent of all bachelor degrees conferred by post-secondary institutions in the U.S. in 2013-14 went to women while in STEM subjects the rate was much lower: 39 percent in physical sciences and science technologies, 18 percent in computer and information sciences, 18 percent in engineering and engineering technologies, and 10 percent in computer engineering.[1] This skewed pattern of gender differences in fields of study naturally determines gender occupational differences in the labor market. For example, only 14% of engineers in the US are women, though this rate is much higher than in the early 1980s, when only 5.8% of engineers in the U.S. were women.[2]

What shapes these gender differences in academic achievements and in university fields of study is the focus of much recent research.[3] In this paper we focus on how teachers' gender role attitudes and stereotypes influence the gender gap by affecting the school environment. We explore the extent to which teachers' gender bias in high school influences students' academic performance in high-stakes exams that determine admission to universities and on students' choice of university field of study. We use data from a large number of high schools in Greece, where performance in these exams is the sole determinant of university admission. Our sample includes female and male teachers, so the analysis reflects potential bias due to the teacher's own gender, allowing us to distinguish the gender bias by teachers gender. We measure teachers' bias as the difference between a student's school exam score in 11^{th} and 12^{th} grade (scored by the student's teacher) and his or her external exam score (taken at the end of 11^{th} and 12^{th} grade and scored nationally).[4] We then define a teachers' bias measure at the class level by the difference

[1]The female share of degrees conferred was 84 percent in health professions and related programs, 69 percent in English language and literature/letters, 58 percent in biological and biomedical sciences, and 43 percent in mathematics and statistics. https://nces.ed.gov/programs/digest/d15/tables/dt15_318.30.asp?current=yes.

[2]STEM Education: Preparing for the Jobs of the Future, A Report by the Joint Economic Committee Chairman's Staff Senator Bob Casey, Chairman April 2012.

[3]Some studies emphasize the role of biological gender differences in determining gender cognitive differences (Witelson 1976, Lansdell 1962, Waber 1976), while others emphasize the social, psychological and environmental factors that might influence this gap. There is limited credible evidence for this debate because it is difficult to disentangle the impact of biological gender dissimilarities from environmental conditions, and because it is difficult to measure stereotypes and prejudices and test their causal implications.

[4]The systematic difference between non-blind and blind assessment across groups as a measure of discrimination

between boys' and girls' average gap between the school score and the national score. Positive values indicate that a teacher is biased in favor of boys in a particular subject. We link teachers over time and are therefore able to get a persistent teacher bias measure based on multiple classes (on average 11 classes per teacher), and the effect is estimated for performance by students in later cohorts. Panel data on teachers relieve concerns that our measure of gender bias may only pick up random (small sample) variation in the unobserved "quality" or non-cognitive skills of the boys vs. girls in a particular single class or any other class-specific dynamics. We find that the same teachers who are biased for one class are biased in the same way for other classes in the same year and in earlier or later academic years. We also find that the correlation between same teacher's biases (when teaching two different subjects) is significantly higher than the correlation between different teachers' biases. This evidence of a "persistent" (average) teacher bias component (across multiple years/classes) reassures us that our bias measure is not picking up random variation in unobserved attributes in the mix of boys and girls.

We use data from high school teachers and students in Greece for the period 2003-2011. Our panel data of teachers include more than 400 teachers from 21 high schools over this entire period. Using this sample, we find that the bias measures derived from the teachers' panel data yield results very similar to those obtained when measuring bias based only on current own class. Perhaps this result is expected, given the higher correlation and persistence of teachers' bias measured in different classes. We also have a sample of an additional 114 schools for which we do not have panel data on teachers. Using this sample, we measure teachers' bias based on grading information from their current own classes and obtain estimates that are similar to the results obtained from the 21 schools sample.

We first estimate the effect of teachers' bias in 11^{th} grade on students' performance in the national exams at the end of 12^{th} grade. We measure the bias in each subject and then average these results over bundles of subjects, as follows: The first bundle includes core subjects that all students are required to study, including modern Greek, history, physics, algebra and geometry. The other three bundles are the three study tracks available to students in 11^{th} and 12^{th} grade: classics, science and exact science. In 11^{th} grade the subjects taught on the classics track include ancient Greek, philosophy and Latin; on the science track they are mathematics, physics, and chemistry, and on the exact science track they include mathematics, physics and computer science. In 12^{th} grade the subjects taught on the classics track are

or stereotypes was pioneered in economics by Blank (1991) and Goldin and Rouse (2000). This approach was first applied to the economics of education in Lavy (2008), to measure gender bias in grading by teachers and it was followed by others, for example, Tyrefors et al. (2011), Hanna and Linden (2012), Cornwell et al. (2013), Burgess and Greaves (2013), and Botelho et al. (2015), who implemented the same methodology using data from other countries and getting overall similar evidence about teachers' stereotypes/biases.

ancient Greek, Latin, literature and history; on the science track they are biology, mathematics, physics and chemistry, and on the exact science track they are mathematics, physics, business administration and computer science. We find that the teachers' biases in all four groups of subjects have a positive effect on boys' and a negative effect on girls' 12^{th} grade external exams scores. All estimates are precise and statistically significantly different from zero, except in the classics track. Based on bias measures derived from teachers' panel data, the effects size (in terms of standard deviation of the test score distribution) is 0.090 for boys and -0.100 for girls, in classics it is 0.185 and -0.051, in science it is 0.211 and -0.109 and in exact science 0.145 and -0.163. We correct for potential sampling error that could bias both our estimates and the standard errors by applying an empirical Bayes technique and a two-step bootstrapping procedure.

Our results are significantly different from the results of placebo exercises, which show no effects on students' later performance. These placebo exercises are based on a random reshuffling of teacher gender bias within each school and across teachers who teach the same subjects. We then re-estimate the effects of teachers' gender favoritism on students' performance. The only difference is that this time we use the bias of teachers who teach the same subjects within the school, but who are not actually teaching this particular class. To do this, we rely on the large number of teachers who teach the same subject within schools for the period 2003-2005. Using these placebo measures of teachers' gender bias, we find (almost) zero and statistically insignificant effects on students' subsequent performances for all groups of subjects. The lack of any discernible effects using the placebo measures of the variable of interest suggest that the estimated effects of the correct measure of treatment are not biased due to omitted unobservable confounders of the effect of interest.

The psychology and sociology literatures provide ample evidence about the potential mechanisms by which teacher's gender-stereotypical attitudes affect students' cognitive and non-cognitive outcomes. For example, teachers are said to treat the successes and failures of boys and girls differently, by encouraging boys to try harder and allowing girls to give up (Dweck et al. 1978 and Rebhorn and Miles 1999). Sadker and Sadker (1985) suggest that teachers give more attention to boys by addressing them more often in class, giving them more time to respond and providing more substantive feedback. Teachers are also found to treat boys and girls differently, in particular with regard to math instruction: Hyde and Jaffe (1998) show that math teachers tend to encourage boys to exert independence by not using algorithms and that boys who adopt this rebellious approach are seen as having a promising future in mathematics. Girls, on the other hand, are controlled more than boys, and are taught mathematics as a set of rules or computational methods. Leinhardt et al. (1979) find that teachers spent more time training girls in reading and less time in math relative to boys. In addition, according to the National Center of Education Statistics (1997) girls are less likely than boys to be advised, counseled and encouraged to take courses in math (Bae and Smith 1997).

In the second part of the paper, we estimate the effect of 11^{th} and 12^{th} grade teachers' bias on post-secondary enrollment status, quality of post-secondary program enrolled, and university enrollment by field of study. We find that teacher gender bias has a significant effect on students' post-secondary options and choices. First, we find that a one standard deviation (sd) increase in a teacher's bias in the core subjects in 11^{th} and 12^{th} grade increases the probability of boys enrolling in some post-secondary program by 2 and 4 percentage points, while it lowers the enrollment probability for girls by 5 and 3 percentage points. We find that teacher bias in all groups of subjects (core, classics, science and exact science) has a significant effect on students' probability of enrolling in post-secondary schooling.

Second, we examine the effect of teacher bias on the choice of university field of study. Girls have a higher enrollment rate in humanities, 34 percent versus 11 percent of boys. In social science and science, enrollment does not differ much by gender. In the exact sciences there are large enrollment disparities by gender: 12 percent of girls and 27 percent of boys. Based on multiple-choice regressions, we find that 11^{th} and 12^{th} grade teachers' bias has a negative and statistically significant effect on girls' choice of program of study. A one sd increase in teacher's bias in 11^{th} grade reduces the probability that a girl will choose that field of study in university by 4 percentage points. The effect of teacher's bias in 12^{th} grade reduces this probability by 5 percentage points. The corresponding estimated effects for boys are positive but small and imprecise, practically not different from zero. We also find significant effects of teacher bias on the quality of the program that students enroll in. The results are in line with all previous results: An increase in teacher bias (a bias in favor of boys) increases the probability of boys' enrolling into programs of higher quality, whereas the effect is opposite for girls. We measure the quality of the program using both the ranking based on the mean performance in the university admission exams of enrolled students in each program and annual program admission cutoffs.

In the third part of the paper, we examine whether there is any association between teachers' gender bias and teachers' quality. We use teachers' value added (TVA) as a measure of teachers' quality (Rockoff 2004, Chetty et al. 2014a, Chetty et al. 2014b). Relying on high persistent teachers' stereotypical behavior across classes, subjects, grades and years that we find in the first part of the paper, we estimate teachers' value added based on 2003-2005 schooling years. Gender bias is computed as the average gender bias a teacher exhibits in the 2006-2011 period. We find that teachers who have no gender bias have significantly higher value added relative to pro-boy or pro-girl teachers. This is the first study to link gender discrimination among teachers and their quality as teachers. There is ample evidence that show negative correlation between individuals' level education and prejudice-discriminatory behavior, for example anti-Black attitudes (Kuppens and Spears 2014, Wodtke 2016), or anti-women views (Sawhill 2014) but there is no evidence of this link for discrimination in the work place.

This paper makes three main contributions. First, it makes a substantive contribution to the literature on gender differences in STEM majors and careers by linking quantitative measures of teacher bias to students' later academic outcomes. To our knowledge, this is the first paper to establish a reliable causal link between high school culture and the prevalence of differential gendered outcomes. Two earlier papers examined the effect of teachers' bias in primary schools on students cognitive performance. Lavy and Sand (2018) analyze teachers' bias in primary schools in Tel Aviv, Israel, and estimate its effect on boys' and girls' scores in math, English and Hebrew in middle school and on high school tests and choices about the level of math and science courses. Terrier (2014) estimates the effect of teachers' bias similarly to Lavy and Sand (2018) and reports similar results, namely that teachers' grading bias in favor of boys in primary schools in France is positively correlated with boys' relative test score achievements. Carlana (2019) shows that teachers' stereotypes affect the gender gap in math, track choice, and self-confidence in own mathematical abilities for girls in middle school.[5] However, these earlier papers lack panel data on teachers, which is necessary to measure teachers' bias 'out of sample'. This drawback is eliminated in the current paper by using panel data on high school teachers in Greece. Using the full record of teaching for all teachers in our sample over several years, we document highly persistent stereotypical biases in teachers and the significant effects on short and longer term outcomes of high school students.

Second, in this paper we are able to examine the effect of teachers' gender favoritism on students' longer-term outcomes: We are able to link students' performance in high school with their university schooling. In particular, we examine whether the biases of those who teach 11^{th} and 12^{th} grade affect students' decisions to enroll in post-secondary institutions, the field of study chosen and the quality of their university–all of which have important implications for students' later occupational choices.

Third, we believe that this is the first paper in the literature on grading biases that relates the persistent pattern of teachers' discriminatory behavior to their quality, as measured by the value added. The important implication of our finding is that improving teacher quality would largely eliminate gender

[5]Field experiments have also been used to study discrimination (Bertrand and Duflo 2016).

stereotypes among teachers.

The rest of the paper is organized as follows. In Section 2, we present our institutional setting and data. Section 3 explains the identification and estimation methodologies. We detail our short run results in Section 4 and longer-term results in Section 5. Section 6 discusses the relationship between teacher gender bias and teacher value added and Section 7 offers conclusions and policy implications.

2 Context and Data

2.1 The Greek University Admissions System

The school education system in Greece is highly centralized (OECD 2018). University admission is also centralized and is administered by the Ministry of Education. The majority of universities in Greece are public, with no tuition fees and admission based solely on national high school exit exams. Most undergraduate degrees in Greek universities are designed to be completed in four years, with the exception of the Polytechnic University in Athens (the most prestigious of the universities that offer engineering) which has a five year undergraduate course. Students have to participate in standardized national tests to be eligible for university admission. All schools that administer these tests follow the same curriculum and offer courses in core and track subjects in accordance with the material covered in the national exams (OECD 2018). Until 2005, students took national exams in 11^{th} and 12^{th} grade. Since 2006 there has only been one national exam, at the end of 12^{th} grade.[6] University admission is based on the results of national exams and school exams that they take throughout the year. Until 2005 the average score used for university admission was also based on students' performance in national and school exams taken in the 11^{th} grade.

The data we use in this study include school and national exam test scores for all students. The Ministry of Education receives the national exams scripts and sends them to examiners across the country with the name and gender of the student concealed. Therefore, we consider national exam scores to be "blind" scores, as the external examiner does not know the name or gender of the student. In contrast, school exams are graded by the student's teacher and therefore are considered to be "non-blind scores", as the identity and gender of the student are obviously not concealed.[7] We use these non-blind test scores in each subject to examine teacher biases. Most students have a different teacher for each subject, but

[6]Since 2006, students also have taken national exams in fewer subjects than previously.

[7]The non-blind scores could be affected by a student's performance in previous class exams in the same term, if there is more than one class exam in class.

some might have the same teacher in two or more related subjects in the same grade. For the period 2003-2005 we observe the blind and non-blind scores of students in both grades (11^{th} and 12^{th}). From 2006 to 2011 the relevant data is the 12^{th} grade national and school test scores.

Even though every student has a first- and second-term school exam score in every subject, we prefer to use the latter for two reasons: (1) the second-term exam is more likely to cover the same material as is included in the end-of-year national exam, and (2) the second-term school exam and the national exam take place around the same time.[8] Schools typically administer the school exam at the very end of the school year, before the national exams. From the student's perspective, both exams are high-stakes: For the purpose of university admission, the final grade that counts towards university admission is a weighted average of the national exam (70%) and the school exam (30%). The blind and non-blind scores are also important for grade completion and high school graduation, as well as for decisions regarding dropout and grade repetition metrics. Both scores are reported in the high school graduation diploma, which is sometimes requested by potential employers.

The assignment of students and teachers to classes within each school is random. In particular, in accordance with a law that is strictly enforced, students are allocated to classes based on a lexicographical order. Using the same data as we use in our main analysis, we provide further evidence for this random assignment in contemporaneous work (Lavy and Megalokonomou 2019). Students are not allowed to switch class for any reason and must remain in their assigned class for all grades in high school. Within schools and across classes, there is no significant variation in student's observed characteristics and abilities.

University applicants submit a list of their preferred post-secondary programs[9] to the Ministry of Education (OECD 2018).[10] The admission score (cutoff) for each post-secondary program is not known when students submit their application and students can apply to several programs conditional on their high school track of study. At the beginning of the 11^{th} and 12^{th} grades, students enroll in one of three study tracks: classics, science, or exact science. All schools offer these three tracks and students must choose one in the 11^{th} grade; most remain in the same track in 12^{th} grade. Each track includes different subjects, but all students in a given track have to take the same subjects. It is conceivable that students' choice of track takes into consideration their aspirations for university field of study. Although students can apply to many programs from all high school tracks,[11] some programs assign a special weight to

[8]However, we note here that we obtain very similar results when we use the first term non-blind scores. These results are not reported in this paper and are available from the authors.

[9]By program we mean the university and field of study.

[10]See Goulas and Megalokonomou (2018) for more details about the admission algorithm.

[11]For example, students from all tracks can apply to economics degrees at the university level.

specific subjects when calculating the university admission grade.[12] In addition to the track subjects, students must take exams in compulsory core subjects. These core subjects are the same for all students, regardless of track.

2.2 Data

In this study we combine information about a large number of high schools in Greece from the schools and other administrative sources. The baseline sample is 11^{th} grade students in 2003-2005 in 21 schools and 12^{th} grade students in 2003-2011 in the same schools. In our main analysis we use information about teachers, students and principals from these 21 high schools. The information on teachers permits us to track individuals through their teaching history in the same high school during 2003-2011. Specifically, for each teacher we have information about the combination of classes, subjects and grades that he/she taught in each sample year. This allows us to construct panel data on teachers and school principals. Teachers' and principals' gender is inferred from their first names. We obtained student-level information about their high school performance from the schools' administrative records.

The student-level information includes identifiers for students and their classes (including class size in both grades), gender, year of birth, study track in high school, absenteeism records in 11^{th} and 12^{th} grade and test scores from the school and national exams in all subjects in 11^{th} and 12^{th} grade. The raw national exam score is scored on a 1-20 scale, which we transform it into z-scores for each year by type of exam and subject, to facilitate comparison over time and interpretation of our findings. The raw school exam is also scored on a 1-20 scale, which we transform it into z-scores for each year, school, type of exam and subject. The sample includes public, private and experimental[13] schools, in large and smaller cities and urban and rural areas. Then, teachers are matched to their classes and students by year, class and subject for the whole nine-year period. To complement our main analysis we also use a sample of another 114 schools for the same time period, for which we obtain only the student-level information.[14]

We then link the students' and teachers' datasets with administrative data from the Ministry of Education. We link our baseline sample and we access information about students' university enrollment status, university admission score,[15] and information about the post-secondary program (university and

[12]For example, for a student who applies to an engineering school, a special weight is given to his/her performance in mathematics in either the science or exact science track.

[13]These are public schools. Admission to these schools is based on a lottery for the years in this study. In 2013 the admissions process was; students now gain admission based on their performance in very competitive admissions exams

[14]For the sample of 135 schools, which is the total number of schools for which we have the student-level information, the sample contains 1,244 11^{th} grade classes and 3,787 12^{th} grade classes.

[15]This is the combination of students' scores in the school and national exams.

department) they enrolled in. We derive each post-secondary program's annual admissions cutoff in two ways. First, we calculate the mean performance of students who enroll in this program for each year; the Ministry's dataset provides us with information about the universe of students who apply to each post-secondary institution and program every year. Second, we can calculate the minimum score for the last-ranked enrolled student; this is the official program admissions cutoff or threshold the Ministry uses.

Table A1 presents descriptive statistics for the full sample of schools. The proportion of female students is 56 percent. The average GPA in 11^{th} and 12^{th} grade is 72 and 77, respectively. 92 percent of students attend public schools, 4 percent and 4 percent attend private and experimental schools, respectively, and 90 percent live in urban areas. Almost 82 percent of students eventually enroll in a university. Students apply on average to 25 programs[16] and on average they study their 8th most preferred program. The proportions enrolled in exact science, science, humanities and social science departments in 2003-2011 are 15 percent, 4 percent, 19 percent and 22 percent. Table 1 presents summary statistics for the full sample and also mean differences between the sample of 114 schools and the sample of 21 schools. The average number of classes in the full sample of 135 schools is 3.90, 3.90 in the sample of 114 schools and 3.92 in the sample of 21 schools. Average class size is 18-20 students in 11^{th} grade and 19-20 students in 12^{th} grade. 37 percent of the students study in the classics track, 28 percent in the science track, and 43 in the exact science track. The differences between the two samples[17] are small and for some variables they are not statistically significant.

3 Methodology and Estimation Framework

3.1 Measuring Teacher Gender Biases

The national exam scripts are collected by the Ministry of Education and are assigned for grading to teachers from other schools, with the name and gender of the student concealed. The student ID does not reveal any information about the student. Therefore, we denote the national exam scores as "blind". In contrast, school exams are graded by the class teacher, who knows the name, gender and other information about the student. Therefore, school exam scores are "non-blind". For each student we observe a set of both national (blind) and school (non-blind) exams scores. Most students have different teachers in each subject, although some may have the same teacher in two or more related subjects in the same year and grade. For the years 2003-2005 we have the data for the blind and non-blind scores for each of the 11^{th}

[16] This is equivalent to submitting 25 program applications.
[17] i.e., the sample of 114 schools and the sample of 21 schools.

and 12^{th} grade students. Starting from 2006, national exams were administered only in 12^{th} grade.[18]

Table 2 presents the means of the blind and the non-blind scores for boys and girls and the gender differences between these test scores in 11^{th} grade for 2003-2005. The gender gap varies by subject and type of exam. Boys outperform girls in physics, geometry and algebra (core subjects) in the blind exams. In all other subjects, girls outperform boys in the blind exams. Gender differences in the non-blind exams scores is always in favor of girls, that is, girls always outperform boys in exams that are graded by the teacher of the class. This girls' advantage is evident even in subjects in which boys outscore girls in externally marked blind exams. These systematic gender differences are interesting as they imply that the achievement gap is always in favor of girls in the non-blind exams– those graded by their own teacher.

Table 3 presents the same descriptive statistics for the 12^{th} grade students. The gender gaps in 12^{th} grade have a similar pattern as in the 11^{th} grade. Boys outperform girls in the blind exams in mathematics and physics (core subjects) and in mathematics, physics and chemistry (science track). The differences between boys' and girls' blind scores are statistically significant in most cases and they vary from 0.36 sd in favor of girls in Modern Greek to 0.16 sd in favor of boys in physics (science track). However, girls obtain higher scores in the non-blind exam in all subjects. The non-blind and blind gender differences vary from 0.47 in favor of girls in Modern Greek to 0.04 sd in science track mathematics in favor of girls. The positive achievement gaps in favor of boys in the blind exams in mathematics and physics are not present in the non-blind exams.

We construct the teacher bias measure in two steps based on each student's test scores in the blind and non-blind exam by subject. We first compute for each student in each exam the difference between her/his non-blind and blind exam scores. We then average these differences for boys and for girls separately for each class and then compute the difference of these two means for each class. That is, we define teacher bias TB of teacher j in class c as the difference between boys' and girls' average gap between the non-blind score (NB) and the blind score (B):

$$TB_{jc} = Mean_c[\sum_{ic}(NB_i - B_i|Male_i)] - Mean_c[\sum_{ic}(NB_i - B_i|Female_i)] \tag{1}$$

We repeat this procedure for every class, subject and grade. This measure takes negative or positive values depending on teachers' stereotypical behavior. Positive (negative) values indicate that a teacher is

[18] The combinations of subjects that students take national exams on also changes slightly. Students now take national exams in six subjects instead of nine. The school exams that students take in all nine subjects are not affected. Students can select the same optional subjects (for example, economics) as in the pre-2006 period in addition to the six compulsory subjects.

biased in favor of boys (girls) in this particular subject and class. Importantly and uniquely in this study, we use panel data for teachers by class, subject, and year, which allow us to compute the persistence of a teacher's stereotypical bias by averaging the bias measure over all of a teacher's classes during the study period. We do however want to exclude from this average the bias in the class for which we want to estimate the impact of the teacher bias, so we construct the average bias of a teacher based on all her/his other classes except the current class for which we are estimating the bias (excluding class i from the mean). In other words, we measure the bias in a particular classroom using the blind and non-blind scores for students in all other classes taught by the same teacher.

To demonstrate, assume we are interested in class c_1, taught by teacher j_1. We measure teacher's j_1 bias in all other classes that she has taught in the sample including classes in earlier years, other classes in the same year and other classes in later years. We use the bias for teacher j_1 that is measured in all these other classes to estimate her impact on class c_1. The intuition behind this is that this measure incorporates information about teacher j_1 with many different groups of students and captures any gender favoritism in different classes. Thus, this measure captures all possible information about the teacher and reflects more reliably her persistent gender biased behavior. Following this approach, of using 'out of sample' data to measure teachers' bias, we alleviate the concern that our teacher bias measure picks up class-level unobserved variation in boys' and girls' behavior or other gender-differential non-cognitive characteristics. The left figures in Figure 5 present the distribution of a teacher's bias measures, the first based on the current own class test scores and the second based on the teacher's other classes in both grades during the study period. The top panel shows the distributions of the two bias measures in 11^{th} grade and the bottom panel in 12^{th} grade. The right figures in Figure 5 present the distribution of a teacher's bias measures, the first based on the current own class test scores and the second based on the other classes a teacher taught only in the same grade and year.

In Figures 1 and 2, we present the histogram for these two measures of teachers' gender bias for 11^{th} and 12^{th} grade. We produce these distributions for core subjects and classics, science and exact science tracks. Different measures of teachers' gender bias are the teacher's bias in the own current class (following the formula in (1)) and the teacher's bias based on all other classes during the study period (average of different teacher bias measures calculated using (1)). We notice that these two distributions show a similar pattern, with both measures indicating considerable variation in teachers' discriminatory behavior. We also show that there is high correlation between the two measures of teacher bias for 11^{th} and 12^{th} grade, which can be seen in the scatter plots in Figures 3 and 4. The points show a high persistency in the two different measures of teachers' gender biases, and there is clustering around the 45-degree line.

In Table 4 we present descriptive statistics for the average number of times a teacher shows up in the sample of 21 high schools. There is one row for each teacher-class-grade-year combination. On average, a teacher appears 16 times in the sample, which means that he/she teaches 16 unique combinations of class, subject, grade, and year. We drop teachers who teach only one class in the whole sampling period, because we cannot construct a measure of teacher bias based on their behavior in other classes. Also, teachers teach on average almost 7 times in the 11^{th} grade over the period 2003-2005. This means that they teach 7 combinations of classes, subjects and years in the 11^{th} grade. On average a teacher teaches 3 combinations of classes and subjects per year, 1.5 different subjects per year and 1.7 different classes per year. There is little variation in these statistics from year to year in 2003-2005 and on average teachers are present in the 11^{th} grade sample in 2.2 years. A 12^{th} grade teacher appears in our sample on average 13 times over the period 2003-2011. She teaches 3.4 combinations of classes and subjects per year, with 1.6 different subjects per year and 1.8 classes per year. Twelfth-grade teachers appear on average in 4.4 years in the sample.

In Table 5 we present descriptive statistics for both measures of teacher bias by subject and grade using the sample of 21 high schools.[19] In column (1) we present the proportion of female teachers per subject and grade. We note a higher proportion of female teachers in humanities-oriented subjects (modern Greek, history, and other classics track subjects) and a higher proportion of male teachers in STEM-subjects (mathematics, physics, science and exact science track subjects). In columns (2)-(3) we present the mean and standard deviation of the measure of teacher bias based on all other classes by subject and grade, in the sample of 21 schools. These descriptive statistics show negative teacher bias throughout, suggesting that on average, teachers are biased in favor of girls across all subjects. The bias in 11^{th} grade is higher in exact science track, highest in computer science (-0.271) and lowest in physics (-0.037). Among 12^{th} grade teachers, the bias is highest in physics in the science track (-0.300) and lowest in economics (-0.085). Columns (4)-(5) show the mean and standard deviation of the measure of teacher bias calculated in own current class, by subject and grade in the sample of 21 schools. Again, teachers on average are pro-girls in all subjects, and the teacher bias is very similar across subjects to those reported in column (2). In columns (6) and (7) we report the difference between columns (2) and (4) and the

[19]To construct these teacher bias measures, we use the non-blind score of the second semester. In Table A2, we present the teacher bias calculated using both the first- and second- semester non-blind score. The pattern in both cases is the same: Teacher bias is on average negative. In most cases, the differences between these two measures is small. As we explain in the text, we use the second semester non-blind score to measure the teacher bias in the main analysis.

standard error. The differences are quite small and not statistically significant. In column (8) we report the simple row correlation between columns (2) and (4). These correlations are high enough to further support our observation that there is high persistency in teachers' gender biases.

In Table 6 we examine the relationship between the teacher's bias as measured in his current class and the bias as measured in all other classes that he taught during the study period. We condition the correlation between these two bias measures on subject, year, and school fixed effects. In particular, the structure of the data is such that there is one row per teacher-school-subject-year-grade and for each row there are two teacher bias measures. One is the bias measured in current class and the other is the bias of the same teacher measured in all other classes she/he ever taught during the study period. In columns 1-2 we present the 11^{th} grade estimates, and in columns 3-4 we show the 12^{th} grade estimates. The estimates in this table provide additional evidence that there is high correlation between the bias measured in current and in all other classes. This is evident in both grades. In the first panel, we present the results based on the full sample of teachers; in the second panel we present the results only for female teachers and in the third panel only for male teachers. For example, the correlation between the 11^{th} grade bias measure in other classes in any year and the bias measure in current own class in 11^{th} grade is 0.714 when estimated in a regression with only subject and year fixed effect. The pattern is clear, high persistency in teacher gender grading bias among both female and among male teachers, and the degree of persistency is very similar for the two genders.

Additional support for our claim that there is high persistency in teachers' discriminatory behavior is documented in Table 7, where we present estimated correlations between teachers' bias in pairs of subjects: once where both subjects are taught by the same teachers and once where they are taught by different teachers. If the teacher gender bias measure we suggest captures teachers' and not students' behavior, we would expect these correlations to be high when both subjects are taught by the same teacher and low when two different teachers teach the two subjects. In columns 1-2, we present the estimated correlation coefficient when the two subjects are taught by two teachers and in columns 3-4 when they are taught by the same teacher. The estimates in columns 1-2 are small–mostly close to zero–and not significant. In columns 3-4 the estimates are positive, large and statistically significant, revealing high correlation between teacher bias when the same teacher is teaching a pair of subjects. These correlations are estimated based not only on school and year but also on class fixed effects. Clearly this evidence indicates that the teacher gender bias we measure captures teachers' and not students' behavior. In Table A3 we present these correlations for various combinations of subjects and report the individual correlations between different pairs of subjects. The pattern is very similar when we look at each pair of subjects separately.

The estimated effects for the same-teachers case in Table 7 are almost identical to the estimated correlation coefficient obtained in Table 6. The estimated effects in columns 3-4 are 0.664 (SE=0.061) in 11^{th} grade and 0.581 (SE=0.064) in 12^{th} grade, while in Table 11 the related estimates were 0.671 (SE=0.050) in 11^{th} grade and 0.628 (SE=0.042) in 12^{th} grade. If this correlation is driven by some classroom-specific unobserved dynamics or other gender-specific unobserved skills or characteristics, then we would expect to find the same pattern in the case of different teachers. These estimated effects are presented in columns 1-2 and correspond to the case in which different teachers instruct students from the same class in two different subjects. The estimates are very close to zero, and are not statistically significant. These results provide further evidence that our teacher bias measures reflect teachers' favoritism behavior.

Using the sample of 114 schools, we are able to measure teacher biases based on their current classes and compare them to the estimates from the sample of 21 schools. These results are presented in online appendix Table A4 and show that the differences between these two sets of bias estimates are small and follow the same pattern across subjects. The combined evidence presented in Tables 5 and A3 shows that the estimates of teacher bias derived from the sample of 21 schools based on all other classes are very similar to the bias measures obtained from the sample of 114 schools based on the current teacher's class. In the next section, we report the results of estimating the effect of various measures of teachers' bias on students' short-term academic performance (i.e., subsequent national exam tests) and on longer-term outcomes–specifically, choice of university field of study and quality of the post-secondary program.

4 Effect of Teacher Biases on High School Outcomes

We estimate the following model to obtain the effect of teacher biases in 11^{th} grade on the performance of students on 12^{th} grade national exams. We note again that test scores on these exams are used by universities to admit students to study programs and therefore these are high stake exams.

$$Y_{icjt} = \alpha + \mu_c + \theta_j + \lambda_t + \gamma X_{icjt} + \pi TB_{cj} + \phi_{cj} + \psi_{icjt} \tag{2}$$

where Y_{icjt} denotes the outcome of student i, in high school or class c, subject j and year t; X_{icjt} is a vector of a student's prior score on the national exam in subject j; μ_c is a high school or class fixed effect; θ_j is a subject fixed effect; λ_t is a year fixed effect, and TB_{cj} is the measure of teachers' biased behavior in school (class) c and subject j. We also control for the gender of the teacher in the regressions. The error term in the equation includes a school- (or class-) and subject-specific random element ϕ_{cj} that allows for any type of correlation within observations of the same school across classes and an individual random element ψ_{icjt}. The coefficient of interest is π and it captures the effect of teacher bias on academic outcomes.

We use the following two correction techniques when estimating equation 2: (1) an empirical Bayes (EB) shrinkage estimation approach to address the sampling error that might result from the fact that some bias measures might be obtained from a small number of observations[20] and (2) a two-step bootstrapping technique to account for the fact that the main variable of interest is a generated regressor.[21] After adjusting the teacher bias measures for estimation error using the EB shrinkage approach, the standard deviations of teacher bias are 0.279 and 0.328 for the core and classics subjects, while before the shrinkage the standard deviations of teacher bias were 0.299 and 0.359. This confirms that only a small part of the variation between teachers is due to sampling noise. It has been shown that applying the empirical Bayes technique to the explanatory variable addresses attenuation bias that would result if we were using standard OLS standard errors (Jacob and Lefgren, 2005). The EB shrinkage adjustment and two-step bootstrapping are used in all of our regressions to estimate the effects of teacher bias on students' subsequent outcomes and choices.

In Table 8 we present the results of estimating equation (2) to obtain the effect of teacher bias in 11^{th}

[20]Our main measure of teacher bias is derived based on a teacher's grading behavior in 16 different classes (Table 4), which makes it highly unlikely for a small number of observations to have a large impact on the outcome variables. Nevertheless, we follow Kane and Staiger (2008) and construct the EB shrinkage factor for teachers in our sample. The EB shrinkage is the ratio of signal variance to signal variance plus noise variance. We assume that the measure of teacher bias includes an error component. Thus, we estimate teachers' effects on students' weighted difference between non-blind and blind scores (where the weights are the inverse proportion of each gender in class, defined positively for boys and negatively for girls). This allows us to distinguish between the signal variance (variance of teachers' effects) and noise variance of teacher i (variance of the residuals for teacher i). The EB estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the EB shrinkage factor. By using this approach the noisy measure of a teacher bias is multiplied by an estimate of its reliability, where the reliability of a noisy measure is the ratio of signal variance to signal variance plus noise variance. Thus, the less reliable estimates of teacher bias (those with a large variation in estimated residuals) are shrunk towards the mean of teacher estimates.

[21]This procedure is done in two steps. Two-step estimations obtain inconsistent standard errors in the second-stage regression, as they fail to account for the presence of a generated regressor (Pagan, 1984). We follow a two-step bootstrapping method to compute standard errors (Ashraf and Galor, 2013). Bootstrapped standard errors are constructed as follows: In the first stage, a random sample with replacement is drawn from each class by the gender of students. Then a new measure of teacher gender bias is calculated using this random sample of students. In the second stage, we estimate the effect of these new teacher gender bias on students' performance in 12^{th} grade national exams and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. Standard deviations in the sample of 1,000 observations of coefficient estimates from the second step are the bootstrap standard errors of the estimates of teacher biases. These standard errors are reported from now on in all tables.

grade as measured based on all classes taught by the teacher except the current class. From now on this is the treatment variable that we use in all regressions as the measure of teachers' biases, unless otherwise stated. The dependent variable in the regression is the blind score in a given subject in 12^{th} grade and we use the sample of 21 schools. We report estimates based on three regression specifications: The first includes subject and year fixed effects, the second includes school fixed effects and the third includes a class instead of a school fixed effect. Standard errors are adjusted for the two-step bootstrapping technique and are clustered at the class level. The estimated effects in all three specifications are positive in the boys' regressions and negative in the girls' regressions. We present estimates on students' performance in the core, classics, science and exact science subjects separately. In the core and exact science subjects, the estimated coefficients for boys and girls are almost identical, but with an opposite sign. A one sd increase in 11^{th} grade core subjects' teacher bias increases boys' test score in 12^{th} grade by 0.09 sd and reduces girls' test score by 0.10 sd. In classics, the effect for boys is 0.19 and smaller and not significant. In science, the effects are larger; a 0.21 increase among boys and a 0.11 decrease among girls. In exact science the effect is large and negative for girls at 0.16, while for boys it is 0.15. Appendix Table A9 is a mirror image of Table 8, but the teacher bias measure is derived from the teacher's current class boys and girls, blind and non-blind test scores. Based on the same sample of 21 schools, the point estimates in this table display the exact same pattern; a positive effect on boys and a negative effect on girls with similar point estimates. The similarity in the estimates presented in Table 9 and Table A8 reflect the high correlation between the two measures of teachers' biases, underscoring the persistence of teachers' stereotypical gender attitudes.

Summarizing the evidence presented in Tables 5-8, we find that teachers who persistently exhibit a pro-boy bias in 11^{th} grade have a positive effect on boys and a negative effect on girls' test score in the 12^{th} grade national exam. The absolute effect size is similar by gender and they are economically meaningful. These results suggest that the various bias measures are highly correlated and the high degree of persistence is similar in male and female teachers.

4.1 Treatment Effect Heterogeneity by Female and Male Teachers and School Principals

In Tables A5 and A6 in the online appendix we present the means and the standard deviations of teacher bias by subject for male and female teachers. Table A5 presents the differences when the bias is measured based on the current class, while Table A6 presents the same statistics when the bias is measured based on all other classes. Teachers in all subjects and both genders appear to be pro-girl, as the mean biases are negative. The difference in the average bias seems to be statistically significant in physics in 11^{th} grade, with male teachers demonstrating significantly more "pro-girl" behavior (-0.164 against -0.068 in Table A5 and -0.212 against -0.076 in Table A6) compared to female teachers. The same pattern persists in economics in 12^{th} grade, in which male teachers seem to have a significantly stronger negative bias compared to female teachers. Additionally, female teachers in mathematics in 12^{th} grade seem to be significantly more "pro-girl" under both measures of the teacher biases. The pattern by which teachers are even more pro-girl in mathematics and physics is intriguing, especially because in Tables 2 and 3 we notice that these are the subjects in which boys perform better than girls in the national exams.

How does teacher gender bias evolve over time for the most experienced teachers in our sample? Figure A7 shows the average teacher bias over time for all teachers and also for male and female teachers only.[22] We focus this analysis on teachers in the sample with 8 or 9 years of experience during the study period. This allows us to follow them for a meaningful period of time and observe their behavior in consecutive years. In this analysis we use average teacher bias in all other classes and subjects, for each year. In the top, middle, and bottom panels of Figure A7, we present the average teacher biases for all teachers and also only for male and female teachers. As expected, the average bias is negative (pro-girl behavior) and stable over time. Male and female teachers exhibit the same pattern.

In Table A7 we present estimates in which we allow the effect of teachers' biases to vary by teacher's gender. We add to equation (2) an interaction term between TB_{cj} and an indicator for female teacher and we also include in the equation the main effect for teacher's gender. The coefficient on the interaction term in the boys' regression is positive in all four groups of subjects (core, classics, science and exact science), but it is not significantly different from zero in all cases. This implies that female teachers' bias has a larger effect on boys than that of male teachers, but we do not have enough power to estimate this difference precisely. The effect of female teachers' bias on girls is smaller than the effect of male teachers' biases. The effect of having pro-boy teachers on girls remains negative and significant in core and exact science subjects and the effect of female teachers is also negative but smaller and not significantly different from the effect of male teachers. Female teachers' grading bias seems more harmful for girls than for boys, but again we cannot draw firm conclusions given the standard error of the estimates. Figure 6 presents the distribution of our main measure of teacher bias for male and female teachers separately.

School principals may act as role models for students, their gender perhaps mitigate or enhance the

[22] A teacher sample is used here.

effect of teachers' gender biases. In Table A8 we allow for teachers' bias effect to vary by the gender of the school principal. Overall, the effect of grading bias on boys is negative and larger in schools with a female principal, but this differential treatment effect is not statistically significant.

4.2 The Effect of Teachers Biased Grading on Students' School Attendance

In Table 9 we examine whether students' absenteeism is affected by teachers' stereotypical behavior toward boys and girls. We view this outcome as a mechanism for the effect of teachers' biases on test scores because it is related to the amount of instructional time students are exposed to in school. We use three measures of school attendance: total number, excused and unexcused hours of being absent from school. The first is the sum of excused and unexcused hours of absence, the second is excused hours of absence and the third is the unexcused hours of absence. Some absences are authorized by parents, often with a note signed by a doctor or parent for short-term illness. An unexcused absence signals the students' reluctance to attend school or a student's suspension.

In Table 9, columns 1-6 and 7-12 we present estimates of the effect of teachers' biases on students' 11^{th} and 12^{th} grade school absences–total, excused and unexcused. We use four different bias measures, each based on a different group of subjects: core, classics, science and exact science. In all specifications we include subject, year, and class fixed effects. The estimated effects on total and unexcused absences are positive in the boys' regressions and negative in the girls' regressions. This means that an increase in a teacher's bias in favor of boys increases boys' class attendance and reduces girls' class attendance. This effect is largest on unexcused absences. For example, a one sd increase in 11^{th} grade bias in core subjects decreases boys' unexcused absences in 11^{th} grade by 0.6 hours and increases girls' unexcused absences by 0.4 hours. Estimates of girls' unexcused absences are statistically significant for all groups of subjects, except classics. Respective patterns are similar in 12^{th} grade. Boys seem to attend class more often when their teacher exhibits pro-boy behavior, while girls seem to attend their class less often when their teacher exhibits pro-boy behavior. A one sd increase in 12^{th} grade science or exact science teacher's bias increases girls' unexcused absences in 12^{th} grade by approximately 1 hour. These results suggest that the effect of teachers' biases on cognitive performance of students in national exams are partly mediated through increasing or decreasing absenteeism from regular school days, making them miss classes and material covered during these days.[23]

[23] There is growing evidence of the effect of instructional time in school on students' test score in primary and high school standardized exams. See for example Lavy (2015) and Lavy (2019) and Rivkin and Schiman (2015).

4.3 Placebo Estimation

Table 10 presents our placebo treatment estimates based on randomly reshuffling teacher gender biases within each school and subject. We examine these effects in each group of subjects (core, classics, science and exact science) separately.

Placebo treatment estimates are very different from the true treatment effect estimates presented in Table 8. For example, placebo estimated effects on boys are -0.004, 0.026, -0.045 and 0.005–all almost zero and not statistically significant–while the respective estimates in Table 8 are 0.090, 0.185, 0.211 and 0.145. The same findings are obtained for girls. The lack of any discernible effects in the placebo regressions suggests that the estimates of the actual bias measures are not biased due to omitted unobservable confounders.

5 Effect of Teachers' Biases on University Enrollment and Choice of Field of Study

5.1 Effect of Teachers' Gender Biases on University Enrollment

We first examine whether teachers' gender bias affects students' probability of enrollment in any post-secondary institution. We again group subjects in core, classics, science and exact science and we estimate the effects of the 11^{th} and 12^{th} grades bias, separately. Table 11 presents estimates from a linear regression model in which the dependent variable is equal to one if a student enrolls in some post-secondary institution and zero otherwise. Given our large sample, the estimates from this model are not different from marginal effects obtained from a probit or logit regression model. Estimates for boys are presented in columns 1 and 3 and for girls in columns 2 and 4. All regressions include subject, year and class fixed effects. Teachers' bias in both grades increase the likelihood of boys' enrollment in any university and they have the opposite effect on girls.

For example, a one sd increase in 11^{th} grade teacher bias in core, classics, science and exact science, increases boys' likelihood of studying in a university by 2, 5, 2, and 2 percentage points. The effects of the same increase in 12^{th} grade is higher and equal to 4, 5, 4, and 3 percentage points. At the same time, the effects for girls are negative. If an 11^{th} grade teacher becomes one standard deviation more pro-boy in core, classics, science and exact science subjects, girls' likelihood of enrolling in any post-secondary program decreases by 5, 1, 3, and 3 percentage points. The effects are similar for females in 12^{th} grade: An increase in teacher bias by one sd decreases their likelihood of enrolling in some post-secondary program by 3, 4, 2, and 3 percentage points.

5.2 Effect of Teacher Bias on University Field of Study

Teachers' gender bias may affect university schooling through two channels. The first is by affecting test scores in exams that are used for admission to universities and various study programs. Higher test scores in these exams enable admission to highly-demanded universities and fields of study. Second, higher test scores in the national exams may increase students' self-confidence and motivation, which can increase their interest in higher education and more challenging and rewarding courses. In this section we will estimate the effect of teachers' bias on students' choice of field of study conditional on enrollment in a university, and on the quality of the university where a student is enrolled.

We group fields of study at the university according to the four study tracks, in the same way as we did for high school. The humanities include liberal arts, literature, psychology, journalism, philosophy, education, Greek language, history, foreign languages, home economics and law. Social sciences includes economics, statistics, business and management, accounting, political science and European studies. Exact science includes mathematics, engineering, physics and computer science. Science includes biology, chemistry, medicine, pharmacy, veterinary studies and dentistry. Table 12 shows that among boys, 3.7 percent enroll in science studies, 22.3 percent in exact science, 21.2 percent in social science and 8.8 percent in humanities. 18.1 percent do not enroll in any post-secondary schooling and 25.8 percent are enrolled in vocational schooling. Among girls, 4.6 percent are enrolled in science studies, 9.9 percent in exact science, 22.5 percent in social science and 27.5 percent in humanities. Of the rest, 18.4 percent of girls do not enroll in any post-secondary schooling and 17.1 enrolled in vocational schooling. Clearly, there are large gender differences in the proportion of enrolled students in exact sciences and humanities. Figure A3 presents the proportion of students enrolled in each field of university study by year and Figure A4 presents the proportion of enrolled boys and girls in each field of university study.[24] We therefore focus our analysis on the effect of high school teachers' biases on the choice of field of study conditional on attending university.

We model students' choices in a linear probability regression in which we stack the four possible choices as the dependent variable for each student against teachers' bias in each of the four areas of study. The dependent variable is a 0/1 indicator, assuming the value of 1 for the observed field of study and 0 for the other three possible choices. We estimate simple linear probability models since a probit or logit model

[24]In Figure A5 we also present the proportion of students enrolled in each STEM field of university study by year and in Figure A6 we present the proportion of enrolled boys and girls in each STEM field of university study.

will yield similar estimates, given that we use very large samples. We estimate three specifications: The benchmark includes year and major fixed effects and the national exam score in 11^{th} or 12^{th} grade, a second specification adds high school fixed effects and in the third we replace the latter with a high school class fixed effect. Standard errors are clustered at the class level.

The teacher bias we assign to each field of study is as follows. For exact science we average the biases in 11^{th} grade in algebra, geometry and physics and the 12^{th} grade biases in mathematics and physics. For science, we average the 11^{th} grade biases in algebra, geometry and physics and the 12^{th} grade bias in biology. For humanities, we the average the biases in 11^{th} and 12^{th} grade in history and Modern Greek. For social science, we average 11^{th} grade biases in Modern Greek and history and 12^{th} grade bias in economics. Figures A1 and A2 present these average bias measures by high school track (Figure A1) and the related annual core subjects bias in each field of university study (Figure A2).

In Table 13 we present the effect of 11^{th} and 12^{th} grade teachers' biases on the choice of university field of study. In the top panel we present evidence from a stacked sample of 11^{th} and 12^{th} grade classes in 2003-2005. These regressions include grade fixed effects in all specifications. In the bottom panel we present evidence based on the sample of 12^{th} grade classes for the whole sample period, 2003-2011.

The absolute size of the estimated effect of 11^{th} grade bias is small and not significant for boys, while for girls the estimates are more precisely measured and are significantly different from zero. The estimated effect on girls is -0.037 with year and school fixed effects and -0.046 when we add class fixed effects. The estimated effects for girls with school or class fixed effects suggest that a one sd increase in bias in favor of boys in a given field lowers the probability of girls' choosing that field of study by 4.6 percent. The respective estimate of the 12^{th} grade bias is similar. The estimated effect for boys is not significant, while for girls it is -0.042 when the preferred specification is used. In Table A10 in the online appendix we present estimates in which the bias is measured based on the current class in a given year. The same pattern emerges with a negative effect of 11^{th} and 12^{th} grade bias on girls and a positive though small and non-significant effect on boys. What is striking is that the effect sizes for girls are practically identical, regardless of how we calculate the bias, based on all other classes taught by a teacher or based only on his current class (-0.046 vs -0.040 for the 2003-2005 sample and -0.042 vs -0.048 for the 2003-2011 sample).

We also estimate a model in which we include in the set of choices the node of vocational education as possible post-secondary schooling. These results are presented in Table A11 in the online appendix. We assign to vocational departments the average bias in 11^{th} and 12^{th} grades in science and exact science. These results show a similar pattern to those reported in Table 13.

5.3 Effect of Teachers' Gender Biases on Quality of Program Enrolled In

We next present estimates of the effect of teachers' gender bias on the quality of post-secondary program students enroll in. We rank universities in each field of study, first based on the average score in national exams of students enrolled either across all years or based on 2003 (the first year in our data) and secondly based on the admissions cutoffs, which we determine using the marginal student enrolled in the program in the average year or in 2003. We then transform the ranking distribution to percentile rank.

In Table 14 we present estimates of the effect of teachers' gender bias on the rank of the students' university and department of study. We use in these regressions the quality of each post-secondary program calculated across all years in the sample. As a measure of teachers' gender bias we use the average bias of teachers in the core subjects that are closest to the university field of study. Estimates in columns 1-4 are from regressions in which we use the 11^{th} grade teachers' gender biases, while in columns 5-8 we use the 12^{th} grade teachers' bias. Overall, we find positive estimates for boys (columns 1 and 3) and negative estimates for girls (columns 2 and 4). For example, a one sd increase in 11^{th} grade pro-boys teachers' bias lowers for girls the rank of the department of study in humanities by 6 percentiles in the admissions cutoff distribution (column 2) or 7 percentiles in the mean score of the admitted students distribution (column 4). Twelfth-grade estimates follow the same pattern although they are smaller for girls but still statistically significant. Estimated effects for boys are positive but imprecise. They imply that a one standard deviation increase in pro-boy teachers' bias in the related subjects leads to lower enrollment of boys in exact science departments by 9 percentiles based on the first quality measure (column 1) or 10 percentiles when the second quality measure is used (column 3). Estimates for girls are negative but not precise. We find similar results when we combine the related fields of study, as shown in the bottom part of Table 14.

In Table 15 we report the effect of 11^{th} and 12^{th} grade teachers' gender bias on the percentile rank of each student's enrolled post-secondary program by field of study. The quality of post-secondary program is calculated only based on the 2003 cohort. The first row presents estimates when the bias variable is the average in core subjects, in the second row in classics track subjects, in the third row in science track subjects and in the fourth row in exact science track subjects. Focusing on the effect of the gender bias based on 11^{th} grade classes (Panel A), all estimates in columns 1 and 3 are positive and all estimates in columns 2 and 4 are negative. The effect in the boys' sample is statistically significant for bias in the core subjects and classics tracks. For girls the effect is statistically significant for bias in core subjects and exact science tracks. In panel B we present estimates when the bias is measured in 12^{th} grade classes and the patterns of signs and significance of estimates are similar to those in panel A.

The results presented in this section add another channel to the arsenal through which teachers' gender stereotypical behavior affect high school students. Since the quality of university schooling impact employment and earnings throughout adulthood, gender based biased grading of teachers clearly impose financial cost on their students. The estimated effects seem larger for boys, especially in science and exact science. Since these fields of study have higher predicted earnings, teachers' grading biases may contribute to the gender wage gap through this channel.

6 Does the Gender Bias Vary with Teacher's Quality?

The results reported above show that teachers' gender bias affects not only the subsequent performance of boys and girls during high school but also their post-secondary choices and decisions and the quality of their university schooling. In this section, we examine who are the teachers that discriminate by exploring the relationship between teacher gender biases and teacher quality. We use teachers' value-added (TVA) as a measure of their quality (Chetty et al., 2014a,b).

We construct TVA for teachers in the sample of 21 schools using the data for the 2003-2005 period. We use students' test score data in 10^{th}-12^{th} grades. We compute TVA using the mean performance of each teacher's class. The random assignment of students and teachers to classes in the Greek high school system guarantees that there is no selection and sorting.[25] However, we still control for the student demographics and lagged test score in the most closely related subject. Table A12 in the online appendix presents summary statistics for these controls which include gender, age, and an indicator for being born in the first quarter of a calendar year. We also control for whether a student expressed an interest in enrolling in given track (classics, science, or exact science). We pool students' test scores in 11^{th} and 12^{th} grades, and as a measure of prior test scores we use the 10^{th} and 11^{th} grade test scores.

Thus, TVA is measured in terms of standard deviations in the test score distribution and is estimated using data from all classes taught by a teacher in all years he appears in the data in our study period. Our sample includes only students who have non-missing values for the control variables we use in the baseline value-added model.[26] We show the distribution of TVA in 2003-2005 in the top panel of Figure 7.

[25]Each school's board decides the assignment of teachers to classes. Specifically, teachers are assigned to classes following a process that schedules their various classes across grades based on the subjects they teach (each teacher who teaches in high school has a specific teaching specialization and teaches specific subjects). According to the law, if there is any disagreement between members of the school board about teachers' assignment to classes, then members of the school authority and the school counselor are asked to attend the meeting and determine the assignment of teachers to classes.

[26]Our baseline VA model includes as controls students' demographics, cubic polynomials of lagged test scores in

We then restrict the sample so that VA is only weighted by the number of teachers (and not students) in the school-year-grade-subject-class-year cell, and we also keep only teachers whose gender bias we can measure, as we describe below.

The teacher gender bias measure we use in this section is the average teacher gender bias overall classes the teacher taught during 2006-2011. We restrict the analysis to this period in order to avoid an overlap between the period in which we measure TVA and the period we use to estimate the correlation between teachers' gender bias and TVA. This restriction is not a limitation at all because of the high persistency in teachers' biased behavior across classes and years as we have shown above. In the bottom panel of Figure 7 we show the distribution of teachers' gender bias using the data for 2006-2011 only.

We start the analysis by presenting descriptive statistics and comparisons of VA for teachers who are pro-boy and those who are pro-girl. We define as pro-boy teachers who have a bias greater than 0.10 and as pro-girl teachers whose gender bias is smaller than -0.10. We consider teachers with a gender bias between -0.10 and 0.10 to be neutral in terms of gender bias. Our sample includes 101 pro-boy teachers, 259 pro-girl teachers, and 58 neutral teachers. In Table 16 we present the means and standard deviations of TVA for these three groups. Column 3 shows the difference and standard errors of the pairwise differences. We note that neutral teachers have a TVA that is not strictly zero, but it is small and statistically not different from zero (equal to 0.053 with sd=0.132).

Pro-boys and pro-girls teachers have lower TVA, -0.037 (sd=0.222) and -0.049 (sd=0.235), respectively, while neutral teachers have high TVA, 0.053 (sd=0.132). These differences, presented in column 3 between neutral teachers and the other two groups, are statistically significant.

In Table 17 we present estimates of the standard coefficient of teachers' gender bias from a TVA regression model. We construct two bias variables as splines indicators. The first is a spline-variable for pro-boy teachers, and it assumes positive values for teacher gender bias, otherwise zero. The second spline is similarly constructed for pro-girl teachers. In particular, the spline-variable for pro-girls teachers assumes negative values for teacher gender bias, otherwise zero. We include these two spline variables in the same regression. All regressions include year, school and grade fixed effects. The estimate of the pro-boy bias variable is negative and statistically significant. Changing the specification, from column 1 to column 4 by adding gradually control variables (teacher's gender, class size and teacher's years of

the same subject, class size, school-level-grade enrollment, gender of the teacher, number of classes taught by the teacher throughout the years (a proxy for a teacher's experience), class and school-grade means of prior-year test scores, and neighborhood income. When a prior test score is missing, we set the prior score equal to 0 and include an indicator for missing data. In Table A12 we show summary statistics for the variables that we used to estimate the TVA models.

teaching experience), does not move the point estimate. This means that the higher the bias in favor of boys, the lower the teacher quality. Estimates of the pro-girl bias variable are symmetrical, indicating that a higher grading bias in favor of girls also leads to lower TVA.

Another way to examine the relationship between TVA and teachers' grading bias is by splitting the teacher bias measure into three ranges – pro-boy, pro-girl, and neutral – and constructing three dummy indicators. We construct these indicators as follows: an indicator for teachers with a bias larger than 0.10 (pro-boy), a second for teachers with a bias smaller than -0.10 (pro-girl), and a third for teachers with a bias between -0.10 and 0.10. These thresholds are somewhat arbitrary, and we examine below how sensitive are the results to varying the thresholds. Of course, only two of these 0/1 indicators can be included in the regression and we choose to omit the neutral group indicator. We present these results in Table 18. Estimated coefficients of the dummies for pro-girl and pro-boy teachers are both negative and statistically significant. Pro-girl and pro-boy teachers are associated with a reduction in TVA by 0.04 sd and by 0.03 sd, relative to neutral teachers. Gradually adding controls to the regression leaves the estimates almost unchanged. This provides further evidence that neutral teachers are on average of higher quality (higher TVA) compared to pro-boy and pro-girl teachers. These estimates are consistent with the findings we report in Tables 15-16.

In contemporaneous work (Lavy and Megalokonomou 2019), we find that there are not statistically significant differences between TVA by the gender of teachers. Males and female teachers seem to be of similar quality, on average, based on their TVA. Additionally, we find that science teachers are of higher quality than exact science and classics teachers. In particular, science teachers have on average positive TVA, while exact science and classics teachers have negative TVA on average. We also find that exact science teachers have lower quality than classics teachers and that more experienced teachers have a higher TVA on average.

The evidence in this section is the first we know that attempt to link gender discriminatory behavior of teachers to their quality. Our data does not include information on teachers' background so we cannot identify who are the teachers that discriminate except based on TVA. However, earlier literature on TVA or other measures of their quality show that these are not correlated with teachers' education, age, experience, and personal status (see for example (Rivkin et al., 2005)). There are however some descriptive studies that examined the correlation between prejudice or discrimination and education or cognition. For example, Kuppens and Spears (2014) find that education reduces explicit self-report measures of anti-Black attitudes, but it is much less related to implicit measures of anti-Black attitudes.

Higher educated people are more likely to be aversive racists, that is, to score low on explicit, but not implicit measures of prejudice. Wodtke (2016) finds that high-ability whites are less likely than low-ability whites to report prejudicial attitudes and more likely to support racial equality in principle, but they are not more likely to support a variety of remedial policies for racial inequality. There is less evidence on the correlation between education and gender based prejudice and discrimination. Sawhill (2014) argues that college-educated men have adapted reasonably well to the feminist revolution but it seems to have bypassed low-income men. This view implies negative correlation between education and gender based discriminatory behavior.

7 Conclusion

In this paper we investigate how teachers' gender biases affect students' school attendance and academic performance in high school, their probability of enrolling in a post-secondary program, the choice of university field of study and the university's national quality rank in the respective study area, in terms of the quality of its admitted students. The measure of teachers' gender-biased behavior we use is based on a comparison of boys' and girls' average class test scores in a non-blind exam that the teacher marks, versus a blind exam graded externally. We use panel data information on teachers' class assignment history throughout the period we study and measure a teacher's grading behavior in each class. We then use the teacher's average gender bias based on all classes except the one on which we are measuring the bias impact. This approach allows us to estimate the effect of the persistent component of teachers' biases, an endeavour that has not previously succeeded due to the lack of panel data. Based on observing teachers in 16 classes, on average, we find that the teachers who are biased in one class are biased in the same way in other classes in the same year and in classes in earlier or later academic years. The very high correlations of within-teacher bias in different classes reveal high persistency in teachers' favoritism or stereotypical behavior.

For identification, we rely on the random assignment of teachers and students to classes in a large number of high schools in Greece. We use novel data from a sample of high schools and compare students who are exposed to teachers who have different patterns of gender stereotypical biases. An important contribution of this paper is the use of gender stereotypical behavior 'out of sample' (that is, in other classes in previous and following years). This enables us to address several threats to the interpretation, and provide further support that our estimates reflect teachers' behavior and not random (small sample) variation in the unobserved quality or non-cognitive skills of the boys vs. girls in a particular class or any other class-specific dynamics. We also construct measures for a teacher's quality using the value added approach and exploiting the panel aspect of the data. An important contribution of this paper is that we investigate the association between teacher quality and teacher gender biases.

We can summarize our results with four broad conclusions. First, the same teachers who are biased for one class are biased in the same way for other classes in the same year and in classes in earlier or later academic years. The very high correlations of within-teacher biases in different classes reveal high persistent stereotypical behavior by some teachers. This finding suggests that the stereotypical biases are deeply rooted, a feature that should be taken into account in any planned remedial interventions. Second, an increase in teachers' bias (more pro-boy behavior) in core and track subjects (classics, social science, science, exact science) has a positive effect on boys' and a negative effect on girls' performance on the end of high school university admissions exams. Female teachers are more pro-girls on average, but the effects of female and male teachers' bias on national exams are not statistically different. Third, teachers' bias in core and track courses affect the likelihood that students will enroll in a post-secondary program and the quality of the program that they enroll in. Additionally, teacher bias has an effect on the related field of study at the university level. This average effect masks large heterogeneity by gender, being larger and statistically significant for girls and not different from zero for boys. Fourth, we find that the most effective teachers (measured by their VA) have neutral attitude towards the two genders, namely they do not exhibit any gender grading biases. This suggest that less effective teachers can harm their students twice, first by being an ineffective teacher and second by discriminating against one of the genders. If this relationship is causal, then it implies that training that improves teachers' quality will likely also reduce gender-based discrimination in schools.

References

Ashraf, Q. and O. Galor (2013). The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development. *American Economic Review 103*(1), 1–46.

Bae, Y. and T. Smith (1997). Women in Mathematics and Science. Findings from" The Condition of Education. *National Center for Education Statistics* (11).

Bertrand, M. and E. Duflo (2016). Field Experiments on Discrimination. Technical Report 22014, National Bureau of Economic Research.

Blank, R. (1991). The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review. *American Economic Review* (81), 1041–1067.

Botelho, F., R. A. Madeira, and M. A. Rangel (2015). Racial Discrimination in Grading: Evidence from Brazil. *American Economic Journal: Applied Economics 7*(4), 37–52.

Burgess, S. and E. Greaves (2013). Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. *Journal of Labor Economics 31*(3), 535–576.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics Forthcoming*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review 104*(9), 2633–79.

Cornwell, C., D. B. Mustard, and J. V. Parys (2013). Non-cognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School. *Journal of Human Resources 48*(1), 236–264.

Dweck, C., W. Davidson, S. Nelson, and B. Enna (1978). Sex Differences in Learned Helplessness: ii. The Contingencies of Evaluative Feedback in the Classroom and iii. An Experimental Analysis. *Developmental Psychology 14*, 268–276.

Goldin, C. and C. Rouse (2000). Orchestrating Impartiality: The Impact of "Blind "Auditions on Female Musicians. *American Economic Review* (90), 715–741.

Goulas, S. and R. Megalokonomou (2018). Which degrees do students prefer during recessions? *Empirical Economics 56*(6), 2093–2125.

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

Hyde, J. and S. Jaffe (1998). Prospectives from Social and Feminist Psychology. *Educational Researcher 27*(5), 14–16.

Jacob, B. A. and L. Lefgren (2005). Principals as Agents: Subjective Performance Measurement in Education. Working Paper 11463, National Bureau of Economic Research.

Kane, T. J. and D. O. Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working Paper 14607, National Bureau of Economic Research.

Kuppens, T. and R. Spears (2014). You don't have to be well-educated to be an aversive racist, but it helps. *Social Science Research 45*, 211 – 223.

Lansdell, H. (1962). A Sex Difference in the Effect of Temporal lobe Neurosurgery on Design Preference. *Nature 194*, 852–854.

Lavy, V. (2008). Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment. *Journal of Public Economics* (92), 2083–2105.

Lavy, V. (2015). Do Differences in School's Instruction Time Explain International Achievement Gaps in Math, Science, and Reading Evidence from Developed and Developing Countries. *Economic Journal 125*(588), F397–F424.

Lavy, V. (2019). Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behavior. *Journal of the European Economic Association.*

Lavy, V. and R. Megalokonomou (2019). Long-Term Effects of Teachers: Evidence from a Random Assignment of Teachers to Students. *Mimeo.*

Lavy, V. and E. Sand (2018). On The Origins of the Gender Human Capital Gap: Short and Long Term Effect of Teachers' Stereotypes. *Journal of Public Economics. 167*, 263–279.

Leinhardt, G., A. M. Seewald, and M. Engel (1979). Learning what's taught: Sex Differences in Instruction. *Journal of Educational Psychology 71*(4), 432–439.

OECD (2018). Education for a Bright Future in Greece. *Reviews of National Policies for Education, OECD. Publishing, Paris.*.

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review 25*(1), 221–247.

Rebhorn, L. S. and D. D. Miles (1999). High-Stakes Testing: Barrier to Gifted Girls in Mathematics and Science? *School Science and Mathematics 99*(6), 313–319.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrica 73*(2), 417–458.

Rivkin, S. G. and J. C. Schiman (2015). Instruction Time, Classroom Quality, and Academic Achievement. *125*(588), 425–448.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review 94*(2), 247–252.

Sadker, M. and D. Sadker (1985). Sexism in the Classroom. *Vocational Education Journal 60*(7), 30–32.

Sawhill, I. V. (2014). *Generation Unbound: Drifting into Sex and Parenthood without Marriage*. Brookings Institution Press.

Terrier, C. (2014). Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students' Achievement. *PSE Working Papers n. 2014-36.*.

Tyrefors, B., E. Hglin, and M. Johannesson (2011). Are Boys Discriminated in Swedish High Schools. *Economics of Education Review 30*(4), 682–690.

Waber, D. (1976). Sex Differences in Cognition: a Function of Maturation Rate? *Science 192*, 572–574.

Witelson, D. (1976). Sex and the Single Hemisphere: Specialization of the Right Hemisphere for Spatial Processing. *Science 193*, 425–427.

Wodtke, G. T. (2016, 01). Are Smart People Less Racist? Verbal Ability, Anti-Black Prejudice, and the Principle-Policy Paradox. *Social Problems 63*(1), 21–45.

Table 1: Descriptive Statistics by Study Sample

Variable	135 Schools Mean	(sd)	114 Schools Mean	(sd)	21 Schools Mean	(sd)	Difference diff	(s.e.)
I. 11^{th} grade								
Number of classes	3.900	(1.199)	3.897	(1.134)	3.923	(1.581)	-0.026	(0.014)
Class size	19.424	(5.050)	19.537	(5.073)	18.653	(4.818)	0.884	(0.059)
School cohort size	75.842	(27.524)	75.868	(26.257)	75.669	(34.945)	0.198	(0.320)
Proportion of students by track:								
Classics	0.366	(0.058)	0.365	(0.058)	0.375	(0.054)	-0.010	(0.001)
Science	0.280	(0.074)	0.282	(0.070)	0.265	(0.095)	0.017	(0.001)
Exact Science	0.344	(0.070)	0.342	(0.068)	0.360	(0.078)	-0.018	(0.001)
Proportion of female students	0.563	(0.496)	0.562	(0.496)	0.573	(0.495)	-0.011	(0.006)
No. of teacher-class-subject bias in year t in own class					1,200			
No. of teacher-class-subject bias in year t in other classes					1,051			
No. of teacher-class-subject bias in any year in all other classes					1,128			
II. 12^{th} grade								
Number of classes	3.866	(1.202)	3.868	(1.143)	3.854	(1.546)	0.014	(0.014)
Class size	19.589	(4.959)	19.675	(4.956)	19.006	(4.924)	0.669	(0.058)
School cohort size	75.853	(27.520)	75.880	(26.252)	75.667	(34.944)	0.213	(0.320)
Proportion of students by track:								
Classics	0.369	(0.060)	0.368	(0.060)	0.376	(0.056)	-0.008	(0.001)
Science	0.159	(0.050)	0.159	(0.050)	0.164	(0.056)	-0.005	(0.001)
Exact Science	0.463	(0.070)	0.463	(0.071)	0.460	(0.064)	0.003	(0.001)
Age	17.902	(0.465)	17.903	(0.451)	17.892	(0.552)	0.011	(0.006)
No. of teacher-class-subject bias in year t in own class					2,290			
No. of teacher-class-subject bias in year t in other classes					1,770			
No. of teacher-class-subject bias in any year in all other classes					2,148			

Notes: Teacher biases are calculated as the difference between boys' and girls' average gap between the non-blind score (NB) and the blind score (B) in the sample period. The *school cohort size* measures the number of students within a grade, school and year. There are three tracks available to students in 11^{th} and 12^{th} grade: classics, science and exact science. The baseline sample consists of 11^{th} grade students in 2003-2005 and 12^{th} grade students in 2003-2011.

Table 2: Mean Scores and Standard Deviations in the National Exam (Blind) and the School Exam (Non-Blind) in 11^{th} Grade, 2003-2005

	National Exam			School Exam		
Variables	Boys (sd) (1)	Girls (sd) (2)	**Difference** (se) (1)-(2)	Boys (sd) (3)	Girls (sd) (4)	**Difference** (se) (3)-(4)
I. Core subjects						
Modern Greek	-0.206	0.167	-0.372	-0.276	0.223	-0.500
	(0.997)	(0.971)	(0.013)	(1.003)	(0.925)	(0.013)
History	-0.116	0.094	-0.210	-0.175	0.142	-0.317
	(0.965)	(1.018)	(0.013)	(1.007)	(0.957)	(0.013)
Physics	0.057	-0.046	0.104	-0.050	0.040	-0.090
	(1.003)	(0.995)	(0.013)	(1.019)	(0.969)	(0.013)
Algebra	0.037	-0.030	0.067	-0.046	0.037	-0.083
	(1.011)	(0.990)	(0.013)	(1.014)	(0.973)	(0.013)
Geometry	0.046	-0.037	0.084	-0.041	0.033	-0.074
	(1.000)	(0.998)	(0.013)	(1.010)	(0.976)	(0.013)
II. Classics Track						
Ancient Greek	-0.233	0.061	-0.294	-0.330	0.087	-0.416
	(1.027)	(0.983)	(0.027)	(1.001)	(0.951)	(0.026)
Philosophy	-0.180	0.047	-0.227	-0.291	0.076	-0.368
	(0.984)	(0.999)	(0.027)	(1.047)	(0.945)	(0.026)
Latin	-0.241	0.063	-0.304	-0.359	0.094	-0.454
	(1.011)	(0.987)	(0.027)	(1.051)	(0.936)	(0.026)
II. Science Track						
Mathematics	0.011	-0.010	0.021	-0.043	0.039	-0.081
	(1.014)	(0.987)	(0.024)	(1.005)	(0.945)	(0.024)
Physics	0.030	-0.027	0.056	-0.023	0.021	-0.043
	(0.996)	(1.003)	(0.024)	(0.996)	(0.955)	(0.024)
Chemistry	-0.002	0.002	-0.004	-0.051	0.046	-0.097
	(1.000)	(1.000)	(0.024)	(0.991)	(0.959)	(0.024)
III. Exact Science Track						
Mathematics	-0.052	0.096	-0.148	-0.107	0.198	-0.304
	(0.992)	(1.007)	(0.022)	(0.980)	(0.950)	(0.022)
Physics	-0.035	0.065	-0.101	-0.088	0.162	-0.250
	(0.997)	(1.002)	(0.022)	(0.990)	(0.941)	(0.022)
Technology and Computers	0.025	-0.046	0.070	-0.095	0.177	-0.272
	(0.971)	(1.051)	(0.022)	(0.984)	(0.948)	(0.022)

Notes: There are three tracks available to students in 11^{th} grade: classics, science and exact science. In 11^{th} grade the subjects taught in the classics track are ancient Greek, philosophy and Latin; in the science track: mathematics, physics, chemistry, and in the exact science track: mathematics, physics and computer science. The national and school exam scores are standardized z-scores. This table presents test scores gender gaps by type of exam (blind and non-blind) and subject in 11^{th} grade. A positive difference means that boys outperform girls, while a negative difference means that girls outperform boys. The non-blind score in each subject is the score in the second term school exam. The estimation is based on the sample of 21 schools.

Table 3: Mean Scores and Standard Deviations in the National Exam (Blind) and the School Exam (Non-Blind) in 12^{th} Grade, 2003-2011

Variables	National Exam Boys (sd) (1)	National Exam Girls (sd) (2)	National Exam **Difference** (se) (1)-(2)	School Exam Boys (sd) (3)	School Exam Girls (sd) (4)	School Exam **Difference** (se) (3)-(4)
I. Core subjects						
Modern Greek	-0.205	0.159	-0.364	-0.263	0.204	-0.467
	(1.002)	(0.969)	(0.008)	(1.034)	(0.906)	(0.008)
History	-0.065	0.052	-0.117	-0.166	0.134	-0.300
	(0.984)	(1.009)	(0.013)	(1.027)	(0.940)	(0.013)
Mathematics	0.035	-0.035	0.070	-0.054	0.055	-0.109
	(1.002)	(0.997)	(0.009)	(1.040)	(0.930)	(0.009)
Physics	0.086	-0.071	0.156	-0.057	0.046	-0.103
	(1.001)	(0.993)	(0.013)	(1.032)	(0.955)	(0.013)
Biology	-0.040	0.024	-0.064	-0.166	0.101	-0.266
	(0.992)	(1.004)	(0.010)	(1.089)	(0.906)	(0.010)
II. Classics Track						
Ancient Greek	-0.223	0.060	-0.283	-0.297	0.080	-0.376
	(1.037)	(0.981)	(0.016)	(1.047)	(0.934)	(0.015)
Latin	-0.242	0.065	-0.307	-0.320	0.086	-0.406
	(1.079)	(0.977)	(0.016)	(1.047)	(0.930)	(0.015)
Modern Literature	-0.240	0.064	-0.305	-0.350	0.094	-0.442
	(1.046)	(0.977)	(0.016)	(1.089)	(0.923)	(0.015)
History	-0.064	0.017	-0.081	-0.189	0.051	-0.240
	(0.990)	(1.002)	(0.016)	(1.053)	(0.950)	(0.015)
III. Science Track						
Biology	-0.002	0.002	-0.004	-0.035	0.026	-0.061
	(1.013)	(0.990)	(0.020)	(1.003)	(0.922)	(0.022)
Mathematics	0.125	-0.087	0.211	0.022	-0.016	0.038
	(1.008)	(0.985)	(0.020)	(0.973)	(0.942)	(0.020)

Continued on next page

	National Exam			School Exam		
Variables	Boys (sd) (1)	Girls (sd) (2)	**Difference** (se) (1)-(2)	Boys (sd) (3)	Girls (sd) (4)	**Difference** (se) (3)-(4)
Physics	0.162	-0.113	0.275	0.022	-0.015	0.037
	(0.986)	(0.993)	(0.020)	(0.968)	(0.947)	(0.020)
Chemistry	0.082	-0.057	0.139	0.003	-0.002	0.006
	(0.978)	(1.010)	(0.020)	(0.972)	(0.947)	(0.021)
IV. Exact Science Track						
Mathematics	-0.026	0.044	-0.071	-0.083	0.139	-0.220
	(1.016)	(0.970)	(0.012)	(1.012)	(0.914)	(0.012)
Physics	0.012	-0.020	0.032	-0.076	0.126	-0.202
	(1.021)	(0.963)	(0.012)	(1.012)	(0.916)	(0.012)
Business Administration	-0.066	0.110	-0.176	-0.142	0.238	-0.380
	(0.996)	(0.996)	(0.012)	(1.028)	(0.850)	(0.012)
Computer Science	0.005	-0.009	0.014	-0.075	0.124	-0.199
	(1.012)	(0.979)	(0.012)	(1.021)	(0.900)	(0.012)
V. Optional						
Economics	-0.023	0.022	-0.046	-0.100	0.096	-0.196
	(0.995)	(1.004)	(0.011)	(1.024)	(0.932)	(0.011)

Notes: There are three tracks available to students in 12^{th} grade: classics, science and exact science. In 12^{th} grade the subjects taught in the classics track are ancient Greek, Latin, literature and history; in the science track: biology, mathematics, physics and chemistry, and in the exact science track: mathematics, physics, business administration and application development. The national and school exam scores are standardized z-scores. This table presents test scores' gender gaps by type of exam (blind and non-blind) and subject in 12^{th} grade. A positive difference means that boys outperform girls, while a negative difference means that girls outperform boys. The non-blind score in each subject is the score in the second term school exam. The estimation is based on the sample of 21 schools.

Table 4: Descriptive Statistics for 11^{th} and 12^{th} Grade Teachers

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Number of classes/subjects/years/grades taught by teacher	15.51	12.92	1	74
11^{th} grade				
Number of classes/subjects/year combinations taught by teacher	6.55	4.54	1	23
Number of classes/subjects taught by teacher by year	3.06	1.83	1	10
2003	3.11	1.89	1	10
2004	2.95	1.76	1	8
2005	3.16	1.85	1	9
Number of different subjects taught by teacher by year	1.48	0.54	1	3
2003	1.51	0.57	1	3
2004	1.50	0.55	1	3
2005	1.44	0.50	1	2
Number of different classes taught by teacher by year	1.70	0.90	1	4
2003	1.72	0.86	1	4
2004	1.57	0.87	1	4
2005	1.83	0.95	1	4
Number of years a teacher teaches	2.15	0.81	1	4
12^{th} grade				
Number of classes/subjects/year combinations taught by teacher	13.25	12.69	1	70
Number of classes/subjects taught by teacher by year	3.35	2.56	1	20
2003	3.50	2.04	1	10
2004	3.84	2.37	1	11
2005	3.79	2.46	1	10
2006	4.35	4.94	1	20
2007	2.73	1.84	1	9
2008	2.63	1.73	1	7
2009	2.64	1.76	1	7
2010	2.65	1.61	1	6
2011	2.83	1.91	1	8
Number of different subjects taught by teacher by year	1.63	0.86	1	6
2003	1.69	0.75	1	4
2004	1.84	0.95	1	6
2005	1.68	0.76	1	5
2006	2.05	1.47	1	6
2007	1.34	0.512	1	3
2008	1.37	0.52	1	3
2009	1.47	0.65	1	3
2010	1.43	0.66	1	3
2011	1.43	0.70	1	4
Number of different classes taught by teacher by year	1.78	1.14	1	7
2003	1.83	1.11	1	6
2004	1.95	1.23	1	7
2005	2.09	1.39	1	7
2006	1.85	1.21	1	6
2007	1.55	0.79	1	4
2008	1.56	0.92	1	5
2009	1.49	0.83	1	5
2010	1.51	0.88	1	5
2011	1.61	1.05	1	5
Number of years a teacher teaches	4.35	2.32	1	9

Notes: The estimation is based on the sample of 21 schools. The sample includes all teachers who teach core or track subjects in 11^{th} and 12^{th} grade. The 11^{th} grade sample is from 2003-2005, while the 12^{th} grade sample is from 2003-2011.

Table 5: Descriptive Statistics for Different Measures of Teacher Bias in 11^{th} and 12^{th} Grade

Variable Bias in	**Prop. of Fem. Teachers** (1)	Teacher Bias measured in other classes (21 schools) **Mean** (2)	**(sd)** (3)	Teacher Bias measured in the own class (21 schools) **Mean** (4)	**(sd)** (5)	**Diff.** (6)	**se** (7)	**Correlation between** (2) and (4) (8)
11th grade (2003-2005)								
Core subjects:								
Modern Greek	0.70	-0.122	(0.376)	-0.130	(0.496)	0.008	(0.034)	0.57
History	0.67	-0.150	(0.365)	-0.192	(0.466)	0.042	(0.030)	0.60
Physics	0.45	-0.157	(0.277)	-0.120	(0.372)	-0.037	(0.026)	0.53
Algebra	0.39	-0.105	(0.281)	-0.103	(0.343)	-0.002	(0.025)	0.52
Geometry	0.37	-0.124	(0.285)	-0.092	(0.348)	-0.033	(0.025)	0.51
Classics Track:								
Ancient Greek	0.62	-0.250	(0.433)	-0.331	(0.455)	0.081	(0.058)	0.61
Philosophy	0.69	-0.165	(0.392)	-0.141	(0.526)	-0.024	(0.061)	0.62
Latin	0.66	-0.153	(0.380)	-0.209	(0.475)	0.056	(0.072)	0.37
Science Track:								
Mathematics	0.49	-0.073	(0.251)	-0.116	(0.470)	0.043	(0.065)	0.59
Physics	0.40	-0.037	(0.276)	-0.018	(0.468)	-0.019	(0.060)	0.63
Chemistry	0.40	-0.108	(0.271)	-0.040	(0.450)	-0.068	(0.088)	0.12
Exact Science Track:								
Mathematics	0.30	-0.076	(0.274)	-0.102	(0.427)	0.026	(0.045)	0.73
Physics	0.39	-0.147	(0.297)	-0.080	(0.379)	-0.067	(0.051)	0.56
Technology and Computers	0.30	-0.271	(0.336)	-0.226	(0.653)	-0.045	(0.101)	0.33
12^{th} grade (2003-2011)								
Core subjects:								
Modern Greek	0.61	-0.145	(0.329)	-0.117	(0.533)	-0.028	(0.025)	0.59
History	0.55	-0.224	(0.366)	-0.189	(0.415)	-0.035	(0.029)	0.64
Biology	0.14	-0.153	(0.506)	-0.165	(0.772)	0.012	(0.048)	0.42
Mathematics	0.25	-0.154	(0.342)	-0.161	(0.547)	0.006	(0.030)	0.49
Physics	0.29	-0.175	(0.313)	-0.206	(0.361)	0.032	(0.025)	0.64
Classics Track:								
Ancient Greek	0.47	-0.171	(0.422)	-0.196	(0.564)	0.025	(0.059)	0.35
Latin	0.68	-0.194	(0.432)	-0.156	(0.526)	-0.038	(0.049)	0.56
Literature	0.54	-0.208	(0.397)	-0.195	(0.645)	-0.013	(0.055)	0.60
History	0.61	-0.213	(0.367)	-0.291	(0.562)	0.078	(0.049)	0.59
Science Track:								
Biology	0.13	-0.210	(0.411)	-0.105	(0.469)	-0.105	(0.066)	0.59
Mathematics	0.09	-0.253	(0.386)	-0.286	(0.525)	0.033	(0.073)	0.41
Physics	0.16	-0.300	(0.254)	-0.284	(0.392)	-0.016	(0.060)	0.38
Chemistry	0.06	-0.200	(0.413)	-0.135	(0.455)	-0.065	(0.092)	0.23
Exact Science Track:								
Mathematics	0.26	-0.156	(0.368)	-0.174	(0.478)	0.018	(0.045)	0.44
Physics	0.19	-0.174	(0.308)	-0.234	(0.476)	0.060	(0.047)	0.35
Business Administration	0.63	-0.147	(0.487)	-0.205	(0.581)	0.058	(0.058)	0.44
Computers	0.35	-0.189	(0.490)	-0.215	(0.475)	0.026	(0.064)	0.17
Optional								
Economics	0.59	-0.085	(0.394)	-0.044	(0.773)	-0.041	(0.047)	0.29

Notes: Negative (positive) bias means that the teacher is pro-girl (boy). The means are weighted by the number of students. The estimation is based on the sample of 21 schools.

Table 6: Correlations Between Different Measures of Teacher Gender Bias

Dependent Variable: Teacher Bias in the Current Class				
	11^{th} grade		12^{th} grade	
	(1)	(2)	(3)	(4)
	Panel A: All Teachers			
Average Bias measured in other classes in any year	0.714	0.671	0.639	0.628
	(0.052)***	(0.050)***	(0.044)***	(0.042)***
Sample Size	1,157	1,157	2,222	2,222
	Panel B: Female Teachers			
Average Bias measured in other classes in any year	0.668	0.652	0.601	0.514
	(0.077)***	(0.079)***	(0.066)***	(0.072)***
Sample Size	576	576	983	983
	Panel C: Male Teachers			
Average Bias measured in other classes in any year	0.764	0.639	0.657	0.618
	(0.061)***	(0.067)***	(0.054)***	(0.055)***
Sample Size	581	581	1,239	1,239
Subjects FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
School FE		✓		✓

Notes: The sample includes all teachers who teach core and track subjects. The estimation is based on the sample of 21 schools. The structure of the data is such that there is one row per teacher-school-subject-year-grade and for each row there are two measures of teacher bias calculated: The bias in the own class-subject-year and the bias in all other classes-subjects-years this teacher taught in the sample. The table reports the estimates of teachers' biases in the current year and in the own class on teachers' biases in any other class and year they taught in the sample. The dependent variable is the teacher *bias measured in the own class* and is calculated as the difference between boys' and girls' average gap between the non-blind score (NB) and the blind score (B) in the sample period. The variable of interest is the teacher *bias measured in all other classes in any year* and is calculated as the difference of boys' and girls' average gap between the non-blind score (NB) and the blind score (B) in all other classes that a teacher taught. All regressions include subjects and year fixed effects, while we add school fixed effects in columns (2) and (4). In Panel A we also control for the gender of a teacher. Panel B includes only female teachers, while Panel C includes only male teachers. Standard errors are clustered by class and are reported in parentheses. *, **, and *** denote significance at the 10%, 5% and 1% level.

Table 7: Correlations between Gender Biases of Same and Different Teachers

Dependent Variable: Teacher bias in the Own Subject

	11^{th} Grade			
	Different Teachers		Same Teachers	
	(1)	(2)	(3)	(4)
Teacher Bias	0.029	0.030	0.665	0.664
	(0.040)	(0.041)	(0.060)***	(0.061)***
Observations	1,198	1,198	444	444
	12^{th} Grade			
	Different Teachers		Same Teachers	
	(1)	(2)	(3)	(4)
Teacher Bias	0.033	0.032	0.581	0.581
	(0.030)	(0.030)	(0.064)***	(0.064)***
Observations	1,691	1,691	608	608
Year FE	✓	✓	✓	✓
School FE	✓		✓	
Class FE		✓		✓

Notes: The table includes stacked observations for the teacher gender biases in the own current class in 11^{th} (top panel) and 12^{th} (bottom panel) grades. The estimation is based on the sample of 21 schools. The (OLS) estimated coefficients in columns 1-2 are the between biases measures of different teachers who instruct students from the same class in two subjects and the (OLS) estimated coefficients in columns 3-4 are the between biases measures of the same teachers who instruct students from the same class in two subjects. Each estimate comes from a separate OLS regression. Standard errors are reported in parentheses and are clustered at the school and year level. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table 8: Effect of 11^{th} Grade Teacher Bias (measured in all other classes) on Blind Score in 12^{th} Grade

Dependent Variable: Blind score in 12^{th} grade national exams						
	Boys			Girls		
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
	0.057	0.037	0.090	-0.060	-0.072	-0.100
	(0.030)	(0.030)	(0.028)***	(0.024)***	(0.024)***	(0.024)***
Sample Size	8,200	8,200	8,200	10,348	10,348	10,348
Classics Track Subjects						
	0.117	0.147	0.185	-0.180	-0.099	-0.051
	(0.077)	(0.076)*	(0.073)**	(0.042)***	(0.042)**	(0.046)
Sample Size	1,401	1,401	1,401	5,499	5,499	5,499
Science Track Subjects						
	0.196	0.150	0.211	0.059	-0.045	-0.109
	(0.056)***	(0.053)***	(0.052)***	(0.050)	(0.050)	(0.050)*
Sample Size	4,460	4,460	4,460	4,910	4,910	4,910
Exact Science Track Subjects						
	0.071	0.043	0.145	-0.071	-0.052	-0.163
	(0.046)	(0.049)	(0.044)***	(0.063)	(0.060)	(0.066)**
Sample Size	5,164	5,164	5,164	3,031	3,031	3,031
Subjects FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: The datasets for the core subjects and each track subjects include stacked observations for each subject/exam. The estimation is based on the sample of 21 schools. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The empirical Bayes shrinkage factor is the ratio of signal variance to signal plus noise variance. We assume that there is a sampling error problem and the measure of teacher gender bias consists of an error component. Estimating teachers' effects on students' weighted difference between "non-blind" and "blind" scores enables us to separate between the signal and the noise variance. The empirical Bayes estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the empirical Bayes shrinkage factor. Standard errors are clustered by class and are reported in parentheses. The standard errors are also corrected using a two-step bootstrapping method. In the first stage, a random sample with replacement is drawn from each class by gender and the corresponding OLS coefficients are obtained. In the second stage, the effect of these new teachers' gender bias measures in 11^{th} grades on students' performance in 12^{th} grade are estimated and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficients estimates from the second stage regression are the bootstrap standard errors of the point estimates of these coefficients. All specifications include the students' blind score and the teacher's gender as controls. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the science track subjects. The forth panel "Exact Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table 9: Effect of 11^{th} and 12^{th} Grade Gender Biases (measured in all other classes) on Students Total, Excused and Unexcused Absences in 11^{th} and 12^{th} Grade

	Dependent Variable: Total, Excused and Unexcused Absences											
	11^{th} Grade						12^{th} Grade					
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
	Total Absences		Excused		Unexcused		Total Absences		Excused		Unexcused	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Bias in core subjects	-0.259	0.299	0.078	-0.316	-0.337	0.615	0.361	0.214	0.969	-0.630	-0.551	0.641
	(0.102)**	(0.130)**	(0.058)	(0.099)***	(0.098)***	(0.096)***	(0.405)	(0.358)	(0.382)***	(0.384)*	(0.253)*	(0.254)**
Sample Size	5,842	7,443	5,842	7,443	5,842	7,443	4,834	5,769	4,533	5,433	4,815	5,729
Bias in classics subjects	-0.297	-0.237	0.123	-0.286	-0.420	0.048	0.319	-0.089	0.954	-0.887	-0.777	0.621
	(0.254)	(0.175)	(0.156)	(0.135)*	(0.195)**	(0.144)	(0.208)	(0.181)	(0.353)***	(0.305)**	(0.221)***	(0.225)***
Sample Size	2,804	4,776	2,804	4,776	2,804	4,776	2,528	3,835	2,385	3,638	2,517	3,816
Bias in science subjects	-0.363	0.223	0.100	-0.163	-0.463	0.387	0.830	0.736	1.381	-0.477	-0.438	0.987
	(0.140)***	(0.094)**	(0.089)	(0.072)**	(0.114)**	(0.092)***	(0.647)	(0.529)	(0.611)*	(0.618)	(0.393)	(0.390)***
Sample Size	4,504	5,492	4,504	5,492	4,504	5,492	2,881	3,340	2,657	3,098	2,869	3,306
Bias in exact science subjects	-0.206	0.225	0.101	-0.211	-0.308	0.436	0.605	0.462	0.861	-0.717	-0.227	0.970
	(0.131)	(0.104)**	(0.080)	(0.085)**	(0.114)***	(0.097)***	(0.459)	(0.454)	(0.429)	(0.529)	(0.278)	(0.327)***
Sample Size	4,071	4,812	4,071	4,812	4,071	4,812	3,996	3,824	3,745	3,568	3,980	3,791
Subjects FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Class FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents estimates for the effects of the bias (measured in all other classes) in the related subjects on students' different types of attendance (in hours). The estimation is based on the sample of 21 schools. The outcome variable is the total number of absences in a year (in hours), the excused number of absences in a year (in hours), and the unexcused number of absences in a year (in hours). The estimates are presented separately for the 11^{th} and 12^{th} grade. All estimates have been calculated using an empirical Bayes estimation strategy. All standard errors (reported in parentheses) are calculated using a two-step bootstrapping technique and are clustered at the class level. In the first panel all core subjects are used. The second panel includes only classics subjects. The third panel includes only science subjects. The forth panel includes only exact science subjects. The scores are standardized z-scores. *,**, and *** denote significance at the 10%, 5%, and 1% level.

Table 10: Placebo Estimation: Randomly Shuffle Biases Across Teachers of the Same Subjects Within Schools

Dependent Variable: Subsequent Blind score in 12^{th} grade						
	Boys			Girls		
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
	-0.010	-0.022	-0.004	0.004	-0.004	0.019
	(0.027)	(0.027)	(0.025)	(0.022)	(0.022)	(0.021)
Sample Size	8,217	8,217	8,217	10,347	10,347	10,347
Classics Track Subjects						
	-0.111	-0.093	0.026	-0.048	0.063	0.030
	(0.083)	(0.084)	(0.090)	(0.043)	(0.050)	(0.068)
Sample Size	1,391	1,391	1,391	5,443	5,443	5,443
Science Track Subjects						
	0.062	0.001	-0.045	0.103	0.062	0.010
	(0.058)	(0.056)	(0.051)	(0.052)	(0.049)	(0.042)
Sample Size	4,492	4,492	4,492	4,932	4,932	4,932
Exact Science Track Subjects						
	0.011	-0.010	0.005	-0.109	-0.075	-0.078
	(0.042)	(0.043)	(0.039)	(0.055)	(0.054)	(0.052)
Sample Size	5,119	5,119	5,119	3,0130	3,030	3,030
Subjects FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: We randomly reshuffle teacher biases within schools across teachers who teach the same subjects. The estimation is based on the sample of 21 schools. The datasets for the core subjects and each track subjects include stacked observations for each subject/exam. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy, for 11^{th} (columns 1-2) and 12^{th} (columns 3-4) grade separately. The empirical Bayes shrinkage factor is the ratio of signal variance to signal plus noise variance. We assume that there is a sampling error problem and the measure of teacher gender bias consists of an error component. Estimating teachers' effects on students' weighted difference between "non-blind" and "blind" scores enables us to separate between the signal and the noise variance. The empirical Bayes estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the empirical Bayes shrinkage factor. Standard errors are clustered using a two-step bootstrapping method. In the first stage, a random sample with replacement is drawn from each class by gender and the corresponding OLS coefficients are obtained. In the second stage, the effect of these new teachers' gender bias measures in 11^{th} grades on students' performance in 12^{th} grade are estimated and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficients estimates from the second stage regression are the bootstrap standard errors of the point estimates of these coefficients. All specifications include the students' blind score as a control. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the science track subjects. The forth panel "Exact Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table 11: Effect of 11^{th} and 12^{th} Grade Gender Biases (measured in all other classes) on Enrollment in Post-Secondary Schooling

Dependent Variable: Dummy variable for Enrollment Status in University

	11^{th} grade		12^{th} grade	
	Boys	Girls	Boys	Girls
	(1)	(2)	(3)	(4)
Bias core subjects	0.020	-0.047	0.036	-0.028
	(0.005)***	(0.005)***	(0.008)***	(0.008)***
Sample Size	6,845	8,640	5,699	6,861
Bias in classics subjects	0.046	-0.014	0.050	-0.037
	(0.008)***	(0.008)*	(0.015)***	(0.014)***
Sample Size	3,288	5,598	2,998	4,675
Bias in science subjects	0.021	-0.028	0.039	-0.021
	(0.006)***	(0.005)***	(0.011)***	(0.010)*
Sample Size	5,219	6,357	3,376	3,950
Bias in exact science subjects	0.015	-0.032	0.030	-0.026
	(0.007)**	(0.006)***	(0.008)***	(0.010)**
Sample Size	4,795	5,587	4,786	4,495
Subjects FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Class FE	✓	✓	✓	✓

Notes: The outcome variable is the post-secondary enrollment status (1 if enrolled, 0 otherwise). In these regressions, we also control for the blind performance a student gets in each grade (11^{th} grade for columns 1-2 and 12^{th} grade for columns 3-4). Standard errors are clustered by class and are reported in parentheses. The datasets for the core subjects and each track subjects include stacked observations for each subject/exam. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy, for 11^{th} (columns 1-2) and 12^{th} (columns 3-4) grade separately. The empirical Bayes shrinkage factor is the ratio of signal variance to signal plus noise variance. We assume that there is a sampling error problem and the measure of teacher gender bias consists of an error component. Estimating teachers' effects on students' weighted difference between "non-blind" and "blind" scores enables us to separate between the signal and the noise variance. The empirical Bayes estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the empirical Bayes shrinkage factor. Standard errors are clustered using a two-step bootstrapping method. In the first stage, a random sample with replacement is drawn from each class by gender and the corresponding OLS coefficients are obtained. In the second stage, the effect of these new teachers' gender bias measures in 11^{th} grades on students' performance in 12^{th} grade are estimated and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficients estimates from the second stage regression are the bootstrap standard errors of the point estimates of these coefficients. All specifications include the students' blind score as a control. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the science track subjects. The forth panel "Exact Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table 12: Descriptive Statistics by University Field of Studies 2003-2011

Field of studies	**Mean Enrolment**								**Difference**	
	Girls				**Boys**					
	Mean (1)	(sd) (2)	**Mean** (3)	(sd) (4)	**Mean** (5)	(sd) (6)	**Mean** (7)	(sd) (8)	(se) (1)-(5)	(se) (3)-(7)
Exact Science	0.099	(0.298)	0.121	(0.326)	0.223	(0.416)	0.272	(0.445)	-0.124	-0.151
									(0.003)	(0.003)
Science	0.046	(0.209)	0.056	(0.230)	0.037	(0.188)	0.045	(0.207)	0.009	0.011
									(0.002)	(0.002)
Social Science	0.225	(0.419)	0.276	(0.448)	0.212	(0.409)	0.259	(0.438)	0.014	0.018
									(0.003)	(0.004)
Humanities	0.275	(0.445)	0.336	(0.472)	0.088	(0.284)	0.108	(0.310)	0.186	0.228
									(0.003)	(0.004)
Vocational Studies	0.171	(0.377)	0.210	(0.407)	0.258	(0.438)	0.315	(0.464)	-0.087	-0.105
									(0.003)	(0.004)
Not Enrolled in Post-Secondary Studies	0.184	(0.387)			0.181	(0.385)			0.003	
									(0.003)	

Notes: The sample includes 37,218 female students and 28,869 male students. In columns (3) and (7) we restrict the sample only to students who enroll in university studies. Humanities include the departments of liberal arts, psychology, journalism, philosophy, education, Greek language, history, foreign languages, home economics and law. Social Science includes the departments of economics, statistics, business and management, accounting, political and European studies. Exact Science includes the departments of mathematics, engineering, physics and computer science. Science includes the departments of biology, chemistry, medicine, pharmacy, veterinary studies and dentistry. Vocational studies include students who enroll in technical education institutes and agricultural studies.

Table 13: Effect of 11^{th} and 12^{th} Grade Teacher Gender Bias on the Choice of University Field of Study by Gender

Dependent Variable: Dummy variable for the Choice of University Study						
	(1)	(2)	(3)	(4)	(5)	(6)
	Boys			Girls		
	A. Stack 11^{th} and 12^{th} grades & Grade FE. (2003-2005)					
	0.010	0.013	0.004	-0.031	-0.037	-0.046
	(0.013)	(0.014)	(0.018)	(0.011)***	(0.011)***	(0.015)***
Sample Size	7,929	7,929	7,929	9,953	9,953	9,953
	B. 12^{th} grade (2003-2011)					
	-0.024	-0.026	-0.030	-0.029	-0.040	-0.042
	(0.014)	(0.014)*	(0.019)	(0.011)**	(0.012)***	(0.016)***
Sample Size	5,161	5,161	5,161	6,559	6,559	6,559
Year FE	✓	✓	✓	✓	✓	✓
Major FE	✓	✓	✓	✓	✓	✓
Track FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. Estimation is based on the sample of 21 schools. The dependent variable is the decision to study in one of the following four fields: Social Science, Science, Exact Science and Humanities. Students not enrolled in any university or students who enroll in vocational studies are not included in the sample. We stack the four possible choices as the dependent variable for each student against the teachers' bias in each of the four areas of university studies. The dependent variable is a 0/1 indicator, assuming the value of 1 for the observed field of study and a value 0 for the other three possible choices. The subjects that we use for each field of study are the following: for exact science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and mathematics and physics in 12^{th} grade. For humanities departments we use the blind score and the bias in history and modern Greek in both 11^{th} and 12^{th} grades. For social science departments we use the blind score and the bias in history and modern Greek in 11^{th}, and economics in 12^{th} grade. For science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and biology and physics in 12^{th} grade. The scores are standardized and have a zero mean and a standard deviation of one. *, **, and *** denote significance at the 10%, 5% and 1% level.

Table 14: Effect of 11^{th} and 12^{th} Grade Teacher Gender Bias on the Average Quality of the Program Students Enrolled In

Dependent Variable: Percentile Quality Rank of Post-Secondary Program								
	Rank based on cutoffs		Rank based on mean perf.		Rank based on cutoffs		Rank based on mean perf.	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	11^{th} grade				12^{th} grade			
Humanities	2.909	-6.407	2.115	-6.509	5.970	-2.357	6.089	-0.321
	(4.410)	(2.256)***	(4.332)	(2.087)***	(6.039)	(2.544)	(5.991)	(2.550)
Sample Size	201	845	201	845	156	668	156	668
Social Science	4.746	-2.380	5.020	-2.157	9.533	-3.274	12.652	-3.872
	(3.021)	(2.365)	(3.381)	(2.849)	(4.056)**	(2.765)	(4.420)***	(3.267)
Sample Size	485	746	485	746	426	675	426	675
Science	10.300	0.517	4.469	-1.685	5.579	-2.913	1.478	-2.333
	(9.671)	(4.327)	(11.973)	(5.555)	(8.434)	(11.805)	(11.522)	(13.159)
Sample Size	137	179	137	179	67	92	67	92
Exact Science	8.916	-4.977	9.444	-8.360	0.751	-1.541	1.576	-1.057
	(4.388)**	(5.017)	(4.831)**	(5.972)	(3.514)	(3.528)	(3.379)	(3.699)
Sample Size	567	343	567	343	978	563	978	563
	11^{th} grade				12^{th} grade			
Humanities or Social Science	6.675	-4.878	6.378	-4.653	9.420	-2.884	11.666	-2.208
	(2.569)***	(1.768)***	(2.792)***	(1.843)***	(3.268)***	(1.848)	(3.495)***	(2.038)
Sample Size	686	1,596	686	1,596	584	1,348	584	1,348
Science or Exact Science	10.211	-0.299	10.688	-3.570	0.223	-1.948	0.815	-1.593
	(3.937)**	(4.757)	(4.266)**	(5.252)	(3.388)	(2.923)	(3.425)	(3.246)
Sample Size	657	468	657	468	1,113	745	1,113	745
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Subject FE	✓	✓	✓	✓	✓	✓	✓	✓
School FE	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Students are assigned the percentile quality rank of the post-secondary program they enroll in. The estimation is based on the sample of 21 schools. These quality measures represent the ranking of each post-secondary program and are measured using two different approaches: 1) the mean test score of enrolled students in each post-secondary program and 2) the admission cutoff or threshold score for each university department over a period of 9 cohorts (2003-2011). We then look at the effects of average teacher biases in the core related subjects that are closely related to area of study (humanities, social science, science, exact science) on the quality of enrolled post-secondary program (using both measures) conditional on the field of study that students chose, while we also control for students' blind score in each subject and the gender of the teacher in each subject. As before there are four fields of study at the university level and these are Humanities, Social Science, Science and Exact Science. In the top panel, we examine the effect of teacher biases in the related subjects on post-secondary quality conditional on each field of study separately. The related bias for those who enroll in Humanities at the university level comes from modern Greek and history, for both 11^{th} and 12^{th} grade. The related bias for those who enroll in Social Sciences at the university level comes from modern Greek and history in 11^{th} grade and modern Greek, history and economics in 12^{th} grade. The related bias for those who enroll in Science at the university level comes from modern algebra and physics in 11^{th} grade and biology and physics in 12^{th} grade. The related bias for those who enroll in Exact Science at the university level comes from algebra and geometry in 11^{th} grade and mathematics, biology and physics in 12^{th} grade. In the bottom panel, we group Humanities together with Social Science and Science together with Exact Science and repeat the same exercise. The baseline sample is 11^{th} grade students in 2003-2005 and 12^{th} grade students in the period 2003-2011. All estimates are adjusted for the empirical Bayes technique. Standard errors reported in parentheses are clustered at the school level and are calculated using a two-step bootstrapping technique. *, **, and *** denote significance at the 10%, 5% and 1% level.

Table 15: Effect of 11^{th} and 12^{th} Grade Teacher Bias on the Quality of the Program Students Enrolled In

Dependent Variable: Percentile Quality Rank of Post-Secondary Program				
	Rank based on mean perf. 2003		Rank based on cutoffs 2003	
	Boys	Girls	Boys	Girls
	(1)	(2)	(3)	(4)
	Panel A: 11^{th} grade			
Teacher Bias in Core	2.644 (1.125)**	-4.246 (0.930)***	2.632 (1.123)**	-4.261 (0.933)***
Sample Size	4,131	5,135	4,131	5,135
Teacher Bias in Classics Track	5.497 (2.408)**	-1.115 (1.323)	5.230 (2.311)**	-1.091 (1.297)
Sample Size	624	2,400	624	2,400
Teacher Bias in Science Track	4.332 (1.702)***	-1.845 (1.268)	4.319 (1.694)***	-1.837 (1.263)
Sample Size	3,156	3,468	3,156	3,468
Teacher Bias in Exact Science Track	1.718 (1.339)	-6.315 (1.706)***	1.735 (1.339)	-6.477 (1.741)***
Sample Size	2,492	1,547	2,492	1,547
	Panel B: 12^{th} grade			
Teacher Bias in Core	4.006 (1.136)***	-0.016 (1.117)	3.934 (1.132)**	-0.219 (1.104)
Sample Size	3,273	3,855	3,273	3,855
Teacher Bias in Classics Track	5.562 (2.112)**	-2.643 (1.338)***	5.160 (1.986)**	-2.555 (1.337)*
Sample Size	912	3,358	912	3,358
Teacher Bias in Science Track	1.723 (2.681)	2.636 (2.131)	1.571 (2.604)	2.508 (2.103)
Sample Size	1,377	1,862	1,377	1,862
Teacher Bias in Exact Science Track	2.636 (1.077)**	-0.086 (1.174)	2.546 (1.065)**	-0.107 (1.200)
Sample Size	4,039	2,457	4,039	2,457
Year FE	✓	✓	✓	✓
Subject FE	✓	✓	✓	✓
Class FE	✓	✓	✓	✓

Notes: The estimation is based on the sample of 21 schools. Year 2003 is excluded from the analysis, as it is used to calculate the quality measures for the post-secondary program students enroll in. These quality measures represent the ranking of each post-secondary program and are measured using two different approaches: 1) the 2003 mean performance of enrolled students in each post-secondary program and 2) the 2003 admission cutoff or threshold score for each university department. This is calculated as the admission score of the last admitted student. We then assign these two measures of program quality to the relevant post-secondary programs and drop the year 2003 from the regressions. We then look at the effects of teacher biases on the quality of enrolled post-secondary program (using both measures), while we control for students' blind score in each subject and the gender of the teacher in each subject. All estimates are adjusted for the empirical Bayes technique. Standard errors reported in parentheses are clustered at the school level and are calculated using a two-step bootstrapping technique. *, **, and *** denote significance at the 10%, 5% and 1% level respectively.

Table 16: Comparisons of Mean Teacher Value Added for Pro-Boys, Neutral, and Pro-Girls Teachers

	(1)	(2)	(3)
		Panel A	
	Neutral /(sd)	Pro-Boy /(sd)	Difference /(se)
Teacher Value Added	0.053	-0.037	0.090
	(0.132)	(0.222)	(0.032)***
N	58	101	159
		Panel B	
	Neutral /(sd)	Pro-Girl /(sd)	Difference /(se)
Teacher Value Added	0.053	-0.049	0.102
	(0.132)	(0.235)	(0.032)***
N	58	259	317

Notes: We pool data on test scores for 11^{th} and 12^{th} grades for the period 2003-2005. The TVA measures are derived following the procedure described in Chetty et al (2014). Pro-boy teacher exhibit bias larger than or equal to 0.10. Pro-girl teachers exhibit a bias smaller than or equal to -0.10. We define as neutral teachers who exhibit bias that is larger than or equal to -0.10 and smaller than 0.10. The teacher bias measures are derived as the average bias across subjects, grades and classes a teacher exhibits in the 2006-2011 sample. Our sample includes only students who have non-missing baseline controls to estimate the VA model. Our baseline VA model controls for a rich set of student demographics and other variables, as well as teacher, class, and school level variables. In particular, our baseline VA model controls for a student's gender, age, a dummy whether the student is born in the first quarter of a calendar year, his/her lagged performance in the same subject, class size, school-level-grade enrollment, a dummy that takes the value of 1 if the teacher is female and 0 otherwise, how many classes each teacher taught in the sample (our proxy for a teacher's experience), students' average performance in all other classes taught by the same teacher in the sample and neighborhood income. When the prior test score is missing, we set the prior score equal to 0 and include an indicator for missing data. In Panel A, we compare the mean TVA of teachers who are neutral (column 1) to the mean TVA of teachers who are pro-boy (column 2). The related standard deviations are reported below the means. Column 3 reports the difference of the means and the respective standard error. In Panel B, we compare the mean TVA of teachers who are neutral (column 1) to the mean TVA of teachers who are pro-girl (column 2). The related standard deviations are reported below the means. Column 3 reports the difference of the means and the standard error.

Table 17: Correlations Between Teacher Bias And Teacher Quality (TVA) for Pro-Girl and Pro-Boy Teachers (Spline Variables)

Dependent Variable: Teacher Quality (Measured by TVA)	(1)	(2)	(3)	(4)
Spline for Pro-Boys Teachers	-0.113	-0.113	-0.112	-0.112
	(0.041)***	(0.040)***	(0.040)***	(0.043)**
Spline for Pro-Girls Teachers	0.049	0.049	0.050	0.050
	(0.026)*	(0.027)*	(0.028)*	(0.029)*
Female Teacher		-0.006	-0.007	-0.007
		(0.018)	(0.018)	(0.019)
Class Size			-0.002	-0.002
			(0.003)	(0.003)
Teacher Experience				-0.0001
				(0.001)
Year FE	✓	✓	✓	✓
School FE	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓
Observations	418	418	418	418

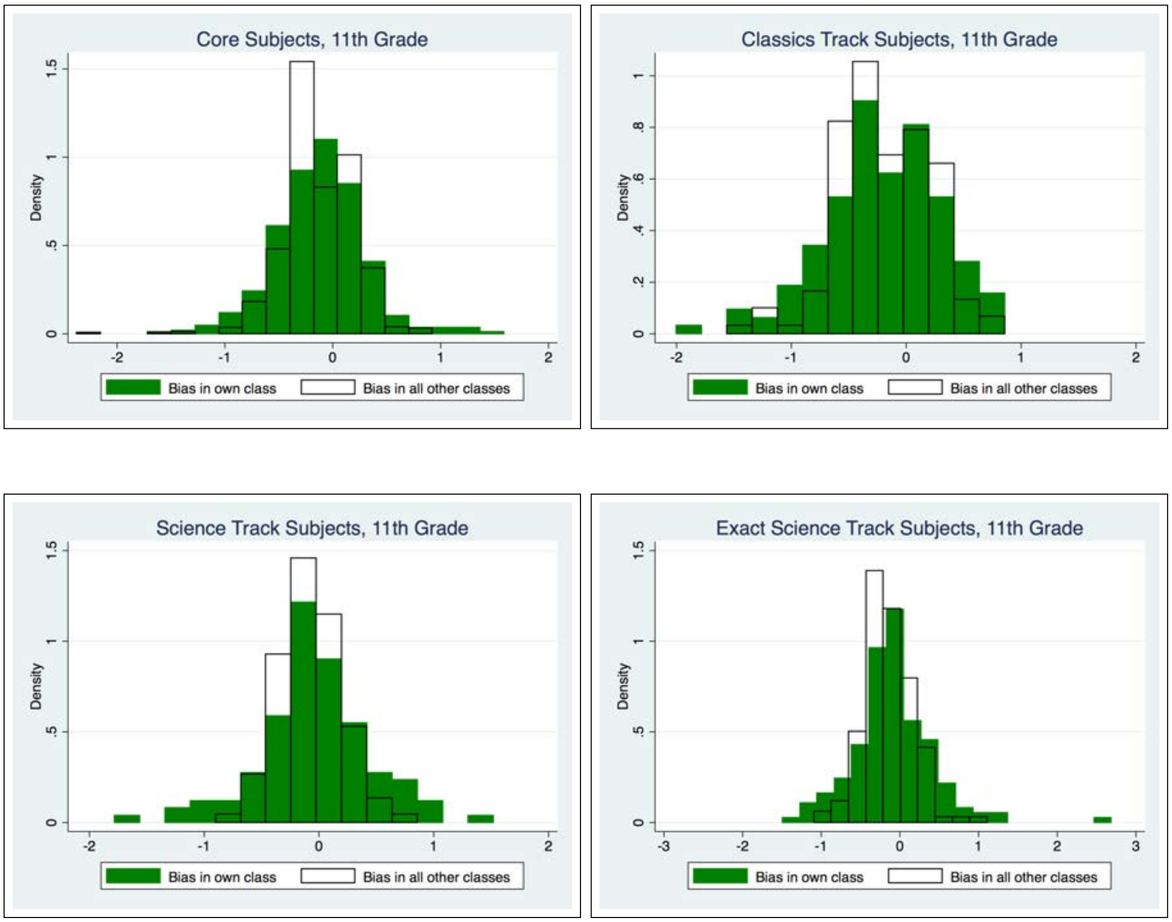
Notes: The "Spline for Pro-Girl Teachers" takes the actual negative teacher bias values, and the value of zero for the positive ones. The "Spline for Pro-Boy Teachers" takes the actual positive teacher bias values, and the value zero for the negative ones. The teacher gender bias measures the average bias a teacher exhibits in different subjects and classes in the 2006-2011 sample. We include the two splines simultaneously in the regression. The outcome variable is the teacher value added derived using the 2003-2005 sample and is described in details in the text. "Teacher experience" measures the different combination of classes and subjects a teacher has taught in 11^{th} and 12^{th} grades in the sample period 2003-2011. Standard errors are clustered by school and year and are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table 18: Correlations Between Teacher Gender Bias And Teacher Quality (Measured by TVA)

Dependent Variable: Teacher Quality (Measured by TVA)				
	(1)	(2)	(3)	(4)
Pro-Girl Teacher Dummy	-0.035	-0.036	-0.037	-0.037
	(0.017)**	(0.017)**	(0.018)**	(0.018)**
Pro-Boy Teacher Dummy	-0.031	-0.032	-0.032	-0.031
	(0.018)*	(0.018)*	(0.018)*	(0.018)*
Female Teacher		-0.007	-0.008	-0.007
		(0.018)	(0.018)	(0.019)
Class Size			-0.003	-0.003
			(0.003)	(0.003)
Teacher Experience				0.001
				(0.001)
Year FE	✓	✓	✓	✓
School FE	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓
Observations	418	418	418	418

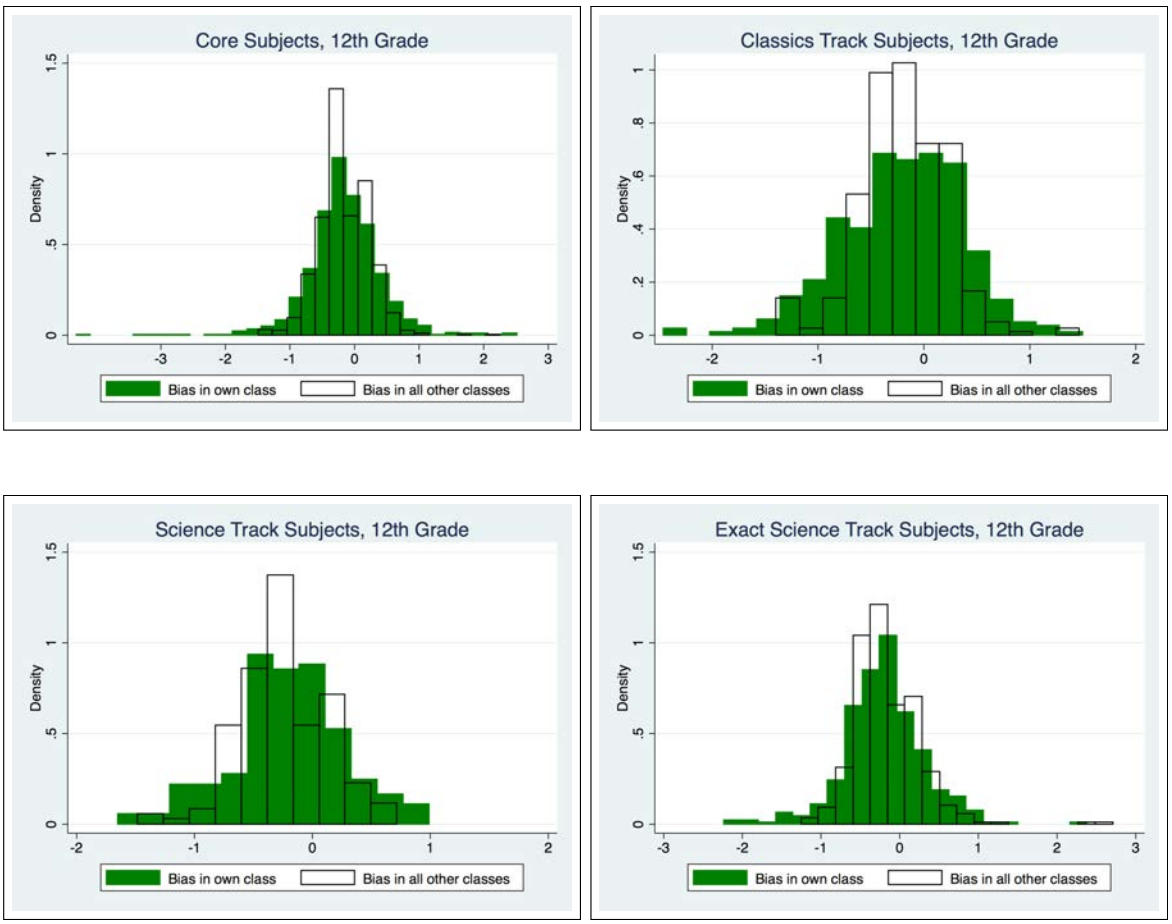
Notes: The "Pro-Girl Teacher Dummy" takes the value of one if the teacher exhibits a bias that is smaller than or equal to -0.10. The "Pro-Boy Teacher Dummy" takes the value of one if the teacher exhibits a bias that is above 0.10. We define as neutral teachers those who have a bias that is larger than or equal to -0.10 and smaller than 0.10. The omitted category in the regression is neutral teachers. The teacher bias is calculated in the sample period of 2006-2011. The outcome variable is the TVA derived using the 2003-2005 sample and described in the text. "Teacher experience" measures the different combination of classes and subjects a teacher has taught in 11^{th} and 12^{th} grades in the sample period 2003-2011. Standard errors are clustered by school and year and are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Figure 1: Histograms of the bias measured in current class and the bias measured in all other classes for Core and Track subjects in 11^{th} grade



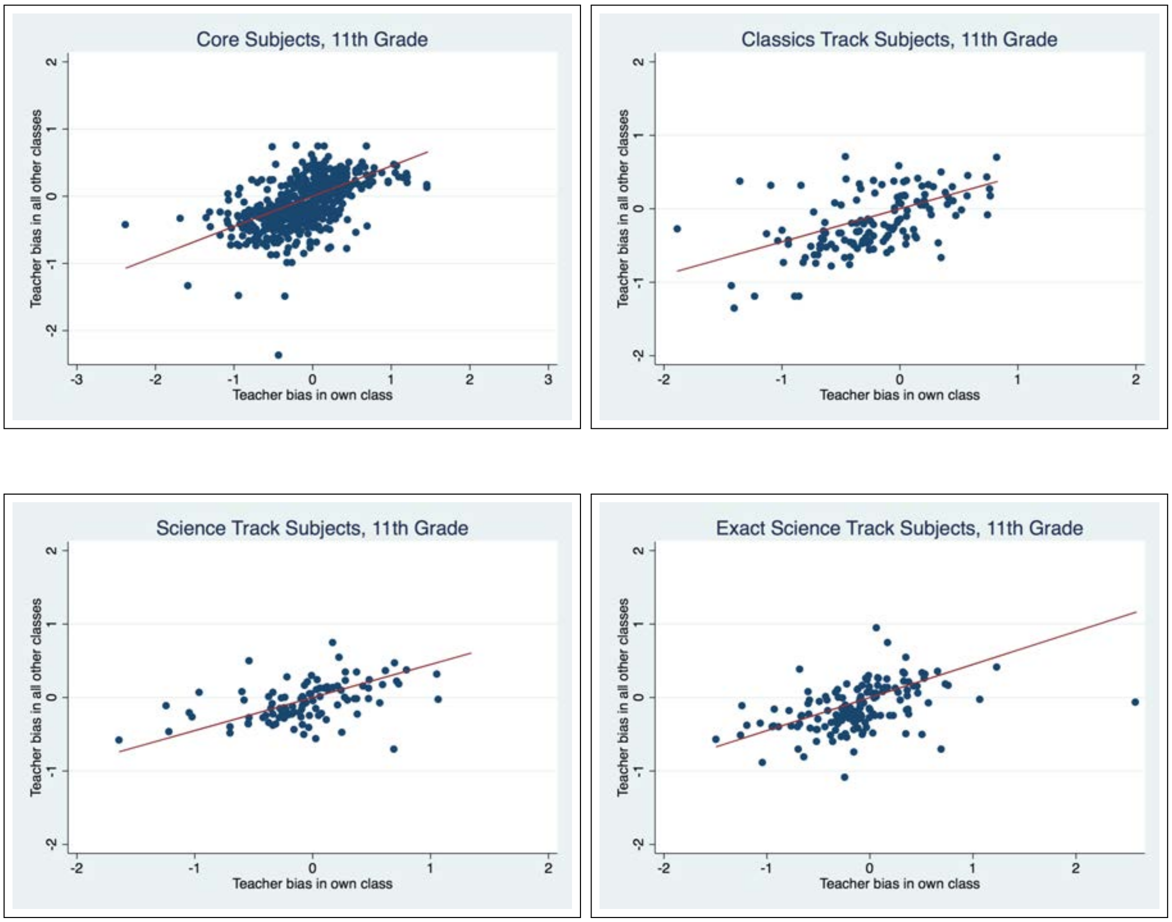
Notes: This figure presents the distributions for the 11^{th} grade teacher gender bias measured in two different ways: a) using the own current class (green histogram) and b) using the bias of a teacher in all other classes except of the current one (white histogram). In 11^{th} grade the core subjects taught are: modern Greek, history, physics, algebra and geometry. There are the following tracks in 11^{th} grade: Classics, Science and Exact Science. In the classics track the 11^{th} grade subjects are: ancient Greek, philosophy and Latin; in the science track: mathematics, physics, chemistry, and in the exact science track: mathematics, physics and computer science.

Figure 2: Histograms of the bias measured in the current class and the bias measured in all other classes for Core and Track subjects in 12^{th} grade



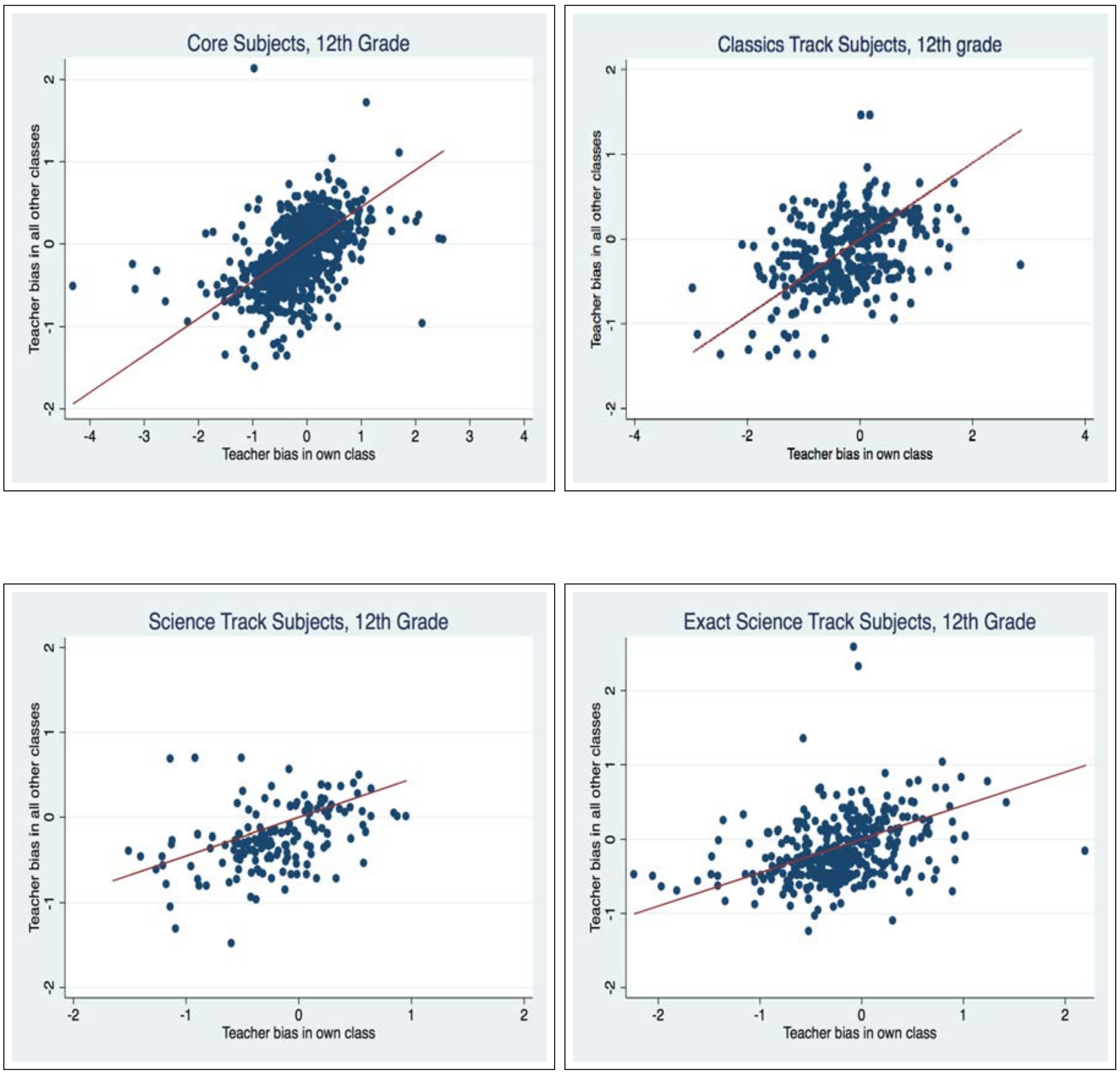
Notes: This figure presents the distributions for the 12^{th} grade teacher gender bias measured in two different ways: a) using the own current class (green histogram) and b) using the bias of a teacher in all other classes except of the current one (white histogram). In 12^{th} grade the core subjects taught are: modern Greek, history, physics, biology and mathematics. There are the following tracks in 12^{th} grade: Classics, Science and Exact Science. In the classics track the 12^{th} grade subjects are: ancient Greek, Latin, literature and history; in the science track: biology, mathematics, physics and chemistry, and in the exact science track: mathematics, physics, business administration and computer science.

Figure 3: Correlations between the gender bias measured in the current class and the bias measured in all other classes for Core and Track subjects in 11^{th} grade



Notes: These scatter plots present the correlations between the two different measures of teacher biases, namely bias in current own class and bias in all other classes, for core subjects (top left panel), classics track (top right panel), science track (bottom left panel) and exact science (bottom right panel). The 45 degree line is also drawn in each figure. In 11^{th} grade the core subjects taught are: modern Greek, history, physics, algebra and geometry. There are the following tracks in 11^{th} grade: Classics, Science and Exact Science. In the classics track the 11^{th} grade subjects are: ancient Greek, philosophy and Latin; in the science track: mathematics, physics, chemistry, and in the exact science track: mathematics, physics and computer science.

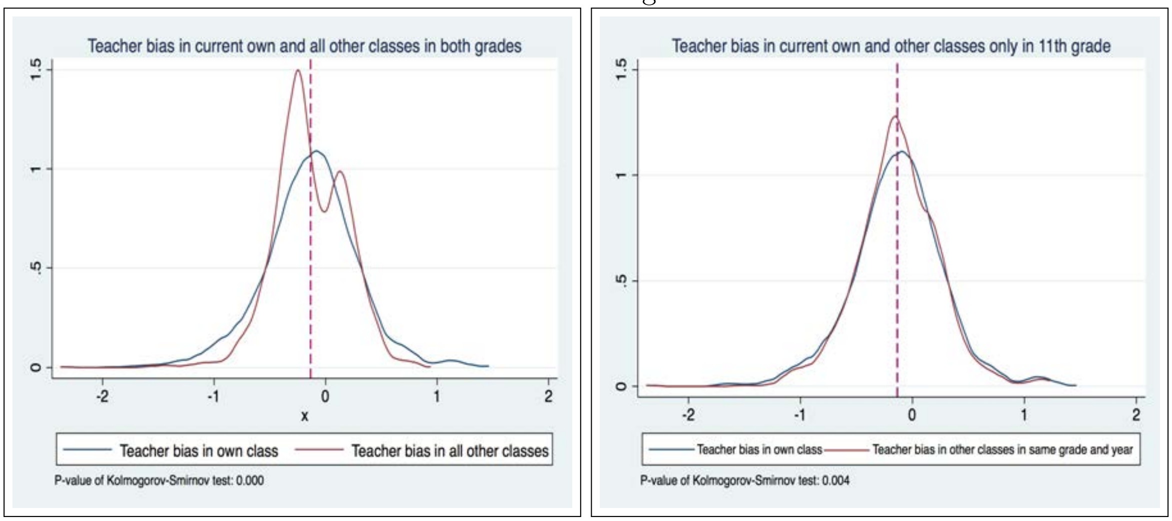
Figure 4: Correlations between the gender bias measured in the current class and the bias measured in all other classes for Core and Track subjects in 12^{th} grade



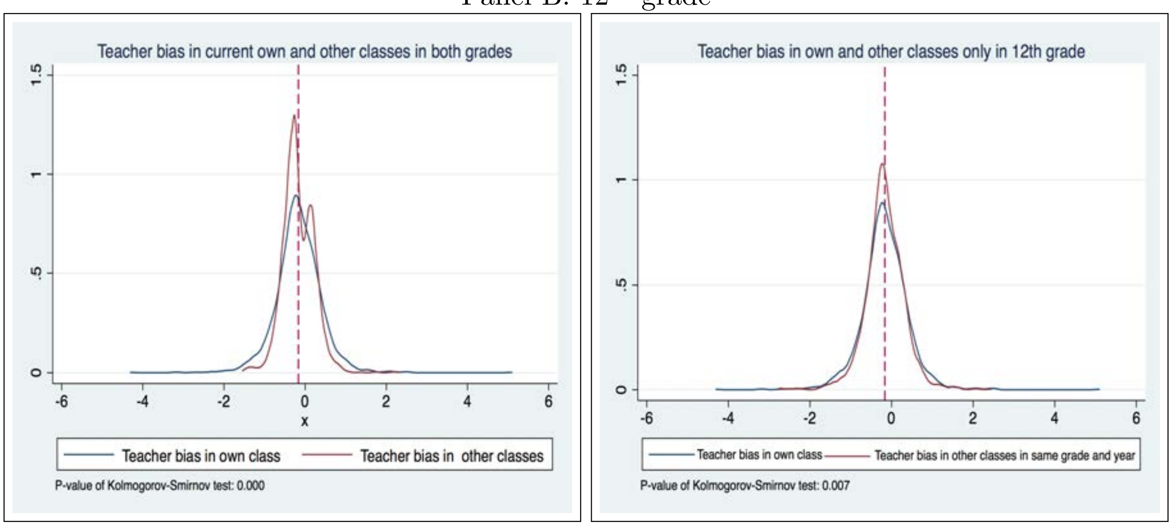
Notes: Notes: These scatter plots present the correlations between the two different measures of teacher biases, namely bias in current own class and bias in all other classes, for core subjects (top left panel), classics track (top right panel), science track (bottom left panel) and exact science (bottom right panel). The 45 degree line is also drawn in each figure. In 12^{th} grade the core subjects taught are: modern Greek, history, physics, biology and mathematics. There are the following tracks in 12^{th} grade: Classics, Science and Exact Science. In the classics track the 12^{th} grade subjects are: ancient Greek, Latin, literature and history; in the science track: biology, mathematics, physics and chemistry, and in the exact science track: mathematics, physics, business administration and computer science.

Figure 5: Distribution of Teacher Gender Bias in 11^{th} and 12^{th} Grade, Sample of 21 Schools

Panel A: 11^{th} grade

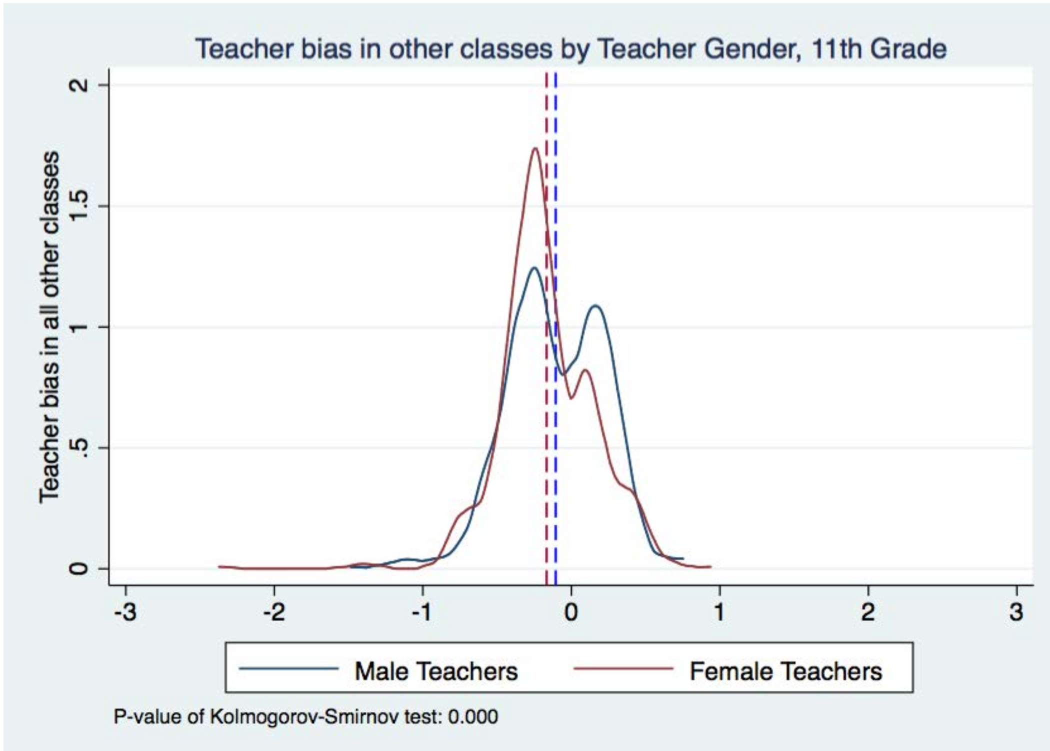


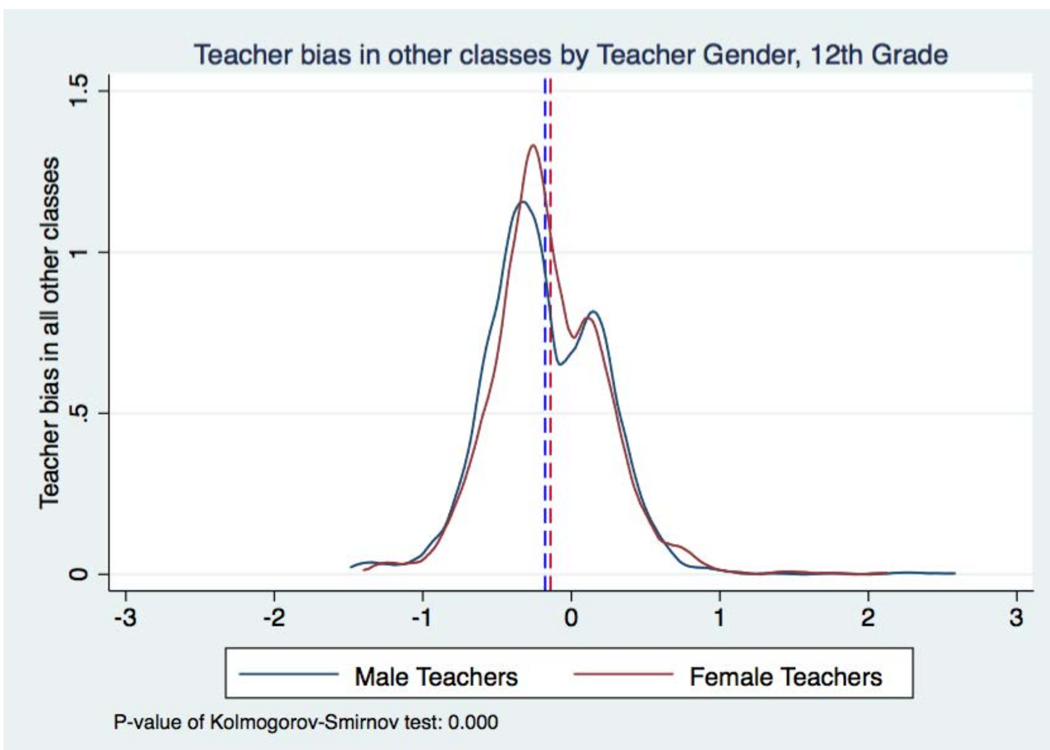
Panel B: 12^{th} grade



Notes: The 11^{th} (Panel A) and 12^{th} (Panel B) teacher bias is measured in the current class, in all other classes in any grade and in all other class in the same grade and year. We use data for the sample period 2003-2011. The left figures show the 11^{th} (Panel A) and 12^{th} (Panel B) teacher bias measured in the current class and in all other classes that a teacher taught in any grade. The teacher bias in the current class is the bias a teacher exhibited in this particular class. The teacher bias in all other classes that a teacher taught in any grade is measured as the average bias that a teacher exhibited in all other classes she/he ever taught in the sample period 2003-2011, irrespective of the grade. The right figures show the 11^{th} (Panel A) and 12^{th} (Panel B) teacher bias measured in the own class and in all other classes that the teacher taught in the same grade and year. The teacher bias in all other classes that a teacher taught in the same grade and year is calculated as the average bias across all other classes a teacher taught in the same grade and year as the current own class.

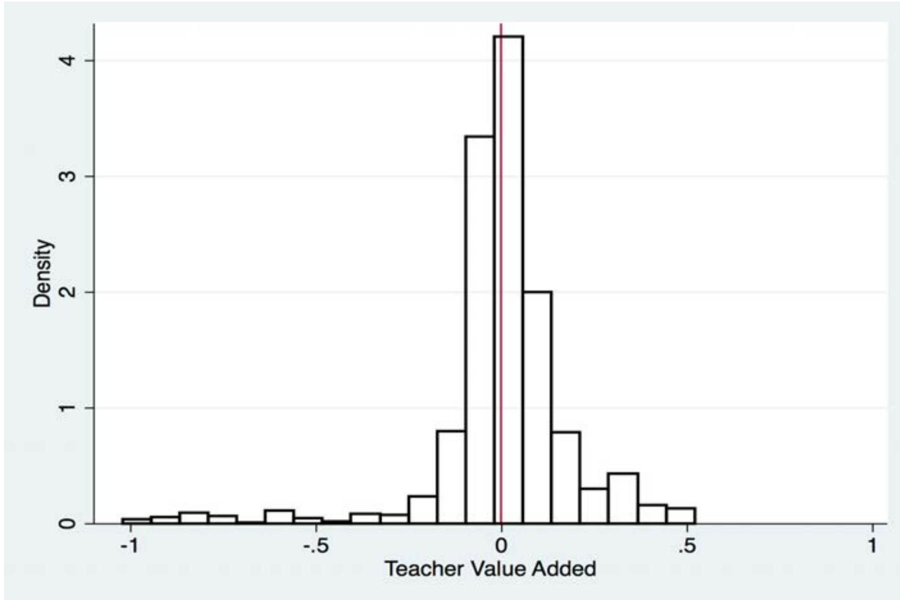
Figure 6: Distribution of Teacher Gender Bias by Teacher Gender, Sample of 21 Schools

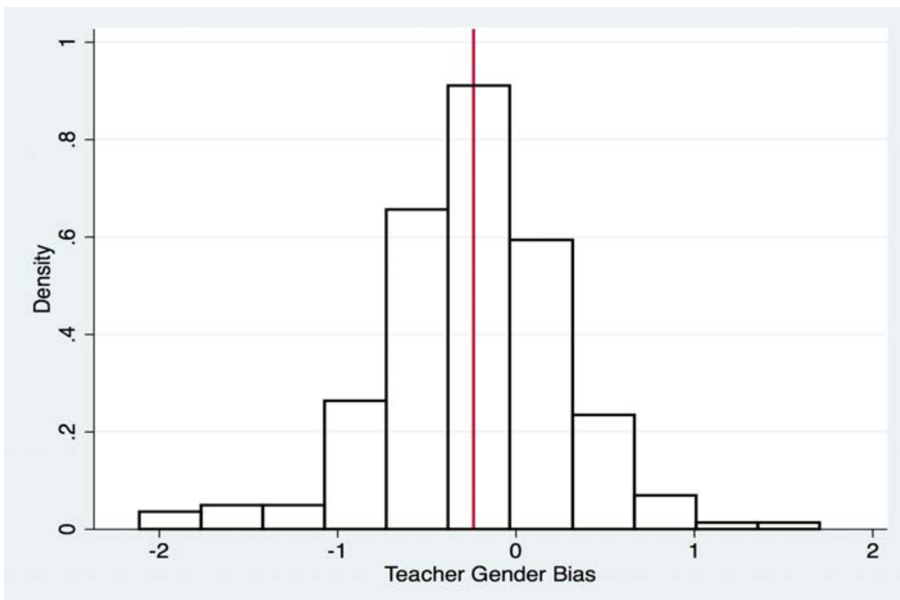




Notes: The teacher bias here is measured based on all other classes in both grades that a teacher taught in the sample period. We report the distribution of the teacher bias separately for male and female teachers in 11^{th} and 12^{th} grades.

Figure 7: Histogram of Teacher Value Added Measure and Average Teacher Bias





Notes: The top panel presents the distribution of the TVA measure, which is weighted by the number of students in the school-year-grade-subject-class year cell. To derive these value added measures we pool the 11^{th} and 12^{th} grade data for the years 2003-2005. We use 10^{th} and 11^{th} grade performance as a prior measure of performance. We follow closely the value added procedure described in Chetty et al. (2014a). This sample includes only students who have non-missing baseline controls to estimate the VA model. TVA is estimated using the baseline control vector, which includes: lagged own-subject scores, student-level characteristics including age, gender, a dummy for being born in the first quarter of the birth year, dummies for whether students expressed a special interest in classics, science or exact science (indicated by the track they have chosen), class size, school-grade enrollment, income as well as school, year, and subject dummies. When prior test scores are missing, we set the prior score equal to 0 and include an indicator for missing data. Student data are from the administrative records of 21 schools in Greece. The structure of the dataset is one observation per teacher-year-grade-subject-class combination. The bottom panel presents the distribution of the average teacher bias measured in all other classes across subjects and classes. To derive a teacher's bias we calculate the average bias a teacher exhibits in all classes they taught in the later years, in particular between 2006 and 2011.

Online Appendix

Not For Publication

Table A1: Descriptive Statistics, Full Sample

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
I. Student Characteristics				
11^{th} grade				
Female	0.563	0.496	0	1
Total absences (in hours per year)	50.850	27.426	1	450
Proportion of students by track:				
Classics	0.366	0.482	0	1
Science	0.280	0.449	0	1
Exact Science	0.344	0.475	0	1
GPA 11^{th} grade	72.335	14.158	0	100
12^{th} grade				
Total absences (in hours per year)	73.442	30.752	1	208
Proportion of students by track:				
Classics	0.369	0.483	0	1
Sciences	0.159	0.366	0	1
Exact Science	0.463	0.499	0	1
GPA 12^{th} grade	76.976	12.526	44	100
Age	17.902	0.465	15	42
II. School Characteristics				
Private School	0.037	0.190	0	1
Experimental School	0.044	0.207	0	1
Public School	0.919	0.275	0	1
Urban	0.896	0.306	0	1
Postcode Income (in 2009 Euro)	22,455	7,945	11,784	66,521
III. University Enrollment Characteristics				
National exams average score	64.987	20.178	10.35	99.3
Retake the national exams	0.113	0.317	0	1
Number of university departments in preference list	25.014	22.071	1	257
Rank in preference list of the actual university attended	8.399	10.616	1	242
Enrollment in university or vocational schooling	0.817	0.386	0	1
Exact Science department	0.153	0.360	0	1
Science department	0.042	0.201	0	1
Humanities department	0.193	0.395	0	1
Social Science department	0.220	0.407	0	1
Vocational schooling	0.209	0.407	0	1

Note: All statistics reported include students who graduate from high school between 2003 and 2011. Total absences are measured in hours per year. *GPA11* and *GPA12* include the average over the school exam scores in the first and second term, in 11^{th} and 12^{th} grade. The full sample of schools is used. There are three types of schools in the sample: public, private, and experimental schools. Experimental are public schools and school admission is based on a lottery for the sample years. A school is located in an urban area, if the area has more than 20,000 inhabitants. Postcode income is expressed in 2009 Euro.

Table A2: Descriptive Statistics for the Teacher Gender Bias Measured in the First and Second Semester (in current class), Sample of 21 schools

	Bias measured in First Semester		Bias measured in Second Semester			
	Mean	s.d.	Mean	s.d.	Difference	s.e.
			Panel A: 11^{th} grade			
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
Modern Greek	-0.077	0.468	-0.130	0.482	0.053	0.012
History	-0.122	0.426	-0.176	0.427	0.054	0.011
Algebra	-0.135	0.300	-0.107	0.317	-0.028	0.008
Geometry	-0.127	0.315	-0.103	0.307	-0.023	0.008
Physics	-0.115	0.360	-0.099	0.368	-0.016	0.009
Classics						
Ancient Greek	-0.160	0.428	-0.201	0.399	0.041	0.011
Philosophy	0.034	0.472	-0.011	0.437	0.045	0.012
History	-0.132	0.351	-0.145	0.367	0.013	0.009
Science						
Mathematics	-0.096	0.326	-0.056	0.331	-0.040	0.009
Physics	-0.019	0.417	-0.083	0.440	0.064	0.011
Chemistry	-0.003	0.344	-0.061	0.350	0.058	0.009
Exact Science						
Mathematics	-0.100	0.282	-0.095	0.314	-0.004	0.008
Physics	-0.033	0.326	-0.072	0.349	0.039	0.009
Technology and Computers	-0.234	0.462	-0.266	0.426	0.032	0.012
			Panel B: 12^{th} grade			
Core Subjects						
Modern Greek	-0.102	0.480	-0.100	0.493	-0.002	0.008
History	-0.178	0.378	-0.174	0.361	-0.004	0.009
Biology	-0.200	0.666	-0.165	0.787	-0.035	0.013
Mathematics	-0.167	0.498	-0.175	0.535	0.007	0.008
Physics	-0.173	0.445	-0.213	0.441	0.040	0.011
Classics						
Ancient Greek	-0.142	0.429	-0.116	0.431	-0.026	0.007
Latin	-0.162	0.364	-0.151	0.378	-0.011	0.006
Modern Literature	-0.176	0.516	-0.119	0.501	-0.056	0.008
History	-0.226	0.383	-0.247	0.398	0.021	0.006
Science						
Biology	-0.121	0.466	-0.051	0.452	-0.069	0.009
Mathematics	-0.172	0.472	-0.182	0.489	0.010	0.008
Physics	-0.207	0.456	-0.244	0.474	0.037	0.008
Chemistry	-0.124	0.475	-0.130	0.449	0.006	0.009
Exact Classics						
Mathematics	-0.178	0.302	-0.151	0.328	-0.026	0.005
Physics	-0.235	0.340	-0.231	0.328	-0.004	0.005
Business Administration	-0.216	0.391	-0.220	0.391	0.004	0.006
Computer Science	-0.225	0.355	-0.205	0.352	-0.020	0.006
Optional						
Economics	-0.130	0.687	-0.092	0.694	-0.038	0.011

Notes: Columns (1)-(2) present summary statistics for the teacher bias measured based on the first semester exam per subject in 11^{th} and 12^{th} grade. Columns (3)-(4) present summary statistics for the teacher bias measured based on the second semester exam per subject in 11^{th} and 12^{th} grade. We use the second semester exam in the main analysis, as it is taken closer in time to the blind exam. Columns (5) and (6) present the differences between the teacher biases measured based on the first and second semester non-blind exams as well as the respective standard error. The means are weighted by the number of students. The baseline sample is 11^{th} grade students in 2003-2005 and 12^{th} grade students in 2003-2011

Table A3: Correlations between Teachers Biases (in current class) by Subjects (Core) of Instruction, 11^{th} and 12^{th} grades

	Different Teachers					Same Teachers				
11^{th} Grade Biases	History	Geometry	Philosophy (Clas.)	Latin(Clas.)	Physics	History	Geometry	Philosophy (Clas.)	Latin(Clas.)	Physics
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Modern Greek	-0.016		0.201	-0.215		0.876		0.735	0.129	
	(0.106)		(0.148)	(0.128)		(0.212)***		(0.256)**	(0.614)	
N	117		120	121		38		24	21	
Algebra		-0.199					0.827			
		(0.222)					(0.096)***			
N		61					91			
Ancient Greek (Classics)	0.085		-0.022	0.220		0.463		0.567	1.195	
	(0.088)		(0.148)	(0.317)		(0.128)***		(0.138)***	(0.201)***	
N	119		111	87		32		25	43	
Math(Science)		-0.024					0.997			
		(0.113)					(0.193)***			
N		102					30			
Physics (Science)					0.038					0.697
					(0.073)					(0.229)**
N					101					35
Math (Exact Science)		0.214					0.093			
		(0.115)*					(0.139)			
N		101					34			
Physics (Exact Science)					0.008					0.013
					(0.095)					(0.241)
N					113					16

	Different Teachers					Same Teachers				
12^{th} Grade Biases	History	Latin (Clas.)	Literature (Clas.)	Physics	Math.	History	Latin (Clas.)	Literature (Clas.)	Physics	Math.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Modern Greek	0.031	0.060	0.300			0.661	0.301	0.369		
	(0.163)	(0.089)	(0.295)			(0.213)***	(0.295)	(0.210)*		
N	76	195	40			54	40	43		
Ancient Greek (Classics)		-0.063	0.158				0.956	0.619		
		(0.145)	(0.161)				(0.120)***	(0.054)***		
N		173	193				69	51		
History (Classics)	0.030	0.241	0.019			0.205	1.060	0.893		
	(0.059)	(0.109)**	(0.089)			(0.151)	(0.101)***	(0.274)***		
N	83	167	177			39	74	70		
Mathematics (Science)					0.169					0.239
					(0.118)*					(0.149)
N					133					28
Physics (Exact Science)				0.069					0.280	
				(0.103)					(0.092)**	
N				84					33	
Mathematics (Exact Science)					-0.037					0.278
					(0.050)					(0.088)***
N					166					70

Notes: The (OLS) estimated coefficients in columns (1)-(5) are the between biases measures of different teachers who instruct students from the same class in two subjects. The (OLS) estimated coefficients in columns (6)-(10) are the between biases measures of the same teachers who instruct students from the same class in two subjects. The top panel corresponds to 11^{th} grade teachers and the bottom panel to 12^{th} grade teachers. Each estimate comes from a separate OLS regression. Standard errors are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A4: Descriptive Statistics for the Teacher Bias (measured in the current class) in 11^{th} and 12^{th} Grade, Samples of 114 and 21 Schools

	Teacher Bias measured in the current class (114 schools)		Teacher Bias measured in the current class (21 schools)			
	Mean	sd	Mean	sd	Diff.	se
Panel A: 11^{th} grade						
	(1)	(2)	(3)	(4)	(5)	(6)
Core subjects						
Modern Greek	-0.145	0.411	-0.130	0.482	-0.016	0.008
History	-0.104	0.384	-0.176	0.427	0.072	0.007
Algebra	-0.133	0.308	-0.107	0.317	-0.026	0.006
Geometry	-0.138	0.306	-0.103	0.307	-0.035	0.006
Physics	-0.173	0.329	-0.099	0.368	-0.074	0.006
Classics Track						
Ancient Greek	-0.155	0.380	-0.202	0.400	0.047	0.007
Philosophy	-0.195	0.435	-0.011	0.437	-0.184	0.008
Latin	-0.168	-0.519	-0.119	0.501	-0.049	0.006
Science Track						
Mathematics	-0.108	0.356	-0.056	0.331	-0.051	0.007
Physics	-0.127	0.368	-0.083	0.440	-0.044	0.007
Chemistry	-0.140	0.372	-0.061	0.350	-0.079	0.007
Exact Science Track						
Mathematics	-0.170	0.347	-0.095	0.315	-0.075	0.007
Physics	-0.183	0.342	-0.072	0.349	-0.111	0.007
Technology and Computers	-0.366	0.444	-0.266	0.426	-0.100	0.009
Panel B: 12^{th} grade						
Core subjects						
Modern Greek	-0.115	0.445	-0.100	0.493	-0.015	0.005
Biology	-0.222	0.665	-0.165	0.787	-0.057	0.009
History	-0.183	0.380	-0.174	0.361	-0.009	0.007
Mathematics	-0.180	0.525	-0.175	0.535	-0.005	0.006
Physics	-0.271	0.357	-0.213	0.441	-0.058	0.007
Classics Track						
Ancient Greek	-0.117	0.366	-0.116	0.431	-0.001	0.005
Latin	-0.118	0.414	-0.151	0.378	0.033	0.005
Literature	-0.168	0.519	-0.119	0.501	-0.049	0.006
History	-0.149	0.418	-0.247	0.399	0.098	0.005
Science Track						
Biology	-0.057	0.536	-0.051	0.452	-0.005	0.008
Mathematics	-0.139	0.439	-0.183	0.487	0.044	0.006
Physics	-0.213	0.480	-0.244	0.474	0.031	0.007
Chemistry	-0.111	0.483	-0.130	0.449	0.019	0.007
Exact Science Track						
Mathematics	-0.154	0.291	-0.151	0.328	-0.003	0.004
Physics	-0.239	0.324	-0.231	0.328	-0.008	0.004
Business Administration	-0.208	0.346	-0.220	0.391	0.012	0.004
Computers	-0.210	0.325	-0.205	0.356	-0.005	0.004
Optional						
Economics	-0.162	0.638	-0.092	0.694	-0.070	0.008

Notes: This table presents the means, standard deviations and differences of teacher biases measured in the current class for the sample of 114 and 21 schools. The teacher biases are presented for each subject in 11^{th} and 12^{th} grades. The means are weighted by the number of students. A negative bias means that the teacher is pro-girl. The baseline sample is 11^{th} grade students in 2003-2005 and 12^{th} grade students in 2003-2011.

Table A5: Descriptive Statistics for the Teacher Bias (measured in current class) by Teacher Gender, Sample of 21 Schools

	Bias measured in current class					
	Male Teachers		Female Teachers			
	Mean	s.d.	Mean	s.d.	Difference	s.e.
	Panel A: 11^{th} grade					
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
Modern Greek	-0.128	0.532	-0.139	0.402	0.011	0.086
Algebra	-0.163	0.462	-0.230	0.477	0.067	0.078
Geometry	-0.112	0.321	-0.107	0.434	-0.005	0.061
Physics	-0.164	0.296	-0.068	0.368	-0.096	0.055
History	-0.115	0.295	-0.076	0.372	-0.039	0.056
Track: Classics						
Ancient Greek	-0.344	0.479	-0.309	0.435	-0.035	0.139
Philosophy	-0.108	0.434	-0.096	0.712	-0.012	0.163
Latin	-0.154	0.474	-0.262	0.459	0.108	0.141
Track: Science						
Mathematics	-0.152	0.296	-0.085	0.553	-0.068	0.149
Physics	-0.051	0.381	0.007	0.510	-0.057	0.150
Chemistry	0.103	0.486	-0.009	0.551	0.111	0.174
Track: Exact Science						
Mathematics	-0.163	0.462	-0.067	0.411	-0.096	0.141
Physics	0.000	0.283	-0.117	0.429	0.117	0.121
Technology and Computers	-0.347	0.480	-0.173	0.675	-0.174	0.207
	Panel B: 12^{th} grade					
Core Subjects						
Modern Greek	-0.138	0.535	-0.067	0.552	-0.071	0.064
History	-0.207	0.442	-0.161	0.372	-0.046	0.073
Physics	-0.024	0.656	-0.226	0.813	0.202	0.106
Biology	-0.222	0.293	-0.202	0.404	-0.020	0.063
Mathematics	-0.207	0.569	-0.144	0.546	-0.063	0.075
Track:Classics						
Ancient Greek	-0.216	0.539	-0.173	0.597	-0.043	0.116
Latin	-0.137	0.490	-0.174	0.586	0.038	0.116
Literature	-0.183	0.616	-0.201	0.691	0.017	0.138
History	-0.310	0.567	-0.270	0.540	-0.041	0.120
Track: Science						
Biology	0.102	0.504	-0.163	0.452	0.265	0.185
Mathematics	-0.286	0.471	-0.286	0.544	-0.001	0.188
Physics	-0.194	0.400	-0.306	0.390	0.112	0.138
Chemistry	-0.256	0.740	-0.133	0.460	-0.124	0.200
Track: Exact Science						
Mathematics	-0.078	0.400	-0.212	0.516	0.134	0.108
Physics	-0.282	0.586	-0.204	0.458	-0.078	0.111
Bus. Administ.	-0.168	0.612	-0.230	0.543	0.063	0.116
Computers	-0.082	0.449	-0.270	0.493	0.188	0.101
Optional						
Economics	-0.135	0.671	0.063	0.869	-0.198	0.096

Notes: This table presents the means, standard deviations and differences of teacher biases measured in the own class by teacher gender for the sample of 21 schools. These teacher biases by teacher gender are presented for each subject in 11^{th} and 12^{th} grades. A teacher sample is used here. A negative bias means that the teacher is pro-girl. The baseline sample is 11^{th} grade students in 2003-2005 and 12^{th} grade students in 2003-2011.

Table A6: Descriptive Statistics for Teachers Biases by Teacher Gender, Sample of 21 schools

	Bias measured in all other classes					
	Male Teachers		Female Teachers			
	Mean	s.d.	Mean	s.d.	Difference	s.e.
	Panel A: 11th grade					
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
Modern Greek	-0.150	0.402	-0.060	0.303	-0.090	0.065
History	-0.185	0.355	-0.080	0.380	-0.105	0.061
Algebra	-0.149	0.223	-0.163	0.315	0.014	0.044
Geometry	-0.155	0.228	-0.074	0.307	-0.081	0.045
Physics	-0.212	0.215	-0.076	0.308	-0.136	0.046
Track:Classics						
Ancient Greek	-0.216	0.440	-0.305	0.428	0.089	0.134
Philosophy	-0.171	0.389	-0.153	0.411	-0.018	0.124
Latin	-0.137	0.365	-0.183	0.419	0.046	0.119
Track: Science						
Mathematics	-0.148	0.175	-0.006	0.291	-0.143	0.084
Physics	-0.035	0.242	-0.039	0.302	0.004	0.094
Chemistry	-0.154	0.203	-0.080	0.306	-0.075	0.100
Track: Exact Science						
Mathematics	-0.149	0.237	-0.045	0.286	-0.104	0.091
Physics	-0.154	0.404	-0.144	0.239	-0.010	0.102
Technology and Computers	-0.318	0.191	-0.250	0.385	-0.068	0.118
	Panel B: 12th grade					
Core Subjects						
Modern Greek	-0.153	0.351	-0.132	0.292	-0.022	0.040
History	-0.238	0.366	-0.203	0.367	-0.035	0.064
Physics	-0.035	0.526	-0.211	0.487	0.175	0.070
Biology	-0.156	0.235	-0.187	0.356	0.031	0.055
Mathematics	-0.159	0.281	-0.153	0.364	-0.007	0.047
Track:Classics						
Ancient Greek	-0.154	0.455	-0.186	0.394	0.031	0.088
Latin	-0.192	0.453	-0.197	0.397	0.004	0.097
Literature -0.181	0.408	-0.251	0.382	0.070	0.087	
History	-0.192	0.360	-0.244	0.382	0.053	0.080
Track: Science						
Biology	0.068	0.296	-0.287	0.409	0.355	0.156
Mathematics	-0.222	0.218	-0.261	0.421	0.039	0.139
Physics	-0.274	0.239	-0.309	0.262	0.035	0.094
Chemistry	-0.254	0.582	-0.181	0.347	-0.073	0.163
Track: Exact Science						
Mathematics	-0.061	0.383	-0.191	0.358	0.131	0.081
Physics	-0.129	0.311	-0.191	0.308	0.062	0.070
Bus. Administration	-0.167	0.459	-0.130	0.516	-0.037	0.100
Computers	-0.153	0.418	-0.209	0.526	0.056	0.106
Optional						
Economics	-0.122	0.329	-0.041	0.457	-0.081	0.049

Notes: This table presents the means, standard deviations and differences of teacher biases measured in the current class by teacher gender for the sample of 21 schools. These teacher biases by teacher gender are presented for each subject in 11th and 12th grades. A teacher sample is used here. A negative bias means that the teacher is pro-girl. The baseline sample is 11th grade students in 2003-2005 and 12th grade students in 2003-2011.

Table A7: Heterogeneity in the Effect of 11^{th} Grade Gender Bias on Blind 12^{th} Grade Score by the Gender of the Teacher

Dependent Variable: Blind score in 12^{th} grade national exams						
	Boys			Girls		
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
Bias	0.045 (0.040)	-0.022 (0.041)	0.048 (0.036)	0.009 (0.031)	-0.048 (0.032)	-0.121 (0.032)***
Bias × Female Teacher	0.026 (0.056)	0.125 (0.057)	0.083 (0.058)	-0.152 (0.048)	-0.051 (0.049)	0.041 (0.049)
Female Teacher	0.008 (0.017)	-0.031 (0.017)	-0.027 (0.016)	-0.008 (0.014)	-0.028 (0.015)	-0.036 (0.014)
Sample Size	8,226	8,226	8,226	10,369	10,369	10,369
Classics Subjects						
Bias	-0.218 (0.138)	0.043 (0.144)	-0.071 (0.131)	-0.044 (0.070)	0.206 (0.076)**	0.157 (0.086)
Bias × Female Teacher	0.470 (0.179)***	0.145 (0.194)	0.361 (0.169)	-0.197 (0.091)*	-0.422 (0.098)***	-0.294 (0.102)**
Female Teacher	0.056 (0.058)	0.090 (0.058)	0.039 (0.061)	0.020 (0.028)	0.050 (0.030)	0.007 (0.032)
Sample Size	1,401	1,401	1,401	5,499	5,499	5,499
Science Subjects						
Bias	0.256 (0.083)***	0.133 (0.069)*	0.218 (0.083)***	0.073 (0.077)	-0.035 (0.069)	-0.115 (0.083)
Bias × Female Teacher	-0.268 (0.168)	0.074 (0.140)	-0.026 (0.155)	-0.064 (0.150)	-0.041 (0.124)	0.024 (0.139)
Female Teacher	0.052 (0.035)	0.034 (0.038)	0.061 (0.047)	0.073 (0.028)***	0.037 (0.027)	-0.010 (0.036)
Sample Size	4,460	4,460	4,460	4,910	4,910	4,910
Exact Science Subjects						
Bias	0.059 (0.072)	0.001 (0.061)	0.100 (0.052)*	-0.070 (0.113)	-0.073 (0.100)	-0.173 (0.093)*
Bias *Female Teacher	0.048 (0.123)	0.159 (0.110)	0.176 (0.121)	-0.005 (0.154)	0.076 (0.158)	0.035 (0.154)
Female Teacher	0.015 (0.031)	-0.046 (0.029)	-0.012 (0.029)	-0.060 (0.038)	-0.050 (0.042)	-0.026 (0.043)
Sample Size	5,164	5,164	5,164	3,031	3,031	3,031
Subjects FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: Notes: Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The empirical Bayes shrinkage factor is the ratio of signal variance to signal plus noise variance. We assume that there is a sampling error problem and the measure of teacher gender bias consists of an error component. Estimating teachers' effects on students' weighted difference between "non-blind" and "blind" scores enables us to separate between the signal and the noise variance. The empirical Bayes estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the empirical Bayes shrinkage factor. Standard errors are clustered using a two-step bootstrapping method. In the first stage, a random sample with replacement is drawn from each class by gender and the corresponding OLS coefficients are obtained. In the second stage, the effect of these new teachers' gender bias measures in 11^{th} grades on students' performance in 12^{th} grade are estimated and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficients estimates from the second stage regression are the bootstrap standard errors of the point estimates of these coefficients. All specifications include the students' blind score as a control. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the science track subjects. The forth panel "Exact Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A8: Heterogeneity in the Effect of 11^{th} Grade Gender Bias (measured in all other classes) on Blind 12^{th} Grade Score by the Gender of the Principal

Dependent Variable: Blind score in 12^{th} grade national exams		
	Boys	Girls
	(1)	(2)
Core Subjects		
Bias	0.047	-0.070
	(0.074)	(0.057)
Bias × Female Principal	0.013	0.071
	(0.103)	(0.103)
Female Principal	-0.051	-0.008
	(0.036)	(0.031)
Sample Size	5,867	7,675
Classics Subjects		
Bias	0.168	-0.239
	(0.110)	(0.062)***
Bias × Female Principal	-0.185	0.094
	(0.162)	(0.092)
Female Principal	-0.256	-0.076
	(0.077)***	(0.040)*
Sample Size	1,080	4,022
Science Subjects		
Bias	0.002	0.048
	(0.095)	(0.081)
Bias × Female Principal	0.240	0.087
	(0.129)*	(0.120)
Female Principal	-0.106	-0.086
	(0.049)***	(0.045)**
Sample Size	2,918	3,364
Exact Science Subjects		
Bias	0.030	-0.025
	(0.076)	(0.087)
Bias × Female Principal	0.101	-0.000
	(0.098)	(0.136)
Female Principal	0.001	-0.008
	(0.036)	(0.040)
Sample Size	3,660	2,327
Subjects FE	✓	✓
Year FE	✓	✓

Notes: Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. Standard errors are clustered using a two-step bootstrapping method. All specifications include the students' blind score as a control. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the science track subjects. The forth panel "Exact Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A9: Effect of 11^{th} Grade Teacher Bias (measured in the current class) on Blind Score in 12^{th} Grade, Sample of 21 Schools

Dependent Variable: Blind score in 12^{th} grade national exams

	Boys			Girls		
	(1)	(2)	(3)	(4)	(5)	(6)
Core Subjects						
	0.203	0.187	0.200	-0.143	-0.157	-0.153
	(0.048)	(0.048)	(0.049)***	(0.044)***	(0.044)***	(0.047)***
Sample Size	8,384	8,384	8,384	10,563	10,563	10,563
Classics Track Subjects						
	0.323	0.276	0.248	-0.228	-0.193	-0.136
	(0.115)***	(0.101)***	(0.099)***	(0.064)***	(0.064)***	(0.066)**
Sample Size	1,884	1,884	1,884	7,399	7,399	7,399
Science Track Subjects						
	0.191	0.134	0.211	-0.004	-0.114	-0.137
	(0.073)***	(0.068)*	(0.077)***	(0.071)	(0.070)	(0.068)**
Sample Size	4,633	4,633	4,633	5,138	5,138	5,138
Exact Science Track Subjects						
	0.090	0.077	0.159	-0.157	-0.162	-0.098
	(0.049)	(0.050)	(0.045)***	(0.068)**	(0.063)***	(0.056)**
Sample Size	9,315	9,315	9,315	5,368	5,368	5,368
Subjects FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: The datasets for the core subjects and each track subjects include stacked observations for each subject/exam. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The empirical Bayes shrinkage factor is the ratio of signal variance to signal plus noise variance. We assume that there is a sampling error problem and the measure of teacher gender bias consists of an error component. Estimating teachers' effects on students' weighted difference between "non-blind" and "blind" scores enables us to separate between the signal and the noise variance. The empirical Bayes estimate for each teacher is a weighted average of the teacher estimated effect and the mean of teacher estimates, where the weight is the empirical Bayes shrinkage factor. Standard errors are clustered using a two-step bootstrapping method. In the first stage, a random sample with replacement is drawn from each class by gender and the corresponding OLS coefficients are obtained. In the second stage, the effect of these new teachers' gender bias measures in 11^{th} grades on students' performance in 12^{th} grade are estimated and the coefficients are stored. This process of two-step bootstrap sampling and estimation is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficients estimates from the second stage regression are the bootstrap standard errors of the point estimates of these coefficients. All specifications include the students' blind score as a control. All scores are standardized z-scores. The first panel "Core Subjects" includes all core subjects. The second panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The third panel "Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the science track subjects. The fourth panel "Exact Science Subjects" includes relevant exams from the core (Algebra, Geometry and physics) and all the exact science track subjects. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Table A10: Effect of 11^{th} and 12^{th} Grade Own Teacher's Bias (measured in the current class) on the Choice of University Field of Study by Gender, Sample of 21 Schools

Dependent Variable: Dummy variable for the choice of University study						
	(1)	(2)	(3)	(4)	(5)	(6)
		Boys			Girls	
	I. Stack 11^{th} and 12^{th} grades, and add grade FE (2003-2005)					
	-0.007	-0.010	0.008	-0.031	-0.036	-0.040
	(0.017)	(0.019)	(0.025)	(0.017)*	(0.017)*	(0.026)
Sample Size	7,947	7,947	7,947	9,973	9,973	9,973
	II. 12^{th} grade (2003-2011)					
	0.012	0.010	0.003	-0.030	-0.033	-0.048
	(0.019)	(0.019)	(0.027)	(0.019)	(0.019)*	(0.028)*
Sample Size	5,201	5,201	5,201	6,616	6,616	6,616
Year FE	✓	✓	✓	✓	✓	✓
Major FE	✓	✓	✓	✓	✓	✓
Track FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

Notes: Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The dependent variable is the choice to study in Social Science, Science, Exact Science or Humanities. Students not enrolled in any university are not included in the sample. The subjects that we use for each field of study are the following: for exact science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and mathematics and physics in 12^{th} grade. For humanities departments we use the blind score and the bias in history and modern Greek in both 11^{th} and 12^{th} grades. For social science departments we use the blind score and the bias in history and modern Greek in 11^{th}, and economics in 12^{th} grade. For science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and biology in 12^{th} grade. For other departments we use the blind score and the bias in algebra and geometry in 11^{th} grade, and mathematics in 12^{th} grade. The scores are standardized and have a zero mean and a standard deviation of one. *, **, and *** denote significance at the 10%, 5% and 1% level.

Table A11: Effect of 11^{th} and 12^{th} Grade Own Teacher's Bias on the Choice of University Field of Study by Gender, Sample of 21 Schools

Dependent Variable: Dummy variable for the choice of University study						
	(1)	(2)	(3)	(4)	(5)	(6)
		Boys			Girls	
	I. Stack 11^{th} and 12^{th} grade, and add Grade FE (2003-2005)					
	-0.001	-0.000	0.003	-0.029	-0.033	-0.037
	(0.012)	(0.013)	(0.018)	(0.011)**	(0.012)***	(0.018)**
Sample Size	9,998	9,998	9,178	12,547	12,547	12,547
	II. 12^{th} grade (2003-2011)					
	-0.028	-0.031	-0.033	-0.038	-0.045	-0.036
	(0.012)*	(0.012)**	(0.019)	(0.010)***	(0.011)***	(0.017)*
Sample Size	6,721	6,721	6,721	8,537	8,537	8,537
Year FE	✓	✓	✓	✓	✓	✓
Major FE	✓	✓	✓	✓	✓	✓
Track FE	✓	✓	✓	✓	✓	✓
School FE		✓			✓	
Class FE			✓			✓

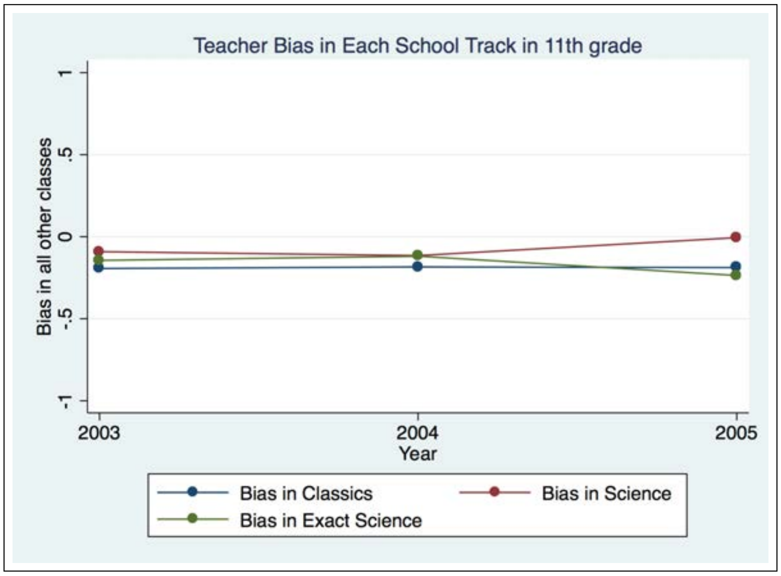
Notes: Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The dependent variable is the choice to study in Social Science, Science, Exact Science, Humanities and vocational departments. Students not enrolled in any university are not included in the sample. The subjects that we use for each field of study are the following: for exact science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and mathematics and physics in 12^{th} grade. For humanity departments we use the blind score and the bias in history and modern Greek in both 11^{th} and 12^{th} grades. For social science departments we use the blind score and the bias in history and modern Greek in 11^{th}, and economics in 12^{th} grade. For science departments we use the blind score and the bias in algebra, geometry and physics in 11^{th} grade, and biology in 12^{th} grade. For other departments we use the blind score and the bias in algebra and geometry in 11^{th} grade, and mathematics in 12^{th} grade. The scores are standardized and have a zero mean and a standard deviation of one. *, **, and *** denote significance at the 10%, 5% and 1% level.

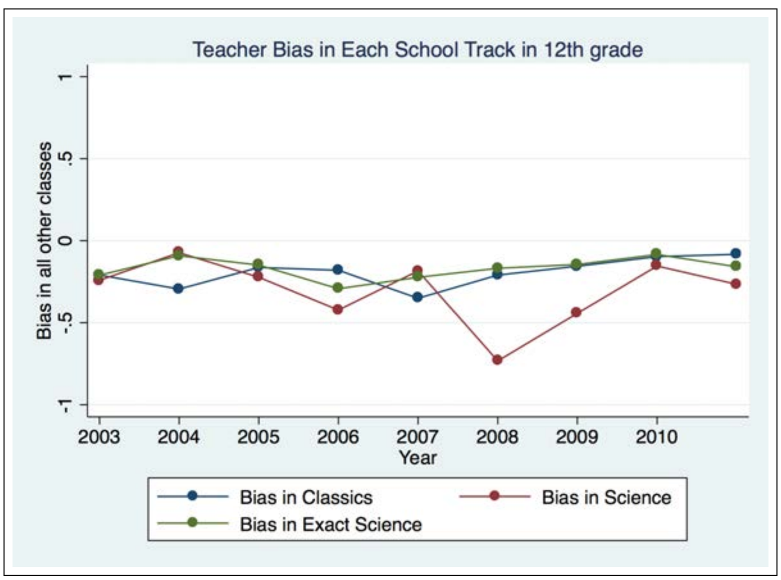
Table A12: Summary Statistics for Sample Used to Estimate Value-Added Models

	Observations	**Mean**	**Std. Dev.**	**Min.**	**Max.**
I. Student Characteristics					
Age (in Years)	50,970	17.363	0.813	16	42
Born in the First Quarter of Birth Year (Yes=1)	50,970	0.240	0.427	0	1
Gender (Female=1)	50,970	0.553	0.497	0	1
Test Scores (s.d.)	50,970	-0.001	0.997	-3.365	2.988
Special Interest in:					
Classics	50,970	0.338	0.473	0	1
Science	50,970	0.235	0.424	0	1
Exact Science	50,970	0.427	0.495	0	1
Number of Subject-School Combinations per Student	50,970	9.766	5.758	1	22
Lagged Test Scores (s.d.)	50,970	0.019	0.965	-6.535	2.867
Missing Lagged Test Score	50,970	0.058	0.233	0	1
II. Class/School/Neighborhood Characteristics					
Class Size	50,970	20.672	4.684	4	31
School Grade Enrollment size	50,970	91.283	46.097	4	187
Neighborhood Income (in Euro)	50,970	19,679	3691	12,266	26,586
Urban Locality (yes=1)	50,970	0.869	0.337	0	1
III. Teacher Characteristics					
Teacher Gender (Female=1)	50,970	0.478	0.500	0	1
No. of other Classes a Teacher Taught (Experience)	20,650	18.331	10.710	2	45

Notes: All statistics reported are for the sample used in estimating the baseline VA model, following the procedure described in Chetty et al 2014. This sample includes only students who have non-missing baseline controls. Student data are from the administrative records of 21 schools in Greece. The sample period of 2003-2005 is used to obtain the VA estimates. All test scores are standardized. Summary statistics (number of observations, mean, s.d., min., and max.) for the baseline variables are reported in the table. Outcome test scores are measured in 11^{th} and 12^{th} grade and the prior test scores are measured in 10^{th} and 11^{th} grade. The variables are weighted by the number of students in the school-year-class-subject-year cell. Only the "Number of other Classes a Teacher Taught " is weighted by the number of teachers in the school-year-class-subject-year cell. The age is measured in years for students the day they start the 11^{th} or 12^{th} grade class. Born in First Quarter of Birth Year is a dummy that takes the value of one if a student is born in the first quarter of the calendar year, and zero otherwise. Students who are born in the first quarter of the calendar year are eligible to enroll a year earlier in the 1st grade. Special interest in Classics, Science or Exact Science is derived by the track they followed in each grade. The school grade enrolment size denotes the number of 11^{th} or 12^{th} graders in a given school and year. The number of subject-school combinations per students is the number of subjects that students take. Each student takes on average 10 subjects. When prior test scores are missing, we set the prior score equal to 0 and include an indicator for missing data. On average, 6% of lagged scores are missing. The total number of observations used here is 50,970. In the last panel that reports teachers' characteristics, a class corresponds to a subject/class/year/grade combination.

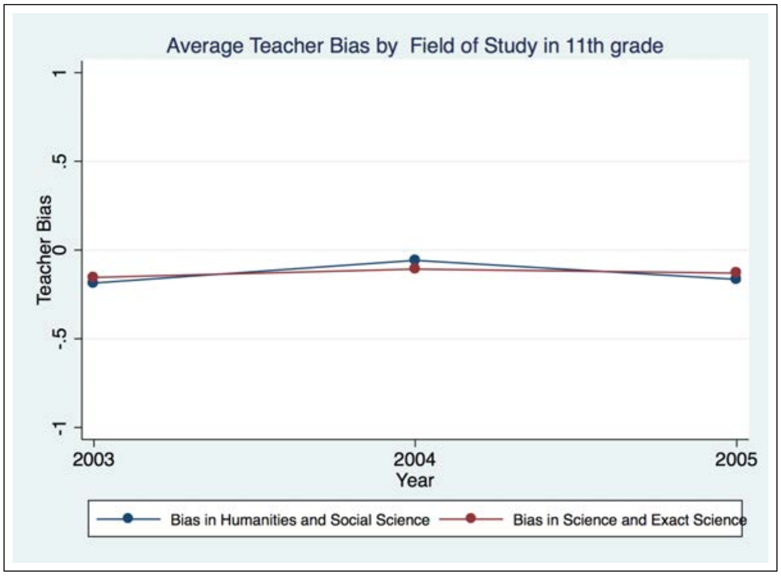
Figure A1: Evolution of Teacher Gender Bias by School Track, Sample of 21 Schools

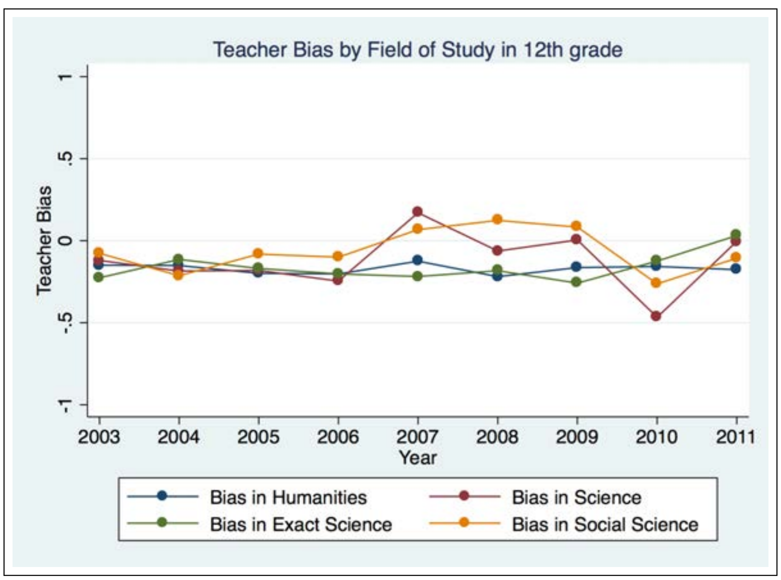




Notes: The 11^{th} grade bias in Classics is the average bias in ancient Greek, philosophy and Latin. The 11^{th} grade bias in Science is the average bias in mathematics, physics and chemistry. The 11^{th} grade bias in Exact Science is the average bias in mathematics, physics and computer science. The 12^{th} grade bias in Classics is the average bias in ancient Greek, Latin, literature and history. The 12^{th} grade bias in Science is the average bias in mathematics, physics,biology and chemistry. The 12^{th} grade bias in Exact Science is the average bias in mathematics, physics, business administration and computer science. All teacher biases that are used here are calculated based on all other classes that a teacher taught in the sample.

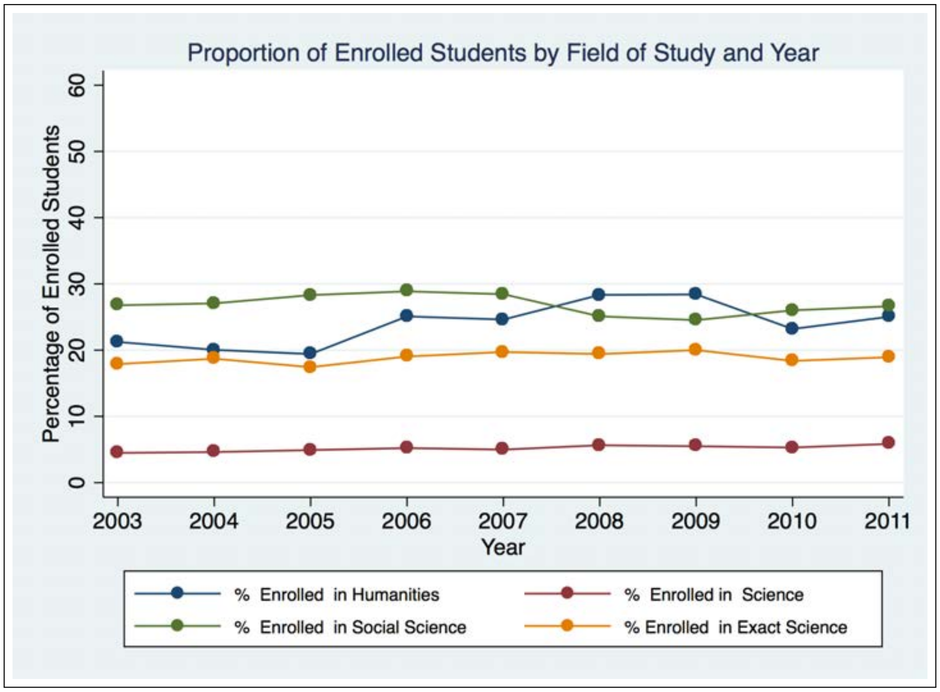
Figure A2: Evolution of Teacher Gender Bias by Field of Study, Sample of 21 schools





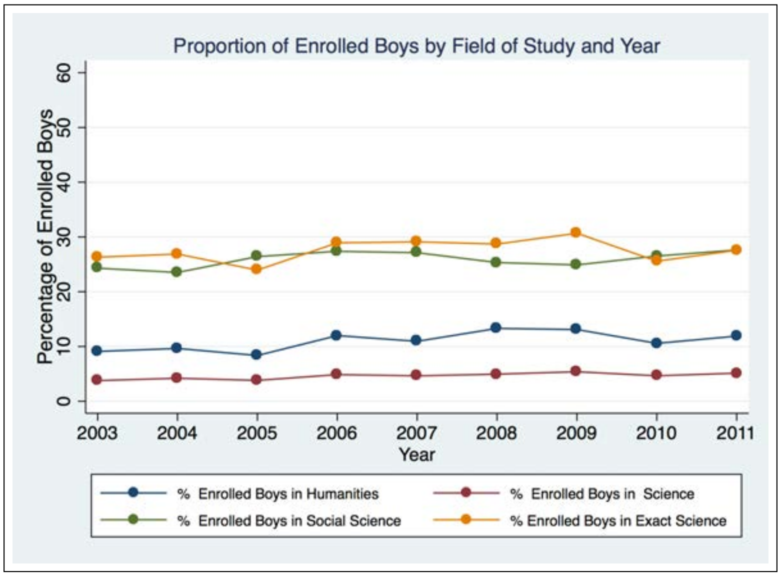
Notes: These figures show the evolution (over time) of the teacher bias in terms of the fields of study at the university level. These fields of study include departments in the following fields: humanities, social science, science and exact science. The 11^{th} grade bias (measured in all other classes) for humanities and social science departments is the average bias in modern Greek and history in 11^{th} grade. The 11^{th} grade bias (measured in all other classes) for Science and Exact Science departments is the average bias in algebra and physics in 11^{th} grade. The 12^{th} grade bias (measured in all other classes) for exact science, science, humanities and social science departments is the bias in mathematics and physics (average), biology, modern Greek and history (average), and economics.

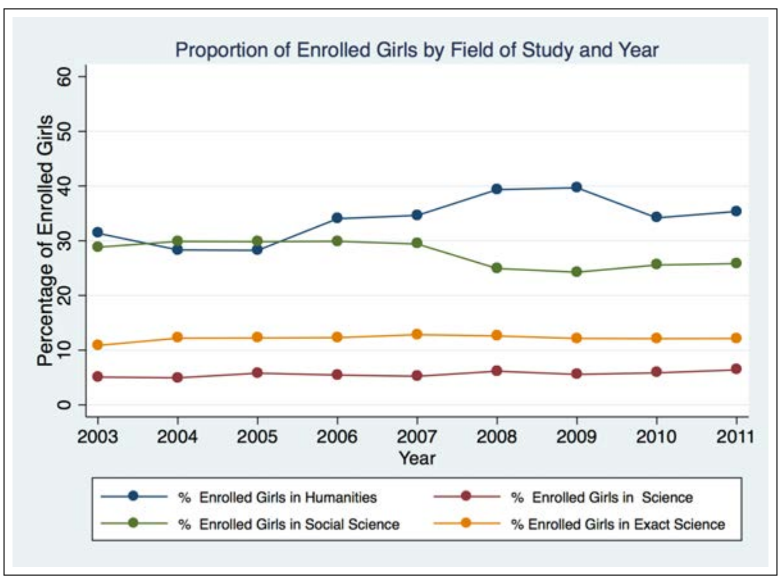
Figure A3: Proportion of University Students Enrolled By Field of Studies, Sample of 21 schools



Notes: This figure presents the proportion of enrolled students in each field of university study and year. We consider four broad categories: humanities, science, social science and exact science. The remaining students enroll in professional studies. Humanities include the departments of Liberal Arts, Psychology, Journalism, Philosophy, Education, Greek Language, History, Foreign Languages, Home Economics and Law. Social Science includes the departments of Economics, Statistics, Business and Management, Accounting, Political and European studies. Exact Science includes the departments of Mathematics, Engineering, Physics and Computer Science. Science includes the departments of Biology, Chemistry, Medicine, Pharmacy, Veterinary Studies and Dentistry.

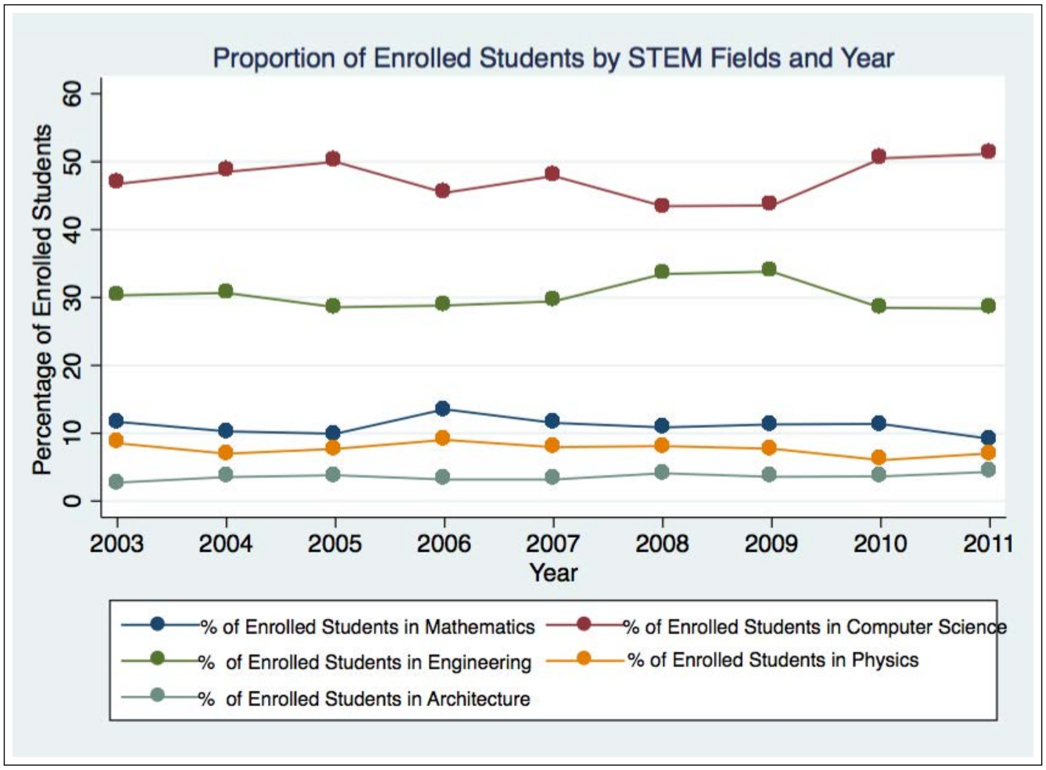
Figure A4: Proportion of University Students, By Field of Studies and Gender, by Year, Sample of 21 schools





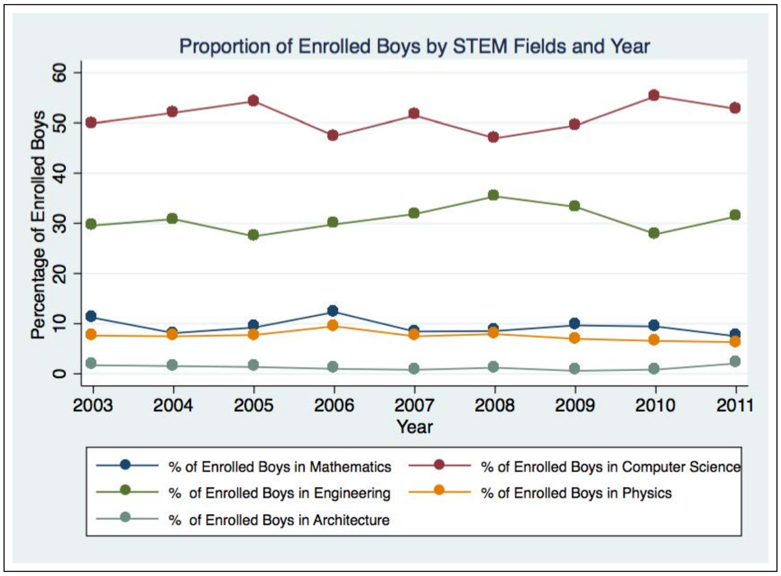
Notes: The top figure presents the proportion of enrolled boys (out of enrolled students) in each field of university study and year. The bottom figure presents the proportion of enrolled girls (out of enrolled students) in each field of university study and year. We consider four broad categories: humanities, science, social science and exact science.

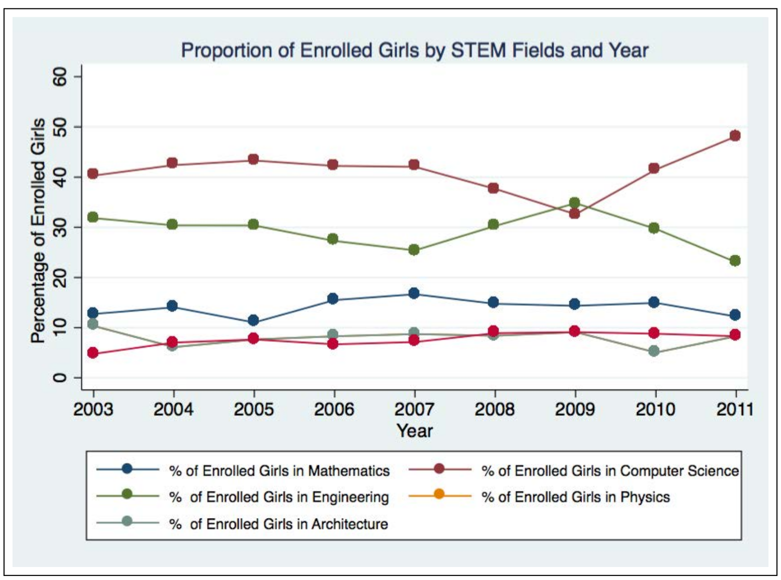
Figure A5: Proportion of University Students Enrolled in STEM by Subject, Sample of 21 schools



Notes: This figure presents the proportion of enrolled students in each stem field and year. We use the following five broad fields: mathematics, engineering, physics, computer science and architecture.

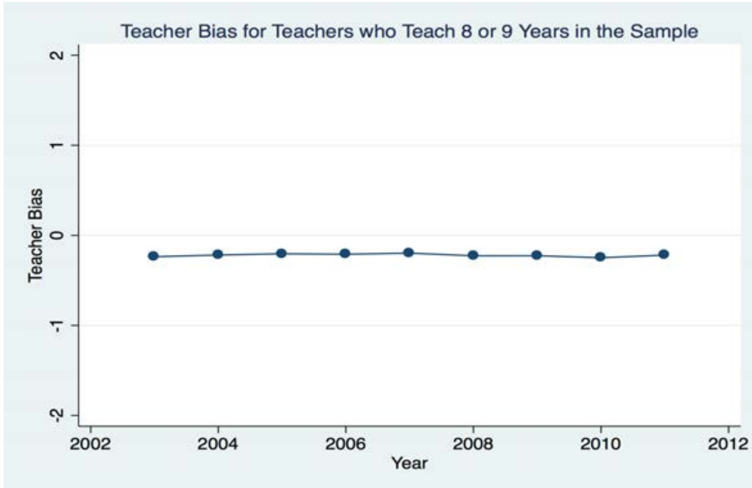
Figure A6: Proportion of University Students in STEM fields by Gender, Sample of 21 schools

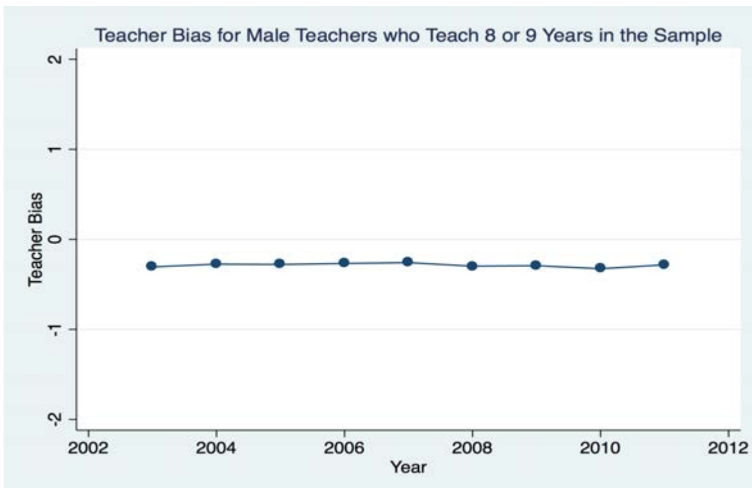


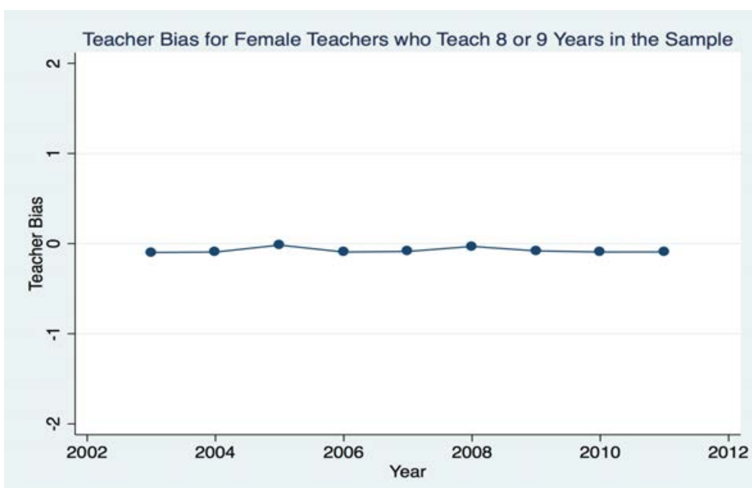


Notes: The top figure presents the proportion of enrolled boys in each STEM field and year. The bottom figure presents the proportion of enrolled girls in each STEM field and year. Five broad fields are used: mathematics, engineering, physics, computer science and architecture.

Figure A7: Evolution of Teacher Bias, All, Male and Female Teachers who Teach 8 or 9 Years in the Sample







Notes: We calculate the annual bias across subjects and classes for all teachers in the sample. We then plot the evolution of teacher bias for male and female teachers who teach 8 or 9 years in the sample. We consider years from 2003 up to and including 2011.