The Short- and the Long-Run Impact of Gender-Biased Teachers

Victor Lavy

Rigissa Megalokonomou

Abstract

We examine the persistence of teachers' gender biases by following teachers over time in different classes. We find a very high correlation of gender biases for teachers across their classes. We find a substantial impact of gender bias on student performance in university admissions exams, choice of university field of study, and quality of the enrolled program. The effects on university choice outcomes are larger for girls, explaining some gender differences in STEM majors. Teachers with lower value-added are also more likely to be gender biased.

Keywords: teacher gender bias, discrimination, university quality, choice of university study, panel information on teachers, teacher value-added

JEL Classification: J24, J21, J16, I24

[1]Victor Lavy: Department of Economics, University of Warwick, Hebrew University of Jerusalem, and NBER (email:victor.lavy@mail.huji.ac.il), Rigissa Megalokonomou: Department of Economics, Monash Business School, CESifo, and IZA (email: rigissa.megalokonomou@monash.edu). This paper was previously circulated under "Persistency in Teachers' Grading Biases and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study". We thank the editor and three anonymous referees for very useful comments and suggestions. The data that support the findings of this study are available from the Ministry of Education and School Authorities but we have no liberty to publish all data online. We also thank seminar participants at the University of Warwick, University of Queensland, Hebrew University, Norwegian School of Economics, University of New South Wales, Queensland University of Technology, Deakin University, University of Sydney, University of Melbourne, Australian National University, Bonn University, University of Naples and Stockholm School of Economics for useful discussions and suggestions. We also thank participants at the 2017 Annual Italian Labor Economics Association conference, 2017 NBER Summer Institute, the 2018 Annual American Economic Association conference, 2018 Applied Econometrics Workshop at the University of Wellington, Royal Economic Society Conference in Sussex, LEER Conference on Education Economics in Leuven, 9th International Workshop on Applied Economics of Education in Catanzaro and XXVII Meeting of the Economics of Education Association in Barcelona. 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 1833757. The data repository is openicpsr-179241 (see Lavy, and Megalokonomou 2023).

1 Introduction

A robust stylized fact established in recent years in many countries is that girls outperform boys in school achievements in primary and secondary school. The gap is larger in school tests that are graded by school teachers and smaller in external exams that are external examiners grade. The gaps in STEM subjects are smaller, often showing boys' advantage. Boys also enroll in college at higher rates in STEM studies, affecting gender occupational differences in the labor market and earnings.[1]

What shapes these gender differences in academic achievements and in university fields of study is the focus of much recent research.[2] There is evidence in psychology and sociology that teachers treat boys and girls differently overall and in math instruction. For example, teachers 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). Additionally, teachers spend more time training girls in reading and less time in math relative to boys (Leinhardt et al. 1979). However, in economics, the evidence is limited.

This paper focuses on how teachers' gender role attitudes and stereotypes influence the gender gap by affecting the school environment. We explore how teachers' gender bias in high school influences students' academic performance in high-stakes exams that determine admission to universities and students' choice of university field of study. We use data from a sample of high schools in Greece, where performance in these exams determines university admission. Our sample includes almost equal proportions of female and male teachers and enables us to distinguish the gender bias by teachers' gender. As has now become a convention, 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 their external exam score in 11^{th} and 12^{th} grade (scored nationally). We then define a teacher bias measure at the class level by the difference between boys' and girls' average gap between the school and national score. Positive values indicate that a teacher is biased in favor of boys.

We follow teachers' teaching assignments over time, on average in 16 classes per teacher, which enables us to assess the persistence of teachers' gender biases. We find a pattern of persistent gender bias among teachers:

[1] 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. 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. Additionally, only 14 percent of engineers in the US are women, though this rate is much higher than in the early 1980s, when only 5.8 percent of engineers in the U.S. were women [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].

[2] Some older 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 (Block 1976; Hoffman 1977; Lewis and Freedle 1972). 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.

those who show gender-biased behavior in a given class reveal the same behavior in their other classes in the same or other academic years. As supporting evidence, we also find that the correlations between teachers' biases in two subjects are significantly higher when the same teacher teaches both subjects than when by two different teachers perform the teaching. This evidence of a "persistent" teacher bias, favoring boys or girls, 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 "out-of-sample" measures (excluding class i) of a teacher's gender bias and estimate its effect on students' performance in class i. This reassures us that our estimates do not reflect random (small sample) variation in boys' unobserved "quality" or non-cognitive skills compared to girls in a particular class or any other class-specific dynamics. We use data for the 2003-2011, consisting of a panel of over 400 teachers from 21 schools. We correct for potential sampling errors that could bias our estimates and standard errors by applying an empirical Bayes technique and a two-step bootstrapping procedure.

We first estimate the effect of 11^{th}-grade teachers' gender bias on students' performance in the national exams at the end of 12^{th} grade. Next, we measure the bias in each subject and then average these results over bundles of subjects, as follows: core subjects that all students are required to study (modern Greek, history, physics, algebra and geometry), classics track subjects (ancient Greek, Latin, literature and history), science track subjects (biology, mathematics, physics, and chemistry), and exact science track subjects (mathematics, physics, computer science and business administration). We also include a pooled specification that includes all available subjects.

We find that the 11^{th}-grade teachers' biases in favor of boys in most groups of subjects positively affect boys' and negatively affect girls' 12^{th}-grade external exam test scores. In terms of standard deviation (SD) of the test score distribution in core subjects, the effect sizes are 0.090 for boys and -0.100 for girls. These effect sizes imply that a 1 SD increase in 11^{th}-grade teachers' bias in core subjects increases boys' test scores in 12^{th} grade by 0.09 SD and reduces girls' test scores by 0.10 SD. The corresponding estimated effects are 0.185 for boys and -0.051 for girls in classics, 0.211 for boys and -0.109 for girls in science, 0.145 for boys and -0.163 for girls in exact science.[3]

In the second part of the paper, we present estimates of the effect of teachers' biases on post-secondary schooling outcomes. We find, for example, that a 1 SD increase in a teacher's bias (in favor of boys) increases

[3]The psychology and sociology literatures provide ample evidence about the potential mechanisms by which teachers' gender-biased attitudes affect students' cognitive and non-cognitive outcomes. Teachers give more attention to boys by addressing them more often in class, giving them more time to respond and providing more substantive feedback (Sadker and Sadker 1985). Teachers treat boys and girls differently in math instruction. They encourage boys to exert independence by not using algorithms, but girls are taught mathematics as a set of rules or computational methods (Hyde and Jaffe 1998). In addition, girls are less likely than boys to be advised, counseled and encouraged to take courses in math (Bae and Smith 1997).

the probability of boys enrolling in any post-secondary schooling by four percentage points, and it lowers that of girls by three percentage points. Teacher bias significantly affects students' probability of enrolling in post-secondary schooling, the quality of the university, and the study program. These effects are similar for boys and girls. However, the effect on the choice of field of study is negative and statistically significant only for girls. A 1 SD increase in a teacher's bias (in favor of boys) in a given subject reduces the probability that a girl will choose that field of study in university by four percentage points. The effect on boys is much smaller and not statistically different from zero.

In the last section of the paper, we examine the association between teachers' gender bias and teachers' effectiveness, using test scores teachers' value-added (TVA) (Rockoff 2004, Chetty et al. 2014a, Chetty et al. 2014b) as a measure of their effectiveness. We believe we are the first to examine the link between the discriminatory behavior of teachers and their effectiveness in the classroom. We find that teachers with no gender bias have significantly higher TVA than pro-boy or pro-girl teachers. This finding is in line with evidence that shows a negative correlation between 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 workplace.

This paper makes four main contributions. The approach of comparing blind and non-blind test scores was introduced in Lavy (2008a) and has since been used in many other studies, for example, Hinnerich, Höglin, and Johannesson (2011), Hanna and Linden (2012), Burgess and Greaves (2013), Cornwell, Mustard, and Parys (2013), Falch and Naper (2013), Botelho, Madeira, and Rangel (2015), Lavy and Sand (2018) and Terrier (2020). In this paper, we address the concern that this measure of teacher bias may pick up differences in unobserved attributes between boys and girls in a particular class. Thus, the estimated gender differences perhaps capture students' behavior and not teachers' behavior. We resolve this concern in this paper by computing the bias based on data from other classes that a teacher taught.

Second, it contributes to the literature on gender differences in STEM majors and careers by linking quantitative measures of teacher biases to students' subsequent academic outcomes. This is among the first papers to establish a reliable causal link between the high school environment and the prevalence of differential gendered outcomes. In addition, having almost an equal proportion of female and male teachers, this is the first study to test heterogeneity in the effect of gender discrimination behavior by the gender of teachers. 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 how math and science courses affect boys' and girls' test scores in middle and high school. Terrier (2020) estimates the effect of teachers' bias similarly to Lavy and Sand (2018) and reports similar results. Teachers' grading bias in favor of boys in French primary schools 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 their mathematical abilities for girls in middle school.[4]

A third important contribution is our focus on high school students and high-stakes exams. The context is very high stakes because the national high school exams are used to determine admission to postsecondary institutions and resolve excess demand for over-subscribed study programs. Therefore, we also examine the effect of teachers' gender biases on university admission and choice of field of study. These are economically meaningful because they affect students' later occupational choices and earnings. While previous studies only suggest that biases can potentially harm students in the long term, we actually demonstrate significant effects on post-high school outcomes. Moreover, the impact on girls of these outcomes is much larger than on boys; this is an important finding because it outweighs the average pro-girls bias among high school teachers. The implication is that gender biases among teachers contribute to gender gaps in the choice of field of study, particularly in STEM, with down-the-road consequences for gender gaps in career and earnings.

Finally, 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 teaching effectiveness, which we measure by their test score value-added. The important implication of our finding is that improving teachers' quality might provide additional benefits in reducing their gender stereotypes.

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 Admission to Greek Universities

Greece's high school and higher education system is highly centralized (OECD 2018), and the Ministry of Education administers university admission. Almost all universities are public, free of tuition fees and admission is based solely on national high school exit exams. Most undergraduate degrees take four years to complete (an exception is the Polytechnic University in Athens, the most prestigious engineering school with a five-year BA degree). Applicants submit a list of their preferred BA programs.[5] Students can apply

[4]Field experiments have also been used to study discrimination (Bertrand and Duflo 2016). Beliefs have been found to affect student achievement. In particular, teachers' beliefs about gender roles are shown to affect student performance (Alan, Ertac, and Mumcu, 2018). Additionally, parental beliefs and the teacher-student gender match may also affect gender differences in scholastic outcomes (Eble and Hu, 2019, 2020).

[5]By program we mean the university and field of study. See Goulas and Megalokonomou (2018) and Goulas, Sofoklis and Megalokonomou, Rigissa and Zhang, Yi (2022) for more details about the admission algorithm.

to several programs, though departments condition eligibility by the high school study track and assign differential weights to specific subjects when calculating the university admission grade. The admission score (cutoff) for each study program is not known to students when submitting their applications.

2.2 High School and National Exams in Greece

Students take standardized national tests by the end of high school. 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 and school exams in 11^{th} and in 12^{th} grade. The weighted average of the national exam scores (70%) and the school exam scores (30%) was used for university admission. However, in 2006 the 11^{th}-grade national exams were cancelled, and since then, university admission has been based only on the test scores in 12^{th}-grade national and school exams.[6] Since then, university admission has been based on their results in national and school exams that students take throughout or at the end of the year in 12^{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 exam scripts and randomly sends them to examiners. Two different examiners grade each exam script. The examiners are teachers who teach the same subject. The school name, student name, and student gender are concealed.[7] Therefore, the national exam score is "blind" because 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, they are "non-blind scores".[8]

All students take the blind exam in a given subject across the country on the same date. The student names and information about their school are covered and thus are unavailable to the external graders. The blind exam is graded by external examiners who teach in a public school in a different school district. This practice eliminates the possibility that teachers will grade their own students' scripts and may thus mark dishonestly. The blind exam is graded in a central state examination centre, and two independent external examiners grade each paper. The final score in the blind exam is the average of the two.

The non-blind exam is usually administered at the end of the school year, only a few weeks before the national exams. The identity and gender of the student are not concealed. The purpose of the school exams is to prepare students for the national exams. The teacher of the student grades the school exam.[9] There

[6]Since 2006, students also take national exams in fewer subjects than previously. Students can select the same optional subjects (for example, economics) as in the pre-2006 period in addition to the compulsory subjects.

[7]Only de-identified numbers are reported for student and school codes.

[8]The school 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.

[9]In each grade, there are multiple classrooms, depending on enrollment. Each classroom has a different combination of teachers. Teachers have specialities and teach subjects only relevant to their specialities. Usually, each subject is taught by a different teacher in each classroom. That means that multiple teachers are teaching in the same subject-by-grade configuration.

is a first and a second term school exam in each subject. We use the second term school exam score for the reasons we describe in Section 2.1. Teachers teaching in each subject-by-grade configuration write the (first and) second-semester school exam questions together (but mark separately). Hence, students across classrooms in the same grade have the same end-of-year school exam questions. In the rare case of more than one second-semester school exam, the average grade of those exams is reported. The school principal must read and approve the school exam questions for each subject in each grade (Government Gazette 2525/1997 A' 188 and 2909/2001 A' 90). The school principal is also responsible for ensuring that the teachers follow the Ministry of Education's grading guidelines for each subject when grading the school exams. The school principal receives the marked exam papers for the school exams from each teacher within five days after the corresponding exam. Then, the regulation requires the principal to read the marked exam papers, approve the marks, write them in the school log, and enter them into the school computer (if available). Through the physical process of reading the exam papers and documenting the marks, the principal ensures they are following the grading guidelines as required by the regulation.

School exams cover identical content and test the same skills and knowledge as the national exams. School exams also have the same format as national exams. Most questions are open in both exams. The duration of the school exam is the same as the duration of the national exam. Both exams take place in students' schools in their regular classrooms. There are no apparent systematic differences in the exam-taking environment that could interact with some characteristics, such as possible higher anxiety levels that may be more pronounced in females. Both this fact and the fact that the exams have the same structure go some way to convince us that the exams test the same skills and cognitive achievements. However, one important difference between the two exams is that all national exams conceal their identity during grading; only their student number appears on the exam notebook. School exams are not anonymous and are graded by the student's teacher, who knows the student's name and gender.

In Figures 1 and 2 in Online Appendix B, we provide an example of a national exam and a school exam from a random public school in 2016. The tested subject is Mathematics in the core. There are four main questions in each exam, with several sub-questions in each question. The distribution of marks across the four questions is the same in both exams. The exam duration is the same in the school and national exams and is 3 hours. Scores range between 0 and 100, with 100 representing the highest performance.

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 covers the same material for all schools, and (2) the second-term school and national exams are administered at the end of the school year, the former before the latter.[10] The school exam grades must be reported to the Ministry of Education before the date of the national exams. School teachers write the school exams, while a national teacher board writes the national exams.

Both scores are reported in the high school graduation diploma, which potential employers sometimes request. Both exams take place in the high school in regular classes; thus, there are no apparent systematic differences in the exam-taking environment. Only a few questions are allowed to be multiple choice, while most questions are open. The purpose of school exams is to prepare students for the national exams. Therefore, the school exams have the same format and content, testing the same skills and knowledge as the national exams.

2.3 Quasi-random Assignment of Students and Teachers to Classes

The assignment of students and teachers to classes within each school is based on a law that states that students should be placed in classrooms based on the lexicographical order of their surname. This leads to a quasi-random match between students and teachers. Students are not allowed to switch classes for any reason and must remain in their assigned class for all grades in high school.[11] We demonstrate in a later section that within schools and across classes, there is no significant variation across classes in students' observed characteristics and abilities.[12]

Assignment of teachers to classes also follows strict regulations. Presidential decree 201 states that the School Board assigns teachers to classes every June before the school year begins. Several rules guide

[10] However, we note that we obtain very similar results when using the first-term school scores. These results are not reported in this paper and are available from the authors.

[11] Students are grouped into classrooms based on the quasi-experimental rule that we describe using last names, and then the same set of students are taught by different teachers throughout the day. However, some subjects, for instance Economics, are optional, and therefore the students enrolled in this subject form a "class" for the economics lesson. Thus, the class composition may vary a bit across subjects. In addition, different teachers teach students assigned to the same classroom throughout the day. Very few students enter or leave a school between 11^{th} and 12^{th} grade. In grade 11, students are also split into tracks, and have different classes for their track subjects.

[12] Students are placed in classes based on their surnames' alphabetical order. This means that students with last name starting with a letter earlier in the alphabet are given a classroom number smaller than the classroom number given to students with last name beginning with a letter later in the alphabet. Assignment based on ability, family background, or other observed characteristics is prohibited. The school principal implements the lexicographic assignment of students to classes, and it is maintained throughout all high school grades. Students are not allowed to switch to another class based on preferences. Law Number 1566 states that schools should be the focal point of integration for students of different backgrounds, gender, and abilities. The same law states that the school should contribute to the "holistic, harmonious and balanced development of the pupils' mental and psychosomatic attributes." The aim is for all students—independent of gender, ethnicity, and ability—-to evolve into complete personalities and develop their skills in a social environment that does not separate students based on any characteristics. This institutional feature of the quasi-random assignment of students and teachers to classrooms has been used in other papers (Goulas, Griselda, and Megalokonomou, 2020a; Lavy and Megalokonomou, 2022; Dinerstein, Megalokonomou, and Yannelis, 2020; Goulas, Griselda, and Megalokonomou, 2021; Kedagni, Krishna, Megalokonomou, and Zhao, 2021). With regards to immigration status, Greece has a low share of immigrants compared with other European countries. In particular, Greece had a 6.79% of immigrants compared to the total population in 2001. Out of them, only 17% were in the 0-14 age group (source: `https://www.migrationpolicy.org/article/greece-history-migration`). Therefore, we are not concerned about students' immigration status being differentially assigned across classes (Einav and Yariv, 2006).

this process. First, it should facilitate teachers' teaching schedules, considering their subject specialization. Several teachers usually teach the same subject in a school, but this also depends on the school enrolment in that particular year. Second, the school should not assign the same teacher to the same class in two consecutive grades. Third, teachers can teach the same class up to twice during the three high school grades. Indeed, in our data, we rarely observe teachers teaching more than twice the same class. Forth, this assignment should be unrelated to teacher status (i.e., permanent/ temporary/hourly teachers) and their teaching experience (years in the profession). In Dinerstein, Megalokonomou, and Yannelis (2020) the authors distinguish between permanent and temporary/hourly teachers and show that their unemployment period before being assigned to their first school and their workload are unrelated to student characteristics in their quasi-randomly assigned classroom for junior temporary and hourly teachers.

The law also states that if there is any disagreement within the school board about the teachers' assignment to classes, then a representative of the School Authority and the School Counsellor has the final word in this decision (Lavy and Megalokonomou, 2022). These institutional rules do not let teachers select classes according to their preferences. The strict implementation of these guidelines is supported by the evidence below of balanced student characteristics and previous test scores against teacher characteristics.[13]

All high schools in Greece offer three study tracks: classics, science and exact science. Students choose one of these at the beginning of 11^{th} grade; most stay on the same track in 12^{th} grade. Each track includes different subjects, which are all compulsory. The curriculum also includes compulsory core subjects, which are the same regardless of the track. Students take national exams in these compulsory subjects as well. Assignment of students to classes within tracks is also based on the same lexicographical rules.

2.4 Data

We use data from various administrative sources for a relatively representative sample of high schools in Greece (Lavy and Megalokonomou, 2023). The sample includes public, private and experimental schools in large and smaller cities and urban and rural areas.[14] In Lavy and Megalokonomou (2022), we provide evidence that compares our sample to the population of high schools in Greece, highlighting the external validity of our sample. The baseline sample includes 11^{th} grade students in 2003-2005 in 21 schools and 12^{th} -grade students in 2003-2011 in the same schools (Greece High School Archives, 2015).[15] To complement our

[13]We note that teachers have no incentives to select a class because they are not evaluated or compensated based on student performance (Dinerstein, Megalokonomou, and Yannelis, 2020; Kedagni, Krishna, Megalokonomou, and Zhao, 2021). Moreover, rankings of schools are not public knowledge, and thus, preferences induced mobility of teachers across schools in the country is uncommon.

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

[15]Figure B3 in the Online Appendix shows the counties where the 21 schools in our sample are located in Greece (Geodata.gov.gr, 2015). Schools in our sample are distributed throughout the country and cover diverse areas. There

primary analysis based on this sample of 21 schools, we also use a sample of 114 more schools with data from 2003-2011. However, the data in this larger sample include only student-level information and not a panel of teaching records of teachers.[16]

The data on teachers allow us to follow them through their teaching history during 2003-2011 in the school they are observed in our data (Greece High School Archives, 2015). The data include the classes, subjects, grades that they taught in each sample year and their gender. Teachers' and principals' gender is inferred from their first names. The data also include a unique student identifier, gender, year of birth, study track, absenteeism records in 11^{th} and 12^{th} grade, test scores from the school, and national exams in all subjects in both grades. The raw exam scores are on a 1-20 scale. To facilitate comparison over time and interpretation of our findings, we transform them into z-scores by year, type of exam and subject.

We obtained records of students' university enrollment, study program (university and department), and university admission scores from the Ministry of Education (Greece Ministry of Education and Religious Affairs, 2011).[17] We use these data to compute each post-secondary program's annual admission cutoff score. We compute two different measures. The first is the mean university admission score of students enrolled in the program. We compute this measure for each year in the period 2003-2011; we use data from the Ministry of Education on all students who applied to each post-secondary institution and program. The second measure is the university admission score of the marginal student admitted to a study program; this is the official study program admission cutoff or the threshold used by the Ministry of Education.

Table A1 in the Online Appendix presents summary statistics for the full sample of 135 schools. The proportion of female students is 56 percent. The average GPA in 11^{th} and 12^{th} grade is 72 and 77 (out of 100), respectively. 92 percent of students attend public schools, 4 percent attend experimental schools, 4 percent attend private schools, and 90 percent live in urban areas. Almost 82 percent enroll in some form of a post-secondary institution. Students apply on average to 25 different programs[18] and on average they enroll in their 8^{th} most preferred program. 82 percent of students gain access to tertiary education. In addition, 15 percent enroll in exact sciences, 4 percent in science, 19 percent in humanities, 22 percent in social sciences and 21 percent in vocational degrees. The remaining 18 percent does not gain access to any postsecondary degree.

Table 1 presents additional summary statistics for the full sample (135 schools) and mean differences

are schools in big cities, including the capital of Athens, and in smaller urban areas and islands. In some counties, there are more than one schools.

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

[17]We note that we do not have access to information regarding students' university degree applications. The university admission score combines students' scores in the school and national exams.

[18]This is equivalent to submitting 25 program applications.

between the samples of 114 and 21 schools samples. The average number of classes is 3.90 in the full sample, 3.90 in the sample of 114 schools and 3.92 in the sample of 21 schools. The average class size is 19-20 students in both grades. Thirty-seven percent of the students are in the classics track in both grades, while a higher percentage of students are in the exact science track than in the science track in both grades (34 vs. 28 in 11^{th} grade and 46 vs. 16 in 12^{th} grade). The differences between the two samples (21 schools versus 114 schools) are small and not statistically significant for some variables.

3 Methodology and Estimation Framework

3.1 Measuring Teacher Gender Biases

Table 2 presents the means of the national and school exams' 2003-2005 11^{th}-grade test scores by gender and their differences. The gender gap varies by subject and type of exam. In the national exams, boys outperform girls in physics, geometry and algebra (core subjects), mathematics and physics (science track), and technology and computers (exact science track). Gender differences in the school exam scores are always in favor of girls. This girls' advantage is evident even in subjects in which boys outscore girls in the national exams.

Table 3 presents the same descriptive statistics for 12^{th}-grade students. The gender gaps in 12^{th} grade follow the same pattern as in 11^{th} grade. The differences between boys' and girls' national scores are statistically significant in most cases, varying from 0.36 SD in favor of girls in Modern Greek to 0.28 in favor of boys in physics (science track). Again, girls have higher scores in the school exams in all subjects.

We construct the teacher bias measure in two steps. We first compute the difference between their school scores and national exam scores for each student in each exam. We then average these differences for boys and girls separately for each class and then compute the difference between 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 school score (NB) and the national 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' gender bias. Positive (negative) values indicate that a teacher favors boys (girls) in this particular subject and class. Figure A1 in the Online Appendix shows the distribution of teacher gender biases.[19] Importantly and uniquely in this study, we use panel data for teachers by class, subject, and year, which allow us to assess how persistent teachers' gender biases are. We show below that

[19]Figures A2 and A3 in the Online Appendix show the distribution of the teacher gender biases for core and track subjects in grades 11 and 12, respectively.

there is significant correlation within teachers in this measure. Therefore, we average the bias using all of a teacher's classes during the study period. However, we 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 effect of the bias (leave-out-means).[20] Naturally, this will be a more accurate measure of a teacher's persistent gender-biased behavior. It alleviates the concern that the bias measure we use picks up class-level unobserved variation in boys' and girls' behavior or other gender-differential non-cognitive characteristics. We view this bias measure as reflecting teachers' perceptions about gender cognitive differences.

One could be concerned that our gender-bias measure is picking up grading practices correlated with gender rather than a gender bias per se. One such concern might be that the school exams may measure somewhat different constructs than national exams, so the bias measure might reflect differences in skills across genders. We provide evidence that the school and the national exams are identical in format, the content of material covered, the type of questions, and skills measured and graded. Another potential concern is that the national exam graders assess different skills or the same skills differently than the graders of the school exam. We note that the graders of national exams are also teachers who grade school exams since they also simultaneously teach in some high schools. So, they are less likely to adopt different grading practices when grading national and school exams. To affirm that our measure is not picking up differential national versus school exam grading practices correlated with gender, we also report evidence from other studies that show a strong positive correlation between our measure of gender grading biases and the implicit association test (IAT). In a recent paper, Avitzour et al. (2020) show a positive correlation between gender discrimination in mathematics grading and assessment and implicit stereotypes. In another recent related study, Alesina et al. (2018) show that teachers give lower grades to immigrant students compared with natives who have the same performance on standardized, blindly graded tests. They then relate differences in grading to teachers' stereotypes elicited through an IAT. They find that teachers with stronger stereotypes give immigrants lower grades than natives of the same performance. In both studies cited above, teachers graded exams more objectively once informed of their stereotypes, as revealed by IAT, suggesting that revealing stereotypes may be a powerful intervention to decrease discrimination. Lavy (2008b) introduced the bias measure that we use in this paper and has shown that it reveals a gender bias in 10 subjects, each subject in exams in three proficiency levels (basic, intermediate, and advanced level). This paper estimates the gender bias for over fifteen subjects and multiple exams that vary by proficiency level. This evidence suggests that the bias measure is not specific to a particular subject, material level, or teaching practice in the courses leading to

[20] For the same reasons, Chetty, Friedman, and Rockoff (2014a) use teacher value-added from other classes and not just the current class.

these national and school exams. However, though unlikely, a remaining teaching practice correlated with gender may lead to a different interpretation of our findings.

In Figure 1, we present the teacher-level distribution of gender bias based on all classes a teacher ever taught. The top panel shows the distribution based on the 11^{th}-grade classes, and the bottom panel is based on 12^{th}-grade classes. The distributions show considerable variation in teachers' gender bias and are very similar. Figures 2 and 3 show the teacher-level distribution of gender bias based on all classes for core and track subjects in 11^{th} and 12^{th} grades, respectively. All these figures indicate that there is considerable variation in teacher gender bias in core and track subjects.

We then show a high correlation between the bias measured in the own class and the empirical Bays of the bias measured in all other classes in 11^{th} and 12^{th} grades in Figures 4 and 5. We do this separately for core, classics, science and exact science subjects. These scatter plots show a highly persistent gender-biased behavior for the two measures, with regression lines having a coefficient very close to 1.

In Table 4, we present descriptive statistics for the average number of times a teacher shows up in the sample of 21 high schools during 2003-2011. There is one row for each teacher-class-grade-year cell. On average, a teacher appears 16 times in the sample, which means they teach 16 unique cells of class-subject-grade-year. We drop teachers who teach only one class in the whole period because we cannot construct a leave-out-mean measure for them.[21]

During 2003-2005, teachers teach on average almost seven different 11^{th} grade classes. On average a teacher teaches 1.5 different subjects per year and 1.7 different classes per year. There is little variation in these statistics from year to year. For example, a 12^{th}-grade teacher appears on average 13 times in our sample over 2003-2011. She teaches 1.6 different subjects per year and 1.8 different classes per year. Twelfth-grade teachers appear on average in 4.4 years in the sample.

In Section 4, 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 test scores) and longer-term outcomes–choice of university field of study and quality of the post-secondary program.

3.2 Evidence on Quasi-random Assignment of Students and Teachers

We present in this section evidence that students are quasi-randomly assigned to classes and teachers. We focus on tests showing balance according to the teacher's gender bias measure. Because we present all of the results separately for boys and girls later, we also show this balance for boys and girls separately.

Table 5 shows that teacher gender biases are uncorrelated with student prior test scores, student age, and class size. Each estimate in this table is generated from a different regression. All six estimated effects are

[21] There are 67 teachers who teach only one class in the sample period.

statistically insignificant and practically zero. In Table 6, we jointly include all three student characteristics in the same regression. At the same time, the outcome variable is the teacher gender bias and we report the F-statistics for the joint significance and the P-value. These estimated effects are statistically insignificant and practically zero, reinforcing the findings of the previous table. This additional evidence further supports the random assignment of students to teachers.

In the Online Appendix Table A2, we show that the remaining teacher characteristics (years of experience, gender, and previous year teacher quality) are uncorrelated with student characteristics (age and 10th-grade test scores) and class size for males and females students, separately.[22] In Table A3, we jointly include all student characteristics and class size in the same regression, while the outcome variable is teacher gender (1=Female) in columns 1-2 or previous year teacher quality in columns 3-4 or teacher experience in columns 5-6. These estimated effects are statistically insignificant and indicate the same pattern as Tables 5-6. All this additional evidence supports the quasi-random assignment of students and teachers to classrooms.

As additional balancing estimation, we also examined the effect of 12^{th}-grade teacher bias on a) 11^{th}-grade national-level scores and b) 11^{th}-grade school-level scores. These results are presented in Tables A4 and A5 in the Online Appendix, respectively, and show estimates for four different specifications (the same we use in the main results). An empirical Bayes estimate is used and the standard errors are corrected using a two-step bootstrapping method that we describe in detail in the following section. None of these estimates are significantly different from zero.

4 Effect of Teacher Biases on High School Outcomes

Since students are randomly assigned to teachers within school and subjects, we structure the empirical design around this randomization. Having multiple grades and years in our data implies that randomization is within a school-by-subject-by-grade-by-year cell. Therefore, the final specification we use in a regression model includes school-by-subject-by-grade-by-year fixed effects. However, we start with specifications that exclude these fixed effects and include subject, year, school, and class fixed effects. We estimate the following model, separately for boys and girls, to obtain the effect of teacher biases in 11^{th} grade on the performance of students on 12^{th}-grade national exams:

$$Y_{icjt} = \alpha + + \beta_{cjt} + \gamma X_{icjt} + \pi TB_{cj} + \psi_{icjt} \quad (2)$$

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; β_{cjt} is a school $\times$ subject $\times$ year fixed

[22] We do not have information about more teacher characteristics, but we note our teacher gender bias measure may be correlated with other teacher characteristics.

effect, and TB_{cj} is the measure of teachers' biased behavior in school (class) c and subject j. There is an individual random element ψ_{icjt}. The coefficient of interest is π, capturing the effect of teacher bias on academic outcomes. We cluster the standard errors at the class level.

We use the following two correction techniques when estimating equation (2): (1) an empirical Bayes (EB) shrinkage estimation approach to address potential sampling error because for some teachers the bias estimates are based on small samples.[23] (2) A two-step bootstrapping technique to account that the main variable of interest is a generated regressor.[24] We note that we use the teachers' bias derived from the empirical Bayes estimates and the two-step bootstrap throughout the paper. We also control for the 11^{th}-grade first-semester school exam test scores and the gender of the teacher in the regressions. The 11^{th}-grade first-semester school exam is the earliest exam for students in grade 11.

Table 7 presents the results of estimating equation (2). The treatment variable of interest is the leave-out-mean gender bias of 11^{th}-grade teachers.[25] We use this measure throughout the paper unless otherwise noted.

[23] We implement an EB shrinkage estimation strategy that accounts for noise in the measurement to deal with the fact that with small samples a few students can have a large impact on test scores (Terrier, 2020; Lavy and Sand, 2018). In particular, to get the EB estimator of teacher gender bias (TB_t^{EB}) for each teacher t, we multiply the estimated (noisy) initial teacher bias measure (TB_t) by an estimate of its reliability (RR_t). This reliability term is the ratio of signal variance [V(θ)] over signal variance plus noise variance [V(ϵ_t)]. In particular,

$$RR_t = \frac{V(\theta)}{V(\theta) + V(\epsilon_t)} = \frac{V(t) - E[V(\epsilon_t)]}{V(t) - E[V(\epsilon_t)] + V(\epsilon_t)}$$

We follow Lavy and Sand (2018) and Terrier (2020) to compute the noise and signal variances. In particular, we estimate–for each teacher–teacher effects on students' differences between blind and non-blind exams using the following regression:

$$Score_{istm} = \gamma_{stm} + \beta_1 Female_i + \beta_2 NB_{istm} + \beta_2 Female_i \times NB_{istm} + \epsilon_{isgct}$$

where $Score_{csg}$ is the score student gets i in school s, grade g (=1 if blind and 0 is non-blind), Female is a binary indicator that indicates whether a student is a female, NB is a binary indicator that is equal to 1 if the exam is non-blind and 0 if it is a blind one, and γ_{stm} is a school $\times$ year $\times$ subject fixed effect. Standard errors are clustered at the student level. For each teacher, we save the coefficient and the standard error of the interaction term, which captures the bias estimate. We calculate the noise variance as the squared standard error of this bias estimate, which is teacher-specific. The signal variance $[V(\theta)]$ is computed by subtracting the mean error variance $E[V(\epsilon_t)]$ (the average of the squared standard error of the estimated teacher gender bias) from the variance of the observed teacher gender bias $[V(t)]$ (Terrier, 2020). In this way, less reliable measures of teacher bias (those with large variation in estimated residuals) are shrunk back towards the mean of the distribution of the teacher gender bias measure. The teacher gender bias measures are normalized to have a zero mean.

[24] This procedure is performed 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; Lavy and Sand, 2018). Bootstrapped standard errors are constructed as follows: In the first stage, a random sample of students is drawn with replacement from each class by the gender of students. Then a new measure of teacher gender bias is calculated using equation (1) and this randomly drawn sample of students. In the second stage, we estimate the effect of this newly created teacher gender bias on students' performance in 12^{th}-grade national exams, and the coefficients are saved. The bootstrapped random samples are used in both parts of the procedure. The two-step bootstrap sampling is repeated 1,000 times. The standard deviations in the sample of 1,000 observations of coefficient estimates from the second step are the bootstrapped standard errors for the estimates of teacher gender biases. These standard errors are reported in all tables.

[25] We match biases in grade 11 to blind performances in grade 12 using the exact same subjects between grades 11 and 12. In the following two cases we match based on closest subjects' correspondences between grades 11 and 12. First, there are no exact correspondences for some subjects (i.e., there is no Biology in grade 11 to match with Biology in grade 12). Second, some subjects are both in the core and the tracks (i.e, Mathematics). Thus, we also consider the additional cases: blind score in Biology in the core in grade 12 against the teacher bias in Physics in the core in

The dependent variable in the regression is the national score in the respective subject in 12^{th} grade and the estimation is based on the sample of 21 schools. We report estimates from four regression specifications: the first includes subject and year fixed effects, the second adds school fixed effects, the third includes a class instead of a school fixed effect, and the fourth includes school $\times$ subject $\times$ year fixed effects. We can use a school fixed effects specification because multiple teachers are in different classes within school-year subject cells. We can also use a class fixed effects specification because classmates have different teachers in different subjects, each with a unique measured bias.

Table 7 Panel A presents results for all teachers, Panel B for teachers of core subjects, Panel C for teachers of classics subjects, Panel D for teachers of science track subjects, and Panel E for teachers of exact science track subjects. All specifications include the students' first-semester test scores in 11^{th} grade and the teacher's gender as controls. Standard errors are adjusted for the two-step bootstrapping technique.

The estimated effects in all four specifications are positive for boys and negative for girls in almost all cases. Across all subjects, a 1 SD increase in 11^{th}-grade teacher bias increases boys' test scores in 12^{th} grade by 0.09 SD and reduces girls' test scores by 0.07 SD in the most augmented specification. For core subjects, the effects are very similar. In particular, a 1 SD increase in 11^{th} grade core subjects' teacher bias increases boys' test scores in 12^{th} grade by 0.10 SD and reduces girls' test scores by 0.05 SD. In classics, the effect for boys is 0.05, while the effect for girls is larger (equal to 0.25) and very precise. In science, the effects are smaller and imprecise for both genders. In exact science, the effect is large and positive for boys at 0.13, while it is smaller and imprecise for girls.[26,27] Estimating these regressions while allowing for an interaction between the teacher bias and student gender yields similar results.[28]

grade 11, blind score in History in the classics track in grade 12 against the teacher bias in History in the core in grade 11, blind score in Ancient Greek in the classics track in grade 12 against the teacher bias in Modern Greek in the core in grade 11, blind score in Latin in the classics track in grade 12 against the teacher bias in Modern Greek in the core in grade 11, blind score in Biology in the science track in grade 12 against the teacher bias in Chemistry in the science track in grade 11, blind score in Mathematics in the science track in grade 12 against the teacher bias in Algebra and Geometry in the core in grade 11, blind score in Biology in the science track in grade 12 against the teacher bias in Algebra and Geometry in the core in grade 11, blind score in Mathematics in the core in grade 12 against the teacher bias in Mathematics in the science or exact science track in grade 11, blind score in Physics in the core in grade 12 against the teacher bias in Physics in the science or exact science track in grade 11, blind score in Business Administration in the exact science track in grade 12 against the teacher bias in Computer Science in the exact science track in grade 11, blind score in Mathematics in the exact science track in grade 12 against the teacher bias in Algebra and Geometry in the core in grade 11.

[26]The estimated effects in Table 7 are unchanged when we replace the teacher's bias measure with one that is based only on classes that the teacher taught in previous years relative to the current one or when we base it only on classes the teacher taught in later years relative to the current class.

[27]As a robustness check for the causal interpretation of our evidence, we reshuffled teacher gender biases within each school and across teachers who teach the same subjects. For instance, a teacher exhibiting positive bias has now been assigned a zero or negative bias. We then re-estimate the effects of teachers' gender reshuffled biases on students' actual performance. The only difference is that we use the bias of teachers who do not teach this particular class but teach the same subject in a different class. 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 suggests that the estimated effects of the correct measure of treatment are not biased due to omitted unobservable confounders of the effect of interest.

[28]We examined whether teachers' gender biases change with years of teaching experience. Figure A4 shows how

As an alternative way of using the track as a benchmark for grouping subjects, we have grouped subjects by STEM and non-STEM subjects. STEM subjects in the core (mathematics and physics) and the science and exact science track subjects are grouped as "STEM". Non-STEM subjects from the core (modern Greek and History) and the classics track are grouped as "Non-STEM". The results are based on these two subsamples in the online appendix Table A6. We use the same four specifications as in the main results. The estimated effects are similar to the main results in Table 7. Male students are positively affected by having a pro-boy teacher in STEM and non-STEM subjects. Female students are negatively affected by having a pro-boy teacher in non-STEM subjects. The effects are less precise for females in STEM subjects, which is not surprising since the estimated effects for females in the science and exact science track were relatively small and insignificant before (using the most augmented specification).

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

In Table A7 in the Online Appendix we present the means and standard deviations of teacher bias by subject for male and female teachers. Teachers of both genders appear to be pro-girl on average in all subjects, as the mean biases are negative. Lavy (2015) has also documented this pattern. In some subjects, male teachers are more pro-girl, while female teachers appear to be more pro-girl in other subjects. In particular, the difference in the average bias is statistically significant in geometry in 11^{th} grade, with male teachers being more "pro-girl" (-0.212 against -0.068 in Table A7) compared with female teachers. This finding aligns with Lavy (2015) and Terrier (2020). Additionally, female teachers in biology (core and science track) in 12^{th} grade are significantly more "pro-girl" based on both measures of teacher biases.

In Table A8, we present estimates that 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 teachers, and we also include the main effect for teachers' gender in the equation. The coefficient of the interaction term in the boys' regression is positive in all four groups of subjects (core, classics, science and exact science) in the most augmented specification, but it is not significantly different from zero. This implies that female teachers' biases have a larger effect on boys than on male teachers, but we do not have enough power to estimate this difference precisely. The effect of female teachers' biases on girls is smaller than that of male teachers' biases. The effect of having pro-boy teachers on girls is negative and significant in science subjects (-0.117). The effect of female teachers is positive but smaller and not significantly different from the effect of male teachers. Figure 6 presents our main measure of teacher bias distribution in each grade for male and female teachers separately.

teacher biases evolve with years of experience for teachers in the sample with 8 or 9 years of experience during the study period. We use the average leave-out-mean bias in all subjects. The bias is stable over time and male and female teachers exhibit the same pattern.

School principals may act as role models for students, but there is a potential for their gender to mitigate perhaps or enhance the effect of teachers' gender biases. In Table A9 in the Online Appendix we allow for teachers' bias effect to vary by the gender of the school principal. Overall, the effect of grading bias on boys is positive and larger in schools with a female principal, while the effect is negative but imprecise for girls. This positive effect on the interaction term for boys is driven by core subjects, while the interactions are insignificant for boys in classics, science and exact science subjects. The gender of the principal in schools with female principals does not seem to affect girls' performance.

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

Table 8 examines whether teachers' biased behavior toward boys and girls affects students' absenteeism. We use three measures of school attendance: total, excused and unexcused hours of absence from school. Excused 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 a school or a student's suspension.

In Table 8, in 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, by type. We use five different bias measures, each based on a different group of subjects: all, core, classics, science and exact science. We include subject, year, and class fixed effects in all specifications. The estimated effects on total and unexcused absences are negative in the boys' regressions and positive in the girls' regressions. This means that a teacher's bias in favor of boys increases boys' class attendance and reduces girls' class attendance. This effect is larger on unexcused absences. For example, a 1 SD increase in 11^{th}-grade bias in core subjects decreases boys' unexcused absences by 0.3 hours and increases girls' unexcused absences by 0.5 hours. Estimates of the effects on girls' unexcused absences are statistically significant for all groups of subjects, except classics. Respective patterns are similar in 12^{th} grade. Boys attend class more often when their teacher exhibits pro-boy behavior, while girls attend class less often when their teacher exhibits pro-boy behavior. A 1 SD increase in 12^{th}-grade exact science teacher's bias reduces boys' unexcused absences in 12^{th} grade by approximately 1 hour, equivalent to a reduction in unexcused absences of around 0.03 SD. These results suggest that the effect of teachers' biases on students' cognitive performance in national exams is partly mediated through increasing or decreasing absenteeism from regular school days, making them miss classes and material covered during these days.[29]

The negative relationship between class attendance and the presence of biased teachers in the school suggests broader effects on student effort and student engagement. This may be a direct effect of grading bias (reduced attendance due to lower grades assigned by the school teacher) or indirect (teachers who are biased graders do other things that discourage students of the negatively impacted gender).

[29] There is growing evidence about the effect of instructional time in school on students' test scores in primary and high school standardized exams. See for example Lavy (2015) and Lavy (2021) and Rivkin and Schiman (2015).

5 Effect of Teachers' Biases on University 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 enrolling in post-secondary institutions. The treatment of interest is the gender bias of high school teachers of core, classics, science and exact science subjects in 11^{th} and 12^{th} grades. Table 9 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.[30] Estimates are shown for 11^{th} and 12^{th} grades, separately. 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' biases in both grades increase the likelihood of boys' enrollment in any university and they have the opposite effect on girls.

For example, a 1 SD increase in 11^{th}-grade teacher bias in all, core, classics, science and exact science increases boys' likelihood of studying in a university by 2, 2, 4, 1, and 1 percentage points, respectively. The effect sizes are similar in grade 12, especially for boys. These effect sizes should be scaled against the mean of 82 percent of all boys who ever attend any post-secondary schooling. The implied effect is that a 1 SD increase in pro-boys behavior of teachers increases post-secondary enrollment for boys by about 5 percent.

At the same time, the effects for girls are negative. If an 11^{th}-grade teacher becomes 1 SD more pro-boy in all, core, classics, science and exact science subjects, girls' likelihood of enrolling in any post-secondary program decreases by 4, 5, 3, 3, and 3 percentage points, respectively. The effects are smaller and less precise for girls in grade 12. Since the percentage of girls who enroll in any post-secondary schooling is identical to that of boys (82 percent), the magnitude of the implied effect of pro-girls behavior by teachers is not much different from that of boys.

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 used for university admission and various study programs. In Table 7 we have shown that teacher gender bias in high school impacts students' performance in subsequent exams. First, higher test scores in these exams enable admission to more highly-demanded universities and fields of study. Second, higher test scores in the national exams may increase students' self-confidence and motivation, increasing 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 the quality of the university where a student is enrolled.

[30] Given our large sample, the estimates from this model are not different from marginal effects obtained from a probit or logit regression model.

The outcome refers to the chosen university field of study and the treatment variable to the teacher gender bias in high school in the respective study track. We group fields of study at the university: Humanities include liberal arts, literature, psychology, journalism, philosophy, education, Greek language, history, foreign languages, home economics and law. Social sciences include economics, statistics, business and management, accounting, political science and European studies. Exact sciences include mathematics, engineering, physics and computer science. Sciences include biology, chemistry, medicine, pharmacy, veterinary studies and dentistry.

Table 10 presents the distribution of students by study area. Among boys, 3.7 percent enroll in science study programs, 22.3 percent in exact science, 21.2 percent in social science and 8.8 percent in humanities. Of the rest, 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.[31] 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 (significantly more boys) and humanities (significantly more girls).[32]

We model students' choices in a linear probability regression in which we stack the six possible university choices as the dependent variable. The treatment variable is the respective high school teacher bias in each of the six areas of high school study. The dependent variable is a 0/1 indicator, equal to 1 for the observed university field of study and 0 for the other five possible choices.[33]

To each university field of study, we assign a teacher leave-out-mean bias measure that is based on the related high school subjects. For exact sciences, we average the 11^{th}-grade biases in algebra, geometry and physics and the 12^{th}-grade biases in mathematics and physics. For sciences, we average 11^{th}-grade biases in algebra, geometry and physics and 12^{th}-grade bias in biology. For humanities, we average 11^{th} and 12^{th}-grade biases in history and Modern Greek. We average 11^{th}-grade biases in Modern Greek and history for social sciences and the 12^{th}-grade bias in economics. For vocational studies, we average the biases in algebra and geometry in 11^{th} grade and the bias in mathematics in 12^{th} grade. For students not enrolled in any university, we use all subjects in grades 11 and 12. Figures A9 and A10 in the online appendix present the evolution of these average bias measures by high school track (Figure A9) and by field of university study (Figure A10).

[31] The gender gap in STEM enrollments in Greece is also documented in Goulas, Griselda, and Megalokonomou (2020b).

[32] Figure A5 presents the proportion of students enrolled in each field of university study by year and Figure A6 presents the proportion of enrolled boys and girls in each field of university study. In Figure A7 we also present the proportion of students enrolled in each STEM field of university study by year and in Figure A8 we present the proportion of enrolled boys and girls in each STEM field of university study.

[33] This model is similar to a multinomial probit or logit and we prefer it given our very large samples which will yield similar estimates from both estimation methods.

We estimate four specifications: The benchmark includes year, major and track fixed effects. A second specification adds school fixed effects. In the third, we replace the latter with class fixed effects. The most augmented specification includes school $\times$ year $\times$ grade fixed effects for Panel A and school $\times$ year fixed effects for Panel B. The reason is that Panel B includes only estimates for a single grade, i.e, grade 12. In all specifications, we also control for the first-semester exam scores in 11^{th} grade. In Table 11 we present the effect of 11^{th} and 12^{th}-grade teachers' biases on 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. 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 the average teacher biases in 11^{th} and 12^{th} grade is small and not significant for boys. In contrast, the estimates for girls are more precisely measured and significantly different from zero. The estimated effect on girls is -0.024 with year and school fixed effects, -0.035 when we replace school fixed effects with class fixed effects and -0.029 when we include school $\times$ year $\times$ grade fixed effects. An 1 SD increase in bias in favor of boys in a given field lowers the probability of girls' choosing that field of study by 2.4, 3.5 or 2.9 percentage points. The respective estimate of the 12^{th}-grade bias is similar. The estimated effect for boys is insignificant, while for girls it is -0.015 when the preferred specification is used. Clearly, the effect of teachers' gender biases on the choice of field of study is much more pronounced for girls and it accounts for some of the gender differences in the choice of field of study. Boys' preferences in the field of studies are more robust and are less or not affected by the gendered biased environment in school, while girls' preferences are more malleable to the gendered environment affected by teachers in high school.

Our findings, namely that teacher biases against girls lower the likelihood that girls are admitted to their most preferred field of study, arise partly because such biases also lower their test scores. However, the gender bias of teachers can also discourage girls from applying to their preferred programs by reducing their self-confidence and beliefs in their ability and success (Kiefer and Shih, 2006; Maass et al., 2008; Heilman, 2012).

In Table 12, we also present the effect of 11^{th} and 12^{th}-grade teachers' biases on the choice of a university department. The structure of this table is similar as to before, but we do not group degrees into broad study fields here. The dependent variable is now the decision to study in one of the following university departments: Biology, History, Mathematics, Physics, Computer Science, Economics, Arts and Language, and Business. We stack each student's university department choices as the dependent variable against the teachers' bias in the most related high school subject. The dependent variable is a binary indicator, assuming the value of 1 for the observed university department and a value of 0 for the other possible choices. We describe the high school subjects used for each university department in the table notes. The estimates reveal the same pattern as that in Table 11. For boys, the effects are small and not different from zero. The estimated effects for girls are negative and statistically significant. In particular, the estimate for girls is -0.025 when class fixed effects are included and -0.026 when school × year × grade fixed effects are included. These estimates mean that 1 SD increase in teacher bias in the closest high school subject increases the probability that female students are more likely to choose a related university department by 2.5-2.6 percent. Boys' decision to choose a particular university department does not seem to be much affected by their teacher gender bias.

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 the post-secondary program of students. We rank universities in each field of study based on students enrolled in the program's average score in national exams. We use the average of this score in 2003 (the first year in our data). Secondly, we use the admissions cutoffs, which we determine using the marginal student enrolled in the program in 2003. We then transform the ranking distribution produced by these two measures into percentile ranks.

In Table A10 in the Online Appendix, we present estimates of teachers' gender bias effects on the percentile and standardized rank of the students' university and department of study. We use the quality of each post-secondary program in these regressions calculated in 2003. We show results separately for all, core, classics, science, and exact science subjects. To measure teachers' gender bias, we use the average bias of teachers in the core subjects closest to each university's field of study.[34]

We find positive estimates for boys (columns 1 and 2) and negative estimates for girls (columns 3 and 4). For example, a 1 SD increase in pro-boy teachers' bias lowers the quality of the enrolled university department of study by 1.8 percentile ranks (column 3) or 6% (column 4) of the quality standard deviation in the admissions cutoff distribution for girls. In classics, the estimate becomes -3.3 for girls. Estimated positive effects for boys indicate that an increase in pro-boy teachers' bias increases the quality of the enrolled university department of study for boys. For example, a 1 SD increase in pro-boy teachers' bias in the related subjects increases the rank of the department of study by 1.7 (column 1) or 6% (column 2) of the quality standard deviation for boys.

In Table 13 we report the effect of teachers' gender bias on the percentile and standardized rank of each student's enrolled post-secondary program by field of study and the likelihood of a student enrolled in a Top 20 or Top 30 % of a university degree. The quality of the post-secondary program is measured based on the 2003 mean performance of enrolled students in each program. The outcome in columns (1) and (5) is the percentile rank, and in columns (2) and (6) is the standardized rank. The first row presents estimated effects when the bias variable is the average in all subjects, the second row the average in classics track subjects, the

[34]In the table notes we describe in detail which subjects we use to calculate the average bias in core subjects that is associated with each field of university study.

third row the average in classics track subjects, the fourth row the average in science track subjects and the fifth row the average in exact science track subjects. Estimates are positive for boys and negative for girls. The effect of the bias in the boys and girls' sample is statistically significant for all subjects. For girls, the effect of the bias is statistically significant for all, core and science track subjects. We also find that being assigned to a pro-boy teacher in core subjects increases boys' likelihood of enrolling in a Top 20% or 30% program by 1.4 or 2 percentage points, respectively. The equivalent effect for girls is a drop in their likelihood of enrolling in a Top 20% or 30% program by 1.8 or 2.2 percentage points, respectively.

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

6 Does the Gender Bias Vary with Teachers' Effectiveness?

The results reported above show that teachers' gender bias affects the subsequent performance of boys and girls during high school and their post-secondary choices and decisions, and the quality of their university schooling. There is also evidence that teacher quality may have long-lasting effects on students' educational decisions and labor market outcomes (Lavy and Megalokonomou, 2022; Chetty, Friedman, and Rockoff, 2014a,b). This section examines whether there is a correlation between teacher gender bias and teacher quality. In particular, we ask the following question: are those teachers who discriminate more likely to be high or low-quality teachers? To explore the relationship between teacher gender biases and teacher effectiveness, we use teachers' value-added (TVA) to measure their quality (Chetty, Friedman, and Rockoff, 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 scores in 10^{th}-12^{th} grades. We compute TVA using the mean performance of each teacher's class. We allow teacher quality to change over time, which is fairly standard in the literature, and thus, we account for drift in teacher quality (Chetty, Friedman, and Rockoff, 2014a). The quasi-random assignment of students and teachers to classes in the Greek high school system guarantees that there is no selection and sorting.[35] We have already provided evidence of the quasi-random assignment of students to teachers in

[35]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

Tables 5, 6, A2 and A3. However, we still control for the student demographics and prior test scores here. Table A11 in the Online Appendix presents summary statistics for the sample used to estimated TVA. It includes controls, such as gender, and an indicator for being born in the first quarter of a calendar year. We control for whether a student is enrolled in a given track (classics, science, or exact science). We pool students' test scores in 11^{th} and 12^{th} grades, and we use the 10^{th} and 11^{th}-grade test scores as a measure of prior test scores.[36]

Thus, TVA is measured in standard deviations in the test score distribution and is estimated using data from all classes taught by a teacher in all years they appear in the data in our study period. Our sample includes only students with non-missing values for the control variables that we use in the baseline value-added model.[37] We show the distribution of TVA in 2003-2005 in the top panel of Figure 7. We then restrict the sample so that TVA is weighted only by the number of teachers (and not students) in the school-year-grade-subject-class-year cell. We also keep only teachers for whom we can measure the gender-biased behavior described below.

We use average teacher gender bias in classes taught during 2006-2011. We restrict the analysis to this period to avoid an overlap between the period we measure TVA and the period we use to estimate the correlation between teachers' gender bias and TVA. This restriction is not a limitation because of the persistent pattern of 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 TVA for pro-boy teachers and those who are pro-girl. We define pro-boy teachers with a gender bias greater than 0.10 and pro-girl teachers with a gender bias smaller than -0.10. We consider teachers with a gender bias larger than -0.10 and smaller than or equal to 0.10 to be neutral in terms of gender bias. Our sample includes 135 pro-boy teachers, 187 pro-girl teachers, and 100 neutral teachers. In Table A12 in the Online Appendix, we present TVA's means and standard deviations for these three groups. Column 3 shows the difference and standard errors of the pairwise differences. Pro-boy and pro-girl teachers have lower TVA, -0.006 (SD=0.153) and -0.045 (SD=0.203), respectively, while neutral teachers have high TVA, 0.035 (SD=0.146). These differences are statistically significant in column 3 between neutral teachers and the other two groups.

In Table A13 in the Online Appendix, we present estimates of teachers' gender bias coefficients from a

to classes.

[36]There are no national exams in 10^{th} grade. Therefore, when we pool students' national exams test scores in 11^{th} and 12^{th} grades, we use the test score in the school exam in 10^{th} grade as the prior year test score.

[37]Our baseline TVA model includes as controls students' demographics, lagged test scores in the same subject, cubic polynomials of lagged test scores in 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), lagged GPA, 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.

TVA regression model. We construct two bias variables as spline 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. We use the EB estimate of teacher gender bias. All regressions include year, school and grade fixed effects. The estimate of the pro-girl bias variable is positive and statistically significant. Changing the specification from column 1 to column 4 by gradually adding control variables (teacher's gender, class size, and years of teaching experience) does not move the point estimate. This means that the lower the bias in favor of girls, the higher the teacher quality. Estimates of the pro-boy bias variable are symmetrical, indicating that a higher grading bias in favor of boys 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 (neutral). These thresholds are somewhat arbitrary, and below, we examine the sensitivity of 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 14. Estimated coefficients of the dummies for pro-girl and pro-boy teachers are negative. The coefficients are more precisely estimated for pro-girl teachers. Pro-girl and pro-boy teachers are associated with a reduction in TVA by 0.04 SD and 0.02 SD, respectively, relative to neutral teachers. Gradually adding controls to the regression leaves the estimates almost unchanged. This provides further evidence that neutral teachers are of higher quality (higher TVA) than pro-boy and pro-girl teachers. These estimates are consistent with the findings we report in Online Appendix Tables A12 and A13.[38].

In contemporaneous work (Lavy and Megalokonomou 2022), we find no statistically significant differences between TVA by teachers' gender. Males and female teachers seem to be, on average, equally productive. Additionally, we find that science teachers are of higher quality than an exact science and classics teachers.

[38]We also compute TVA based on students' long term outcomes, and specifically, the quality of the enrolled university post-secondary degree. In particular, we use the student enrolled post-secondary program quality (percentile rank increasing in quality). Figure A11 in the Online Appendix shows the distribution of TVA based on student program's quality rank. There is considerable variation in teacher quality. Table A14 in the Online Appendix shows comparisons of TVA for neutral teachers and pro-boy or pro-girl teachers. Neutral teachers have a higher TVA than pro-girl or pro-boy teachers. In particular, pro-boy and pro-girl teachers have negative TVA, -0.908 (SD=9.557) and -1.560 (SD=10.077), respectively, while neutral teachers have high TVA, 2.043 (SD=9.672). These differences, presented in column 3 between neutral teachers and the other two groups, are statistically significant. Table A15 examines the relationship between TVA (based on the quality of the enrolled university post-secondary degree) and teachers' grading bias. Our results are less precise now, since 2003 is used to construct the quality of degree, and thus, excluded from the analysis. Overall, the estimated effects point to the same direction as when the TVA is based on test scores (Table 14).

In particular, science teachers have, on average positive TVA, while exact science and classics teachers have negative TVA. We also find that exact science teachers have lower quality than classics teachers and that more experienced teachers have on average, higher TVA.

The evidence in this section is the first we know that attempts to link the gender-biased behavior of teachers to their productivity. Our data do not include teachers' backgrounds, so we cannot identify who the teachers that discriminate except based on TVA are. However, earlier literature on TVA or other quality measures shows that these are not correlated with teachers' education, age, experience, and personal status (Rivkin, Hanushek, and Kain (2005)). Nevertheless, some descriptive studies examine 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 is much less related to implicit 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. Still, 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 this adaptation seems to have bypassed low-income men. This view implies a negative correlation between education and gender-based biased 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 compares boys' and girls' average class test scores in a school exam that the teacher marks versus a national exam graded externally. We use panel data on teachers' class assignment history throughout the study period 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 for which we measure the impact. This approach allows us to estimate the effect of the persistent component of teachers' biases, an endeavor that has not previously succeeded due to the lack of panel data. Observing 16 classes per teacher, 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' gender-biased behavior.

For identification, we rely on the quasi-random assignment of teachers and students to classes in high schools in Greece. We use novel data from a reasonable sample of high schools and compare students exposed to teachers with different patterns of gender-based biases. An important contribution of this paper is the use of gender discriminatory 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 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 TVA 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 some teachers’ high persistent gender-biased behavior. This finding suggests that the biases are deeply rooted, which should be considered in any planned remedial interventions. Second, an increase in teachers’ bias (more pro-boy behavior) in core and track subjects (classics, social science, science, and 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’ biases on national exams are not statistically different. Third, teachers’ biases in core and track courses affect the likelihood that students will enroll in a post-secondary program and the quality of the program in which they enroll.

Additionally, teacher bias affects the related field of study at the university level. Third, 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 TVA) have a neutral attitude towards the two genders; they do not exhibit gender grading biases. This suggests that less effective teachers can harm their students twice, first by being ineffective and second by discriminating against one of the genders. Assuming that the causality runs from a teacher’s quality to her gender bias, it implies that training that improves teachers’ quality will likely also reduce gender-based discrimination in schools.

References

Alan, S., S. Ertac, and I. Mumcu (2018). Gender Stereotypes in the Classroom and Effects on Achievement. *Review of Economics and Statistics 100*(5), 876–890.

Alesina, A., M. Carlana, L. F. Eliana, and P. Pinotti (2018). Revealing stereotypes: Evidence from immigrants in schools. *Working Paper, National Bureau of Economic Research*.

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.

Avitzour, E., A. Choen, D. Joel, and V. Lavy (2020). On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes. *Working Paper,National Bureau of Economic Research*.

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.

Block, J. H. (1976). Issues, Problems, and Pitfalls in Assessing Sex Differences: A Critical Review of the Psychology of Sex Differences. *Merrill-Palmer Quarterly of Behavior and Development 22*(4), 283–308.

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 134*(3), 1163–1224.

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.

Dinerstein, M., R. Megalokonomou, and C. Yannelis (2020). Human Capital Depreciation and Returns to Experience.

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

Eble, A. and F. Hu (2019). How Important are Beliefs about Gender Differences in Math Ability? Transmission Across Generations and Impacts on Child Outcomes. *CDEP-CGEG Working Paper 53*.

Eble, A. and F. Hu (2020). Child Beliefs, Societal Beliefs, and Teacher-Student Identity Match. *Economics of Education Review 77*, 101994.

Einav, L. and L. Yariv (2006, March). What's in a Surname? The Effects of Surname Initials on Academic Success. *Journal of Economic Perspectives 20*(1), 175–187.

Falch, T. and L. R. Naper (2013). Educational Evaluation Schemes and Gender Gaps in Student Achievement. *Economics of Education Review 36*, 12–25.

Geodata.gov.gr (2015). Greece Maps. https://geodata.gov.gr/en/dataset/periphereies-elladas (accessed September 21, 2021).

Goulas, S., S. Griselda, and R. Megalokonomou (2020a). Comparative Advantage and Gender Gap in STEM. IZA Discussion Papers 13313, Institute of Labor Economics (IZA).

Goulas, S., S. Griselda, and R. Megalokonomou (2020b). Comparative Advantage and Gender Gap in STEM.

Goulas, S., S. Griselda, and R. Megalokonomou (2021). Compulsory Class Attendance versus Autonomy. IZA Discussion Papers 14559, Institute of Labor Economics (IZA).

Goulas, S. and R. Megalokonomou (2018). Which Degrees do Students Prefer During Recessions? *Empirical Economics 56*(6), 2093–2125.

Goulas, Sofoklis and Megalokonomou, Rigissa and Zhang, Yi (2022). Female Neighbors, Test Scores, and Careers. *CESifo Working Paper No. 10112*.

Greece High School Archives (2015). Student and Teacher Data, 2001-2011. Authors' personal archive, restricted access.

Greece Ministry of Education and Religious Affairs (2003-2011). Twelfth Grade Student Outcomes, Institute of Educational Policy [distributor]. http://www.iep.edu.gr/en.

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

Heilman, M. E. (2012). Gender Stereotypes and Workplace Bias. *Research in Organizational Behavior 32*, 113–135.

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

Hoffman, L. W. (1977). Changes in Family Roles, Socialization, and Sex Differences. *American Psychologist 32*(8), 644.

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

Kedagni, D., K. Krishna, R. Megalokonomou, and Y. Zhao (2021). Does Class Size Matter? How, and at What Cost? *European Economic Review 133*(C).

Kiefer, A. and M. Shih (2006). Gender Differences in Persistence and Attributions in Stereotype Relevant Contexts. *Sex Roles 54*(11-12), 859–868.

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

Lavy, V. (2008b). Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment. *Journal of Public Economics 92*(10), 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. (2021). 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 (2022). Long-Term Effects of Teachers: Evidence from a Random Assignment of Teachers to Students. *Mimeo*.

Lavy, V. and R. Megalokonomou (2023). Replication Data for: The Short- and the Long-Run Impact of Gender-Biased Teachers, American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. openicpsr-179241.

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.

Lewis, M. and R. Freedle (1972). Mother-infant Dyad: The Cradle of Meaning. *ETS Research Bulletin Series 1972*(1), 1–43.

Maass, A., C. D'Ettole, and M. Cadinu (2008). Checkmate? The Role of Gender Stereotypes in the Ultimate Intellectual Sport. *European Journal of Social Psychology 38*(2), 231–245.

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. (2020). Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement. *Economics of Education Review 77*, 101981.

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

	135 Schools		114 Schools		21 Schools			
	Mean	Std.Dev	Min	Max	Mean	Std.Dev	Diff	s.e.
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
I. 11^{th} **Grade**								
Number of Classes	3.900	1.199	3.897	1.132	3.923	1.580	-0.026	0.014
Class Size	19.424	5.050	19.537	5.073	18.658	4.826	0.880	0.059
School Cohort Size	75.846	31.453	75.868	30.264	75.698	38.581	0.170	0.366
Prop. of Students in Classics	0.366	0.058	0.365	0.058	0.375	0.054	-0.010	0.001
Prop. of Students in Science	0.280	0.074	0.282	0.071	0.265	0.095	0.017	0.001
Prop. of Students in 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
II. 12^{th} **Grade**								
Number of Classes	3.866	1.203	3.868	1.143	3.854	1.546	0.014	0.014
Class Size	19.589	4.959	19.675	4.959	19.007	4.921	0.667	0.058
School Cohort Size	75.855	31.440	75.880	30.258	75.682	38.529	0.197	0.366
Prop. of Students in Classics	0.369	0.060	0.368	0.060	0.376	0.055	-0.008	0.001
Prop. of Students in Science	0.159	0.051	0.158	0.050	0.164	0.056	-0.005	0.001
Prop. of Students in 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

Notes: 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 and the School Exam in 11^{th} Grade 2003-2005

	National Exam						School Exam					
	Boys		Girls		Difference		Boys		Girls		Difference	
	Mean	Std.Dev	Mean	Std.Dev	(1)-(3)	se	Mean	Std.Dev	Mean	Std.Dev	(7)-(9)	se
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
I. Core Subjects												
Algebra	0.032	1.002	-0.026	0.984	0.059	0.013	-0.047	1.018	0.038	0.979	-0.085	0.013
Geometry	0.042	0.993	-0.034	0.990	0.075	0.013	-0.043	1.014	0.035	0.983	-0.078	0.013
History	-0.119	0.958	0.096	1.009	-0.215	0.013	-0.175	1.015	0.141	0.960	-0.316	0.013
Modern Greek	-0.215	0.985	0.174	0.963	-0.389	0.013	-0.274	1.014	0.221	0.927	-0.495	0.013
Physics	0.055	0.998	-0.044	0.985	0.099	0.013	-0.050	1.023	0.041	0.975	-0.091	0.013
II. Classics Track												
Ancient Greek	-0.222	0.995	0.058	0.965	-0.280	0.026	-0.352	1.037	0.093	0.959	-0.444	0.027
Latin	-0.222	0.983	0.058	0.969	-0.280	0.026	-0.366	1.073	0.096	0.947	-0.462	0.026
Philosophy	-0.169	0.958	0.044	0.978	-0.214	0.027	-0.311	1.077	0.082	0.952	-0.392	0.027
III. Science Track												
Chemistry	0.006	0.978	-0.005	0.972	0.011	0.024	-0.052	1.012	0.047	0.969	-0.099	0.024
Mathematics	0.015	0.983	-0.013	0.967	0.028	0.024	-0.047	1.015	0.042	0.966	-0.089	0.024
Physics	0.031	0.974	-0.028	0.974	0.059	0.024	-0.028	1.006	0.025	0.977	-0.053	0.024
IV. Exact Science Track												
Mathematics	-0.054	0.975	0.100	0.981	-0.154	0.022	-0.107	0.997	0.199	0.955	-0.306	0.022
Physics	-0.036	0.982	0.066	0.974	-0.102	0.022	-0.087	1.003	0.161	0.953	-0.247	0.022
Technology and Computers	0.023	0.959	-0.042	1.016	0.065	0.022	-0.092	1.000	0.171	0.957	-0.263	0.022

Notes: This table presents test score gender gaps by type of exam (national and school) and subject in 11^{th} grade. The national and school exam scores are standardized z-scores. A positive difference means that boys outperform girls, while a negative difference means that girls outperform boys. 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 technology and computers. The school 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 and the School Exam in 12^{th} Grade 2003-2011

	National Exams						School Exams					
	Boys		Girls		Difference		Boys		Girls		Difference	
	Mean	Std.Dev	Mean	Std.Dev	(1)-(3)	se	Mean	Std.Dev	Mean	Std.Dev	(7)-(9)	se
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
I. Core Subjects												
Biology	-0.040	0.992	0.024	1.004	-0.064	0.010	-0.166	1.089	0.100	0.906	-0.266	0.010
History	-0.065	0.984	0.052	1.009	-0.117	0.013	-0.166	1.027	0.134	0.940	-0.300	0.013
Mathematics	0.035	1.002	-0.035	0.997	0.070	0.009	-0.054	1.040	0.055	0.930	-0.109	0.009
Modern Greek	-0.215	0.994	0.167	0.957	-0.382	0.008	-0.263	1.034	0.204	0.906	-0.466	0.008
Physics	0.086	1.001	-0.071	0.993	0.156	0.013	-0.057	1.032	0.046	0.955	-0.103	0.013
II. Classics Track												
Ancient Greek	-0.222	1.012	0.059	0.959	-0.281	0.015	-0.297	1.073	0.080	0.934	-0.376	0.015
History	-0.068	0.970	0.018	0.978	-0.086	0.015	-0.189	1.053	0.051	0.950	-0.240	0.015
Latin	-0.239	1.024	0.064	0.954	-0.303	0.015	-0.320	1.079	0.086	0.930	-0.406	0.015
Modern Literature	-0.235	1.019	0.063	0.956	-0.298	0.015	-0.350	1.089	0.094	0.923	-0.444	0.015
III. Science Track												
Biology	-0.011	0.967	0.007	0.940	-0.018	0.019	-0.036	1.003	0.026	0.922	-0.061	0.022
Chemistry	0.062	0.948	-0.043	0.951	0.105	0.019	0.003	0.972	-0.002	0.947	0.006	0.021
Mathematics	0.101	0.960	-0.070	0.939	0.171	0.019	0.022	0.973	-0.016	0.942	0.038	0.020
Physics	0.138	0.949	-0.096	0.941	0.234	0.019	0.022	0.968	-0.015	0.947	0.037	0.020
IV. Exact Science Track												
Business Administration	-0.070	0.980	0.117	0.974	-0.188	0.012	-0.142	1.028	0.237	0.850	-0.380	0.012
Computer Science	0.000	0.995	-0.000	0.960	0.000	0.012	-0.075	1.021	0.124	0.900	-0.199	0.012
Mathematics	-0.031	0.997	0.051	0.953	-0.081	0.012	-0.083	1.012	0.139	0.914	-0.222	0.012
Physics	0.006	1.001	-0.010	0.949	0.016	0.012	-0.076	1.012	0.126	0.916	-0.202	0.012
V. Optional												
Economics	-0.030	0.984	0.028	0.979	-0.058	0.011	-0.100	1.024	0.096	0.931	-0.196	0.011

Notes: This table presents test score gender gaps by type of exam (national and school) and subject in 12^{th} grade. A positive difference means that boys outperform girls, while a negative difference means that girls outperform boys. The school score in each subject is the score in the second term school exam. 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. The estimation is based on the sample of 21 schools.

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

	Mean	Std.Dev	Min	Max
	(1)	(2)	(3)	(4)
No.of Classes/Subjects/Years/Grades Taught by Teacher	15.511	12.919	1	74
11^{th} Grade				
No.of Classes/Subjects/Years Combination Taught by Teacher	6.545	4.540	1	23
No.of Classes/Subjects Taught by Teacher By Year	3.065	1.832	1	10
2003	3.113	1.885	1	10
2004	2.953	1.762	1	8
2005	3.158	1.849	1	9
No.of Different Subjects Taught by Teacher By Year	1.483	0.542	1	3
2003	1.510	0.569	1	3
2004	1.500	0.554	1	3
2005	1.436	0.497	1	2
No.of Different Classes Taught by Teacher By Year	1.697	0.896	1	4
2003	1.716	0.857	1	4
2004	1.566	0.869	1	4
2005	1.829	0.948	1	4
No.of Years a Teacher Teaches	2.142	0.806	1	4
12^{th} Grade				
No.of Classes/Subjects/Years Combination Taught by Teacher	13.255	12.690	1	70
No.of Classes/Subjects Taught by Teacher By Year	3.348	2.560	1	20
2003	3.496	2.041	1	10
2004	3.840	2.371	1	11
2005	3.793	2.457	1	10
2006	4.359	4.941	1	20
2007	2.735	1.838	1	9
2008	2.632	1.735	1	7
2009	2.642	1.763	1	7
2010	2.648	1.612	1	6
2011	2.829	1.909	1	8
No.of Different Subjects Taught by Teacher By Year	1.628	0.861	1	6
2003	1.685	0.751	1	4
2004	1.844	0.947	1	6
2005	1.680	0.759	1	5
2006	2.050	1.465	1	6
2007	1.336	0.512	1	3
2008	1.373	0.520	1	3
2009	1.467	0.652	1	3
2010	1.430	0.665	1	3
2011	1.427	0.700	1	4
No.of Different Classes Taught by Teacher By Year	1.778	1.136	1	7
2003	1.835	1.114	1	6
2004	1.954	1.229	1	7
2005	2.089	1.391	1	7
2006	1.851	1.212	1	6
2007	1.555	0.787	1	4
2008	1.561	0.925	1	5
2009	1.492	0.833	1	5
2010	1.508	0.877	1	5
2011	1.614	1.046	1	5
No.of Years a Teacher Teaches	4.347	2.324	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: Random Assignment: Balancing Test of Student Characteristics and Teacher Gender Bias by Student Gender

	Females			Males		
	Student Prior Performance	Student Age	Class Size	Student Prior Performance	Student Age	Class Size
Teacher Gender Bias	0.029	0.086	0.151	-0.034	0.098	0.409
	(0.036)	(0.083)	(0.188)	(0.049)	(0.088)	(0.299)
Observations	10,351	10,351	10,351	7,622	7,622	7,622
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table presents the estimated effects from separate regressions of each of the student pre-assignment characteristics and prior test scores on teacher gender bias. Each estimate in this table is generated from a different regression. The scores are standardized and have a zero mean and a standard deviation of one. All regressions condition on subject fixed effects, year fixed effects, grade fixed effects, and class fixed effects. Robust standard errors clustered at the class level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 6: Random Assignment: Balancing Test of Teacher Gender Bias and Student Characteristics by Student Gender

	Females	Males
	(1)	(2)
Lagged Test Score	0.002	-0.002
	(0.003)	(0.004)
Age	0.014	0.012
	(0.014)	(0.011)
Class Size	0.005	0.012
	(0.006)	(0.008)
Observations	10,351	7,622
Mean of Y	-0.04	-0.04
Grade FE	Yes	Yes
Year FE	Yes	Yes
Class FE	Yes	Yes
F-Statistic	0.69	1.23
P-value of F-model	0.56	0.30

Notes: All estimates in each column are generated from the same regression. The table reports OLS estimates from separate regressions of teacher gender bias on student lagged test score, age and class size, by student gender. Estimated effects for female students are shown in column (1) and for male students in column (2). All regressions include class fixed effects, year fixed effects and grade fixed effects. Robust standard errors clustered at the class level are reported in parentheses.

Table 7: Effect of 11^{th} Grade Teacher Bias on National Score in 12^{th} Grade

Dependent Variable: Test score in 12^{th}-grade national exams								
	Boys				Girls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: All Subjects								
	0.072	0.064	0.067	0.090	-0.057	-0.064	-0.093	-0.067
	(0.034)	(0.032)	(0.032)	(0.047)	(0.029)	(0.028)	(0.028)	(0.042)
Sample Size	18,503	18,503	18,503	18,474	21,119	21,119	21,119	21,111
Panel B: Core Subjects								
	0.131	0.094	0.095	0.101	-0.007	-0.049	-0.097	-0.052
	(0.043)	(0.042)	(0.041)	(0.047)	(0.037)	(0.036)	(0.037)	(0.041)
Sample Size	8,241	8,241	8,241	8,241	10,380	10,380	10,380	10,373
Panel C: Classics Subjects								
	0.066	0.103	0.018	-0.060	-0.180	-0.095	-0.055	-0.249
	(0.097)	(0.085)	(0.066)	(0.189)	(0.049)	(0.045)	(0.039)	(0.087)
Sample Size	1,381	1,381	1,381	1,343	5,462	5,462	5,462	5,462
Panel D: Science Subjects								
	0.172	0.173	0.179	0.039	0.121	0.058	-0.009	0.068
	(0.088)	(0.085)	(0.091)	(0.128)	(0.074)	(0.070)	(0.075)	(0.112)
Sample Size	4,467	4,467	4,467	4,461	4,913	4,913	4,913	4,903
Panel E: Exact Science Subjects								
	0.096	0.062	0.095	0.066	-0.088	-0.066	-0.177	0.118
	(0.063)	(0.060)	(0.065)	(0.094)	(0.072)	(0.072)	(0.078)	(0.115)
Sample Size	8,999	8,999	8,999	8,998	5,130	5,130	5,130	5,126
Subject FE	✓	✓	✓		✓	✓	✓	
Year FE	✓	✓	✓		✓	✓	✓	
School FE		✓				✓		
Class FE			✓				✓	
School-by-Subject-by-Year FE				✓				✓

Notes: The datasets for the core subjects and each of the 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 standard errors are also corrected using a two-step bootstrapping method. All specifications include the students' 11^{th} grade first-semester school exam scores and the teacher's gender as controls. All scores are standardized z-scores. The first panel "All Subjects" includes all core and track subjects. The second panel "Core Subjects" includes all core subjects. The third panel "Classics Subjects" includes relevant exams from the core (history and modern Greek) and all the classics track subjects. The fourth panel "Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the science track subjects. The fifth panel "Exact Science Subjects" includes relevant exams from the core (algebra, geometry and physics) and all the exact science track subjects.

Table 8: Effect of 11^{th} and 12^{th} Grade Gender Biases 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 All Subjects	-0.273	0.024	0.006	-0.332	-0.278	0.356	-0.304	0.011	0.474	0.185	-0.688	-0.108
	(0.208)	(0.200)	(0.122)	(0.157)	(0.144)	(0.141)	(0.199)	(0.073)	(0.229)	(0.224)	(0.276)	(0.176)
Sample Size	9,612	12,049	9,612	12,049	9,612	12,049	4,306	4,935	3,933	4,508	4,288	4,894
Bias in Core Subjects	-0.201	0.060	0.062	-0.471	-0.263	0.531	-0.163	0.044	0.401	0.291	-0.464	-0.117
	(0.234)	(0.241)	(0.153)	(0.176)	(0.164)	(0.160)	(0.171)	(0.100)	(0.284)	(0.325)	(0.354)	(0.216)
Sample Size	5,975	7,589	5,975	7,589	5,975	7,589	2,388	2,782	2,211	2,573	2,380	2,762
Bias in Classics Subjects	-0.396	-0.161	0.054	-0.354	-0.450	0.193	-0.051	-0.037	0.401	-0.237	-0.542	0.127
	(0.341)	(0.249)	(0.194)	(0.195)	(0.259)	(0.199)	(0.246)	(0.135)	(0.294)	(0.403)	(0.187)	(0.289)
Sample Size	2,921	5,002	2,921	5,002	2,921	5,002	1,629	2,306	1,519	2,149	1,622	2,293
Bias in Science Subjects	-0.240	0.233	0.025	-0.180	-0.240	0.414	-0.009	-0.001	-0.010	0.199	-0.542	0.127
	(0.143)	(0.110)	(0.050)	(0.086)	(0.150)	(0.128)	(0.300)	(0.124)	(0.356)	(0.493)	(0.187)	(0.289)
Sample Size	5,887	7,113	5,887	7,113	5,887	7,113	5,955	7,075	5,503	6,521	5,945	7,032
Bias in Exact Science Subjects	-0.188	0.252	-0.006	-0.200	-0.182	0.452	-0.231	-0.071	0.709	-0.106	-0.866	0.102
	(0.179)	(0.126)	(0.044)	(0.103)	(0.167)	(0.143)	(0.256)	(0.214)	(0.353)	(0.458)	(0.367)	(0.247)
Sample Size	7,070	7,057	7,070	7,057	7,070	7,057	7,147	7,022	6,690	6,474	7,130	7,000
Subject 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 variables are 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 total number of absences in a year equal the excused number of absences in a year and the unexcused number of absences in a year. 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. In the first panel all core and track subjects are used. In the second panel all core subjects are used. The third panel includes only classics subjects. The fourth panel includes only science subjects. The fifth panel includes only exact science subjects. The scores are standardized z-scores.

Table 9: Effect of 11^{th} and 12^{th} Grade Gender Biases 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 in All Subjects	0.021	-0.039	0.022	0.010
	(0.007)	(0.006)	(0.010)	(0.010)
Sample Size	11,348	14,100	5,240	6,049
Bias in Core Subjects	0.020	-0.047	0.021	0.005
	(0.008)	(0.007)	(0.014)	(0.013)
Sample Size	7,000	8,809	2,857	3,358
Bias in Classics Subjects	0.040	-0.027	0.021	0.003
	(0.012)	(0.012)	(0.025)	(0.022)
Sample Size	3,412	5,847	1,951	2,821
Bias in Science Subjects	0.013	-0.028	0.002	0.012
	(0.007)	(0.007)	(0.015)	(0.015)
Sample Size	5,415	6,558	1,516	1,777
Bias in Exact Science Subjects	0.008	-0.034	0.024	0.004
	(0.010)	(0.010)	(0.016)	(0.015)
Sample Size	6,264	6,456	2,264	1,852
Subject FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Class FE	✓	✓	✓	✓

Notes: The outcome variable is the post-secondary enrollment status (1 if enrolled, 0 otherwise). We report the estimated effects of the high school bias in all, core, classics, science and exact science subjects in 11^{th} and 12^{th} grades, separately. In these regressions, we also control for the national 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 of the 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 standard errors are also corrected using a two-step bootstrapping method. 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.

Table 10: Descriptive Statistics by University Field of Studies 2003-2011

	Mean Enrollment											
	Girls				Boys							
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev	Diff. (1)-(5)	s.e.	Diff. (3)-(7)	s.e.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Exact Science	0.099	0.299	0.121	0.327	0.223	0.416	0.272	0.445	-0.124	0.003	-0.151	0.003
Science	0.046	0.209	0.056	0.230	0.037	0.189	0.045	0.208	0.009	0.002	0.011	0.002
Social Science	0.225	0.418	0.276	0.447	0.212	0.409	0.259	0.438	0.014	0.003	0.018	0.004
Humanities	0.275	0.446	0.336	0.473	0.088	0.284	0.108	0.310	0.186	0.003	0.228	0.004
Vocatonal Studies	0.171	0.377	0.210	0.407	0.258	0.438	0.315	0.465	-0.087	0.003	-0.105	0.004
Not Enrolled in Post-Secondary Studies	0.184	0.388	0.000	0.000	0.181	0.385	0.000	0.000	0.003	0.003	0.000	0.000

Notes: The sample includes 37,218 female students and 28,869 male students. In columns 3-4 and 7-8 we restrict the sample only to students who enroll in university studies. Humanities include the departments of liberal arts, literature, 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 science 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 11: 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)	(7)	(8)
	Boys				Girls			
A. Stack 11^{th} and 12^{th} grades & Grade FE. (2003-2005)								
Teacher Bias	0.000	0.000	0.004	0.000	-0.021	-0.024	-0.035	-0.029
	(0.010)	(0.011)	(0.016)	(0.014)	(0.008)	(0.009)	(0.014)	(0.012)
Sample Size	15,400	15,400	15,400	15,400	19,182	19,182	19,182	19,182
B. 12^{th} grade (2003-2011)								
Teacher Bias	-0.008	-0.008	-0.011	-0.009	-0.012	-0.013	-0.022	-0.015
	(0.010)	(0.011)	(0.018)	(0.013)	(0.008)	(0.009)	(0.014)	(0.010)
Sample Size	11,806	11,806	11,806	11,806	15,366	15,366	15,366	15,366
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Major FE	✓	✓	✓	✓	✓	✓	✓	✓
Track FE	✓	✓	✓	✓	✓	✓	✓	✓
School FE		✓				✓		
Class FE			✓				✓	
School × Year × Grade FE				✓				✓

Notes: This table shows the estimated effects of teacher gender bias in high school subjects on the related field of study at the university level. The dependent variable is the decision to study in one of the following six fields: Social Sciences, Sciences, Exact Sciences, Humanities, vocational departments or not attending any post-secondary education. We stack the six possible choices as the dependent variable for each student against the teachers' bias in each of the four areas of high school 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 treatment variable is the respective high school teacher bias in each of the four areas of high school study, as they are described below. The high school subjects that we use for each university field of study are the following: for exact science departments we use 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 bias in history and modern Greek in both 11^{th} and 12^{th} grades. For social science departments we use the bias in history and modern Greek in 11^{th}, and economics in 12^{th} grade. For science departments we use the bias in algebra, geometry and physics in 11^{th} grade, and biology and physics in 12^{th} grade. For vocational departments, we use the bias in algebra and geometry in 11^{th} grade and mathematics in 12^{th} grade. For students not enrolled in any university, we use the bias across all subjects in 11^{th} and 12^{th} grades. The scores are standardized and have a zero mean and a standard deviation of one. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The standard errors are also corrected using a two-step bootstrapping method. The first panel includes school × year FE , while the second panel includes × year × grade FE in the most augmented specification. Estimation is based on the sample of 21 schools.

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

Dependent Variable: Dummy variable for the Choice of University Department								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Boys				Girls			
	A. Stack 11^{th} and 12^{th} grades & Grade FE. (2003-2005)							
Teacher Bias	0.002	0.006	-0.000	-0.004	-0.018	-0.022	-0.025	-0.026
	(0.006)	(0.006)	(0.007)	(0.008)	(0.008)	(0.008)	(0.010)	(0.012)
Sample Size	17,817	17,817	17,817	17,798	19,947	19,947	19,947	19,927
	B. 12^{th} grade (2003-2011)							
Teacher Bias	-0.005	-0.001	-0.006	-0.008	-0.017	-0.016	-0.020	-0.023
	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)	(0.009)	(0.011)
Sample Size	15,577	15,577	15,575	15,548	18,656	18,656	18,655	18,614
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Major FE	✓	✓	✓	✓	✓	✓	✓	✓
Track FE	✓	✓	✓	✓	✓	✓	✓	✓
School FE		✓				✓		
Class FE			✓				✓	
School × Year × Grade FE				✓				✓

Notes: This table shows the estimated effects of teacher gender bias in high school subjects on the related university department. The dependent variable is the decision to study in one of the following university departments: Physics, History, Economics, Mathematics, Biology, Computer Science, Chemistry, Arts and Language, and Business department. We stack the university department choices as the dependent variable for each student against the teachers' bias in each of the related subjects of high school studies. The dependent variable is a 0/1 indicator, assuming the value of 1 for the observed university department and a value 0 for the other possible choices. The treatment variable is the respective high school teacher bias in each of the subjects of high school study, as they are described below. The high school subjects that we use for each university department of study are the following: for physics departments we use the bias in physics in the core and tracks (Science and Exact Science) in grades 11 and 12. For history departments we use the bias in history in the core and the classics track 11^{th} and 12^{th} grades. For economics departments we use the bias in economics in 12^{th} grades. For mathematics departments, we use the bias in algebra and geometry in the core and mathematics in the tracks (Science and Exact Science) in 11^{th} grade and mathematics in the core and the tracks (Science and Exact Science) in 12^{th} grade. For biology departments we use the bias in biology in the core and the science track in 12^{th} grades. For computer science departments we use the bias in technology and computers in the exact science track in grade 11 and computer science in the exact science in grade 12. For chemistry departments we use the bias in chemistry in the core and the science track in grade 11 and the bias in physics in grade 12. For arts and language departments we use the bias in modern Greek and history in grades 11 and 12. For business departments we use the bias in economics and modern Greek in the core in grade 12 and the bias in modern Greek in the core in grade 11. Each row presents estimates from a separate regression using an empirical Bayes estimation strategy. The standard errors are also corrected using a two-step bootstrapping method. The first panel includes school × year FE , while the second panel includes × year × grade FE in the most augmented specification. Estimation is based on the sample of 21 schools.

Table 13: Effect of Teacher Gender Bias on the Quality of the Program Students Enrolled In

	Boys				Girls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Rank	Stand Rank	Top 20	Top 30	Rank	Stand Rank	Top 20	Top 30
Bias in All Subjects	1.709	0.060	0.014	0.020	-1.835	-0.064	-0.018	-0.022
	(0.953)	(0.033)	(0.008)	(0.011)	(0.811)	(0.028)	(0.008)	(0.010)
Sample Size	8,452	8,452	10,942	10,942	10,125	10,125	12,614	12,614
Bias in Core Subjects	1.268	0.043	0.011	0.015	-3.293	-0.114	-0.033	-0.041
	(1.043)	(0.036)	(0.009)	(0.012)	(0.941)	(0.032)	(0.010)	(0.012)
Sample Size	4,986	4,986	6,426	6,426	6,053	6,053	7,545	7,545
Bias in Classics Subjects	2.636	0.092	0.022	0.029	0.203	0.007	0.002	0.002
	(1.775)	(0.062)	(0.015)	(0.020)	(1.544)	(0.054)	(0.013)	(0.018)
Sample Size	2,668	2,668	3,459	3,459	4,134	4,134	5,252	5,252
Bias in Science Subjects	0.481	0.017	0.004	0.006	-2.012	-0.069	-0.022	-0.026
	(1.329)	(0.046)	(0.012)	(0.016)	(1.204)	(0.041)	(0.013)	(0.016)
Sample Size	3,637	3,637	4,497	4,497	4,342	4,342	5,216	5,216
Bias in Exact Science Subjects	1.249	0.044	0.009	0.012	-1.217	-0.042	-0.011	-0.014
	(1.539)	(0.054)	(0.011)	(0.015)	(1.206)	(0.042)	(0.011)	(0.014)
Sample Size	4,264	4,264	5,673	5,673	4,199	4,199	5,291	5,291
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Subject FE	✓	✓	✓	✓	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓	✓	✓	✓	✓
Class FE	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Year 2003 is excluded from the analysis, as it is used to calculate the quality measures for the post-secondary program students enroll in. "Rank" represents the ranking of each post-secondary program based on the 2003 mean performance of enrolled students in each post-secondary program. "Stand Rank" is the standardized measure of "Rank" which is normalised to have a zero mean and a standard deviation of 1. Students not enrolled in any university are not included in the calculation of "Rank" and "Stand Rank". "Top 20" and "Top 30" measure the likelihood a student is admitted to a Top 20 or Top 30 selective university program. Top 20 or Top 30 are binary indicators determined based on the 2003 mean performance of enrolled students in each post-secondary program. Students not enrolled in any university are included when considering "Top 20" and "Top 30" and we assign the value of 0 for them. We then assign these four 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. All specifications include the students' first semester 11^{th} grade performance and the teacher's gender as controls. 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.

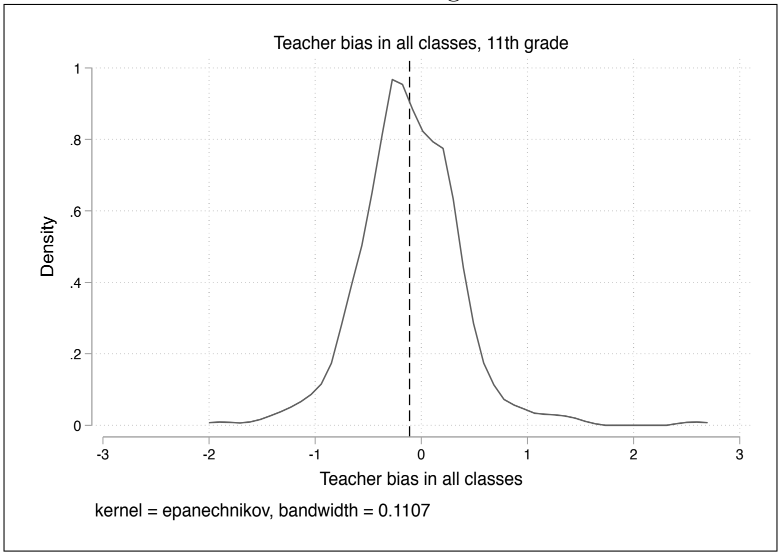
Table 14: Correlations Between Teacher Gender Bias And Teacher Quality (Measured by TVA)

Dependent Variable: Teacher Quality (Measured by TVA)				
	(1)	(2)	(3)	(4)
Pro-Girl Teacher Indicator	-0.044	-0.043	-0.043	-0.042
	(0.013)	(0.013)	(0.013)	(0.013)
Pro-Boy Teacher Indicator	-0.010	-0.010	-0.009	-0.008
	(0.014)	(0.014)	(0.014)	(0.015)
Female Teacher		0.009	0.008	0.009
		(0.010)	(0.010)	(0.010)
Class Size			-0.002	-0.002
			(0.002)	(0.002)
Experience				-0.001
				(0.002)
Obs.	422	422	422	422
Year FE	✓	✓	✓	✓
School FE	✓	✓	✓	✓
Grade FE	✓	✓	✓	✓

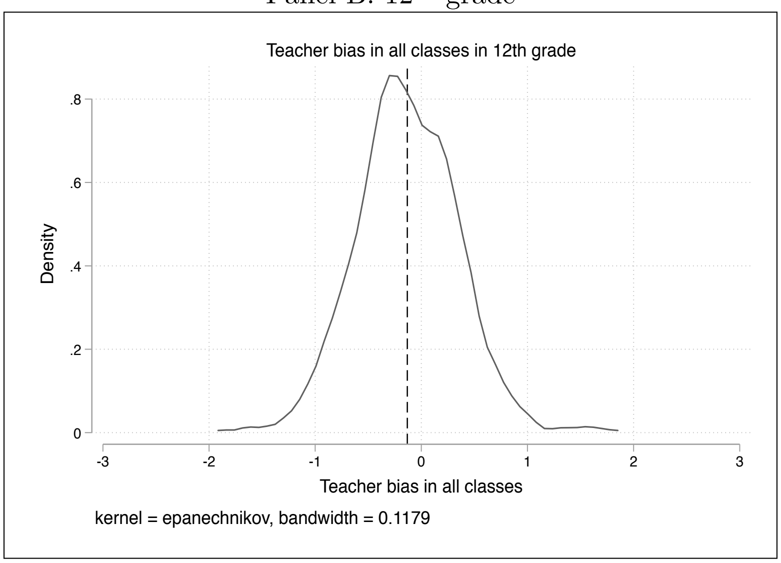
Notes: The "Pro-Girl Teacher Indicator" takes the value of one if the teacher exhibits a bias that is smaller than or equal to -0.10. The "Pro-Boy Teacher Indicator" takes the value of one if the teacher exhibits a bias that is above 0.10. We define as neutral teachers those who have a bias that is larger than -0.10 and smaller than or equal to 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. "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. The empirical Bayes estimates of teacher gender biases are used. Standard errors are clustered by school and year and are reported in parentheses.

Figure 1: Distribution of Teacher Gender Bias in 11^{th} and 12^{th} Grade, Sample of 21 Schools

Panel A: 11^{th} grade

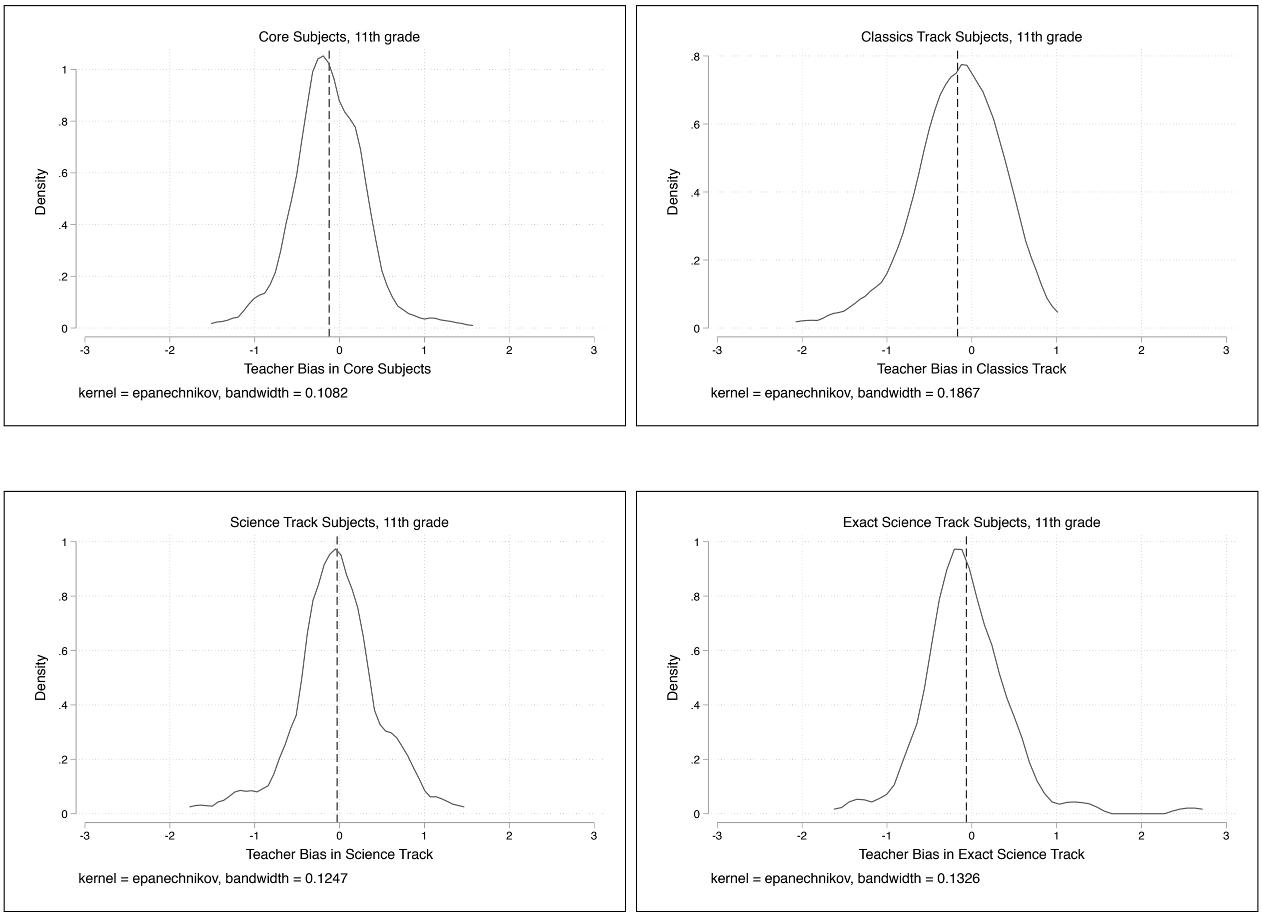


Panel B: 12^{th} grade



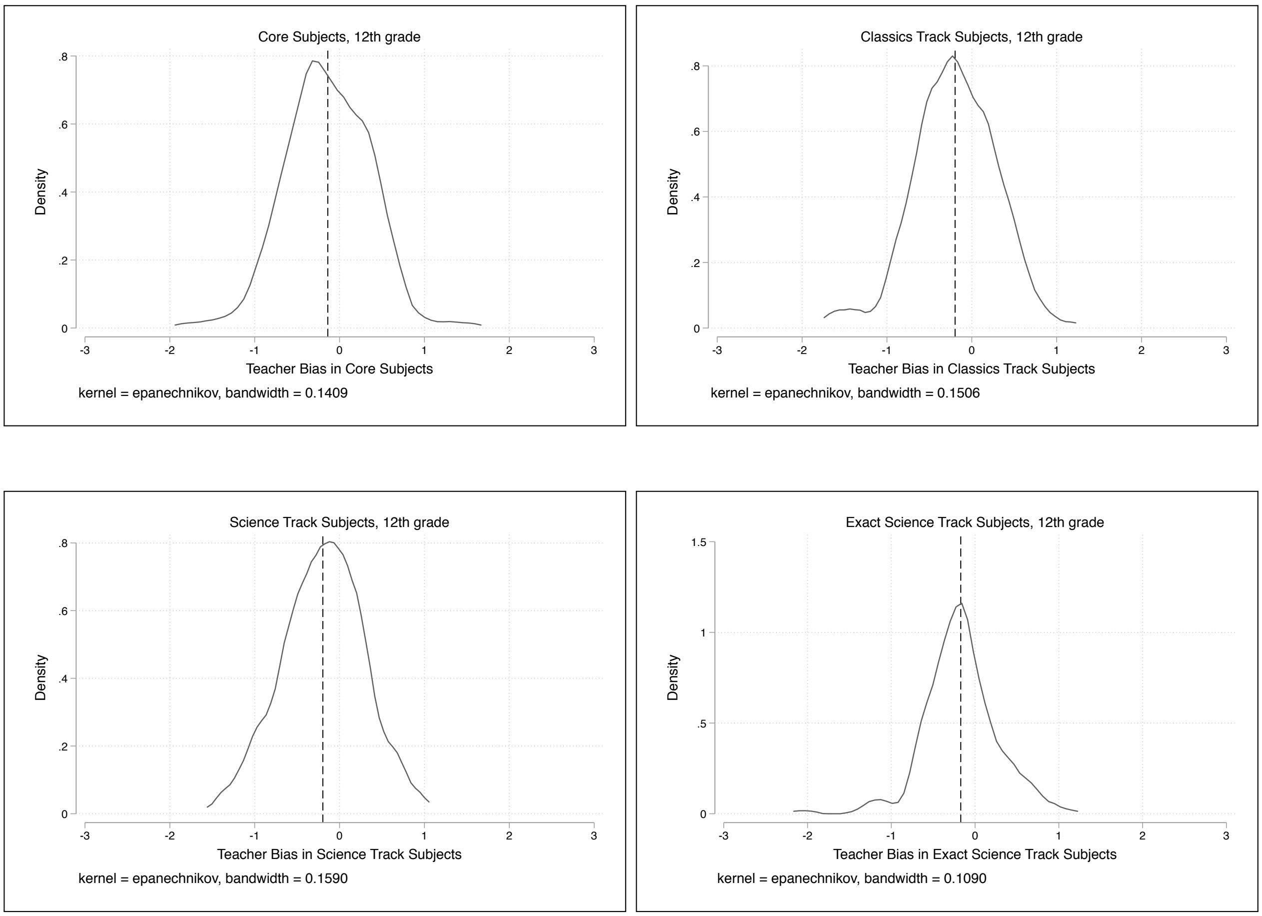
Notes: Panels A and B show the teacher-level distribution of bias based on all classes (including the own) in grades 11^{th} and 12^{th}, respectively. We use data for the sample period 2003-2011. The teacher bias in all classes that a teacher taught is measured as the average bias that a teacher exhibited in all classes she/he ever taught in the sample period 2003-2011 in each grade.

Figure 2: Histograms of the Bias Measured in All Classes in Core and Track Subjects in 11^{th} Grade



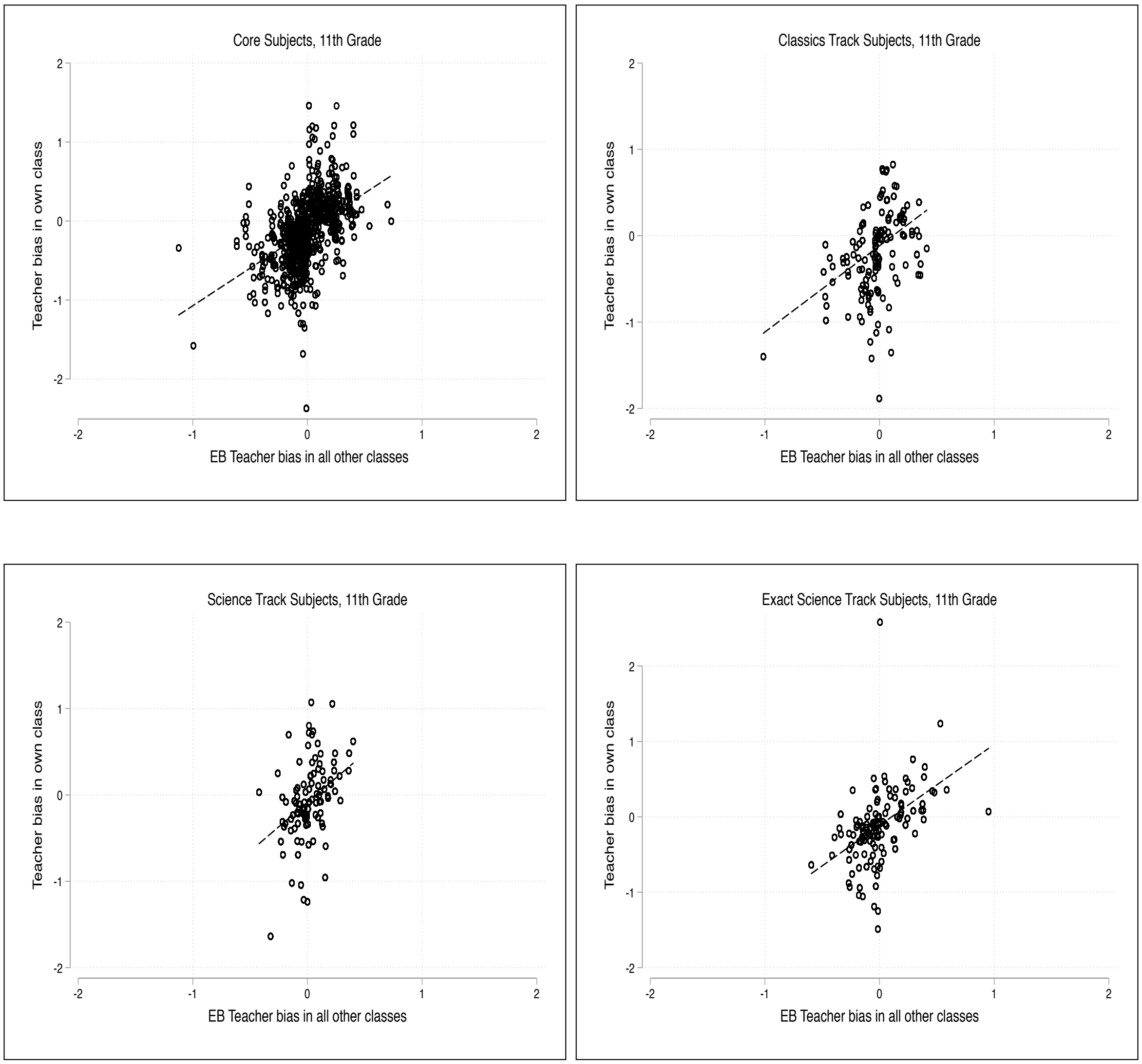
Notes: This figure presents a teacher-level distribution of bias based on all classes (including the own) for core and track subjects. In 11^{th} grade the core subjects taught are: modern Greek, history, physics, algebra and geometry. There are the following tracks in 11^{th} grade: Classics, Science 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 technology and computers. The mean (s.d.) of the teacher-level measure of the gender bias based on all classes is -0.123 (0.418), -0.166 (0.514), -0.029 (0.502) and -0.065 (0.516) in core, classics track, science track, and exact classics track subjects, respectively.

Figure 3: Histograms of the Bias Measured in All Classes in Core and Track Subjects in 12^{th} Grade



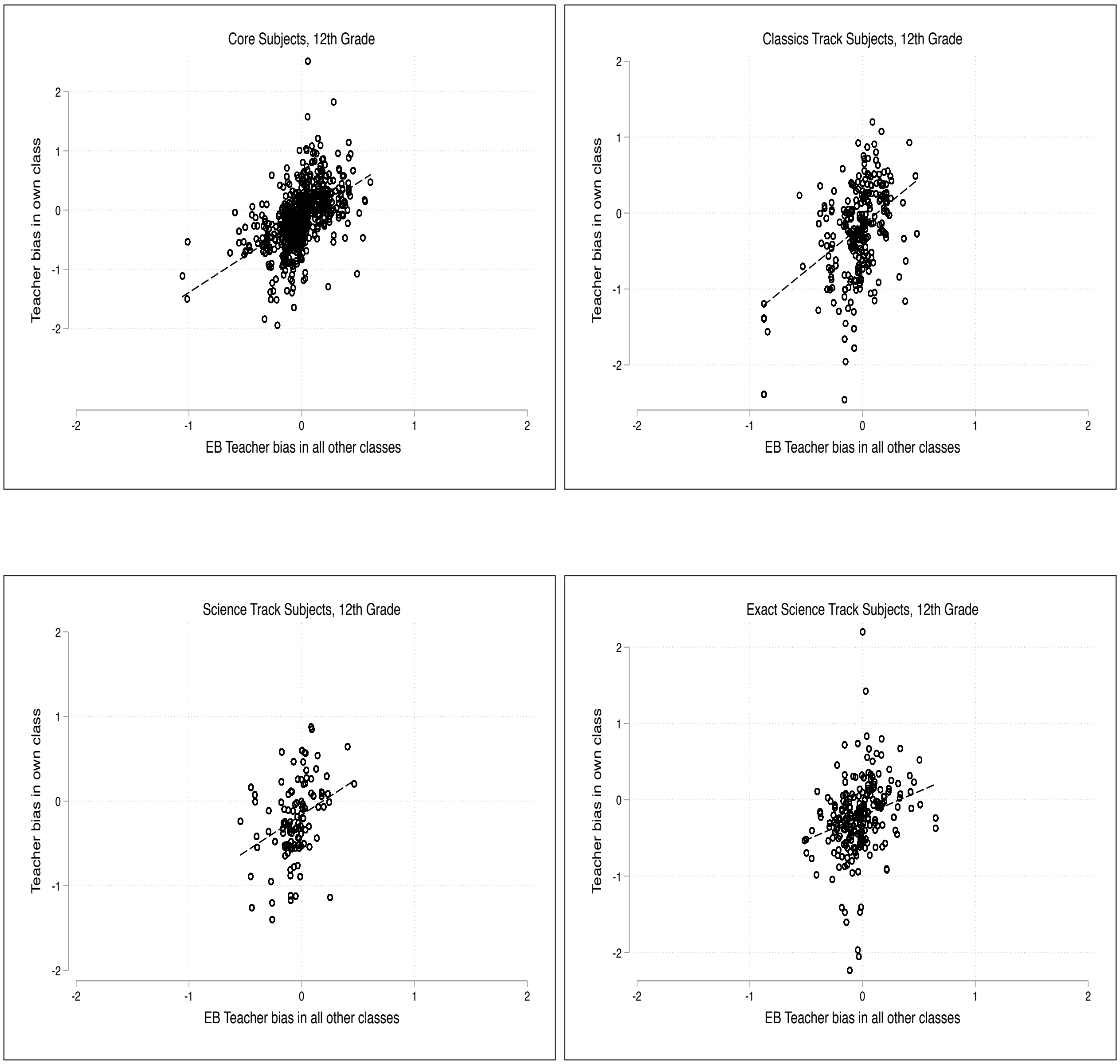
Notes: This figure presents a teacher-level distribution of bias based on all classes (including the own) for core and track subjects. In 12^{th} grade the core subjects taught are: modern Greek, history, physics, biology and mathematics. There are the following tracks in 12^{th} grade: Classics, Science 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. The mean (s.d.) of the teacher-level measure of the gender bias based on all classes is -0.139 (0.491), -0.197 (0.478), -0.197 (0.472) and -0.171 (0.430) in core, classics track, science track, and exact classics track subjects, respectively.

Figure 4: Correlations Between the Teacher Bias Measured in the Current Class and the Teacher Bias Measured in all Other Classes for Core and Track Subjects in 11^{th} Grade



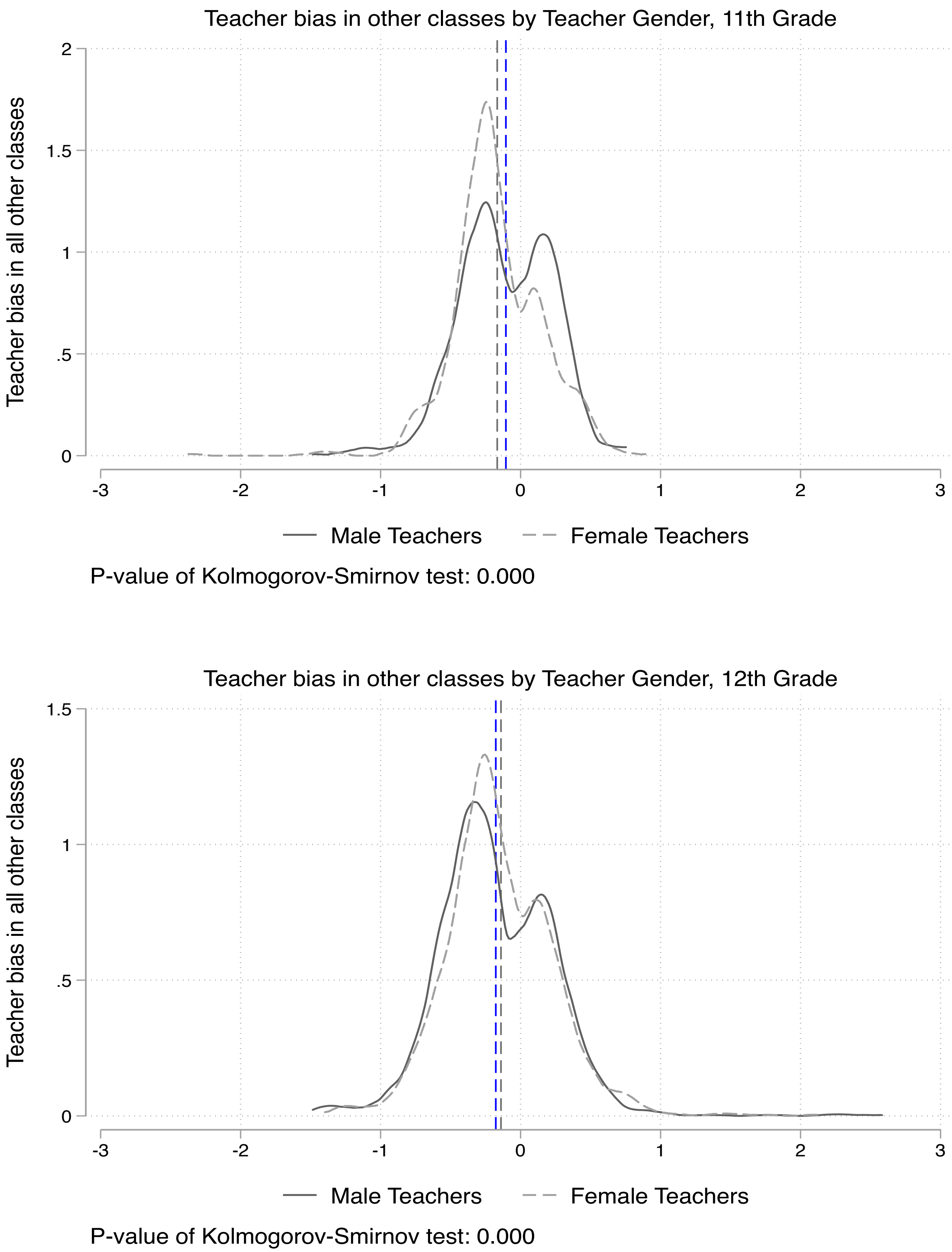
Notes: These scatter plots present the correlations between the two different measures of teacher biases, namely bias in current own class and bias in all other classes, for core subjects (top left panel), classics track (top right panel), science track (bottom left panel) and exact science (bottom right panel) subjects. We have also plotted the regression line from a linear regression of the teacher bias in the own class on teacher bias in all other classes 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 technology and computers.

Figure 5: Correlations Between the Teacher Bias Measured in the Current Class and the Teacher Bias Measured in all Other Classes for Core and Track Subjects in 12^{th} grade



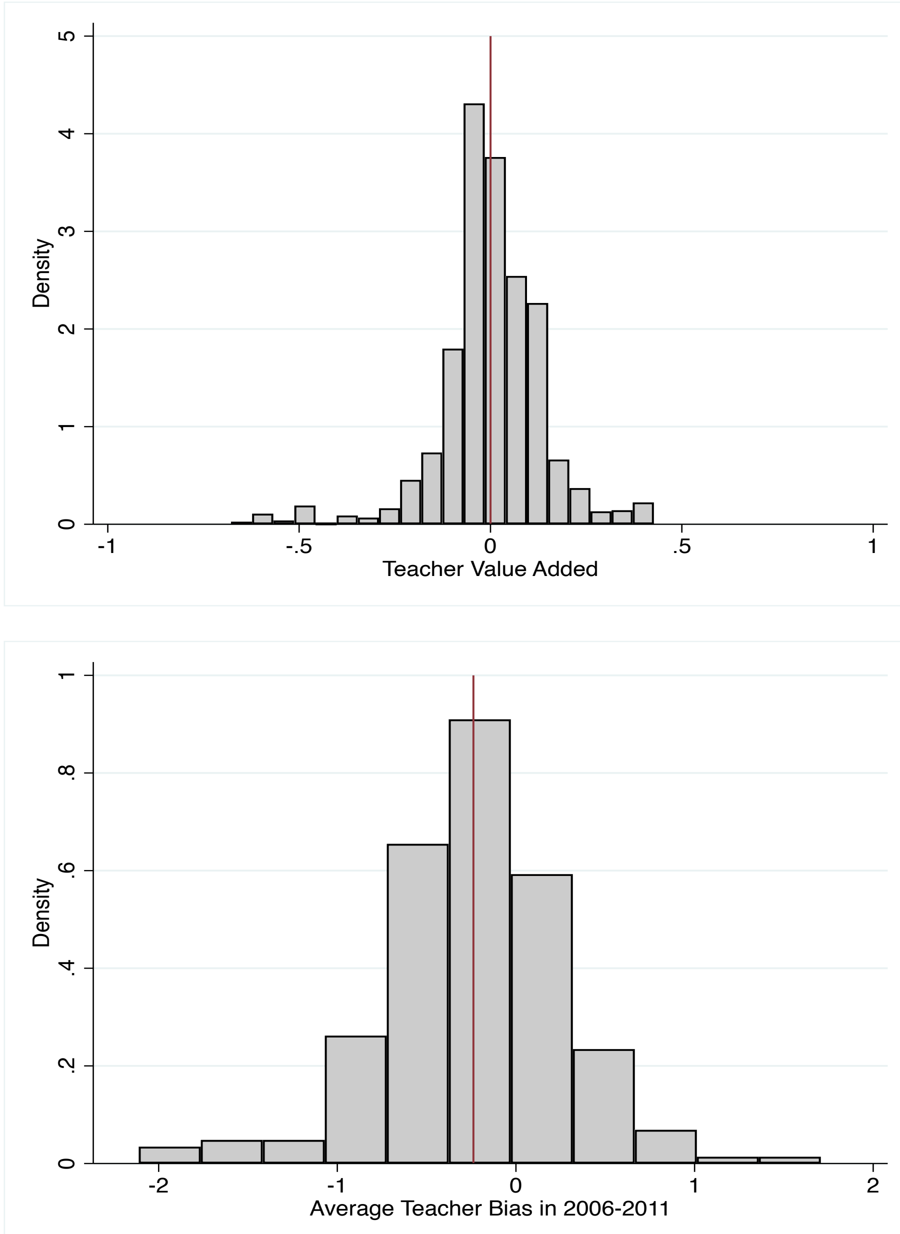
Notes: These scatter plots present the correlations between the two different measures of teacher biases, namely bias in current own class and bias in all other classes, for core subjects (top left panel), classics track (top right panel), science track (bottom left panel) and exact science (bottom right panel) subjects. We have also plotted the regression line from a linear regression of the teacher bias in the own class on teacher bias in all other classes 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 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, Friedman, and Rockoff (2014a). This sample includes only students who have non-missing baseline controls to estimate the TVA 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 cell. 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 use the empirical Bayes estimate and we calculate the average bias a teacher exhibits in all classes between 2006 and 2011.