Bank of Israel **Research Department**

On the Origins of Gender Gaps in Human Capital: Short and Long Term Consequences of Teachers' Biases[1]

Victor Lavy* Edith Sand**

Discussion Paper 2016.02
February 2016

Research Department, Bank of Israel – E-mail http://www.boi.org.il
* Victor Lavy, University of Warwick, Hebrew University and NBER
E-mail: v.lavy@warwick.ac.uk ; Victor Lavy, msvictor@huji.ac.il
** Edith Sand, Bank of Israel – E-mail: Edith.sand@boi.org.il Phone: 972-6552624
[1] We thank the Education Department of Tel-Aviv-Yafo Municipality and Yossef Shub, the CEO of Optimal Scheduling Systems, for making the data available for this study, Israel's Ministry of Education for allowing restricted access to secondary schooling data in the Ministry online protected research lab, and Israel's National Insurance Institute (NII) for allowing restricted access to data at its protected research lab. We benefitted from comments and suggestions from Naomi Hausman, Shulamit Kahn, Larry Katz, Kevin Lang, Yoram Mayshar, Analia Schlosser, Moses Shayo, Sarit Weisburd, Assaf Zussman, participants in seminars and conferences at Hebrew University, Tel-Aviv University, Ben Gurion University, Paris School of Economics, University of Warwick, NBER 2015 Summer Institute Education conference, CEPR 2015 Public Economics Annual Symposium, LAGV 2015 Conference in Public Economics, COSME 2016 Gender Economics Workshop, the Barcelona 2015 Summer Forum and the 2016 Applied Family Economics conference in Honk Kong. The first author acknowledges financial support from the European Research Council through ERC Advance Grant 323439, from the Falk Institute and from the Israeli Science Foundation.

חטיבת המחקר, בנק ישראל ת"ד 780 ירושלים 91007
Research Department, Bank of Israel. POB 780, 91007 Jerusalem, Israel

על מקורות הפערים המגדריים בהון אנושי: ההשלכות של הטיות סטריאוטיפיות של מורים בטווח הקצר ובטווח הארוך

ויקטור לביא ואדית זנד

תקציר

בעבודה זו אנו מעריכים את השפעת ההטיות המגדריות של מורים בבתי ספר יסודיים על הישגיהם הלימודיים של בנים ובנות בחטיבת הביניים ובבית הספר התיכון ועל בחירתם במגמות מוגברות במתמטיקה ובמקצועות המדעיים בתיכון. לצורך זיהוי הסיבתיות, אנו מתבססים על ההשמה האקראית של מורים ותלמידים בכיתות הלימוד בבתי הספר היסודיים. מהתוצאות שלנו עולה כי ההטיות של המורים לטובת הבנים משפיעות באופן אסימטרי על בנים ובנות – השפעתן חיובית על הישגי הבנים ושלילית על הבנות. הטיות מגדריות כגון אלה משפיעות גם על המשך דרכם של התלמידים בלימודים מוגברים במתמטיקה בתיכון – על בנים באופן חיובי ועל בנות באופן שלילי. התוצאות שלנו מצביעות על כך שההתנהגות המוטית של מורים בשלב מוקדם של הלימודים במערכת החינוך נושאת בחובה השלכות ארוכות טווח על בחירת העיסוק ועל ההשתכרות בחיים הבוגרים, מפני שלימודים מתקדמים במקצועות המתמטיקה והמדעים בתיכון הם דרישת קבלה מוקדמת להשכלה גבוהה בהנדסה, במדעי המחשב וכד׳. השפעה זו היא הטרוגנית, והיא רבה יותר עבור ילדים ממשפחות שבהן האב משכיל מן האם ועבור בנות מרקע סוציו-אקונומי נמוך.

On the Origins of Gender Gaps in Human Capital: Short and Long Term Consequences of Teachers' Biases

Victor Lavy and Edith Sand

Abstract

We estimate the effect of primary school teachers' gender biases on boys' and girls' academic achievements during middle and high school and on the choice of advanced level courses in math and sciences during high school. For identification, we rely on the random assignment of teachers and students to classes in primary schools. Our results suggest that teachers' biases favoring girls have an asymmetric effect by gender — a positive effect on girls' achievements and negative effect on boys' and vice versa. Such gender biases also impact students' enrollment in advanced level math courses in high school – girls positively and boys negatively. These results suggest that teachers' biased behavior at early stages of schooling has long run implications for occupational choices and earnings at adulthood, because enrollment in advanced courses in math and science in high school is a prerequisite for post-secondary schooling in engineering, computer science and so on.

1. Introduction

Over the past decades there has been a large increase in female human capital investment and labor force participation. The ratio of male to female college graduates has decreased consistently, to the extent that it has even reversed in many countries – in some countries there have been more female than male graduates in recent years (Goldin et al. (2006), Becker et al. (2010) and Goldin 2014). This trend is partly due to more women graduating in what used to be male-dominated fields such as math, science and engineering. The math and science test score gender gap is of special interest because it is a good predictor of future income (Murnane et al. (1995) and Paglin and Rufolo (1990), Brown and Corcoran 1997) and because there is still a considerable gender gap in employment in these fields. For example, evidence based on recent PISA testing[1] shows that in most countries, girls outscore boys in reading while being outscored in math (Machin and Pekkarinen (2008)). This gap is shown to grow during early years of schooling (Fryer and Levitt (2010)), and is larger at the upper tail of the test scores distribution (Ellison and Swanson (2010), Hyde et al. (2008)). Striking evidence from the UK shows that in 2012 about 80% of those who took A level physics were male[2], and that men were awarded 85% of engineering and technology degrees and 82% of computer science degrees, while in the same year, 83% of medical degrees and 79% of veterinary science degrees went to women.[3] The related employment gaps are even larger, as females are only 6% of the engineering workforce, 5.5% of engineering professionals and 27% of engineering and science technicians.[4]

What explains these gender disparities in cognitive performance and in math and science scores is still an open question. Some emphasize the role of biological gender differences in determining gender cognitive differences,[5] while others emphasize the social, psychological and environmental factors that might influence this gap. For example, some argue that gender role attitudes and stereotypes influence the gender gap by shaping the way parents raise their

[1] Programme for International Student Assessment (PISA), which surveyed 15-year old students from OECD countries in 2003, 2006 and 2009.
[2] Joint Qualification Council, quoted in The State of Engineering, Engineering UK 2013. HESA, 2010/11, quoted in WISE statistics 2012.
[3] HESA, 2010/11, quoted in WISE statistics 2012.
[4] These statistics on women in engineering compiled by Women's Engineering Society revised February 2014, Joint Qualification Council, quoted in The State of Engineering, Engineering UK 2013. See also Friedman (1989) and Wilder and Powell (1989) for reviews of the literature.
[5] This approach suggests that the difference in chromosomal determinants (Vandenberg (1968)), hormone levels (Benbow (1988) and Collaer and Hines (1995)) and brain structure (Witelson (1976), Lansdell (1962), Waber (1976)) can explain the evidence that men perform better in spatial tests, whereas women do better in verbal tests.

children[6]; by affecting the environment at school and teachers' attitudes; and by determining social and cultural norms.[7] 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.

In this paper we focus on the effect of gender bias in a schooling environment. Stereotypical attitudes of teachers towards boys and girls in class have been widely documented in the psychology and sociology literature, and have been argued to substantially influence students' self-image and educational outcomes. For example, teachers are said to treat the successes and failures of boys and girls differently, by encouraging boys to try harder and allowing girls to give up (Dweck et al. (1978) and Rebhorn and Miles (1999)). Sadker and Sadker (1985) suggest that teachers give more attention to boys by addressing them more often in class, giving them more time to respond and providing them with more substantive feedback. Teachers are also found to treat boys and girls differently, in particular with regard to math instruction: Hyde and Jaffe (1998) show that math teachers tend to encourage boys to exert independence by not using algorithms and that boys who pursue this rebellious approach are seen as having a promising future in mathematics; girls, on the other hand, are controlled more than boys, and are taught mathematics as a set of rules or computational methods.

---

[6] Different parental treatment and expectations are manifested in several ways, such as a different attitude from birth—boy babies are handled more than girl babies, whereas girl babies are spoken to more than boy babies (Lewis and Freedle, 1973)—to later stages of childhood (boys receive more encouragement for achievements and competition (Block 1976), and are trained to be more independent (Hoffman 1977); in addition, parents engage in a more positive attitude when children engage in gender-appropriate behavior (Block 1976), and instruct their sons and daughters in the different behaviors expected of them by providing them with different toys: boys' are "moveable and active and complex and social;' whereas girls' are "the most simple, passive, and solitary" (Brooks-Gunn and Lewis 1979).

[7] Social norms and beliefs are said to shape the perception of the appropriate division of roles in the home and family, paid employment and the political sphere (Inglehart and Norris 2003). Guiso et al. (2008) try to assess the relative importance of biological and cultural explanations, by exploring gender differences in test performances across countries. Their identification strategy relies on the fact that biological differences between sexes are much less likely to vary compared to the cultural environment. They show that there is a positive correlation between gender equality and the gender gap in mathematics achievements according to data from OECD's international tests (PISA 2003) and data that measure gender equality taken from the World Economic Forum's Gender Gap Index (GGI). Moreover, they show that these results are not driven by biological differences across countries (which was based on a measure developed by Spolaore. and Wacziarg, 2009), by using a genetic distance measurement between the populations. Pope and Sydnor (2010) and Fryer and Levitt (2010) replicate this methodology for different sets of countries. Also related is Alesina et al. (2013) who examine the historical origins of existing cross-cultural differences in beliefs and values regarding the appropriate role of women in society.

Leinhardt, Seewald and Engel (1979) find that teachers spent more time training girls in reading and less time in math, relative to boys. In addition, according to the National Center of Education Statistics (1997) girls are less likely than boys to be advised, counseled and encouraged to take courses in math.

Using all of these mechanisms through which gender biases of teachers potentially affect their students' educational outcomes, we build a quantitative measure of primary school teachers' gender biases and estimate the impact on boys' and girls' academic achievements during middle and high school, and on the selection of advanced level courses in math and sciences during high school. We measure teachers' gender biased behavior by comparing their average marking of boys' and girls' papers in a "non-blind" exam to the gender means in an anonymously marked "blind" national exam. We take this measure of grading bias to reflecting teachers' perceptions about gender cognitive differences and use it as a proxy for their level of prejudice and discriminatory behavior in class.[8] We show that there is large variation within schools on this measure, and that it has a significant effect on the academic achievements of both genders during middle school and high school in math, science and language and the difficulty of the math and science courses chosen in high school. These high stakes choices determine whether a student will meet requirements for admission to science and math studies at university.

We address threats to the interpretation of our findings. First, we show that there is meaningful variation in gender based grading biases by teachers within schools and within classes: it is often the case that within school and same subject or within a class, teachers will have opposite gender biases. The within school variation and within school by subject variation in our treatment variable permits using a school (/school by subject) fixed effect estimation strategy. This protects the interpretation of our findings against a variety of alternatives that rely on gender specific differences in behavior and characteristics, even if they are subject specific. For example, our findings cannot be explained by the possibility that girls (boys) do better or worse in external exams generally or in specific subjects such as math or science. Our findings cannot be caused by teachers who take into account the good behavior or higher effort of girls (boys), or the higher popularity of boys (girls) among peers in grading internal assessments. Nor can our findings reflect higher average cognitive ability

[8] This measure of teacher bias might capture conscious discriminatory behavior of teachers as well as unintentional biases of teachers that they themselves are not aware of. See Bertrand et al (2005) for a discussion of the concept of 'implicit discrimination', ways of measuring it and possible ways of limiting its prevalence.

or socioeconomic status of girls (boys), or that girls (boys) do worse if new material (say geometry) is introduced in 6th grade, or that there is a systematic measurement error in the internal or the external score. Second, the within class variation in teachers' bias allows for a class fixed effect estimation strategy, so we can dismiss the possibility that our measure of gender bias of teachers may simply pick up random (small sample) variation in the unobserved "quality" or "non-cognitive" skills of the boys vs. girls in a particular class. Remarkably, all three of these alternative model specifications – school fixed effects, school by subject fixed effects and class fixed effects – yield very similar estimates.

Second, we provide direct evidence that our estimates reflect teachers' behavior and not students' characteristics or behavior. For example, we show that teachers' biases are correlated with their characteristics and that the correlation between same teacher's biases (when teaching two different subjects) is significantly higher than the correlation between different teachers' biases. Finally, we provide evidence for robustness by adding controls to the regression that rule out alternative interpretations of the teachers' bias measure, such as class level differences in average students' ability or non-cognitive skills.

The systematic difference between non-blind and blind assessment across groups as a measure of discrimination or stereotypes was pioneered in economics by Blank (1991) and Goldin and Rouse (2000).[9] This approach was first applied to the economics of education in Lavy (2008), to measure gender bias in grading by teachers and it was followed by others, for example, Björn, Höglin, and Johannesson (2011), Hanna and Linden (2012), Cornwell, Mustard, Van Parys (2013), Burgess and Greaves (2013), and Botelho, Madeira and Rangel (2015), who implemented the same methodology using data from other countries and getting overall similar evidence about teachers' stereotypes/biases.[10] In the present paper, however,

[9] Blank (1991) shows that the probability of papers being accepted by economic journals depends on authors' affiliation. Goldin and Rouse (2000) examine sex-biased hiring patterns in orchestras by comparing blind and non-blind auditions. See Bertrand and Duflo (forthcoming) for a recent survey of the literature on discrimination, which reviews the existing field experimentation literature on the prevalence of discrimination, the consequences of such discrimination and possible approaches to undermine it.

[10] Lavy (2008) finds that in high schools, male students are being discriminated against in all subjects. Based on evidence from primary school in the U.S, Cornwell et al. (2013) found that boys who preform equally well as girls are graded less favorably by their teachers, but that this favorable treatment vanishes once students' non-cognitive skills are taken into account. Other papers using a similar methodology examine the existence of racial discrimination: Burgess and Greaves (2013) find that in English public schools, black Caribbean and black African students are under-assessed relative to their white peers while other minority groups (such as Indian, Chinese and Asian) are over-assessed. Botelho et al (2015) find that black students are being

we go beyond measuring teachers' biased behavior and focus on the implications of this behavior for gender differences in human capital formation. We think this paper is the first to highlight teachers' biased behavior as a source of the gender gap in human capital, in particular regarding gender differences in math and science studies.[11] To this end, we focus on boys' and girls' choices about the difficulty of the math and science courses they select in high school. In Israeli higher education, as in many other countries, these choices have important implications for occupational choices at adulthood, because advanced courses in math and science in high school are a prerequisite for post-secondary schooling in engineering, computer science and so on. We test whether teachers' biases towards one of the sexes, as reflected by a more positive evaluation on the "non-blind" tests relative to the "blind" tests of this group, influence this group's future achievements and affect their orientation toward enrollment in advanced math and science studies in high school.

Our data enables us to evaluate the impact of teachers' gender biases on students' test scores in later years by following three cohorts of 6th grade students between the years 2002–2004 in Tel-Aviv, Israel. By tracking students from primary school to the end of high school, we are able to measure students' exposure to teachers' gender biases in primary school, and to estimate the effect on both 8th grade (middle school) test scores in national tests as well as on the high stakes matriculation exam scores at the end of high school, more than six years after the exposure to biased behavior. In addition, we are able to examine whether this measure of teachers' biases is correlated with certain teachers' characteristics, such as age, ethnicity, marital status and gender composition of own children.

Our results suggest that teachers' more positive assessment of boys in primary school in a specific subject has a positive and significant effect on boys' achievements in that subject in

---

discriminated against relative to their white classmates in Brazilian schools. Björn et al. (2011) report a similar attitude towards students from foreign backgrounds in Swedish high schools.

[11] In a recent paper, Leslie et al. (2015) argue that women are underrepresented in disciplines whose practitioners believe that innate talent is the main requirement for success, controlling for the disciplines' characteristics. This correlation is argued to be partly driven by the negative stereotype against women on this dimension, which is measured based on survey questionnaires. Also related is Reuben et al. (2014) who study the effect of stereotypes in an experimental market, where subjects were hired to perform an arithmetic task that, on average, both genders perform equally well. They find that when the employer had no information other than candidates' physical appearance, women were only half as likely to be hired as men, while revealing information on the candidate's arithmetic ability reduced the degree of discrimination, but did not eliminate it completely. Terrier (2015) uses a similar idea to the one we pursue in an earlier draft (Lavy and Sand 2014) and in this paper. Consistently with our finding she shows that in primary schools in France there also exists a positive correlation between teachers' grading bias in favor of boys in a specific subject and the progress of boys relative to girls in class in that subject.

middle school and high school national tests, and it has an asymmetric significant negative effect on girls. In addition, we find that the favoring of boys over girls by primary school math teachers also affects the successful completion of advanced courses in math and science in high school. Teachers' biases that favor boys encourage boys to enroll in advanced math courses while doing the opposite for girls. Since these courses are prerequisites for admission to higher education in these subjects, teachers' stereotypical biases contribute to the gender gap in qualifications in fields like engineering and computer science, and therefore to the gender gap in related occupations. These impacts on human capital outcomes by the end of high school have meaningful economic consequences for quantity and quality of post-secondary schooling and for earnings in adulthood. We also find large spillover effects of stereotypical biases of teachers across subjects, implying that a teacher's biases against girls or boys in one subject can have a broader influence on students' achievements in other subjects. In addition, we show that these effects have interesting patterns of heterogeneity by parental years of schooling, parental education gap, ethnicity and birth order of children.

The rest of the paper is organized as follows. In Section 2, we present our data. Section 3 explains the identification and estimation methodologies. We detail our results in Section 4, and Section 5 offers conclusions and policy implications.

2. Data

In this study we use data from the school authority for the municipality of Tel-Aviv. The baseline sample is sixth-grade students in the city's schools in 2002–2004. Each record contains an individual identifier, a school and class identifier in the sixth grade and students' test scores from exams in three subjects (math, English and Hebrew) held in the midterm of 6th grade. These tests were graded by the students' teachers[12] ("non-blind" assessments) and were created and administered by Tel-Aviv municipality for monitoring purposes. These data were merged with Israel Ministry of Education students' registry files that include students' demographic information (gender, ethnicity, number of siblings, and parents' education). We combined this dataset with data from four additional sources:

1) The first is GEMS records (Growth and Effectiveness Measures for Schools - *Meizav* in Hebrew) for the three cohorts that we study. The GEMS records were created and

[12] Students were tested in these three subjects only, and the tests in each subject (Hebrew, math and English) were graded by the class teacher for the subject.

administered by the Division of Evaluation and Measurement of the Ministry of Education.[13] The students' GEMS includes test scores of fifth and eighth graders for a series of tests (in math, Hebrew and English), which were transformed into z-scores for each year and for each subject to facilitate interpretation of the results. The GEMS tests were administered during the midterm of each school year to a representative 1-in-2 sample of all elementary and middle schools in Israel, so that each school participated in GEMS tests once every two years. GEMS tests were graded blind by an independent agency: the identity of the student is never revealed. The proportion of students tested is above 90 percent.

2) The second is matriculation exam scores and credits from the Israel Ministry of Education for the three cohorts that we study. Matriculation exams are national exams in core and elective subjects, taken between the tenth and twelfth grades. Students choose to be tested at various levels of proficiency, with each test awarding from one to five credits per subject, depending on difficulty. Some subjects are mandatory, and for many the most basic level is three credits. Advanced level subjects carry four or five credits. A minimum of 20 credits is required for a matriculation certificate, which is a prerequisite for university admission. The average scores in the matriculation certificate, which are calculated by the Higher Education Council, are weighted based on the number of taken (advanced level subjects are also given bonuses: four credits are awarded a bonus of 12.5 points and five credits are awarded 25 points). All schools in the sample offer an academic track leading to a matriculation diploma. We focus on the following matriculation exam outcomes: test score in math, English and Hebrew (transformed into z-scores by each year and each subject), the probability of matriculating, the number of successfully completed exams, and the number of successfully completed units in English and in science related subjects (math, physics and computer science).[14]

[13] For more information on the GEMS, see the Division of Evaluation and Measurement website (in Hebrew): http://cms.education.gov.il/educationcms/units/rama/odotrama/odot.htm.

[14] The matriculation exams in math, English and Hebrew are mandatory: the number of credits required in Hebrew is two, and in math and English students are allowed to choose between the most basic level (three credits) and the advanced level (four or five credits). On the other hand, matriculation exams in computer science and physics are optional and students can take a maximum of 5 credits in these subjects. Regarding the scores in the mandatory matriculation exams (in math, Hebrew and English), students received zero values if they did not take the exam and did not receive a matriculation certificate (the proportion of students who received a matriculation certificate is 20%, and the rate of attrition is about 10%).

3) The third dataset is teachers' GEMS questionnaires. The teachers' GEMS questionnaires were addressed to almost all teachers in schools for which we have students' GEMS scores in the relevant years (except from the first year of the sample for which we have only partial data because it was also the first year that the GEMS was administered).[15] Although all teachers were asked to fill in these questionnaires, we could only merge the information of homeroom teachers with our teachers' bias measure, since other teachers were not asked which classes they teach. Thus, the information we gathered from these files referred to teachers' identifier, if they are homeroom teachers, and if so the class they teach and their subjects of instruction. Since we could link only homeroom teachers to their class of instruction, and only a few schools appear in the sample twice, we could not track homeroom teachers over the years.

4) The fourth is data from the Population Registry at the National Insurance Institute (NII) on the demographic background of teachers and students.[16] We merged teachers' identifiers with data from the Population Registry, which enables us to observe homeroom teachers' demographic background such as gender, age, marital status, ethnicity and number and gender of children. Additional demographic information was also obtained for the students (the place of birth of the student's grandparents and the birth order of children).

To construct the teachers' biased behavior measure we combined the scores from the "blind" exam with those of "non-blind" exam. Specifically, this measure was defined at the class level by the difference between boys' and girls' average gap between the school score (non-blind) and the national score (blind). The GEMS test is a "blind" assessment since the GEMS exams are graded by an independent agency and the identity and gender of the student are never revealed. In contrast, the other exam, which is graded by the student's teacher, contained the name of the student and therefore is a "non-blind" assessment. We assume that this measure of teacher's stereotypes captures her/his overall perception about gender cognitive differences and we use it as a proxy for her/his level of prejudice and discriminatory behavior in class. We then test the effect of this measure on boys' and girls' academic achievements during middle school (GEMS exams scores in the 8th grade) and high school (matriculation exam scores) and on the choice of advanced level courses in math and sciences during high school.

[15] For 2002 only 13 homeroom teachers from 33 classes were identified: in 2003 and 2004 more than 30 are identified.
[16] We accessed this data at the protected research lab of the National Insurance Institute.

In addition to the different ways these two tests are administered, they differ in other aspects as well: First, the structure of these tests is different. While some of the questions in the GEMS are multiple choice, most of the “non-blind” tests questions are open responses. Although the way students answer these types of questions may differ across gender, the fact that the “non-blind” tests consist mostly of open questions give teachers more freedom in grading those tests and enables scores to reflect other possible factors besides students’ knowledge. Second, the material being evaluated in both tests might not completely overlap, although most of the topics covered should be comparable. This results from the fact that the time gap between these two tests is less than one year (one is given in the midterm of 5th grade and the other at the midterm of 6th grade), and the fact that students' educational environment remains almost unchanged throughout these two consecutive years (teachers in both 5th and 6th grades are usually the same teachers[17], and students stay in the same classes and have the same curriculum). Third, both tests are low stakes tests because they are not used for matters important directly to students and are mainly used for monitoring purposes.[18] In addition, since they are both created and administered by external agencies (the Division of Evaluation and Measurement of the Ministry of Education and Tel-Aviv municipality) all tests in each subject were the same and teachers did not have the ability to construct their own tests. [19] Fourth, the timing of these tests differs. Since GEMS test are administered at the mid-term of the 5th grade, only three or four months after the teachers have begun instructing the class, we posit that their prejudice and discriminatory behavior in class only marginally affects students’ GEMS test scores, while 6th grade test scores are influenced much more by the behavior of the teachers (as well as by preferential grading attitude which shows up only in the internal assessments). [20] Furthermore, the fact that internal scores are revealed to students only after the GEMS test eliminates the possibility that GEMS scores are affected directly by preferential grading of their teachers.

[17] Teachers in Israel elementary schools are generally assigned to the same classes for two years consecutively.
[18] Niederle and Vesterlund (2007) show that men and women of the same ability differ in their selection into a competitive environment: women shy away from competition and men embrace it.
[19] We also note that only the school means of GEMS tests results are sent to schools. This implies that the main mechanism through which teachers’ biases affect students' future achievements is not through the preferential grading channel but by implementing discriminatory teaching practices in class which encourage one group more than the other and improve this group’s future educational attainments.
[20] If we were to assume that teachers’ biases affected to some extent also 5th grade external scores, it would have biased our teacher measure towards zero (i.e., underestimating the magnitude of teachers’ biases effects).

Although we cannot completely rule out that the differences in what is being evaluated by these two exams are not gender-neutral, our results are not sensitive to these cross gender differences, since the teachers' bias measure, which relies on the average scoring of boys and girls in these exams at the class level, would have been affected in a similar way in all the classes of the sample.

The final merged dataset includes the national external test scores (blind) in the 5th grade, the school test scores (non-blind) in the 6th grade, GEMS surveys questions in 5th and 6th grade, national exam GEMS test scores in 8th grade, matriculation exam scores and units of study at the end of high school for 2001–2008, 2002–2009 and 2003–2010, and student characteristics. In addition, we also observe teachers' characteristics for a sub-sample of teachers (homeroom teachers).

Table 1 presents descriptive statistics, and information about sample size, number of schools, and number of classes for the three sixth-grade cohorts that we use: 2002, 2003 and 2004. The panel data includes 20 secular elementary schools and 5 secular middle schools each year.[21] There are on average two classes in each elementary school (3 schools have only one classroom). The sample includes 867 students (in 33 classes) from the 2002 cohort, 1,127 students (in 41 classes) from the 2003 cohort, and 1,017 (38 classes) from the 2004 cohort. The table indicates that the three cohorts' samples are similar across all background variables: mean parental education, average family size, and ethnicity.

Appendix Table A1 presents descriptive statistics for the sub-sample of teachers, homeroom teachers for whom we have additional demographic information, sample size, and subject of instruction. The sample includes 13 math teachers, 29 Hebrew teachers and 36 teachers who teach both math and Hebrew. English teachers are not part of this sample because none of them also served as a homeroom teacher. We note that all identified teachers in our sample are female. This is expected as most teachers in primary school in Israel are female.[22] Although we do not have information on the other teachers in our sample, we note that each student in primary school usually studies 10 to 15 different subjects, with the same group of peers (tracking in primary school is forbidden by law). Homeroom teachers are the instructors of a quarter to a third of these subjects, while the other subjects are taught by

[21] The number of middle schools presented in the table refers only to middle schools with GEMS test scores, which participate in GEMS once every two years. The overall number of middle schools in the sample is 12.

[22] According to a recent publication of the Ministry of Education, 92 percent of primary school teachers in Israel (in secular Jewish schools) in 2007/8 were female. This statistic is from: http://meyda.education.gov.il/files/MinhalCalcala/facts-and-figures_v2_2014.pdf.

subject-specific teachers, such as English, who might work in more than one school. Thus, homeroom teachers generally teach one class, while subject-specific teachers might teach several classes.[23]

Table 2 presents the means of the “non-blind” and “blind” test scores, and the mean of the difference between them, separately for boys and girls. We also present in column 7 the difference between boys' “non-blind” and “blind” exams scores (column 3) less the difference between girls' “non-blind” and “blind” exams scores (column 6).

The gender gap in test scores varies substantially by type of exam (“non-blind” versus “blind”) and by subject. Girls in primary schools outscore boys in the Hebrew “non-blind” and “blind” exams. This implies that there is no teachers’ gender grading bias in Hebrew. In math we see a different pattern—girls outscore boys in the “blind” exam and boys outscore girls in the “non-blind” exam, implying that teachers assess boys more positively than girls. In English girls outscore boys in both types of exam, and they are more positively assessed relative to boys.

Next we examine whether the apparent gap between “non-blind” and “blind” test scores of boys relative to girls (column 7) is statistically significant, using the estimation framework from Lavy (2008). We assume that the students' test scores depend on gender, type of test (non-blind test=1) and their interaction term. Appendix Table A2 presents estimates based on this basic specification. We first run a regression that includes individuals’ characteristics and year, subject and class fixed effects, and then a second regression that includes year, subject and students fixed effects. The estimated coefficient of the interaction term, which measures the difference between the “blind” and “non-blind” scores of boys relative to that of girls (similar to the measure presented in the last column of Table 2), is positive in math, negative in English, and practically zero in Hebrew. While the estimates in Hebrew and English are not statistically different from zero in both regressions, the positive estimate in math is statistically different from zero in the first regression (OLS), and positive but not significantly different from zero in the second (student fixed effect specification). These results imply that there exists some degree of bias against girls in the marking of math exams (teacher biases in favor of girls in English are of similar magnitude to the biases in favor of boys in math, but these biases are not significantly different from zero). We note that these results differ from Lavy (2008) who finds that in a sample that includes all high school students in Israel, male

[23] See the Director General’s Circular regarding the syllabus regulation of primary schools in Israel: (http://cms.education.gov.il/EducationCMS/Applications/Mankal/EtsMedorim/3/3-1/HoraotKeva/K-2006-3a-3-1-25.htm).

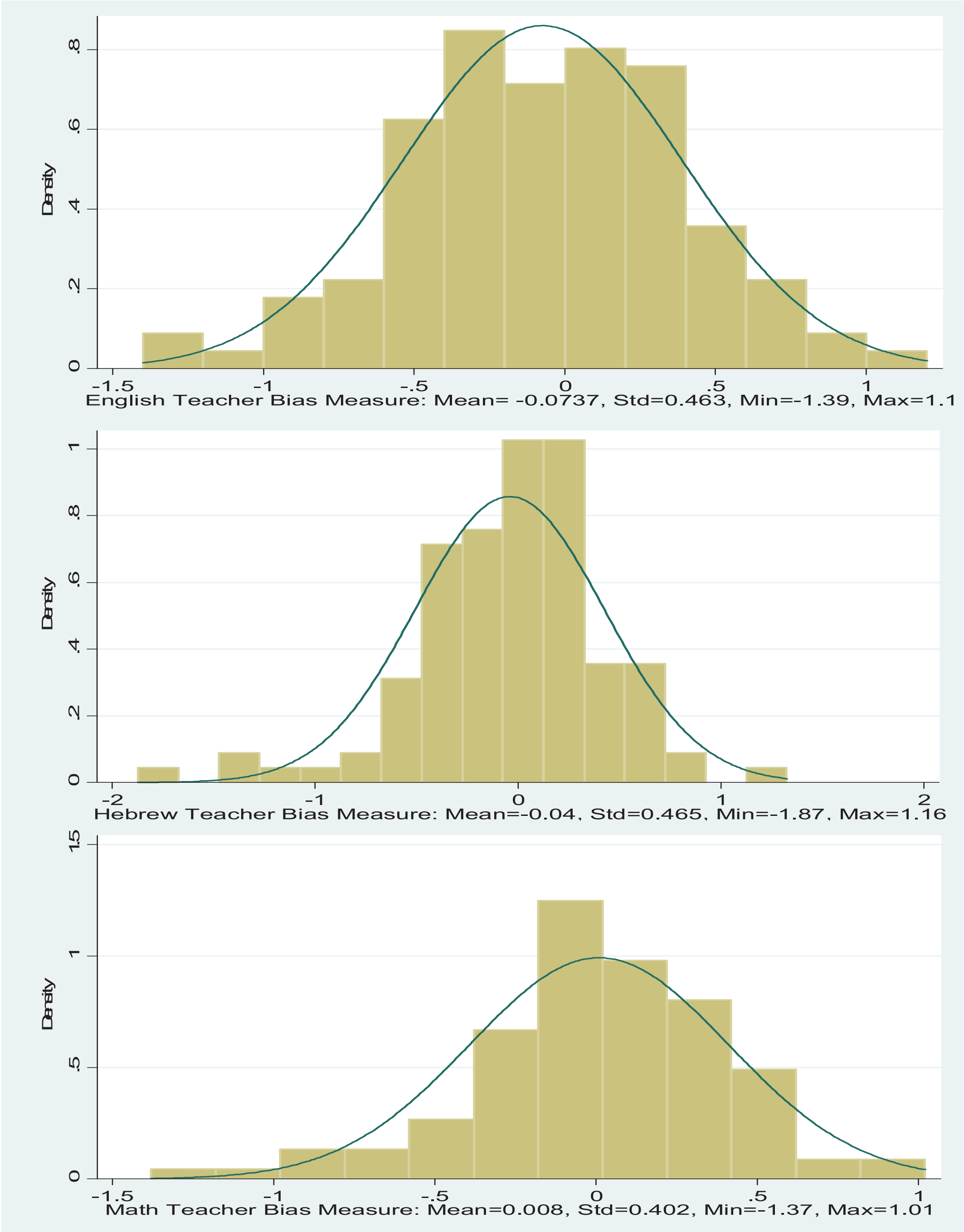
students face discrimination when “blind” and “non-blind” scores of matriculation examinations are compared.

Table 3 presents the means of both middle school and high school test scores in external exams, separately for boys and girls. The gender gap in favor of girls in Hebrew and English external exams persists to a large extent in middle school and high school: the gap in Hebrew is 0.3 in middle school and 0.2 in high school, while the gap in English is about 0.175 in middle school and 0.02 in high school. At the same time, the gender gap in favor of girls in math external exams in primary school is reversed in middle school and in high school, from a gap of 0.028 in favor of girls, to a gap of 0.024 in favor of boys in middle school and a gap of 0.08 in favor of boys in high school.

Appendix Table A3 presents the distribution of students across matriculation exam units of study, for boys and girls separately. Although girls have a higher probability of receiving a matriculation diploma and outnumber boys in the number of completed matriculation exam units, boys outnumber girls in English and in science oriented advanced courses. The proportion of boys and girls who successfully completed the advanced 5 credit course in English is 60.5.8% and 58.2% respectively. The proportion of boys who successfully completed the 5 credit course in math is 21.1%, while the proportion of girls is 14.1%. In science courses this gender gap is even larger: 15% of boys successfully completed advanced physics and 11.1% advanced computer science, while the rates for girls are only 4.8% and 3.2%, respectively. In the remaining part of the paper we will test whether these differences in achievements, especially in math scores, and in successful completion of advanced math and science courses, can partly be explained by exposure to teachers' gender biases during earlier stages of schooling.

The teachers' biased behavior measure is defined at the class level by the difference between boys' and girls’ average gap between the school score (non-blind) and the national score (blind). This bias measure is calculated in each subject (Hebrew, math and English) for each one of the 112 classes in our sample. It is estimated uniquely for each subject and the higher it is, the higher the stereotypical bias in favor of boys and against girls. The distributions of this measure by subject are presented in Figure 1. English teachers in primary school assess girls more positively than boys (mean is -0.074) and the same pattern is seen for Hebrew teachers (mean is -0.041). Math teachers' assessment in primary school, on the other

Figure 1: The Distributions of Teachers' Biases Measure, by Subject



Notes: The teachers' biases measure is defined at the class level by the difference between boys' and girls' average gap between the school exams scores (non-blind) and the national exams scores (blind), by subject.

hand, is on average gender neutral (0.01).[24] However, these means hide a large heterogeneity among teachers. The range and the standard deviation of the stereotypical bias measures, which are similar across subjects (SD= 0.45, min= -1.87, max= 1.16), reveal that there is considerable variation in gender biased behavior among teachers (65% of this variation is within school). We will exploit this significant variation and test whether teachers' stereotypes have short and long term effects on students' test scores. The short period between the timing of the two tests increases the likelihood that nothing else occurs between tests to affect student achievement differentially by gender, other than the tests being graded differently (blind vs. non-blind). This is not a very strong assumption, particularly since the class has the same teacher during the two tests so there is no change in teacher quality, no change in class composition – implying no change in social interactions, no changes in curriculum, and the period elapsed between the two tests is short enough that one can exclude the possibility of differential trends in student achievement. Note also that if any such change does occur it will affect our interpretation of the teachers' bias only if it has a differential effect by gender.

3. Identification and Estimation

The main goal of this paper is to investigate how teachers' biases towards a gender influence this group's future achievements and affect educational choices. Our data allows us to track students from primary school, where students were exposed to different teachers' gender stereotypes, through middle and then high school. Thus we can examine the implications of this exposure for their human capital formation, in particular test scores in middle school and high school national standardized tests, and choices about math and science studies in their final years of high school. Our main identification strategy relies on the random assignment of students and teachers to classes within a school. Using within-school analysis (primary school fixed effect framework), we compare students that study in the same primary school but were randomly exposed to different teachers, and potentially different gender biased behavior. In robustness exercises, we first expand the within school analysis by interacting the school fixed effect with subject fixed effect, thus comparing students that are randomly assigned to different teachers of the same subject in the same school, and then replace the school by subject fixed effects with class fixed effects to further test the impact of variation in teachers' biases on students within the same class.

[24] We note that the means of teachers' biases measured by subject of instruction are different from the difference between the blind and non-blind scores of boys versus girls (Table 2, column 7), since the first is calculated at the class level while the second is calculated at the student level.

The randomness of class composition results from the fact that students' assignments into class based on ability, family background or any other characteristics of the students are forbidden by law in Israel and this law is strictly enforced.[25] In order to test explicitly for the randomness of class composition in our sample, we perform a series of Pearson Chi-Square (χ^2) tests that check whether the student's characteristics and the class assignment are statistically independent. Based on 37 elementary schools (with two or more classes) and eight characteristics (gender, four ethnicity groups, number of siblings, and level of parents' education) we find that out of 296 p-values, only 14 were equal to or lower than 5 percent. This implies that for only 5% of the classes we cannot reject that there is non-random assignment. In addition, of the 37 elementary schools in our sample, the p-value was equal or lower than 5% in only two schools. We therefore conclude that in our sample of schools and classes there is no evidence of systematic non-random formation of classrooms with respect to students' characteristics.[26] The implication of this evidence is that since there is no difference in all classes within a school in terms of average students' ability or any unobserved characteristics, teachers' assignments to classes are also unrelated to observed and unobserved students' backgrounds.

In the empirical model we assume that the test scores and the choice of advanced level courses by pupils in middle/high school are determined by the following equation:

$$(1)\quad y_{icjt} = \alpha + \beta_s + \delta_j + \gamma_t + \lambda X_{icjt} + \beta_1 DS_{icjt} + \beta_2 CS_{icjt} + u_c + \varepsilon_{icjt}$$

where y_{icjt} denotes the outcome of student *i*, from primary school class c, subject *j* and year *t*; X_{icjt} are the student characteristics; β_s is a primary school fixed effect; δ_j is a subject fixed effect; γ_t is a year fixed effect; DS_{icjt} is the measure of teachers' biased behavior in subject *j* (direct-subject effect); CS_{icjt} is a measure of the average teachers' biased behavior in the

[25] The 1968 Integration Law in Israel clearly states that schools should be the focal point of integration of different socioeconomic and ethnic groups in Israeli society. Therefore tracking students in primary or middle schools based on students' characteristics is prohibited. Numerous publications of the Director General's Circulars at the Ministry of Education note that a specific committee at the Ministry is responsible for the implementation of the integration policy. This committee monitors periodically the integration process between and within schools. (See for example the Director General's Circular publication regarding the integration policy of Ethiopian students: http://cms.education.gov.il/EducationCMS/applications/mankal/arc/sd9ak3_7_47.htm). See also the Bank of Israel Report No. 2014.07 which examined whether the allocation of students to classes by socio-demographic characteristics was random during the years 2001-2010 and found very little segregation within schools in Israel.

[26] See also Lavy (2011) and Lavy and Sand (2014) for evidence that suggests no systematic nonrandom formation of classrooms in primary and middle schools in Israel.

other subjects (other than j) and we denote its effect as a cross-subject effect. The error term in the equation includes a class-specific random element u_c that allows for any type of correlation within observations of the same school across classes and an individual random element ε_{icjt}.[27]

The coefficients of interest are β_1 and β_2. The first captures the direct-subject effect of teacher's biases and the second captures the cross-subject effect of teachers' biases. We will also consider a specification where we include a measure of teachers' average biases in all three subjects, and we denote its effect as the average-subject effect instead of the two separate measures of the direct-subject effect and cross-subject effect. In that case we assume that the test scores and the choice of advanced level courses by pupils in middle/high school are determined by the following equation:

$$(2)\quad y_{icjt} = \alpha + \beta_s + \delta_j + \gamma_t + \lambda X_{icjt} + \beta_1 AS_{icjt} + u_c + \varepsilon_{icjt}$$

where AS_{icjt} is the average of the teachers' biases in all three subjects. The coefficient of interest in this case is β_1 which captures the average-subject effects of teacher's biased behavior in all three subjects on the outcome in subject j.[28] The average-subject effect captures the overall stereotypical biased environment that students are exposed to in primary school.

For the purpose of comparison, we will present first estimates based on a regression specification that includes only year and subject dummies as controls, a second specification that also includes primary school fixed effect and a third specification where we will include also pupil's characteristic (including the mother's and father's years of schooling, number of siblings, immigration status, and ethnic origin) as controls. These various specifications will provide indirect evidence about whether our measure of teachers' stereotypical biases is correlated with students' predetermined characteristics.

[27] Changing the standard errors level of the clusters to class by subject clusters has only marginal effects on the significance levels of the results. Furthermore, since homeroom teachers often teach their class more than one subject the standard errors in the baseline model are clustered by class.

[28] We note that this coefficient is by construction exactly the sum of the direct-subject coefficient and the cross-subject coefficient in a simple OLS regression without controls.

4. Results: Effect of Teachers' Biases

A. Main Results

The estimated effect of teachers' gender biased behavior on students' academic achievements, based on estimating equations 1 and 2, is shown in Table 4. We present the estimates of the direct-subject effect of teachers' stereotypical biases and of their cross-subject effect where both are included jointly in a regression. The estimates based on the sample of boys are presented in columns 1-2 and on the sample of girls in columns 3-4. In columns 5-6 we present the estimated coefficient of the average over the three subjects of the teachers' biases. Each regression includes subject and year fixed effects. Panels A and B show results of the estimated effect of teachers' biases on 8th grade GEMS test scores and on matriculation test scores respectively. In both panels, test scores in all three subjects (math, English, and Hebrew) are pooled.[29] Panel C reports the estimated effect of teachers' stereotypical biases on both 8th grade test scores and matriculation test scores, where the scores in all three subjects and in all tests (8th grade test scores and matriculation test scores) are pooled together and a dummy variable for type of test (GEMS or matriculation tests) is added to the regression. Panel D and Panel E report the estimated effect of teachers' stereotypical biases on both 8th grade test scores and matriculation test scores, but include school by subject and class fixed effects, instead of by school fixed effects. All test scores are standardized scores, by year and subject. In Table 5, we present evidence of the effects of teachers' gender biases on several long term educational outcomes: Panel A of Table 5 reports the average-subject effects on the probability of receiving a matriculation diploma; and Panel B of Table 5 reports the average-subject effects on the total number of successfully completed matriculation unit exams. These exams are taken at the end of 12th grade, more than 6 years after 'exposure' to teachers' gender biases in primary school.

Short term effects

In Panel A of Table 4, we report results from three different specifications. The simple OLS estimates (first row) are positive for boys for the direct-subject effect (column 1), the cross-subject effect (column 2) and the average-subject effect (column 5); for girls, these estimates are considerably lower. All of these estimates are not significantly different from zero. Adding primary school fixed effects to the regressions (second row) does not change the estimates for boys to a large extent, but it reduces the estimated standard errors and as a result

[29] The number of observations in Panel B is double the number of observations in Panel A because each school participated in GEMS tests only once every two years.

the most estimated effects are now significantly different from zero. The estimates for girls in this second specification are all negative and most estimated effects are now also significantly different from zero. Remarkably, adding students' characteristics leaves the estimates for boys and for girls unchanged, implying that pupil's characteristics are not correlated with the teacher's stereotypical bias measure once we control for primary school.

The estimated effect of teachers' biases on boys' outcomes is positive—this indicates that teachers' more positive assessment of boys' test scores improves their achievements at a later age. The estimate of the direct-subject effect for boys is 0.098 (SE=0.058), the estimate of the cross-subject effect is 0.180 (SE=0.088) and the average-subject effect is 0.278 (SE=0.123). Calibrating the effect size, increasing a teacher's stereotypical bias in a specific subject from zero (no gender bias) to one (the maximal value observed in the sample), will increase boys' test scores in that subject by 0.098 of a standard deviation. Increasing the average biased behavior measures in the other two subjects, from zero to one, will similarly increase boys' test scores in that subject by 0.18 of a standard deviation. If, in this scenario, we change the exposure of a male student in all three subjects from no gender bias to the highest stereotypical bias observed in the data, his test score in that subject will improve by 0.278 of a standard deviation.

The estimated effects of all biased behavior measures on girls' test scores in 8^{th} grade are negative but only two of the three are precisely measured. The estimated direct-subject effect is -0.061 (SE=0.069), the estimated cross-subject effect is -0.215 (SE=0.095) and the average-subject effect is -0.277 (SE=0.131). These estimates suggest that the overall classroom stereotypical biased environment has a much broader impact on girls' achievements than just the specific subject teacher's bias. In terms of effect size, these estimates indicate that increasing the average stereotypical bias against girls from zero to its maximal value of one will reduce girls' outcomes by 0.277 of a standard deviation.

For both genders, the estimated cross-subject effects are almost twice the size of the direct effect and the estimated average subject effects are even higher than that. This might be explained by the time that students are exposed to the teachers' stereotypical biases. For example, an increase in a students' stereotypical biased environment from zero to one reflects an increase in the average teachers' biases measures from zero to one for all three teachers, while an increase in teacher's stereotypical bias in a specific subject refers to an increase from zero to one in the bias of only one teacher. Thus, the differences in the coefficients of the direct, cross and average subject effects implies that the more time the students spend with discriminating teachers, the stronger is their influence on their students' future achievements.

Long term effects

In Panel B of Table 4, we present evidence of the effects of teachers' gender biases on test scores in the high school matriculation exams in the three subjects. Similar to the pattern we found in Panel A, the OLS estimates are not significantly different from zero, though the estimates for boys are positive, while for girls they are negative. The within school estimation reduces again the estimated standard errors, which makes most of the estimates statistically significant. Adding student characteristics as additional controls in the regressions again leaves the estimated effect almost unchanged.

Comparing the estimates based on the third specification in Panel B to those in Panel A of Table 4 reveals that the effects of the stereotypical bias measures persist through high school, since most of the point estimates are similar. For example, the effects of the stereotypical classroom environment on boys' matriculation scores (0.245) and on boys' GEMS test scores (0.281) are very similar. The effects on girls' matriculation scores have a slightly different pattern: the direct-subject effect of teachers' biases on girls' matriculation scores is -0.102 and significantly different from zero (SE=0.038), whereas the cross-subject effect on girls is smaller, -0.061, and less precisely measured (SE=0.064). The overall effect is -0.163 (SE=0.082).

In Panel C we take advantage of the estimates in Panels A and B being similar and report estimates based on pooling the middle school GEMS test scores and the high school matriculation scores data. In this pooled short- and longer-term outcome analysis, we use the third regression specification, which includes school fixed effects and students characteristics as controls. The estimates from this regression are approximately the average estimated short-term effects (Panel A) and longer-term effects (Panel B).[30]

In Panel D we report estimates from a regression where we replaced the school fixed effects with school by subject fixed effects. The within school estimation is based here on

[30] We preformed several robustness checks that we present in the online appendix: 1) since the size of our sample of teacher biases measure is relatively small, we replicated the analysis restricting the values of our teacher biases measure to be between the interval [-1,1]. The estimates which are presented in Appendix Table A4 with this modification remain very similar to those in the baseline results (Table 4 Panel C). 2) Appendix Table A5 reports the estimates of the direct-subject effect of teachers' stereotypical biases and of their cross-subject effect from two separate regressions. The estimates are very similar to those reported in Table 4 Panel C, suggesting that the direct-subject and cross-subject effects are not very correlated, though, as will be discussed in Table 6, these correlations depend on whether it is the same teacher or different teachers instructing multiple subjects. 3) in order to test the sensitivity of our results to the possible influence of some type of tracking in grade 8 or high school, we also included in Appendix Table A6 high school by subject fixed effects, and show that it has only marginal effects on our estimates.

variation within subject in each school. This is a more 'demanding' specification because it requires teachers of a given subject in a school to display sufficient grading bias to precisely measure its treatment effect on test scores. It means that students in different classes in a school are taught math, for example, by two different teachers who have enough difference in their grading bias to allow precise estimation of the effect of grading bias on students' performance. To this end we introduce here as well the 8th and 12th grade test scores in each of the three subjects. The point estimates presented in Panel D are remarkably similar to those presented in Panel C. For example, the overall effect on boys is 0.264 (se=0.067) versus 0.263 (se=0.066) in the school fixed effect estimation and the effect on girls is -0.161 (se=0.076) versus -0.160 (se=0.076) in the school fixed effect estimation.

We also estimated the models above with the outcomes being presented as the percentile ranking of students in each subject instead of the z-scores. The transformation to percentile ranking is done by subject, type of test (GEMS test or matriculation test) and year. These results are presented in Appendix Table A7. The estimates in this table are consistent with those presented in Table 4.

In panel E of Table 4 we present estimates from a specification with class fixed effects instead of school fixed effects. This specification enables us to compare students in the same class that were randomly exposed to teachers in different subjects with different stereotypical biases, thereby eliminating all stable (across-subject) class characteristics. This specification also addresses possible measurement errors in teacher's biases measure which are related to class composition variation. We note that in this specification we can uncover only the direct effect because the indirect effect does not vary within a class.[31] Note also that there could be no bias in the direct effect estimates as a result of not estimating the indirect effect because the latter is accounted for by the class fixed effect. Since the variation of our treatment variable is much lower within classes, we prefer to pool in one sample the observations of boys and girls and include an interaction term between student's gender and the treatment variable. To increase sample size and power we stack again in one sample the GEMS and matriculation test scores. We note that this sample pooling is justified since we have shown above that the short and long term effects of teachers' biases are very similar when estimated separately in a school fixed effect model. The estimated effects reported in panel E are

[31] We include only the direct-subject effect in the class fixed effects specification, since including both direct and cross-subject effects jointly (as well as estimating the average-subject effect) implies estimating the effect of a linear combination of the same treatment variable for all teachers in a given class.

positive for boys and negative for girls and they are remarkably similar to the corresponding estimates from the school fixed effects specifications. For example, in Panel C, the direct-subject effect on boys is 0.090 (SE=0.032) in the regression that includes school fixed effects, and 0.065 (SE=0.039, p-value=0.104) in the regression that includes class fixed effects, and the respective estimates for girls are -0.078 (SE=0.039) and -0.107 (SE=0.048). We also compared the panel E estimates to estimates with school fixed effects based on pooling boys' and girls' observations and allowing the treatment effect to be interacted with gender. The estimated effect on boys in this model is 0.091 (se=0.037) and on girls, -0.077 (se=0.043). These two estimates are almost identical to those presented in panel C.[32]

We conclude this section by reporting in Table 5 the effect of teachers' grading biases on two additional high school outcomes, the probability of receiving a matriculation diploma (Panel A) and the number of successfully completed matriculation exam units (Panel B). In both panels, the estimated average-subject effects are positive for boys and negative for girls and are almost all precisely measured.[33] The overall teachers' biases effect on boys' probability of receiving a matriculation diploma is 0.079 (SE=0.049, p-value=0.110), and the overall effect on their number of successfully completed matriculation exam units is 2.739 (SE=1.233). Girls' outcomes are affected in the opposite way (-0.090, SE=0.042 and -2.954, SE=1.101 respectively). These results suggest that the overall stereotypical biased environment in the classroom that students are exposed to in primary school increases boys' probability of receiving a matriculation diploma and their total number of successfully completed matriculation exam units while reducing that of girls. We also note that these two outcomes feature prominently in the admission criteria of students to universities, in particular to highly demanded fields of study, and therefore they can have far reaching implications for students' careers. We discuss these implications in more detail in Section C below.

[32] Comparing these estimates to direct-subject effects estimates from similar girls' and boys' pooled regressions with school fixed effects, which includes both direct and cross-subject effects, yields similar results.

[33] Table 5 Panel A reports the estimated effect on the probability of receiving a matriculation diploma based on a linear probability regression. We also estimated logit regressions and we present the estimated marginal effects from this model in Appendix Table A8. Since these marginal effects are similar to the estimates obtained from the linear probability regressions, we focus our discussion here on the latter estimates presented in Table 5.

B. Interpretation of the Results

Does the Grading Bias Capture Teachers' Behavior?

In this section we provide direct evidence that our gender bias measure captures teachers' and not students' behavior. In the second part of this section we provide evidence that rules out alternative interpretations of the teachers' bias measure, such as class level differences in average students' ability or non-cognitive skills.

We first examine the within classroom correlation coefficient between the bias measure in math and Hebrew when these two subjects are taught by the same teacher and compare it to the within classroom correlation coefficient when the two subjects are taught by two different teachers. We exclude English teachers from this analysis because they do not teach math or Hebrew. The majority of teachers in this sample are homeroom teachers who teach the class several subjects, including math and Hebrew. Since we identify the classes and subjects of instruction of those homeroom teachers, we are able to divide our sample into homeroom teachers who teach their classroom both math and Hebrew, and those who teach only one of these subjects. If our gender bias measure indeed captures teachers' and not students' behavior or classroom characteristics, we expect the correlation between the math and Hebrew bias measures of the same teacher to be higher relative to the case where there are two different teachers for these subjects. We next examine the correlations between teachers in different subjects from the overall sample of teachers (without restricting it to homeroom teachers). In this analysis we also expect the correlation between math or Hebrew bias and the English bias to be lower than the correlation between math and Hebrew biases, because English teachers do not teach math or Hebrew, while math and Hebrew are often taught by the same teacher.

Table 6 presents the correlations between biases of teachers by subjects of instruction: The estimates in each row in columns 1-2 are the correlation coefficients between bias measures using the sample of all teachers (same or different teachers for the two subjects), estimated in separate OLS regressions. The estimated coefficients in each row in columns 3-4 are from regressions that include primary school fixed effects. The (OLS) estimates in column 5 are based on a sub-sample of classes where the same teacher teaches both math and Hebrew while the (OLS) estimates in column 6 are based on the sample of classes where math and English are taught by two different teachers.

Comparing the estimates in columns 1-4 reveals that the correlation between the math and the Hebrew teachers' bias measures is statistically higher than the correlation between the bias shown by English teachers and the teachers of the other two subjects. Furthermore, once we add as a control primary school fixed effects, the correlation between the math and the Hebrew biases is positive and statistically significant, whereas the correlations between math/Hebrew bias and the English bias are both not significantly different from zero. Since most math teachers instruct Hebrew as well, and no English teachers instruct the other two subjects (Hebrew/math), this finding reinforces our interpretation that the teachers' bias measures do not capture students' or classes' behavior. This conclusion is reinforced further by the correlation coefficient estimate that is based on a sample of classes where the same teacher teaches math and Hebrew: the estimate is positive, large and statistically significant (0.783, SE=155). In contrast, the respective estimates based on a sample of classes with different teachers of math and Hebrew is much smaller and not significantly different from zero (0.171, SE=0.192).

We find further evidence linking the bias measure to teachers' behavior by relating it to teachers' characteristics. We propose that if the bias measures captured students' and not teachers' behavior, they should not be correlated with any of the teacher's personal characteristics. We find the opposite, however. Using administrative data from NII we are able to examine the characteristics of a sub-sample of homeroom teachers. In Table 7 we present the estimated correlations between several teachers' characteristics and teachers' bias measures. The estimates are from a separate regression for each of the teachers' characteristics that we have, using a simple OLS regression with year and subject fixed effects. Teachers' characteristics include age, ethnicity, marital status and number of children and their gender. We note that all the identified teachers in the sub-sample are female, thus we could not test this aspect in our analysis.[34]

Older and single teachers have a statistically significant pro-boys grading bias: the estimated effect is larger among teachers older than 50 years (0.206, SE=0.104). This effect is also larger among single teachers though this estimate is only marginally significant (0.315, SE=0.202). Teachers with a European-North American origin have a grading bias in favor of girls: the estimate is negative and significantly different from zero (-0.204, SE=0.113). The effects of the other three teacher's characteristics that we examine are not precisely measured:

[34] Although the issue of the correlation of teachers' gender with the measure of teachers' stereotypical biases is irrelevant in our context since all the teachers in our sample are women (as it is also in many developed countries), the literature has documented different patterns of discriminatory behavior across gender. Dee (2005) presents evidence that gender and race matches between students and teachers influence the teacher's subjective evaluations of student. Fershtman and Gneezy (2001) find a lower level of discriminatory behavior among females towards minority groups, while Reuben et al. (2014) report that both males and females tend to discriminate among job candidates based on their gender in a similar way.

being married (positive but insignificant) and the number of children and the proportion of daughters[35] (negative but not significantly different from zero). Although these findings suggest that teachers' bias is correlated with characteristics that are not randomly assigned to teachers, they support our claim that the bias measure captures teachers' behavior. It is difficult to provide reasonable explanations that link students' behavior to this pattern of correlations between teachers' bias measures and demographic characteristics.

Does the Grading Bias Measure Capture Variation in Student Characteristics Such as Ability or Non-Cognitive Skills?

It can be argued that teachers may take into account in determining the "non-blind" scores factors other than the actual performance in the tests which are not necessarily related to teachers gender biases. For example, teachers may know the true ability of students better than their external assessments might reveal, or teachers may 'award' bonus marks to students with good behavior, to those who are popular among peers or to those who make more effort in school. If so, the question that remains to be answered is whether the reason for these grading patterns is gender based or not (for example, these grading patterns might result from systematic gender differences in popularity level or good behavior). Our *first* reply to this 'threat' is that the class fixed effect estimates that we presented in the previous section are 'immune' to this concern because the within class differences estimation controls for any such class level variation in potential 'typical' gender based behavioral differences.

Moreover, we also argue that such 'threats' are unlikely to bias the school (/school by subject) fixed effect estimates. In order for such a concern regarding our interpretation of the evidence to be valid, it must be that such gender differences vary across classes and across (/within) subjects in school because our school (/school by subject) fixed effect estimation relies on within school variation across teachers (/within subjects). For example, for this argument to hold, it must be the case that in one class in school the girls have higher ability, are better behaved or make more effort and that their teacher rewards these attributes in terms of higher non-blind scores, while in the other class in the same school the boys have higher ability or are better behaved or make more effort and the teacher reward them for these attributes by giving them higher non-blind scores. So the argument against the school (/school by subject) fixed effects estimates cannot rely on gender specific behavioral differences. On

[35] Psychologists and recently also economists have shown that parenting daughters increases feminist sympathies. For example, Washington (2008) has demonstrated that the propensity to vote liberally among legislator fathers on reproductive rights increases significantly with their proportion of daughters.

the contrary, they should vary in a convoluted way across gender, classes and subjects within a school in order to be consistent with our results.

The same rationale for rejecting this alternative interpretation of the school (/school by subject) fixed effect estimates holds for many other possible explanations. Suppose that the math curriculum in 6th grade includes new material, for example geometry, and that girls do not do as well as boys in geometry questions in the internal assessments in 6th grade. If that is the source of the gender marking bias in math, then it must lead to the same marking bias for all math teachers in Tel Aviv because the teaching curriculum is identical in all primary schools in the city. But we find that some of the teachers are pro-boys and some are pro-girls and that this variation holds within school. So this potential alternative for our school (/school by subject) fixed effect specification findings is irrelevant. Equally so is the suggestion that girls do worse under pressure of external exams, or that girls do worse under the stereotypical threat environment of internal assessments, girls are more prone to peer pressure that leads to underperformance when test scores are not anonymous[36], and so on with other explanations that are based on gender specific characteristics or behavior.

Since alternative interpretations cannot rely on gender behavioral differences in general, it might be argued that the teacher bias measure reflects random variation in boys' versus girls' cognitive or behavioral outcomes in class. We present several pieces of empirical evidence in an attempt to rule out such alternative interpretations.

We first test the sensitivity of our result to adding several class level controls. Of course the class fixed effect model controls for such class level (stable across-subject) differences so we examine the sensitivity of the school fixed effect estimates to adding such controls. We first test the sensitivity of our result to adding the difference between boys' and girls' violent behavior at the class level (from students' 5th grade GEMS questionnaires) to the regressions as a control variable.[37] We also consider as a control variable the classroom proportion of boys, since previous studies have shown that it has a positive causal effect on boys' classroom misbehavior. To test for the possibility that the teacher's marking bias reflects student's ability, we add the 5th grade GEMS external test score in each subject as a control in the regression. We first add this ability measure as an individual level control and secondly, as an

[36] Burnsztyn and Jensen (2015) find that when academic effort is observed by peers, students may conform to the prevailing norms in the peer group.

[37] The level of violence was based on students' reports on classroom environment, available from the GEMS questionnaire survey administered at the midterm of 5th grade. Students were asked the extent to which they agree with the following statement: "I was involved in violence (physical fights) in school many times this year".

alternative control, we use the class mean difference between the means of boys' and girls' external exam test scores.

Table 8 presents the estimated effect of teachers' stereotypical biases on test scores when classroom level controls are added to the regressions. We report only the estimated average-subject effect of teachers' stereotypical biases, separately for boys and girls. The test scores in 8^{th} grade and in the matriculation exams in all three subjects (math, English, and Hebrew) are stacked together, and each regression includes students' characteristics and a dummy for type of exam (GEMS exam or matriculation exam), and year and primary school fixed effects as controls. The controls added to the regressions are the following: In the first row, the control variable is the difference between boys' and girls' violent behaviors in class; in the second row, the control variable is the proportion of boys in class; in the third row, the control variable is the difference in each subject between boys' and girls' 5th grade GEMS test scores in class, in the fourth row, 5th grade GEMS test score in each subject at the individual level is added as a control variable, and in the last row, all of the above control variables are added jointly in the regression. The estimated effects of the average stereotypical bias measure on boys' and on girls' outcomes are very similar to the estimates of our preferred specification (the effect of teachers' bias on both 8^{th} grade GEMS test scores and matriculation test scores are presented in Table 4, Panel C). Adding classrooms' and students' level controls to the regression leads to only minor changes in the estimates. We think that these results provide direct evidence that the measure of teachers' biased behavior is not correlated with individual or classroom characteristics.

In Table 9 we check the robustness of the class fixed effect estimates of the direct effect to adding controls that exhibit across subject variation within a class. These include the average difference between boys and girls in 5^{th} grade GEMS scores by subject and the student level test score in each of these 5^{th} grade GEMS subjects. We add each of these controls separately and jointly. The point estimates are positive for boys and negative for girls and they are significantly different from zero.

The within class variation in teachers' bias permits dismissing the possibility that our measure of gender bias of teachers may just pick up random (small sample) variation in the unobserved cross-subject stable characteristics of boys vs. girls in a particular single class. Controlling for the average difference between boys and girls in 5^{th} grade GEMS scores by subject accounts for any other subject specific variation in achievements. Yet, we also directly test the alternative interpretation that "non-blind" test scores reflect a more accurate evaluation of student's ability than "blind" test scores do. This might be a plausible alternative interpretation because teachers are able to observe students throughout the school year while the external evaluation is based on one exam only. In order for that interpretation to hold, one would expect a higher correlation between "non-blind" test scores and future test scores than between "blind" test scores. Moreover, by the same reasoning, one would expect that being a homeroom teacher (who teaches the student more than one subject) would further increase teachers' evaluation precision. We address the first question by examining the correlation between both test scores and future test scores. We use a regression that includes both "non-blind" and "blind" tests scores jointly and controls additionally for subject, type of test (GEMS or matriculation exam) and year fixed effects. Test scores in both types of tests (8^{th} grade GEMS test scores and matriculation test scores) and in all three subjects (math, English, and Hebrew) are pooled. Comparing the estimate coefficients of both tests reveals that the estimate coefficient of "blind" test scores is statistically higher than that of "non-blind" test scores (0.384, SE=0.02 versus 0.302, SE=0.019), which means that teachers' grading does not capture more information on students' ability relative to the external evaluation. In order to test the second hypothesis, that being a homeroom teacher increases teachers' evaluation precision, we include a dummy variable that indicates whether the "non-blind" test score is being graded by the homeroom teacher or not, and an interaction term between this dummy variable and "non-blind" test scores. The coefficient of the interaction term is negative and not significantly different from zero (-0.026, SE=0.022), suggesting that homeroom teachers are not better at predicting student's future academic outcomes.

Additional falsification tests

To further assess the possibility that our results are driven by the variation in boys' versus girls' cognitive ability in class, we performed the following falsification test. We assumed alternatively that the difference between non-blind and blind scores reflects teachers rewarding perceived effort (or encouragement) of low (/high) versus high (/low) achievers in class, the latter determined based on students' predetermined external scores (5^{th} grade GEMS test scores).[38] We defined a similar measure of teachers' attitude toward low/high achievers in class instead of our previous bias measure based on students' gender. This measure is defined at the class level by the difference between high performing students' and low performing students' average gap between the school score (non-blind) and the national score (blind).

[38] Teachers might differ in the way they encourage their students in class: while some teachers might want to encourage more low achievers in class for exerting effort relative to high achievers, others might prefer to encourage high achievers more.

Higher/lower achievers are defined as students with higher/lower scores in GEMS 5^{th} grade than the class average score (i.e., their mean scores in all three subjects are higher/lower than the average scores in all subjects in class). Thus, this alternative measure presumably reflects teachers' encouraging attitude toward low versus high achievers in class. Furthermore, having a teacher with a more encouraging attitude toward low (/high) achievers relative to high (/low) achievers in class might affect the students' performances in later years.

Appendix Figure A1 presents the distribution of the teachers' grading bias by high and low achievers. This evidence suggests meaningful variation across teachers in being pro-high or pro-low achieving students (61.7% of this variation is within schools).[39] Appendix Table A9 presents the estimated effect of this alternative measure of grading biases on test scores, separately for low and high achievers. Test scores in 8th Grade GEMS exams and in the matriculation exams in all three subjects (math, English, and Hebrew) are stacked together. We present here estimates based on a specification that includes students' characteristics, a dummy for type of exam (GEMS exam or matriculation exam), and year and primary school fixed effects as controls. The estimates of the direct-subject effect of teachers' attitude and of their cross-subject effect based on the sample of high performing students are presented in columns 1-2 and for the low performing students, in columns 3-4. In columns 5-6 we present the estimated coefficient of the average over the three subjects of teacher's attitude, for high and low performing students. The estimates in Appendix Table A9 are all close to zero and not significantly different from zero, both based on the low achiever and the high achiever samples. The implication of these results is that students are not affected by the meaningful grading biases for these two groups, at least as far their exam performance 3 to 6 years after initial exposure is concerned. Repeating the same analysis with a measure of teachers' grading biases based on students' parental education as an alternative measure of students' unobserved cognitive ability, yields similar results (presented in Appendix Table A10).[40] These results are in sharp contrast to the comparable evidence regarding the teachers' gender

[39] Replicating the same analysis leaving out the most extreme outliers (restricting their values to be between the interval [-1,1]) of teacher biases yields very similar results.

[40] Appendix Table A10 reports the effect of this alternative grading bias on students' later educational outcomes, separately for students with high and low parental education (higher/lower than the average parental education in class). As reported in the table, in all the four categories the estimates are small and all of them are not different from zero (except for the teacher biases' direct effect on students with high parental education). Although we do find a significant overall variation in the teachers' grading biases with respect to students' parental education (76 percent of total variation was within school) these grading biases do not carry meaningful implications for the students' later cognitive performance.

grading bias, which might imply that the main driving forces behind our results are not only the direct effect of teachers' encouragement toward a specific student but mainly teachers' overall attitude towards a well-defined group that students feel a part of. Moreover, these findings provide further evidence that the effect of teachers' gender bias measure on their students' future educational outcomes does not result from differences in boys' and girls' cognitive abilities in class.

C. Additional Results

Estimated Effects by Subject

In this section we present and discuss results of estimating the effect for each subject separately. In Table 10, we present evidence based on estimating a separate regression for each subject, using the specification of a regression that includes students' characteristics, year and primary school fixed effects. As before, we present the estimates of the direct-subject effect of teachers' biases and of their cross-subject effect from one joint regression, for boys and for girls separately. In the last two columns we present the estimated coefficient of the average effect of teachers' biases in all three subjects from separate regressions for boys and girls. In Panel A the dependent variable is 8th grade GEMS test scores whereas in Panel B the dependent variable is matriculation test scores.

In Panel A of Table 10, the estimates of the direct subject effect of teachers' biases are relatively small and not significantly different from zero for both genders, except for boys in math. The estimated effect of math teachers' biases on boys' 8th grade math test scores is the largest direct-subject effect and is positive and significant, 0.374 (SE=0.142). The estimated effect of math teachers' stereotypical biases on girls' math test scores is also relatively large, though it is not statistically larger than other direct effect estimates nor precisely measured at -0.135 (SE=0.143). This result suggests that the 8th grade students' test scores in math are mainly affected by their math teachers' biases. In Hebrew and English, the cross-subject effects are larger than the direct effect on students' 8th grade test scores for both genders.[41]

[41] These patterns might be explained by the time of exposure to each one of the teachers. Since more time is generally dedicated to math instruction in 6th grade than the instruction of the two other subjects, we postulate that it might be the reason for the stronger effect of math teachers on their students' test scores relative to the effects of the two other teachers. This might explain both the magnitude of the direct-subject effect of math teachers on their students' math test scores and the magnitude of the cross-subject effects in Hebrew and English. Another explanation for these patterns could be the different teaching practices towards boys and girls in class that are documented in the psychology and sociology literature. These are especially pronounced in math

The average-subject effect is significant in three of the six estimates, indicating that the overall stereotypical biased environment is positive and significant for boys in math and Hebrew (p-values are 0.06 and 0.07, respectively), while the opposite is true for girls in English (p-value=0.04).

In Panel B of Table 10, we present the estimated (long term) effect on the matriculation score by subject. Focusing on the overall effect presented in columns 5-6, we note that the estimated effects of teachers' biases on boys' test scores in all subjects are positive and significantly different from zero at the 5 percent level of significance. The respective effect on girls is negative in all subjects and it is significantly different from zero for the Hebrew test scores. These estimates indicate that increasing the average stereotypical bias against girls from zero to its maximal value of one will increase boys' average test scores in all three subjects by about 0.25 standard deviations.

Effects on Choice of Advanced Courses in Math and Science

The evidence for the effects of teachers' stereotypical biases on students' successful completion of advanced courses in science, math and English in high school (equivalent to honors classes in the US) is presented in Table 11 and Table 12. Table 11 presents the estimated effect of teachers' biases on the probability of successfully completing such courses and Table 12 presents the estimated effect of teachers' biases on the total number of matriculation credits a student gains in each of these advanced courses. We note that an advanced class yields 5 matriculation credits and a basic class yields only 3 matriculation credits. In science we included advanced computer science and physics courses.[42] Both tables present evidence based on estimating a separate regression for each subject, using the specification that includes students' characteristics, year and primary school fixed effects. As in earlier tables, we present the estimates of the direct-subject effect of teachers' stereotypical biases and of their cross-subject effect separately for boys and girls. We note that the direct-subject effect of teachers' biased behavior on students' science test scores (in both computer science and physics courses) refers to the effect of math teachers' stereotypical behavior on science test scores. In columns 5-6 we present the estimates for the overall exposure in all three subjects.

---

instruction: math teachers use different teaching strategies towards boys and girls (Hyde and Jaffe (1998)) and give more attention in class to boys (Leinhardt, Seewald and Engel (1979)).

[42] These subjects are chosen since they constitute the basic requirement for university admission to STEM studies in most universities in Israel (including also Chemistry, which is also required in few universities leads to only minor changes in the estimates).

Table 11 presents estimates from linear probability regressions. We also estimated logit regressions and we present the marginal effects estimated from this model in Appendix Table A8. Since these marginal effects are very similar to the estimates obtained from the linear probability regressions, we focus our discussion here on the latter estimates presented in Table 11. The estimated effect of math teachers' biased behavior on the probability of successfully completing advanced studies in math (4 or 5 credits) is positive and significant for boys (0.093, SE=0.05) and negative and marginally significant for girls (-0.075, SE=0.046, p-value=0.104). The respective estimates in English and science are not precisely measured, though they are in most cases positive for boys and negative for girls. The estimated average-subject effects are positive and significant for boys in English and in math, and they are negative but not different from zero for girls. In order to assess the magnitude of the effect for boys, we simulate a situation where a group of boys is moved from a neutral teachers' stereotypical biased environment to one with a boys' bias of one. This will increase the completion rate of boys in advanced math studies by 11 percentage points and in an advanced English program by 6.3 percentage points.

Table 12 presents the estimated effect of teachers' stereotypical biases on students' total number of matriculation credits gained in these study programs. The average-subject estimated effects on math credits are significant for both boys and girls. For boys the estimated effect is also significant for physics and English matriculation units. As before, we can simulate the impact of moving from a neutral teachers' stereotypical biased environment to one with a boys' bias of one. Such change will increase boys' number of matriculation units in math by 0.360, and decrease girls' number of math units by 0.305. Similarly, it will increase boys' number of matriculation units in physics by 0.415 and in English by 0.277.

The estimated effects of teachers' biases on math test scores are of special interest because of the considerable gender gap in math achievements at the end of high school and the impact on future labor market outcomes. [43] Our results suggest that students' math test scores and advanced math studies' completion rates are affected mainly by their math teachers' biases. These results are in line with the different teaching practices towards boys and girls in class.[44] To shed light on the effect size of these estimates, we examine how

---

[43] Several papers which have documented the correlation between students' math test scores and their future labor market income, suggest that the gender gap in math test scores in later stage of high school leads to the underrepresentation of women in STEM careers and that this sorting might be one of the reasons for gender differences in adult wages (Paglin and Rufolo (1990), Brown and Corcoran (1997)).

[44] See footnote 39 above.

eliminating teacher gender bias against girls in math affects the gender gap in math achievements. Based on our evidence in Appendix Table A2, a simulated 0.07 decrease in a math teacher biased behavior[45] will decrease boys' math achievements in middle school by 0.026 standard deviations. Such effect size will eliminate the positive gender gap in favor of boys in math achievements in middle school (0.024). It will also decrease boys' advanced math studies' completion rate in high school by 0.7 percentage point and will increase girls' completion rate by 0.5 percentage point. As a result, the gender gap in studying math at the highest level in high school would decline from 3 to 1.8 percentage points. A more drastic decline in math teachers' biases, say a decrease of one standard deviation in the math bias (0.4), will reverse the gender gap in math achievements in middle school from a positive gap of 0.024 SD in favor of boys to a negative gap of 0.126 SD in favor of girls. A similar change will also impact the gender gap in completion rates of advanced math studies from 3 percentage points in favor of the boys, to 3.6 percentage points in favor of girls.

The long term effects on high school matriculation programs and test scores have meaningful economic consequences for quantity and quality of post-secondary schooling and on earnings at adulthood. In Appendix Table A11 we present results of regressions of three key matriculation exams' outcomes on post-secondary enrollment and attainment and on earnings at age 30, based on a sample of older cohorts of Tel Aviv high school graduates. Each of the three outcomes is a good predictor of the various outcomes at adulthood. All three matriculation exams' outcomes are positively and significantly correlated with enrollment and attainment of post-secondary schooling in general and with quality (university schooling, academic colleges and other). They are also positively correlated with annual earnings at age 30; for example, each credit unit is associated with a gain of NIS 1,270 ($343) per annum and having a matriculation certificate is associated with a gain of NIS 15,648 ($4,230).

Pursuing a similar question to that raised in the current paper and in an earlier draft (Lavy and Sand 2014) and using a comparable methodology, Terrier (2015) presents similar evidence on the effects of teachers' gender biases using French data. She relies on a similar definition of teachers' gender biases and tests for the direct-subject effect of teachers' gender biases on the gender gap in achievements in class. She finds that the classes in which teachers present a high degree of discriminatory in favor of girls are also classes in which girls tend to progress significantly more than boys and choose a high level of general training at a higher probability compared to boys. The reported estimated effect of teachers' biases on boys'

[45] We note that although the mean of the teachers' biases measure is close to zero, the evidence presented in Appendix Table A.2 suggests that girls are being discriminated against in math.

versus girls' relative probabilities of choosing a high level of general training in Math and French are 0.153 (SE=0.044) and 0.163 (SE=0.048) respectively. These estimates are of the same magnitude as our reported estimates. For example, according to Table 11, simulating a situation where boys and girls move from a math teacher with neutral stereotypical biases to one with a boys' bias of one would increase the gap between the completion rate of boys and girls in advanced math studies by 16.7 percentage points.

Heterogeneous Treatment Effects of Teachers' Biases

To gain further insight into the effects of teachers' gender biased behavior on students' academic success we explore heterogeneous effects across several dimensions. In Table 13 we present the estimated effect of the average over the three subjects of the teacher's stereotypical bias on test scores for boys and for girls separately, based on different stratifications of the full sample. In Panel A we present the estimates of the average-subject effect on both GEMS and matriculation test scores and in Panel B we present the estimates of the average-subject effect only on GEMS test scores.[46] We use the specification that includes students' characteristics, year and primary school fixed effects.[47] In Panel A, the first part reports the heterogeneous treatment effects of teachers' biases by parental education level (whether the average parental years of schooling is above the median of 12 years)[48], and the second reports the heterogeneous treatment effects by the gap in parental education (referring to cases were mothers are more educated than fathers and vice versa). In Panel B we report the heterogeneous treatment effects by ethnicity (whether grandparents' place of birth is Asia/Africa) and the last reports the heterogeneous treatment effects by the child birth order (whether the student is a firstborn child).

The first part of Panel A reports the estimated effects of the overall stereotypical biased environment, based on stratifying the sample by parental level of schooling. According to the relevant sociology and psychology literature, the mother's level of education and employment status is correlated with a more egalitarian attitude towards gender roles.[49] We thus posit that

[46] Panel B addresses additional demographic information from the population registry that was available only at National Insurance Institute lab and could not be merged with the matriculation test scores (which were available at the Ministry of Education lab).

[47] The specification in Panel A includes also a dummy variable for type of exam (GEMS or matriculation exams). We note that the results obtained from separated regressions by types of tests yields similar results.

[48] We note that stratifying the sample by mothers' or fathers' education levels yields similar results.

[49] See, for example, Hoffman (1977) and Herzog et al. (1983).

students of educated mothers should be less influenced by teachers' stereotypical biases. The table indicates that the average-subject effect for girls with low parental education is indeed significantly stronger (p-value=0.014) than that for girls with high parental education (-0.281 (SE=0.109) versus 0.052 (SE=0.107)). However, boys from both groups are similarly affected by teachers' stereotypical biases (0.238 (SE=0.116) for high education background versus 0.257 (SE=0.107) for low education background).

Following a similar line of thought we also consider the heterogeneous treatment effects of teachers' stereotypical biases based on a slightly different stratification of the sample, where we group students based on the within-family parental education gap. We postulate that children from families where the mothers are more educated than the fathers might also be less prone to the influence of gender stereotypes at school. The treatment effects of teachers' stereotypical attitudes by parental education gap are presented in the second part of Table 13. The table indicates that the overall stereotypical biased environment has similar effects on boys and girls according to the parental education gap in their family.

Panel C in Table 13 presents the heterogeneous treatment effects by ethnicity—ethnic origin Asia-Africa versus all others, which includes mainly European-North American origin. This division proxies a division by income and wealth as well as other socioeconomic background characteristics. For example, the Asia-Africa ethnic group has a much lower level of parental education in comparison to the Europe-North America ethnic group. We also note that Jewish families from Asia-Africa ethnicity tend to be more patriarchal with an enhanced role for the male in family matters and decision making. Growing up in such an environment might lead children to be more susceptible to gender biases at school. Indeed the estimated effects by ethnicity have a similar pattern to those by parental education. The estimated average-subject effect is negative and significant for girls of Asia-Africa ethnicity (-0.556, SE=0.215), while the opposite is true for boys from other ethnic groups (0.289, SE=0.135).

Panel D of Table 13 reports estimated treatment effects by birth order. This stratification is not common, but much has been argued about the impact of birth order on children's personalities and behavior, especially with regard to firstborn children who are said to be a more sociable, dependent and conforming.[50] We therefore posit that a stereotypical biased environment could affect children differently by their birth order. Interestingly, the results suggest that firstborn children of both sexes tend to be slightly less influenced by teacher biases though the differences are not significantly different from zero: the average-subject

[50] See Adams (1972) for a review of the literature.

effect is significant only among non-firstborn children. The estimated effect on non-firstborn boys is positive and significant (0.313, SE=0.185)), whereas the estimated effect on non-firstborn girls is negative and significant (-0.417, SE=0.189)).

5. Conclusions

In this paper we investigate how primary school teachers' biases toward one of the genders reinforce this group's future academic achievements and orientation toward enrollment in advanced math and science studies in high school. We base the measure of teachers' gender-biased behavior on a comparison of primary school classroom boys' and girls' average test scores in a "non-blind" exam that the teacher marks, versus a "blind" exam marked externally. We then estimate the impact of this measure of teachers' stereotypical biases on the academic achievements of students in standardized national exams during middle school and high school, and on completion of higher level courses in math and sciences during high school.

For identification, we rely on the random assignments of teachers and students to classes within a given primary school. We compare students in the same primary school who are exposed to different teachers, who might have different patterns of gender stereotypical biases. This identification strategy enables us to address several threats to the interpretation of our findings and demonstrates that our estimates reflect teachers' behavior and not students' characteristics or behavior. In addition we also present evidence based on a class fixed effect model, which rules out the potential threats to estimates from a school fixed effects model.

The results we present suggest that teachers' more positive relative assessment of boys in a specific subject has a positive and significant effect on boys' overall future achievements in that subject, while having a significant negative effect on girls. We also provide evidence that suggests that spillover effects from biased behavior of teachers of different subjects can also impact students' achievements in other subjects. These effects persist through middle school and high school and actually have dramatic implications for matriculation exam scores and on the probability of receiving a matriculation diploma. Interestingly, we find that teachers' biases have a greater influence on girls with low parental schooling or from an Asia/Africa ethnicity, as well as on students who are the youngest among their siblings.

We also find that favoritism of boys among math and science teachers has an especially large and positive effect on boys' math test scores and on their successful completion of advanced math and science studies in high school: the respective effect on girls is negative. The estimates of the direct-subject effect in math are of special interest because of the considerable gender gap in math achievements and its impact on future labor market outcomes. Moreover, since this gap in math achievement partly results from teachers' stereotypical biases against girls in math, eliminating these biases will go a long way toward reducing the math achievement gender gap, and it will also decrease the gender gap in enrollment in advanced math studies. The impact on the various end-of-high-school matriculation outcomes carries meaningful economic consequences, because these high stakes outcomes sharply affect the quantity and quality of post-secondary schooling as well as impacting earnings in adulthood.

6. References

Alesina, A., P. Giuliano, and N. Nunn. 2013. "On the Origin of Gender Roles: Women and the Plough", *Quarterly Journal of Economics* 128(2), 469-530.

Adams, B.N., 1972. "Birth Order: A Critical Review", Sociometry, 35(3), 411-439.

Bae, Y. and T.M. Smith, 1997. "Women in Mathematics and Science". Findings from "The Condition of Education", *National Center for* Education Statistics 1997, no. 11.

Becker, G.S., W.H. Hubbard and K.M. Murphy, 2010. "Explaining the Worldwide Boom in Higher Education of Women", Journal of Human Capital 4, 203-241.

Benbow, C.P., 1988. "Sex-Related Differences in Precocious Mathematical Reasoning Ability: Not Illusory, not Easily Explained", Behavioral and Brain Sciences 11, 217-232.

Bertrand, M., D. Chugh and S. Mullainathan, 2005, "Implicit Discrimination", American Economic Review, 95(2), 94-98.

Bertrand, M. and E. Duflo, "Field Experiments on Discrimination", Forthcoming in: A. Banerjee and E. Duflo (Eds.), Handbook of Field Experiments.

Björn, T.H., Höglin, E. and M. Johannesson, 2011. "Are Boys Discriminated in Swedish High Schools?", Economics of Education Review 30(4), 682-690.

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

Blass, N., Tsur, S. and N. Zussman, 2014. "Segregation of Students in Primary and Middle Schools", Bank of Israel Discussion Paper No. 2014.07.

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, 283-308.

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

Brown, C. and M. Corcoran, 1997. "Sex-Based Differences in School Content and the Male-Female Wage Gap", Journal of Labor Economics 15, 431-465.

Burgess, S. and E. Greaves, 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities", Journal of Labor Economics 31, 535-576.

Burnsztyn, L. and R. Jensen, 2015, "How Does Peer Pressure Affect Educational Investments?", The Quarterly Journal of Economics, 130(3), 1329-1367.

Collaer, M.L. and M. Hines, 1995. "Human Behavioral Sex Differences: a Role for Gonadal Hormones during Early Development?", Psychological Bulletin 118, 55.

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

Dee, T. S., 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" American Economic Review, 95(2), 158-165.

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

Ellison, G. and A. Swanson, 2010. "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions", Journal of Economic Perspectives 24, 109-128.

Fershtman, C. and U. Gneezy, 2001. "Discrimination in a Segmented Society: An Experimental Approach", *The Quarterly Journal of Economic* 116(1), 351-377.

Friedman, L., 1989. "Mathematics and the Gender Gap: A Meta-Analysis of Recent Studies on Sex Differences in Mathematical Tasks", Review of Educational Research 59, 185-213.

Fryer, R.G. and S.D. Levitt, 2010. "An Empirical Analysis of the Gender Gap in Mathematics", American Economic Journal: Applied Economics 2, 210-240.

Gneezy, U., M. Niederle and A. Rustichini, 2003. "Performance in Competitive Environments: Gender Differences", The Quarterly Journal of Economics 118, 1049-1074.

Goldin, C., "A Grand Gender Convergence: Its Last Chapter", American Economic Review. Forthcoming 104.

Goldin, C., L.F. Katz and I. Kuziemko, 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap", Journal of Economic Perspectives 20, 133-156.

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

Guiso, L., F. Monte, P. Sapienza and L. Zingales, 2008. "Culture, Gender, and Math", Science 320, 1164- 1165.

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

Herzog, A.R., Bachman, J.G., and L.D., Johnston, 1983. "Paid Work, Child Care, and Housework: A National Survey of High School Seniors' Preferences for Sharing Responsibilities Between Husband and Wife", Sex Role, 9(1), 109-135.

Hoffman, L. W., 1977. "Changes in Family Roles, Socialization, and Sex Differences", American Psychologist, 32(8), 644-657.

Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., and Williams, C.C, 2008. "Gender Similarities Characterize Math Performance", Science 321(5888), 494–495.

Hyde, J.S., and S. Jaffe, 1998. "Perspective from Social and Feminist Psychology", Educational Research 27 (5), 14-16.

Inglehart, R. and P. Norris, 2003. "Explaining the Rising Tide of Gender Equality", In Inglehart, R. and P. Norris, Rising Tide: Gender Equality and Cultural Change Around the World. (Cambridge University Press)

Lansdell, H., 1962. "A Sex Difference in Effect of Temporallobe Neurosurgery on Design Preference", Nature 194, 852-854.

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

Lavy, V., 2011. "What Makes an Effective Teacher? Quasi-Experimental Evidence", Forthcoming CESifo Economic Studies.

Lavy, V. and E. Sand, “On the Origins of the Gender Human Capital Gap: Short and Long Term Effect of Teachers’ Stereotypes”, Draft, Applied Micro Seminar, Department of Economics, Hebrew University of Jerusalem, July 2014.

Leinhardt, G., A.M. Seewald and M. Engel, 1979. "Learning What's Taught: Sex Differences in Instruction", Journal of Educational Psychology 71, 432-439.

Leslie, S.J., A. Cimpian, M. Meyer and E. Freeland, 2015. “Expectations of Brilliance underlie Gender Distributions across Academic Disciplines”, Science 347, 262-26.

Lewis, M. and J. Brooks-Gunn, 1979. "Towards a Theory of Social Cognition: The Development of Self", New Directions for Child and Adolescent Development, 4, 1-20.

Lewis, M. and R. Freedle, 1972. "Mother-Infant Dyad: The Cradle of Meaning", In P.K., Pliner and T. Lester Alloway (Eds), "Communication and Affect: Language and Thought", Oxford England: Academic Press.

Machin, S. and T. Pekkarinen, 2008. "Global Sex Differences in Test Score Variability", Science 322, 1331-1332.

Murnane, R.J., J.B. Willett and F. Levy, 1995. “The Growing Importance of Cognitive Skills in Wage Determination”, Review of Economics and Statistics 77, 251-266.

Niederle, M. and L. Vesterlund, 2007. “Do Women Shy Away from Competition? Do Men Compete Too Much?” The Quarterly Journal of Economics 122, 1067-1101.

Paglin, M. and A.M. Rufolo, 1990. “Heterogeneous Human Capital, Occupational Choice, and Male-Female Earnings Differences”, Journal of Labor Economics 8, 123-144.

Pope, D.G. and J.R. Sydnor, 2010. “Geographic Variation in the Gender Differences in Test Scores”, The Journal of Economic Perspectives 24, 95-108.

Reuben, E., Sapienza P. and L. Zingales, 2014. ‘How Stereotypes Impair Women's Careers in Science,” Proceeding of the National Academy of Science, Forthcoming.

Sadker, M. and D. Sadker, 1986. "Sexism in the Classroom: From Grade School to Graduate School", Phi Delta Kappan 67, 512-515.

Spolaore, E. and R. Wacziarg, 2009, “The Diffusion of Development”, The Quarterly Journal of Economics 124, 469-529.

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

Vandenberg, S. G. 1968. "Primary Mental Abilities or General Intelligence? Evidence from Twin Studies", In J.M. Thoday and A.S. Parkers (Eds), "Genetics and Environmental Influences on Behaviour", New York: Plenum.

Voyer, D., S. Voyer and M.P. Bryden, 1995. "Magnitude of Sex Differences in Spatial Abilities: A Meta-Analysis and Consideration of Critical Variables", Psychological Bulletin 117, 250.

Waber, D.P., 1976. "Sex Differences in Cognition: a Function of Maturation Rate?", Science 192, 572-574.

Wilder, Gita Z., and K. Powell, 1989. "Sex Differences in Test Performance: A Survey of Literature". No. 89. New York: College Entrance Examination Board.

Washington, E. L., 2008. “Female Socialization: How Daughters Affect Their Legislator Fathers”, The American Economic Review 98(1), 311-332.

Witelson, D.F., 1976. "Sex and the Single Hemisphere: Specialization of the Right Hemisphere for Spatial Processing", Science 193, 425-427.

Table 1: Summary Statistics of Students' Characteristics by Cohort

	2002 (1)	2003 (2)	2004 (3)
Mean Father's Education	13.477	13.339	12.992
	(3.391)	(3.468)	(3.482)
Mean Mother's Education	13.614	13.610	13.287
	(3.073)	(3.115)	(3.116)
Mean Number of Siblings	2.190	2.336	2.259
	(0.996)	(1.039)	(1.130)
Proportion of Asia/Africa Ethnicity	0.114	0.110	0.103
	(0.318)	(0.313)	(0.304)
Proportion of Europe/America Ethnicity	0.171	0.182	0.189
	(0.376)	(0.386)	(0.392)
Proportion of Israel Ethnicity	0.611	0.615	0.601
	(0.488)	(0.487)	(0.490)
Proportion of Former Soviet Union	0.081	0.063	0.083
	(0.273)	(0.244)	(0.276)
Number of Students	867	1127	1017
Number of Elementary Schools	17	20	20
Number of Elementary Classes	33	41	38
Number of Middle Schools	5	7	5
Number of Middle School Classes	26	32	31

Notes: Each column is based on a different cohort of sixth grade students. Number of middle schools and middle school classes refers only to middle school with GEMS test scores. Standard deviations are reported in parentheses.

Table 2: Means and Standard Deviations of National and Primary School Exams Scores and Difference Between them, by Gender

	Boy			Girl			Difference Between School and National Exams Scores of Boys and Girls
	School Score Exams	National Score Exams	Difference Between School and National Exams Scores	School Score Exams	National Score Exams	Difference Between School and National Exams Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Hebrew	-0.098	-0.136	0.037	0.168	0.139	0.029	0.009
	(1.000)	(1.022)	(1.047)	(0.951)	(0.952)	(0.925)	
Math	0.052	-0.014	0.066	0.003	0.014	-0.011	0.076
	(0.985)	(1.034)	(0.960)	(0.971)	(0.963)	(0.903)	
English	-0.035	-0.046	0.010	0.111	0.047	0.064	-0.054
	(1.002)	(1.035)	(0.999)	(0.941)	(0.963)	(0.933)	
Number of Students	4245	4246	4245	4122	4123	4122	8367

Notes: The national exams scores and the primary school exams scores are standardized scores. The number of students refers to the number of students in all three subjects. The last column (column 7) equal to the difference between boys' school and national exams scores (column 3) less the difference between girls' school and national exams scores (column 6). Standard deviations are reported in parentheses.

Table 3: Means and Standard Deviations of National Exams Scores in Middle School and High School at the Student Level, by Gender

	Boy	Girl	Boy	Girl
	National Score Exams	**National Score Exams**	**National Score Exams**	**National Score Exams**
	(1)	(2)	(3)	(4)
	Middle School		**High School**	
Hebrew	-0.147	0.151	-0.099	0.095
	(1.061)	(0.908)	(1.053)	(0.936)
Math	0.012	-0.012	0.041	-0.040
	(1.043)	(0.952)	(1.040)	(0.957)
English	-0.081	0.085	-0.011	0.011
	(1.036)	(0.952)	(1.023)	(0.976)
Number of Students	1676	1618	3889	4041

Notes: The national exams scores are standardized scores. The number of students refers to the number of students tested in all three subjects. Matriculation test scores are weighted based on the number of credit units taken, as computed by the Ministry of Education. Standard deviations are reported in parentheses.

Table 4: Estimated Effect of Teachers' Biases on Test Scores

	Boy		Girl		Boy	Girl
	Direct-Subject Effect	Cross-Subject Effect	Direct-Subject Effect	Cross-Subject Effect	Average-Subject Effect	Average-Subject Effect
	(1)	(2)	(3)	(4)	(5)	(6)
A. 8th Grade GEMS Test Scores						
OLS	0.090	0.113	-0.007	-0.041	0.203	-0.048
	(0.081)	(0.131)	(0.092)	(0.144)	(0.189)	(0.215)
6th Grade School Fixed Effects	0.096	0.165*	-0.074	-0.247**	0.261**	-0.321**
	(0.060)	(0.091)	(0.073)	(0.099)	(0.129)	(0.140)
6th Grade School Fixed Effects and Student Characteristics	0.098*	0.180**	-0.061	-0.215**	0.278**	-0.277**
	(0.058)	(0.088)	(0.069)	(0.095)	(0.123)	(0.131)
Number of Observations	1593	1593	1510	1510	1593	1510
B. Matriculation Test Scores						
OLS	0.088	0.190	-0.072	-0.029	0.278	-0.101
	(0.061)	(0.117)	(0.065)	(0.105)	(0.169)	(0.156)
6th Grade School Fixed Effects	0.086**	0.188***	-0.134***	-0.145**	0.274***	-0.280***
	(0.036)	(0.061)	(0.042)	(0.069)	(0.080)	(0.088)
6th Grade School Fixed Effects and Student Characteristics	0.083**	0.163***	-0.102***	-0.061	0.245***	-0.163**
	(0.034)	(0.055)	(0.038)	(0.064)	(0.071)	(0.082)
Number of Observations	3704	3704	3788	3788	3704	3788

Notes: The table reports the estimates of teachers' biases on 8th grade GEMS test scores (Panel A) and on matriculation test scores (Panel B) respectively. In both Panel A and B test scores in all three subjects (math, English, and Hebrew) are pooled. These test scores are standardized scores, by year and subject. The first specification is a simple OLS regression with subject and year fixed effects; the second specification includes also primary school fixed effects; the third specification includes additionally students' characteristics (gender, parental education, number of siblings, and dummies for four ethnicity groups). The direct-subject effect is the effect of teacher bias in a specific subject on the test scores in the same subject; the cross-subject effect is the impact of the average teacher bias in the other subjects on the test scores in the referred subject. Average-subject effect is the impact of the average teacher bias in all subjects. The estimates in each row in columns 1-2 are each from a joint regression and so are the estimates in columns 3-4. The estimates in each row in columns 5-6 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 4: Estimated Effect of Teachers' Biases on Test Scores- Continued

	Boy		Girl		Boy	Girl
	Direct-Subject Effect	**Cross-Subject Effect**	**Direct-Subject Effect**	**Cross-Subject Effect**	**Average-Subject Effect**	**Average-Subject Effect**
	(1)	(2)	(3)	(4)	(5)	(6)
C. Pooled 8th Grade GEMS and 12th Grade Matriculation Test Scores						
6th Grade School Fixed Effects and Student Characteristics	0.090***	0.173***	-0.078**	-0.082	0.263***	-0.160**
	(0.032)	(0.048)	(0.039)	(0.060)	(0.066)	(0.076)
Number of Observations	5297	5297	5298	5298	5297	5298
D. Pooled 8th Grade GEMS and 12th Grade Matriculation Test Scores with School by Subject Fixed Effects						
6th Grade School by Subject Fixed Effects and Student Characteristics	0.114***	0.150***	-0.058*	-0.103*	0.264***	-0.161**
	(0.034)	(0.045)	(0.032)	(0.057)	(0.067)	(0.076)
Number of Observations	5297	5297	5298	5298	5297	5298
E. Pooled 8th Grade GEMS and 12th Grade Matriculation Test Scores with Class Fixed Effects						
6th Grade School by Subject Fixed Effects and Student Characteristics	0.065		-0.107**			
	(0.039)		(0.048)			
Number of Observations	10595		10595			

Notes: In all panels tests scores in all three subjects (math, English, and Hebrew) and in all tests (8th grade GEMS and Matriculation exams) are pooled. These test scores are standardized scores, by year and subject. In Panel C, the specification includes students' characteristics (gender, parental education, number of siblings, and dummies for four ethnicity groups), a dummy variable for type of exam (GEMS or Matriculation exams) and subject, year and primary school fixed effects. In Panel D, the specification is the same as in Panel C, but includes school by subject fixed effects instead of school fixed effects. In Panels E, girls and boys test scores are pooled. The specification includes the direct-subject effect and an interaction term between the direct-subject effect and the gender of the student. In addition, the specification includes primary class school fixed effects instead of primary school fixed effects and also controls for students' characteristics, type of test (8th grade GEMS or Matriculation exams), year and subject fixed effect. The direct-subject effect is the effect of teacher bias in a specific subject on the test scores in the same subject; the cross-subject effect is the impact of the average teacher bias in the other subjects on the test scores in the referred subject. Average-subject effect is the impact of the average teacher bias in all subjects. The estimates in each row in columns 1-2 are each from a joint regression and so are the estimates in columns 3-4. The estimates in each row in columns 5-6 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 5: Estimated Effect of Teachers' Biases on Other Educational Outcomes

	Boy	Girl
	Average-Subject Effect	**Average-Subject Effect**
	(5)	(6)
A. Probability of Receiving a Matriculation Diploma		
OLS	0.117	-0.067
	(0.077)	(0.059)
6th Grade School Fixed Effects	0.091*	-0.123***
	(0.051)	(0.040)
6th Grade School Fixed Effects and Student Characteristics	0.079	-0.090**
	(0.049)	(0.042)
Number of Observations	1242	1270
B. Total Number of Successfully Completed Matriculation Exams' Units		
OLS	3.156	-2.343
	(1.953)	(1.547)
6th Grade School Fixed Effects	3.177**	-4.087***
	(1.312)	(1.112)
6th Grade School Fixed Effects and Student Characteristics	2.739**	-2.954***
	(1.233)	(1.101)
Number of Observations	1242	1270

Notes: The table reports the estimates of teachers' biases on other educational outcomes: Panel A shows results of the estimated effect of teachers' biases on the probability of receiving a matriculation diploma and Panel B shows results of the estimated effect of teachers' biases on the total number of successfully completed matriculation exams units. The first specification is a simple OLS regression with subject and year fixed effects; the second specification includes also primary school fixed effects; the third specification includes additionally students characteristics (gender, parental education, number of siblings, and dummies for four ethnicity groups). Average-subject effect is the impact of the average teacher bias in all subjects. The estimates in each row in columns1-2 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 6: Correlations between Biases of Teachers by Subjects of Instruction

	Overall		Within School		Same Teachers	Different Teachers
	Teachers' Biases in Hebrew	**Teachers' Biases in Math**	**Teachers' Biases in Hebrew**	**Teachers' Biases in Math**	**Teachers' Biases in Hebrew**	**Teachers' Biases in Hebrew**
	(1)	(2)	(3)	(4)	(5)	(6)
Teachers' Biases in Math	0.589***		0.338***		0.783***	0.171
	(0.095)		(0.121)		(0.155)	(0.192)
Teachers' Biases in English	0.289**	0.270***	0.075	0.161		
	(0.092)	(0.079)	(0.114)	(0.105)		
Number of Observations	112		112		36	42

Notes: The table presents the estimated correlation coefficient of teachers' biases measures by subjects of instruction. The estimates in each row in columns 1-2 are the correlation coefficients between bias measures using the sample of all teachers (same or different teachers for each two subjects), from separate OLS regressions. The estimated coefficients in each row in columns 3-4 are similar to those in columns 1-2, but primary school fixed effects are included in the regressions. The (OLS) estimated coefficient in column 5 is between biases measures of the same teachers who instruct students from the same class both math and Hebrew; and the (OLS) estimated coefficient in column 6 is between biases measures of different teachers who instruct students from the same class in both math and Hebrew. Both last estimated coefficients are from using separate OLS regressions. Standard errors are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 7: Correlation of Teachers' Biases Measure with Characteristics of Teachers

	Age Dummy (dummy=1 if Older than 50 Years Old)	Ethnicity Europe/America	Married	Single	Number of Teachers' Offspring	Proportion of Daughters among Teachers' Offspring	At Least one Daughter among Teachers' Offspring
	(1)	(2)	(3)	(4)	(7)	(5)	(6)
OLS	0.206*	-0.204*	0.032	0.315	-0.034	-0.047	-0.090
	(0.104)	(0.113)	(0.141)	(0.202)	(0.046)	(0.186)	(0.173)
Number of Teachers	114						

Notes: The table presents the estimated correlation between several teachers' characteristics and teachers' stereotypical bias measure. Each regression includes subject and year fixed effects. The estimates in each column in columns 1-6 are from a separated regression. Standard errors are reported in parentheses. Significance level of regressions are reported as follows: "***"=1% level, "**"=5% level, and "*"=10% level.

Table 8: Estimated Effect of Teachers' Biases on Test Scores, Controlling for Several Classroom's Characteristics

	Boy	Girl
	Average-Subject Effect	Average-Subject Effect
	(5)	(2)
Difference Between Boys' Violent Behavior and Girls' Violent Behavior in Class	0.272***	-0.187**
	(0.077)	0.079
Proportion of Boys	0.258***	-0.122
	(0.073)	(0.076)
Difference Between Boys' and Girls' 5th Grade GEMS Scores	0.250***	-0.151**
	(0.065)	(0.070)
5th Grade GEMS Scores	0.264***	-0.133**
	(0.075)	(0.055)
All Above Controls	0.372***	-0.093
	(0.088)	(0.070)
Number of Observations	4118	4159

Notes: See Table 4 Panel C. The scores in all three subjects (math, English, and Hebrew) and in all tests (GEMS and matriculation exams) are pooled together. Each regression includes students' characteristics, type of test (8th grade GEMS or matriculation exams), primary school, year and subject fixed effect. The first regression includes as a control the difference between boys' and girls' violent behaviors in primary school class; the second regression includes as a control the proportion of boys in primary school class; the third regression includes the differences between boys' grade to girls' grades in 5th grade national exams in each subject as a control variable; the fourth regression includes the 5th grade national exams at the student level in each subject as a control variable; and in the last row of the table all above controls are included jointly in the regression. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: "***"=1% level, "**"=5% level, and "*"=10% level.

Table 9: Estimated Effect of Teachers' Biases on Test Scores from Regressions with Class Fixed Effects, Controlling for Several Classroom's Characteristics

	Boy	Girl
	Direct-Subject Effect	Direct-Subject Effect
	(1)	(2)
Difference Between Boys' and Girls' 5th Grade GEMS Scores	0.052	-0.119**
	(0.049)	(0.057)
5th Grade GEMS Scores	0.172***	-0.134***
	(0.041)	(0.040)
Both	0.210***	-0.095
	(0.054)	(0.058)
Number of Observations	9953	9953

Notes: Notes: See Table 4 Panel E. The scores in all three subjects (math, English, and Hebrew) and in all tests (GEMS and matriculation exams) are pooled. Girls and boys test scores are also pooled. Each regression includes the direct-subject effect and an interaction term between direct-subject effect and the gender of the student, and also controls for students' characteristics, type of test (8th grade GEMS or matriculation exams), primary class school, year and subject fixed effect. The first regression includes additionally the differences between boys' grade to girls' grades in each subject in 5th grade national exams as a control variable; the second regression includes the 5th grade national exams at the student level in each subject as a control; the third regression includes both previous controls in the same regression. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 10: Estimated Effect of Teachers' Biases on Test Scores, by Subject

	Boy		Girl		Boy	Girl
	Direct-Subject Effect	**Cross-Subject Effect**	**Direct-Subject Effect**	**Cross-Subject Effect**	**Average-Subject Effect**	**Average-Subject Effect**
	(1)	(2)	(3)	(4)	(5)	(6)
A. 8th Grade GEMS Test Scores						
Hebrew	0.038	0.282	-0.019	-0.209	0.325*	-0.230
	(0.125)	(0.183)	(0.180)	(0.160)	(0.177)	(0.179)
Math	0.374**	0.029	-0.135	-0.139	0.367*	-0.267
	(0.142)	(0.145)	(0.143)	(0.170)	(0.193)	(0.178)
English	0.076	0.079	-0.021	-0.313**	0.161	-0.315**
	(0.108)	(0.142)	(0.094)	(0.153)	(0.167)	(0.149)
						-2.107
B. Matriculation Test Scores						
Hebrew	0.035	0.216*	-0.110	-0.094	0.246**	-0.205**
	(0.094)	(0.122)	(0.085)	(0.110)	(0.095)	(0.097)
Math	0.173	0.113	-0.099	-0.064	0.281**	-0.164
	(0.123)	(0.118)	(0.091)	(0.124)	(0.116)	(0.107)
English	0.096**	0.113*	-0.024	-0.093	0.209***	-0.119
	(0.048)	(0.060)	(0.072)	(0.086)	(0.070)	(0.087)

Notes: See Table 4 Panel A and Panel B. Each row present estimates from separate regressions for each subject. Each regression includes students' characteristics, primary school, year and subject fixed effect. The estimates in each row in columns 1-2 are each from a joint regression and so are the estimates in columns 3-4. The estimates in each row in columns 5-6 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: “***”=1% level, “**”=5% level, and “*”=10% level.

Table 11: Estimated Effect of Teachers' Biases on Students' Probability of Successfully Completing Advanced Courses in High School

	Boy		Girl		Boy	Girl
	Direct-Subject Effect	**Cross-Subject Effect**	**Direct-Subject Effect**	**Cross-Subject Effect**	**Average-Subject Effect**	**Average-Subject Effect**
	(1)	(2)	(3)	(4)	(5)	(6)
English (dummy=1 if # units=5|4)	0.026	0.036	-0.013	-0.033	0.063*	-0.045
	(0.022)	(0.028)	(0.030)	(0.035)	(0.035)	(0.035)
Math (dummy=1 if # units=5|4)	0.093*	0.025	-0.075	-0.010	0.114*	-0.086
	(0.050)	(0.052)	(0.046)	(0.059)	(0.061)	(0.057)
Physics/Computer Science (dummy=1 if units=5)	0.017	0.004	0.020	-0.003	0.021	0.018
	(0.054)	(0.057)	(0.030)	(0.035)	(0.045)	(0.028)

Notes: See Table 4. Each row presents estimates from separate linear probability regressions for each subject (English /Math/Science oriented subjects). The dependent variables are discrete and equals one if the number of matriculation credits exceeds a certain level. Each regression includes students' characteristics, primary school and year fixed effect. The estimates in each row in columns 1-2 are each from a joint regression and so are the estimates in columns 3-4. The estimates in each row in columns 5-6 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: "***"=1% level, "**"=5% level, and "*"=10% level.

Table 12: Estimated Effect of Teachers' Biases on Students' Total Number of Successfully Completed Units in Science, Math and English Courses in High School

	Boy		Girl		Boy	Girl
	Direct-Subject Effect	Cross-Subject Effect	Direct-Subject Effect	Cross-Subject Effect	Average-Subject Effect	Average-Subject Effect
	(1)	(2)	(3)	(4)	(5)	(6)
English	0.033	0.250**	-0.195	0.044	0.277**	-0.155
	(0.073)	(0.120)	(0.124)	(0.148)	(0.117)	(0.140)
Math	0.210	0.154	-0.066	-0.241	0.360**	-0.305*
	(0.163)	(0.167)	(0.122)	(0.179)	(0.170)	(0.156)
Computer Science	-0.055	0.013	0.017	-0.080	-0.040	-0.062
	(0.163)	(0.170)	(0.111)	(0.145)	(0.184)	(0.108)
Physics	0.251	0.171	-0.059	0.072	0.415*	0.011
	(0.244)	(0.254)	(0.110)	(0.109)	(0.213)	(0.117)
Sum of Number of Units in Math, Physics and Computer Science	0.465	0.568	-0.276	-0.017	1.026**	-0.299
	(0.545)	(0.527)	(0.282)	(0.377)	(0.438)	(0.318)

Notes: See Table 4. Each row present estimates from separate OLS regression for each subject (English /Math/Science oriented subjects). The dependent variables in each row are continuous and equals to the total number of matriculation units students' gained in each of these study programs. Each regression includes students' characteristics, primary school and year fixed effect. The estimates in each row in columns 1-2 are each from a joint regression and so are the estimates in columns 3-4. The estimates in each row in columns 5-6 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: "***"=1% level, "**"=5% level, and "*"=10% level.

Table 13: Estimated Average-Subject Effect of Teachers' Stereotypes on Test Scores, by Sub-Groups

	Boy	Girl	Boy	Girl
	(1)	(2)	(3)	(4)
A. Pooled 8th Grade GEMS and 12th grade Matriculation Test Scores				
	Low Parental Education		**High Parental Education**	
	0.238**	-0.281**	0.257***	0.052
	(0.116)	(0.109)	(0.082)	(0.107)
Number of Observations	2471	2644	2826	2654
	Mothers are More Educated than Fathers		**Mothers are Less/Equally Educated than Fathers**	
	0.135	-0.209	0.290***	-0.183*
	(0.167)	(0.133)	(0.079)	(0.096)
Number of Observations	1621	1575	3676	3723
B. 8th Grade GEMS Test Scores				
	Ethnicity Asia/Africa		**Other Ethnic Groups**	
	0.242	-0.556***	0.289**	-0.223
	(0.237)	(0.215)	(0.135)	(0.186)
Number of Observations	502	495	918	822
	Firstborn Children		**Non-Firstborn Children**	
	0.179	-0.189	0.313*	-0.417**
	(0.229)	(0.237)	(0.185)	(0.189)
Number of Observations	527	496	893	819

Notes: The table presents the estimated average-subject effect of teachers' stereotypical attitude on test scores, by sub-groups. Each regression includes pupil's characteristics, primary school, year and subject fixed effect. In Panel A the scores in all three subjects (math, English, and Hebrew) and in all tests (8th grade GEMS and matriculation exams) are pooled and the regression includes a dummy for the type of exam (8th grade GEMS versus matriculation exams). High parental education is defined as more than 12 years of average parental schooling. In Panel B the scores are only 8th grade GEMS test scores since the data were obtained from the protected research lab of the National Insurance Institute and could not be merged with matriculation test scores. The scores are pooled in all three subjects (math, English, and Hebrew). Ethnicity Asia/Africa is defined by grandparents' place of birth. The estimates in each row in columns 1-4 are each from separate regressions. Standard errors are clustered by class and are reported in parentheses. Significance level of regressions are reported as follows: "***"=1% level, "**"=5% level, and "*"=10% level.