Journal of Public Economics 92 (2008) 2083–2105

Contents lists available at ScienceDirect

Journal of Public Economics

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

Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment ☆

Victor Lavy *

Department of Economics, Hebrew University, Israel
Department of Economics, Royal Holloway, University of London, United Kingdom
NBER, United States
CEPR, United Kingdom

ARTICLE INFO

Article history:
Received 9 November 2006
Received in revised form 12 February 2008
Accepted 25 February 2008
Available online 12 March 2008

Keywords:
Gender stereotypes
Discrimination
Natural experiment

ABSTRACT

Schools and teachers are often said to be a source of stereotypes that harm girls. This paper tests for the existence of gender stereotyping and discrimination by public high-school teachers in Israel. It uses a natural experiment based on blind and non-blind scores that students receive on matriculation exams in their senior year. Using data on test results in several subjects in the humanities and sciences, I found, contrary to expectations, that male students face discrimination in each subject. These biases widen the female–male achievement difference because girls outperform boys in all subjects, except English, and at all levels of the curriculum. The bias is evident in all segments of the ability and performance distribution and is robust to various individual controls. Several explanations based on differential behavior between boys and girls are not supported empirically. However, the size of the difference is very sensitive to teachers' characteristics, suggesting that the bias against male students is the result of teachers', and not students', behavior.

1. Introduction

Schools and teachers are often said to be a source of stereotypes that harm girls. Bernard (1979), Dusek and Joseph (1983), Madon et al. (1998), and Tiedemann (2000) are only a few of the many scholars who have claimed that teachers occasionally rely on stereotypes in forming perceptions about their students.[1] Examples of such stereotypical perceptions are that boys excel in math and science and girls excel in other subjects, or that boys have more talent and that girls compensate by working hard (Deaux and LaFrance, 1998). Girls are then encouraged, on the basis of these stereotypes, to pursue traditionally female studies instead of mathematics, science, and other traditionally male subject areas, from as early as first grade (Carr et al., 1999) and women are steered toward certain occupations, as evidenced by studies of college students (Glick et al., 1995), PhD holding research students, (Rowsey, 1997) and others (Deaux and LaFrance, 1998). Another claim about stereotypes is that beliefs are manifested through teachers' evaluation of students. This claim is supported by evidence from a survey of 1st grade teachers (Fennema et al., 1990), the AAUW (1992) report, which surveyed girls from kindergarten to 12th grade, a survey of mothers and their 11–12 year old children

☆ I would like to thank Daron Acemoglu, Joshua Angrist, David Author, Abhijit Banerjee, Raquel Fernandez, Oded Galor, Esther Duflo, Guy Michaels, Sendhil Mullainathan, Analia Schlouser, Michael Steinberger for discussion and suggestions and seminar participants at CEPR conference, Hebrew University, MIT, Tel-Aviv University and the referees and editor of this journal for useful comments. Alex Levkov and Katherine Eyal provided excellent research assistance. I also thank the Israeli Education Ministry officials for assisting with the data. All errors are my own.

* Department of Economics, Hebrew University, Mt. Scopus, Jerusalem 91905, Israel.
E-mail address: msvictor@mscc.huji.ac.il.

[1] These conclusions result from studies with widely different samples, such as 240 male and female high-school teachers (Bernard, 1979), a meta study of 75 different studies (Dusek and Joseph, 1983), 2000 American 7th graders in public school math classes (Madon et al., 1998) and 600 German elementary school students (3rd and 4th graders and their parents) (Tiedemann, 2000).

0047-2727/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2008.02.009

(Jacobs and Eccles, 1992), and others (Ben-Zvi Mayer et al., 1995; Hildebrand, 1996).[2] The bottom line of the literature on gender stereotypes is that they are partly responsible for one of the alleged forms of discrimination against women and that they may have far reaching implications and consequences regarding gender differences in human capital investment and outcomes. However, there is very little convincing evidence to date that substantiates these claims and this study attempts, using a unique empirical context, to fill some of this deficiency.

This paper tests for the existence of gender stereotyping and discrimination by public high-school teachers in Israel. It uses a natural experiment based on blind and non-blind scores that students receive on matriculation exams in their junior and senior years. The natural experiment arises from rules that are used in Israel to determine scores in matriculation subjects (these rules are explained in detail in Section 3.1). This testing protocol elicits two scores, a blind and a non-blind score, both of which are meant to measure the student's knowledge of the same material. Due to this testing method, we may safely assume that the blind score is free of any bias that might be caused by stereotyped discrimination on the part of the external examiner. The non-blind score, however, may reflect biases occasioned by teachers' gender stereotypes.

As long as the two scores are comparable, i.e., as long as they measure the same skills and cognitive achievements (grounds for this assumption are discussed further in Section 3.1), the blind score may be used as the counterfactual measure to the non-blind score, which may be affected ("treated") by stereotyping and discrimination. This identification framework is similar to that used by Goldin and Rouse (2000) and by Blank (1991). I use the difference between the boys' and girls' gaps between the blind and the non-blind test scores as a measure of a potential gender bias.

Using data on all matriculation scores of several cohorts of high-school seniors in Israel, I applied this natural-experiment framework to test for the existence of such a gender bias in nine subjects—four in the humanities (English, history, literature and biblical studies), mathematics, and four in science (biology, chemistry, computer science, and physics). The distributions of the blind and non-blind scores in many of these subjects are very similar and, in many cases, are identical. The basic results of these quasi-experiments show that, contrary to expectations, the potential bias is against male students. The sign of this difference is the same in all nine subjects examined and in all tests in cases where there is more than one exam per subject. The extent of the potential bias varies by subject and test, ranging from 0.05 to 0.25 of the standard deviation of the blind-score distribution. This gap against male students, on average, doubles the gender-score difference because female students outperform male students on state external exams in all subjects except English. The results are not sensitive to various student-level controls because the identification strategy is based on differences-in-differences at the student level, for which reason individual fixed effects are assumed away. In some subjects the difference is largest for low achievers and in some subjects it is actually largest for the most proficient male students.

The basic results withstand several specification checks. Overall, they do not support the hypotheses that the gender difference in the non-blind score reflects statistical discrimination against male students. For example, limiting the sample to schools where boys outperform girls on average, overall or in specific subjects, leaves the results basically unchanged. The variance in performance of boys is higher on average and in every subject than that of girls, suggesting that statistical discrimination against boys may also occur due to "noisier" signals in boys' test scores. The data, however, do not support this interpretation because the gender potential bias is not different in schools where girls demonstrate more variability in performance on average. I also examined the possibility that the results mainly reflect the effect of the specific pattern in the timing of the exams where the state follows the school exam. Using data from a second chance state-level exam in English that was taken 4 months after the school-level exam led to almost identical estimates, suggesting that the short time interval between the state and school exams cannot explain our basic results.

An interesting and obvious question in this context is whether this estimated gender difference represents students' behavior or teachers' behavior. An example of students' behavior that may explain this potential bias is differential pattern of mean reversion by gender, e.g., due to a time-varying gender difference in knowledge or due to girls not performing as well in the state exams because they may represent a more 'pressured environment' for girls. The evidence suggests, if anything, stronger mean reversion for girls than for boys, namely girls tend to improve their scores more if they perform below average at the school exam (either relative to the class mean, the class mean by gender or the prior-self in earlier exams). Another possibility is that the state and school exams do not share the same content and/or do not measure the same skills. For example, some systematic gender differences in within-class behavior (discipline, attitude, absenteeism) may end up impacting grading at the school level and not at the state level. Unfortunately, the data do not include information that would help in directly addressing this source of concern but various pieces of indirect evidence do not support this hypothesis either. For example, the basic evidence is relatively robust to controlling for lagged school scores for other exams in the same or other subjects.

The paper also examines explanations based on empirical insights gained from experiments in social psychology. One such important insight is that a "stereotype threat"–the threat of being perceived as a negative stereotype or the fear of a poor performance that would confirm the stereotype–may be powerful enough to shape the intellectual performance and academic identities of entire groups of people. The implication in our context is that the difference in favor of girls of the gap between the non-blind and blind-test scores reflects the inferior performance of girls in non-blind tests because it involves a stereotype threat and superior performance in blind tests that conceal their gender. The evidence provided in this paper does not allow us to state that this mechanism explains the negative male difference. Instead, the data support the hypothesis that teachers' behavior is responsible for this difference. I show that the gender potential bias in various subjects is very sensitive to teachers' characteristics

[2] Distortions in teachers' evaluation of their students were also discussed recently in another context, that of test-based accountability systems in education. For example, Jacob and Levitt (2003) have shown that teachers cheat on standardized tests (using data from Chicago Public Schools and students from grades 3 to 8), and that the frequency of cheating appears to respond strongly as incentive for high test scores increase even mildly.

such as gender, age, years of experience, and even family size. There is no reason to expect this pattern of sensitivity of the difference to teachers' characteristics to reflect students' behavior. Based on the evidence presented in the paper, I conclude that the estimated gender differences can be interpreted as an anti-male bias.

The paper is organized as follows: The next section discusses the literature of stereotypes and discrimination, and in particular evidence about teachers' stereotypical behavior culled from experiments in social psychology. Section 3 explains the design of the study, the Israeli high-school matriculation exam system, and the comparability of school- and state-exam scores, and presents a graphical exposition of the blind–non-blind test score difference by gender. The basic results of the experiment are presented in Section 4. Sections 5 and 6 discuss alternative interpretations and explanations and Section 7 concludes.

2. Teachers' stereotypes and behavior

Psychologists who have studied stereotypes have assumed traditionally that beliefs about social groups are a powerful determinant of attitudes and behaviors toward members of these groups (Fiske, 1998, in a meta study of hundreds of articles). Stereotyping is thought to promote prejudice, which promotes discrimination (meta-analysis by Dovidio et al., 1996). In other words, beliefs about members of a social group are assumed to arouse like or dislike for the group, and these, in turn, dictate behavior toward group members. Recent experimental evidence shows that stereotypes and prejudice do appear to be positively associated. However, there is little evidence about the link between stereotypes and discrimination. Research on the relationship between stereotypes and racial discrimination, for example, has found only a modest relationship between whites' stereotypes of American blacks and measures of discriminatory actions, or degrees of discrimination (Dovidio et al., 1996). Thus, very little research demonstrates that stereotypes cause discrimination.

Several studies document how stereotypes affect teachers' behavior in the classroom. Teachers give boys more praise, criticism, encouragement, and permission for strategy use than they give girls. Teachers often view boys' rebellious invented strategies as signs of a promising future in math and unconsciously control girls more than boys. They apply this control by encouraging strategy use or accepting girls' lack of participation. These teacher characteristics are noted by Hyde and Jaffee, in their 1998 meta-analysis. Carr, Jessup, and Fuller (1999) in a study of 92 1st graders and their female teachers, argue that in teacher–student interactions girls and boys are influenced to develop different skills, knowledge, and motivation. For example, interaction of teachers with boys often increases boys' understanding and self-concepts in math and interaction of teachers with girls does not have this outcome. When examining the results of middle schoolers SATM tests, Rebhorn and Miles (1999) found that teachers often call on boys and praise them but give girls more time for easy tasks. Middleton and Spanias (1999), in a meta-analysis of studies of mathematics students, report that teachers reinforce learning helplessness among girls: when teachers encounter girls who do not want to succeed, they are liable to allow them to persist in their apathetic view of mathematics. According to the National Center for Education Statistics (1997), females are less likely than males to be advised, counseled, and encouraged to take mathematics courses. Fennema and Hart (1994), in a meta review of articles published in the JRME, found that teachers tend to structure their classrooms in ways that favor male learning. Hallinan and Sorensen (1987) (using longitudinal data for 4th to 6th graders in segregated and desegregated schools) found that intra-class ability groupings in school were influenced more by pupils' gender than by their performance in math and that boys tended to be assigned to higher ability groups. Girls with strong mathematical aptitude were less likely to be assigned to a high set than boys with similar aptitude. Intra-class grouping was found to have no effect on pupil performance compared with whole-class teaching.

It was also found that parents, especially mothers, often entertain stereotypic gender-role beliefs about their children's abilities that interact with and reinforce teachers' stereotypes. For example, parents often overestimate sons' abilities in math and science and underestimate the abilities of their daughters, in both childhood and adolescence (Eccles et al., 1990). Other related studies of 11 to 12 year olds found that parents' perceptions of the value of math and its difficulty for their children colored both the children's confidence in their ability and the likelihood that they would enroll in advance math classes (Jacobs and Eccles, 1992).

3. Study design

3.1. The Israeli high-school matriculation exam system

Israeli post-primary education consists of middle school (grades 7–9) and high school (grades 10–12). High-school students are enrolled either in an academic track leading to a matriculation certificate (*bagrut* in Hebrew)[3] or in a vocational track leading to a high-school diploma. Students complete the matriculation process by passing a series of state exams in core and elective subjects beginning in tenth grade and additional tests in eleventh grade and, in greater part, in twelfth grade. Students choose to be tested at various levels of proficiency, each test awarding one to five credit units per subject, depending on difficulty. Some subjects are mandatory and, for many, the most basic level of study is three credits. A minimum of twenty credits is required to qualify for a matriculation certificate. About 52% of high-school graduates in 2002 and 46% of members of the relevant age cohort received matriculation certificates.[4]

[3] The matriculation certificate is a prerequisite for university admission and is one of the most economically important education milestones. Many countries and some American states have similar high-school matriculation exams, e.g., the French Baccalaureate, the German Certificate of Maturity (*Reifezeugnis*), the Italian Diploma di Maturità, the New York State Regents examinations, and the recently instituted Massachusetts Comprehensive Assessment System.

[4] See the Israel Ministry of Education web site (www.education.gov.il) and Lavy (2002, 2004).

The final matriculation score in a given subject is the mean of two intermediate scores. The first of the two is based on state exams that are "external" to the school because they are written and scored by an independent agency. The scoring process for these exams is anonymous; the external examiner is not told the student's name and gender. The second intermediate score is based on a school-level ("internal") exam that is called a *matkonet*, derived from the word *matkon*, recipe, meaning that the school exam follows the "recipe" of the state exam. The school exam not only follows the exact format of the state exam, but it also draws its questions from the same bank of questions used for the state exam. The school exam is scheduled normally a week to three weeks before the state exams and each student is informed about his score in this exams. The school scores must be reported to the Ministry of Education before the date of the state exam. Both exams take place in the school in the regular classes so that there are no apparent systematic differences in the exam-taking environment that could interact with some female characteristics, such as possible higher anxiety levels. Both this fact, and the fact that the exams use the same bank of questions, and have the same structure, goes some way to convince us that the exams are testing the same skills and cognitive achievements. However, there is one important difference between the two exams: all state tests conceal the student's identity during grading; only his or her I.D. number appears on the exam notebook. School-level "matkonet" tests, in contrast, are not anonymous, and are graded in school by the student's own teacher, who of course knows the student's name and gender.

There are two exam "seasons"–winter (January) and summer (June)–and all students are tested in a given subject on the same date. The state exams are graded by expert teachers in a central state examination center; each exam by two independent external examiners and the final external score is the average of the two. This protocol eliminates the possibility of teachers grading their own students' exams and, thereby, reduces the possibility that teachers will inflate their students' scores dishonestly. To discount the possibility of sample selection propelling the results, it is important to know that half of the exams are not optional (Hebrew, Math, English and Bible Studies exams). Thus for these tests at least we can expect similar test taking rates for both boys and girls.

Students are admitted to post-secondary programs mainly on the basis of their matriculation scores.[5] Therefore, higher-education institutions and the National Council for Higher Education, which monitors the level of difficulty of all exams and their comparability from year to year, scrutinize these tests closely. To assure long-term consistency in the state exams, they are written under contract by an outside independent nonprofit organization that has been performing this task for many years.

To assure comparability between school-exam and state-exam scores, the Ministry of Education, since 1996, has been using a scheme of close supervision of the gaps between the school and the state scores and a set of rules about sanctions when large gaps are found. Schools and teachers are well aware and informed about this scheme (called "Differential Weighting") because the sanction in cases where the gaps recur in several subjects and in a few consecutive years involve revoking altogether the school privilege to administer school matriculation scores.[6] The purpose of this system is "to insure the validity of the final matriculation scores so that they may be viewed as a unified and credible criterion in the process of admission to higher-education institutions" (Ministry of Education, High School Division web site). The Ministry guidelines to schools indicate that the school score should be submitted to the Ministry before the state matriculation exams and should reflect the student's knowledge of the subject. The exam on which the school score is based is administered several weeks before the state exam.[7]

3.2. The data

The data used in this study pertain to the school years 2000–2002. The micro student files included the full academic records of each student on matriculation exams taken during high school (grades 10–12) and student characteristics (gender, parents' schooling, family size, and recent-immigrant status, where applicable) for the three cohorts of high-school seniors in 2000–2002. The information about each matriculation exam included its date, subject, applicable credits, and score. Members of the Ministry of Education staff and experts from an independent agency write each matriculation exam.

The school data provide information about the ethnic (Jewish or Arab) nature of each school and the religious orientation (secular or religious) of Jewish schools. In this study I used a sample that includes only Jewish secular schools, which account for about 60% of all schools and enrollment countrywide. I excluded Jewish State-Religious schools and Arab schools because many of them are either all-male or all-female and in many others classes are segregated by gender. This unique feature may be correlated and confounded with different patterns of gender stereotyping and discrimination in comparison with secular schools.

[5] Each higher-education institution undertakes to rank applicants according to the same formula, thus producing an index based on a weighted average of the student's average score on all his/her matriculation exams and the average score on the national exams in three core subjects (math, English, and Hebrew). Students may replace the second component in the formula with an ATS-type score from an examination administered by the National Testing Center.

[6] A Ministry of Education document describes the rules of the Differential Weighting scheme in detail. If the average school score in a given exam is higher than the school average score in the state test by 20 points or more or if it is lower by 10 points or more, the case is considered an outlier. If the probability of such an event is 1:10,000, the weights of the two scores are adjusted to 30% and 70%, respectively, instead of 50% each. If the probability of such an event is 1:1,000,000, the two scores are weighted at 10% and 90%, respectively. If outliers are defined in 8% or more of the exams, in at least two subjects and in two of three consecutive years, disciplinary actions are taken against the school. In particular, a ministerial committee may prohibit the school from submitting school scores and advertise its decision in the national print media. Since this decision implies significant consequences for the school, it has to be approved by the Ministry director general.

[7] The Ministry of Education's formal guidelines about the school score state specifically that it should measure only the student's level of knowledge in the subject of study, and the school-level exam is intended to measure just that. However, schools are allowed to deviate from the score on the school exam to reflect the student's performance on previous exams. As I show below, the distributions of the school and state scores are very similar and are even identical on many tests. Casual evidence also suggest that teachers tend to rely completely on the school exam in determining the school score to avoid parental pressure and to protect themselves against the "differential weighting" sanctions discussed above.

Table 1
Means and standard deviations of national and school-score exams during senior year

Subject type	Subject	Level of test	Number of observations	Mean school score	Mean national score	*T* test for the difference in means
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sciences	Chemistry	Basic	964	84.9	76.5	7.5
				(12.2)	(17.2)	
		Advanced	3616	85.7	78.4	14.1
				(11.3)	(14.5)	
	Computer science	Advanced	3546	83.5	73.8	14.2
				(13.7)	(17.7)	
	Math	Basic 1	16,647	71.1	67.5	6.2
				(20.5)	(29.2)	
		Basic 2	11,862	88.6	91.2	7.2
				(18.1)	(19.1)	
		Intermediate 1	7346	80.8	80.7	0.3
				(15.5)	(21.1)	
		Intermediate 2	4490	80.9	75.2	8.6
				(15.1)	(18.7)	
		Advanced 1	3983	87.2	81.3	12.1
				(13.1)	(20.2)	
		Advanced 2	2197	85.9	84.4	2.3
				(12.8)	(15.7)	
	Physics	Advanced 1	4854	84.0	78.2	11.6
				(13.2)	(20.0)	
		Advanced 2	1916	86.2	82.8	4.1
				(13.2)	(16.4)	
		Advanced 3	4927	84.7	79.4	10.3
				(13.5)	(19.9)	
		Advanced 4	3023	89.6	88.1	3.6
				(9.8)	(10.2)	
Humanities	Bible studies	Basic	26,450	75.0	63.6	27.8
				(15.4)	(21.0)	
	Biology	Advanced 1	5274	83.4	83.1	0.6
				(11.2)	(14.9)	
		Advanced 2	5277	82.9	77.0	18.9
				(11.4)	(13.5)	
		Advanced 3	5276	82.7	74.8	15.4
				(11.6)	(15.4)	
		Advanced 4	5101	85.0	86.1	3.2
				(10.8)	(9.9)	
		Advanced 5	5079	85.0	79.8	11.8
				(10.8)	(12.8)	
	English	Basic 1	8130	77.2	80.1	8.0
				(14.1)	(15.5)	
		Basic 2	8986	73.1	64.9	18.1
				(14.2)	(20.4)	
		Intermediate	11,466	74.1	65.4	27.3
				(10.9)	(14.7)	
		Advanced	13,843	82.4	75.1	27.5
				(10.7)	(13.8)	
	History	Basic	19,647	77.0	69.1	20.1
				(15.1)	(17.2)	
	Literature	Basic	32,908	76.1	69.9	25.6
				(14.3)	(15.8)	
		Advanced	3097	82.1	75.7	20.0
				(11.3)	(10.7)	

Notes: Dependent variables are standardized scores. The *T* statistic in Column (7) reflects estimated standard errors that are corrected for clustering by school.

3.3. Comparability of school- and state-exam scores

Table 1 presents the mean and the standard deviations of the school and state test scores for a representative sample of tests in nine subjects: six math exams (two at the basic level, two intermediate, and two advanced), four English exams (two at the basic level and one apiece at the intermediate and the advanced level), five in biology (all in the advanced program), two in Hebrew literature, two in chemistry, three in physics, and one each in computer science, history, and bible studies.[8] The mean gap between

[8] These exams are grouped with the science subjects (math, chemistry, physics, and computer science) first, followed by the humanities subjects. This grouping is retained in the following tables. Girls traditionally do better in humanities subjects, and this grouping should aid to determine if the gap between blind and non-blind scores is consistent across these two types of subjects or not.

the school and state test scores is in most cases positive and significantly different from zero (T tests for the difference are presented in Column 7 of Table 1). The standard deviation of the school score is also generally 10–15% smaller than in the state exam. In several subjects, however, the internal score is lower than the external score; there are also cases in which the extent of score variation is similar in both cases.

Figs. 1–8 present kernal-density estimates of the school and state exams of some of the tests presented in Table 1. The distributions of the state and the school test scores (the first graph in each row) are very different across subjects and tests. Some do not resemble a normal distribution shape. A striking feature in all the figures, however, is the overall similarity between the distribution of the internal and the external scores for each given test in comparison with the large differences between and within subjects. In other words, the distributions of the internal and external scores in each exam change in a similar way from one test to another. In most cases, the school-score distribution is a constant shift, to the right, of the state-exam test score distribution. The direction of the shift is expected after one observes the pattern in differences in means, as shown in Table 1. The figures also reveal that the internal score distribution in many exams is "thinner" at the left tail of the distribution. In many tests, however, the two distributions are almost completely identical, a feature that we exploit in the analysis presented in the sections to come.

3.4. The blind–non-blind test score difference by gender: graphical exposition

Figs. 1–8 also present kernal-density estimates of the school and state test score distributions by gender (second and third graphs in each row). Both score distributions for each test are similar for male and female students, respectively. However, the leftward shift of the school-score distribution relative to the external-score distribution is always larger for female students than for male students. Even in cases where the shift factor is negative, i.e., where the school-score distribution is to the left of the external-score distribution (as occurred in three biology tests, several math exams, and several cases in which the two distributions intersected), the difference is less negative for female students than for male students. The next section estimates these differences and their statistical confidence intervals more precisely and subjects the main results to several specification tests.

4. Estimating gender-stereotyped differences

I took advantage of the blind and non-blind nature of the scoring procedures for the state and school exams, across subjects and tests, to identify the effect of the procedure of anonymous evaluation of cognitive ability on the likelihood that male or female abilities would be systematically under- or over-valued. The score of student i on test j is a function of gender (M) and the anonymous or non-anonymous nature of the evaluation (NB). The repetitive structure of two scores in each subject (and across various tests in each subject), one blind (subscript $b=0$) and the other non-blind (subscript $b=1$), made it possible to use a differences-in-differences estimation strategy. Assuming a linear specification, the score equation may be written as

$$S_{ijb} = \alpha + \lambda M_i + \delta \mathrm{NB}_{ijb} + \gamma\left(M_i \times \mathrm{NB}_{ijb}\right) + u_{ijb}. \tag{1}$$

The coefficients for M and NB identify the effects of being male and a non-blind scoring procedure, respectively, on the test score. The parameter of interest is that pertaining to the interaction of M and NB, γ, which measures the difference between the non-blind scores of male students and those of female students, given the respective difference in the blind score. The differences-in-differences nature of the estimation of Eq. (1) implies that individual and school fixed effects are implicitly assumed away in this model with regard to the estimated coefficient of interest, γ, as long as they have the same effect on the blind and non-blind scores. The coefficient of interest, γ, could also be estimated from the following difference equation and its estimate would be algebraically identical to that estimated from Eq. (1).

$$S_{ijb} - S_{ijnb} = \alpha + \gamma M_i + u_{ij} \tag{1'}$$

where S_{ijb} is the blind score, and S_{ijnb} is the non-blind score. However, using the specification of Eq. (1) has the advantage of revealing the estimates of the other parameters. It should also be noted that the differences-in-differences nature of Eq. (1) or (1′) implies that any student fixed effect is accounted for in both equations with regard to the estimate of γ. For Eq. (1) it also implies that any within student correlation between the random term in the blind and the non-blind scores is netted out and indeed the standard errors estimated using Eqs. (1) and (1′) are identical.

Table 2 presents the estimated parameters for Eq. (1) in nine subjects–chemistry, computer science, math, physics, bible studies, biology, English, Hebrew literature, and history,–obtained from data for 2001. In each subject, the data set is a stacked file including the internal and external scores for each of the exams in the respective subject. All test scores were standardized to a distribution with zero mean and a unit standard deviation. This procedure was applied within subjects to each test separately. The sample-size varied by subjects, including most students in compulsory subjects but far fewer in elective subjects. Table 2 reports all the parameter estimates of Eq. (1). The standard errors reported in Table 2 are adjusted for clustering at the school level, using formulas set forth by Liang and Zeger (1986).

4.1. Empirical results

Overall, female high-school students in Israel had higher achievements on the state matriculation exams (blind tests) in all subjects presented in Table 2 except for English. Girls had advantages of 0.24 of a standard deviation of the external-score

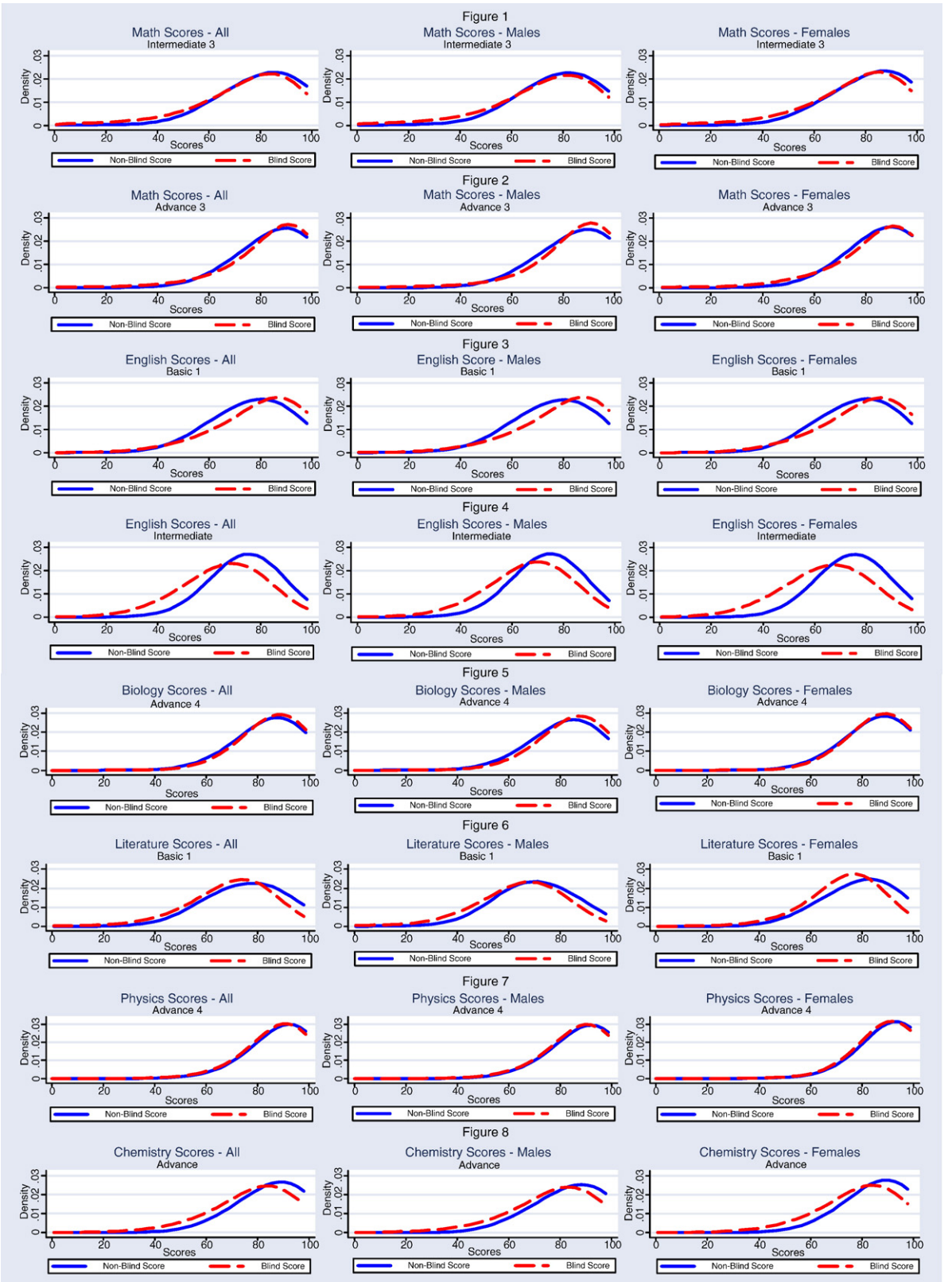
Figure 1
Math Scores - All
Intermediate 3
Math Scores - Males
Intermediate 3
Math Scores - Females
Intermediate 3
Figure 2
Math Scores - All
Advance 3
Math Scores - Males
Advance 3
Math Scores - Females
Advance 3
Figure 3
English Scores - All
Basic 1
English Score - Males
Basic 1
English Scores - Females
Basic 1
Figure 4
English Scores - All
Intermediate
English Scores - Males
Intermediate
English Scores - Females
Intermediate
Figure 5
Biology Scores - All
Advance 4
Biology Scores - Males
Advance 4
Biology Scores - Females
Advance 4
Figure 6
Literature Scores - All
Basic 1
Literature Scores - Males
Basic 1
Literature Scores - Females
Basic 1
Figure 7
Physics Scores - All
Advance 4
Physics Scores - Males
Advance 4
Physics Scores - Females
Advance 4
Figure 8
Chemistry Scores - All
Advance
Chemistry Scores - Males
Advance
Chemistry Scores - Females
Advance
Density
Scores
Non-Blind Score
Blind Score

Figs. 1–8.

Table 2
Estimated gender bias in the non-blind scores by subject

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.096	0.097	0.045	−0.012	0.107	0.269	0.041	0.051	0.214
	(0.044)	(0.052)	(0.023)	(0.035)	(0.029)	(0.030)	(0.020)	(0.030)	(0.022)
Male	−0.122	−0.037	−0.105	−0.015	−0.243	−0.139	0.114	−0.117	−0.475
	(0.040)	(0.042)	(0.015)	(0.035)	(0.019)	(0.022)	(0.014)	(0.021)	(0.018)
Non-blind score	−0.025	0.004	0.020	0.041	0.037	−0.059	0.048	0.025	0.024
	(0.036)	(0.047)	(0.015)	(0.033)	(0.017)	(0.022)	(0.018)	(0.025)	(0.015)
Male x (non-blind score)	−0.058	−0.122	−0.086	−0.130	−0.083	−0.125	−0.180	−0.075	−0.053
	(0.033)	(0.042)	(0.012)	(0.024)	(0.013)	(0.023)	(0.013)	(0.016)	(0.014)
Number of observations	9562	8006	109,928	29,992	58,676	52,888	84,850	41,758	75,568
Number of schools	196	237	363	242	325	190	359	248	328

Notes: Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The number of observations is twice the number of exam takers, since the datasets are stacked (for each student there are two observations per exam code, one for the school score and one for the external score).

distribution in bible studies, 0.14 in biology, 0.12 in chemistry, 0.12 in history, 0.48 in literature and 0.11 in math. In physics (0.02), and computer science (0.04), the advantage of female students was positive but not statistically different from zero. The advantage of male students in English was 0.11 and statistically different from zero. The male indicator coefficients may reflect the selective distribution of students among elective subjects and among levels of study (basic, intermediate, and advanced) in compulsory subjects; therefore, these estimates may be biased. However, as I show below, the advantage of female students recurred at all levels of study, suggesting that the selection bias may not have been very important. For example, girls had on average higher math scores in the blind test at all three levels of study: basic, intermediate, and advanced.

The mean differences between the (non-blind) school scores and the (blind) state scores, conditional on gender and on the interaction between gender and non-blind testing, were very small and seldom significantly different from zero. The largest differences were in biology (− 0.059) and in English (0.048) and in both of these cases they were significantly different from zero. These results are important because they imply that the large and significant differences between the blind and the non-blind mean scores seen in Table 1 disappear once the controls included in Eq. (1) are taken into account.

The main parameter of interest is the estimated coefficient of the interaction between the gender indicator for male students and the non-blind test indicator. These estimates were negative and significantly different from zero in all nine subjects. The highest estimate was in English (− 0.180), the lowest was in literature (− 0.053), and in four of the nine subjects the estimate exceeded one-tenth of a standard deviation in the respective anonymous test score distribution.[9] The range of the estimates for humanities subjects versus science subjects was very similar, if the large English estimate is excluded, from − 0.053 to − 0.125 (humanities) and − 0.058 to − 0.130 (science). Including the English estimate reveals a larger gender gap for the humanities (where girls traditionally do better), which implies some support for the interpretation that stereotype bias drives these results.

The signs of the estimated coefficients for the interactions of the male and non-blind test indicators–the focal points of this study–differ from common perceptions and beliefs. Insofar as these coefficients reflect a disparity in the perception of cognitive ability by students' gender, due to stereotypes or other sources of discrimination, then the evidence suggests that such stereotypes act against male students and not female students. These results may suggest that teachers favor female students by consciously or unconsciously "inflating" their scores on non-blind tests. The direction of the difference enhances the female students' advantage on the blind test. In English, the difference in favor of female students more than offset these students' "real" disadvantage, as reflected on the blind test (0.180 versus 0.114).

To account for individual characteristics (X), the following equation was also estimated:

$$S_{ijb} = \alpha + \gamma M_i + \delta NB_{ijb} + \gamma\left(M_i \times NB_{ijb}\right) + X_i\theta + u_{ijb}. \quad (2)$$

The student characteristics included in X are mother's and father's years of schooling, number of siblings, immigration status, and a set of dummy variables for ethnic origin–Asia/Africa, America/Europe, former Soviet Union, Ethiopia, and Israel-born–and two average measures of achievements on external matriculation tests taken in tenth and eleventh grades (lagged outcomes). The two measures are the mean credit-weighted average score on all previous matriculation exams (coding zeros for those who had taken no exams) and the respective overall number of credits. These measures are powerful predictors of students' success on each matriculation exam in twelfth grade in each subject (Angrist and Lavy, 2004). As noted above, adding student, school, or specific-exam fixed effects should lead to exactly the same estimates of the coefficients of the interaction between gender and the non-blind test indicator because the basic specification of Eq. (1) saturates all these fixed effects, since the difference-in-differences

[9] I also estimated Eq. (1) while using the percentile-rank score as a dependent variable instead of the standardized score. The evidence from this specification is very similar to the results presented in Table 2.

Table 3
Estimated gender bias in the non-blind scores by subject with included controls for student characteristics and lagged outcomes

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	−0.061	−0.059	−0.072	0.092	−0.233	−0.057	0.117	−0.081	−0.483
	(0.032)	(0.040)	(0.014)	(0.028)	(0.015)	(0.020)	(0.014)	(0.018)	(0.015)
Non-blind score	−0.025	0.004	0.020	0.041	0.037	−0.059	0.048	0.025	0.024
	(0.036)	(0.047)	(0.015)	(0.033)	(0.017)	(0.022)	(0.018)	(0.025)	(0.015)
Male x (non-blind score)	−0.058	−0.122	−0.086	−0.130	−0.083	−0.125	−0.180	−0.075	−0.053
	(0.033)	(0.042)	(0.012)	(0.024)	(0.013)	(0.023)	(0.013)	(0.016)	(0.014)
Number of observations	9562	8006	109,928	29,992	58,676	52,888	84,850	41,758	75,568
Number of schools	196	237	363	242	325	190	359	248	328

Notes: Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls students' lagged outcomes and background characteristics: father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin (Asia/Africa, America/Europe, Israel, Soviet Union Republics, Ethiopia and a category for students with missing data on ethnic origin). The number of observations is twice the number of exam takers, since the datasets are stacked (see note in Table 2).

occurs at the student level within each subject and exam. These results are presented in Table 3. Only minor changes were observed in the male estimated coefficients, while the estimates on the non-blind score and the male and non-blind score interactions were left unchanged.[10]

In the following two sections, we check the robustness of the basic results and attempt to assess empirically the validity of possible explanations. However, we should note a concern regarding the validity of the scores in the state tests as a blind score: the scoring of state exams may be less "blind" if examiners can guess student' gender by his or her handwriting. Girls' handwriting is clearly if not perfectly distinguishable from boys'. Furthermore, since Hebrew verbs are gender-marked, the verbs used on the examination forms reveal the examinee's gender. This, of course, may be the case in the context of this study. Tests in some subjects, however–such as English, math, biology, physics, and computer science–are composed mainly or exclusively of multiple-choice questions or writing of numbers without any text.

As I showed above, the potential bias against male students in these subjects was not very different from that of other subjects in which girls' tests could more easily be identified through their handwriting. Examples of such tests are also mentioned below in the analysis of additional empirical work reported in the following sections.

5. Checking the robustness of the estimates

5.1. Do the blind and non-blind scores measure different skills?

A material issue of relevance in interpreting the estimated potential anti-male bias is the possibility that the blind and non-blind tests measure different abilities or that one of the two reflects attributes not included in the other. These differences in content or in what is being evaluated by these two exams are a real concern if they are not gender-neutral. An obvious example would be that the estimated differences reflect non-conformist behavior by male students that teachers do not like, e.g., absenteeism or discipline problems in class. If teachers adjust scores to reflect such undesired behavior, even though the scores should reflect only the students' knowledge of the subject, a discrepancy between male students' blind scores and non-blind scores would occur. Although this interpretation may be consistent with the above evidence, other results presented below suggest that the likelihood of its being the source of the potential bias is very small.

To check for the possibility that the non-blind score reflects male students' behavioral problems, the best way would be to control for such behavior. Since data on students' behavior are not available, a reasonable alternative is to include in Eq. (2) a variable that accounts for students' behavior. If students' behavior in twelfth grade correlates with behavior in earlier grades, then the non-blind state score on a matriculation exam taken earlier in high school may reflect this consideration as well. For most students, the math program also entails a matriculation exam at the end of eleventh grade, with the same evaluation format including a blind and a non-blind score. Assuming that the behavior of students in eleventh and twelfth grades is highly correlated, the inclusion in Eq. (2) of the eleventh-grade math non-blind score as an explanatory variable will control for variations in students' behavior. Table 4 reports the results when this control is added in two alternative specifications: In Column 2, the non-blind eleventh-grade math score is added as a main effect as well as an interaction with the twelfth-grade non-blind test indicator. In Column 3, two other interaction terms are added; the eleventh-grade non-blind score interacted with the male indicator and also with the male and the twelfth-grade non-blind test indicators. Since the sample used in Columns 2–3 is smaller than the full sample used in Table 2, Column 1 in Table 4 presents the results of an estimation of the basic model that does not include these

[10] The results obtained from the 2000 and 2001 data are very similar to the findings presented in Tables 2 and 3; therefore, they are not presented in this paper.

Table 4
Estimated 12th grade gender bias in the non-blind scores, while controlling for the 11th grade math non-blind score

	(1)	(2)	(3)
Male	−0.093	−0.048	−0.048
	(0.022)	(0.020)	(0.020)
Non-blind score	0.346	0.274	0.264
	(0.557)	(0.544)	(0.542)
Male x (non-blind score)	−0.118	−0.104	−0.103
	(0.021)	(0.021)	(0.021)
Non-blind score in 11th grade	–	0.362	0.367
		(0.016)	(0.021)
Non-blind score x (non-blind score in 11th grade)	–	0.094	0.080
		(0.017)	(0.020)
Male x (non-blind score in 11th grade)	–	–	−0.010
			(0.023)
Male x (non-blind score) x (non-blind score in 11th grade)	–	–	0.029
			(0.022)
Number of observations	23,170	23,170	23,170
Number of schools	270	270	270

Notes: Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls all the level variables used in all the interactions (father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin), the number of matriculation credits achieved in 11th grade, average score in 11th grade.

additional controls. As Table 4 shows, the sign and the implied size of the estimated differences in Columns 2–3 are just slightly lower than the estimated effect in Column 1 of Table 4.[11, 12]

Another insight that can help determine whether the results above reflect a difference in the abilities measured by the two scores can be gained by restricting the empirical analysis to exams for which the distribution of blind and non-blind scores seems absolutely identical. Several of the distributions in Figs. 1–8 meet this criterion: Intermediate Math 3, Advanced Math 2 and 3, Basic Literature 1, Advanced Physics 2 and 4, and Intermediate and Advanced English. Table 5 presents the estimates of the effect of the interaction between the gender and blind-test indicators in each exam in this subset. The results from this sample of exams are very similar to results based on the pooling of all exams into one sample in each subject: negative and significant estimates of interaction between the male-gender and the non-blind score indicators. This suggests that even in cases where the two score distributions are unquestionably identical, in which it may safely be assumed that they measure the same outcome, the potential bias is in the same direction and of a similar magnitude.

5.2. Potential male bias in non-blind scores with control for student ability

The differences-in-differences nature of the estimates in Tables 2–5 means that account is taken of any ability measure that has a similar effect on the anonymous and non-anonymous scores. It may well be, however, that ability affects each of the scores differently and, for this reason, directly affects the difference of the two scores. This model implies the following modification of Eq. (1′):

$$S_{ijb} - S_{ijnb} = \alpha + \gamma M_i + \delta A_i + u_{ij} \tag{3}$$

where S_{ijb} is the blind score, S_{ijnb} is the non-blind score, and A_i is student ability. The average score in all previous matriculation exams taken during tenth and eleventh grades can measure student ability. However, the credit-weighted average score on all previous blind exams may be thought of as a more precise reflection of student ability. For the purpose of comparison, however, it is also of interest to compute the credit-weighted average score of all previous non-blind scores. This score may reflect other unobserved attributes or behavior of the student that is reflected in all subjects of study. An example, again, is a bad attitude and discipline problems in school that teachers may "punish," perhaps unconsciously, by lowering the student's non-blind score.

Table 6 reports results from the estimation of Eq. (3) with each of these two controls added, sequentially, to the regressions of each of the nine subjects. Since not all students have lagged matriculation scores, the samples used in Table 6 are slightly smaller than those used in Table 3. Therefore, Column 1 in Table 6 reports, for the purpose of comparison, the results of the estimation of the basic regression (a specification that does not include the aforementioned ability measures) using the new samples. The table reports only the estimated coefficient of the interaction between the non-blind score and the male indicators.

[11] Additional evidence about student behavior may be revealed by the propensity to disqualification or to suspected cheating on the matriculation exams. No systematic differences by gender in this propensity were revealed: in some subjects, a higher proportion of male students were caught cheating in exams, in others, girls had the upper hand in this respect.

[12] The 11th grade matriculation test scores provide the twelfth-grade math teachers ample opportunity to become highly familiar with each student in their class. It is interesting to note that the potential male bias persists even when this information is available to the twelfth-grade teacher. The persistence of the gap in the face of signals of student's true ability resembles the confirmatory bias discussed and modeled in Rabin and Schrag (1999). A person suffers from confirmatory bias if he tends to misinterpret ambiguous evidence as confirming his current hypothesis about the world. Rabin and Schrag frequently used as an example the situation where teachers misread performance of pupils as supporting their initial impressions of those pupils or as supporting their stereotypes about groups to which these pupils belong.

Table 5
Estimated gender bias in the non-blind scores, using only exams for which the blind and non-blind score distributions are identical

	Math	Math	Biology	Biology	English	English	English	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	−0.099	0.181	−0.089	−0.029	0.074	0.179	0.160	−0.489
	(0.038)	(0.047)	(0.026)	(0.027)	(0.025)	(0.023)	(0.019)	(0.015)
Non-blind score	−0.016	−0.029	−0.105	−0.101	−0.004	0.043	0.084	0.027
	(0.044)	(0.050)	(0.031)	(0.035)	(0.024)	(0.026)	(0.024)	(0.016)
Male x (non-blind score)	−0.062	−0.147	−0.162	−0.216	−0.078	−0.224	−0.272	−0.057
	(0.034)	(0.048)	(0.037)	(0.037)	(0.021)	(0.022)	(0.018)	(0.015)
Number of observations	8980	4394	10,202	10,158	16,260	22,932	27,686	65,816
Number of schools	167	127	178	177	338	334	306	317

Notes: Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls students' lagged outcomes and background characteristics: father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin (Asia/Africa, America/Europe, Israel, Soviet Union Republics, Ethiopia and a category for students with missing data on ethnic origin). The number of observations is twice the number of exam takers, since the datasets are stacked (see note in Table 2).

The estimates in Column 1 of Table 6 are almost identical to those presented in Tables 2 and 3, indicating that the sample of students with lagged scores is a close representative of the full sample used in Tables 2 and 3. Column 2 reports the estimates when the mean of the lagged blind scores is added to the regression to control for ability. Column 3 presents the results when the average lagged non-blind scores are used to control for ability. The estimates in the table depict an interesting pattern: adding the mean blind score leads to negligible changes in the estimated differences but adding the mean of the non-blind lagged scores does lead in some subjects to a significant reduction of the potential male bias estimates. In biology, for example, when the ability measure is based on the blind score, the estimated coefficient of the interaction between the non-blind score and the male indicators is − 0.101, as against − 0.059 when ability is measured by the mean of the non-blind scores. Both estimates are lower than the estimate obtained (− 0.122) when neither of the two ability measures is included. Similar dramatic changes are found in chemistry and literature, but in the six other subjects the potential male bias estimates are basically unchanged when the proxy for ability is added as a control to the basic regression.

Columns 4–5 in Table 6 report again the results of adding ability measures as a control. This time, however, they were computed distinctly for the nine subjects, excluding in each case the lagged scores from exams in that specific subject. The estimates in Columns 4–5 closely resemble those in Columns 2–3. The overall results in Table 6 reinforce the interpretation that the negative sign on the interaction between the non-blind and the male indicators does not reflect the omission of unobserved student attributes, in particular cognitive ability or non-cognitive characteristics such as behavioral problems reflected in the non-blind scores.

5.3. Variation of the potential male bias in the non-blind score with student ability

An interesting question is whether the negative estimated difference in scores occurs across the entire distribution of students' ability or focuses mainly on some segment of the ability distribution. To address this question, I again measured students' ability by

Table 6
Estimated gender bias in the non-blind scores while controlling for student's ability

	The basic model	Controlling for average lagged score in all subjects		Controlling for average lagged score in all subjects but of that of the relevant subject	
		Using blind score	Using non-blind score	Using blind score	Using non-blind score
	(1)	(2)	(3)	(4)	(5)
Chemistry	−0.055	−0.051	−0.017	−0.053	−0.019
n=4780	(0.033)	(0.033)	(0.031)	(0.033)	(0.031)
Computer science	−0.126	−0.127	−0.110	−0.127	−0.098
n=3949	(0.042)	(0.043)	(0.043)	(0.043)	(0.044)
Math	−0.085	−0.086	−0.073	−0.089	−0.067
n=54,286	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Physics	−0.131	−0.130	−0.101	−0.132	−0.078
n=14,965	(0.023)	(0.023)	(0.024)	(0.024)	(0.024)
Bible studies	−0.087	−0.091	−0.084	−0.099	−0.084
n=29,195	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)
Biology	−0.122	−0.101	−0.059	−0.099	−0.043
n=26,362	(0.025)	(0.025)	(0.025)	(0.026)	(0.026)
English	−0.183	−0.182	−0.168	−0.179	−0.148
n=41,450	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
History	−0.071	−0.079	−0.063	−0.084	−0.066
n=20,821	(0.016)	(0.016)	(0.017)	(0.016)	(0.016)
Literature	−0.054	−0.054	−0.038	−0.055	−0.033
n=37,605	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)

Notes: Standard errors are corrected for school-level clustering and are presented in parentheses.

Table 7
Estimated gender bias in the non-blind scores by quartiles of the ability distribution

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1st quartile	−0.120	−0.211	−0.081	−0.145	−0.064	−0.138	−0.111	−0.153	0.073
	(0.060)	(0.117)	(0.026)	(0.074)	(0.025)	(0.042)	(0.023)	(0.036)	(0.026)
2nd quartile	−0.029	−0.142	−0.107	−0.190	−0.070	−0.109	−0.196	−0.045	−0.059
	(0.062)	(0.079)	(0.018)	(0.050)	(0.023)	(0.033)	(0.021)	(0.028)	(0.021)
3rd quartile	−0.097	−0.047	−0.094	−0.122	−0.094	−0.075	−0.211	−0.059	−0.112
	(0.048)	(0.055)	(0.019)	(0.037)	(0.020)	(0.035)	(0.018)	(0.024)	(0.018)
4th quartile	0.005	−0.080	−0.068	−0.044	−0.104	−0.077	−0.182	−0.055	−0.098
	(0.042)	(0.043)	(0.015)	(0.033)	(0.017)	(0.033)	(0.019)	(0.019)	(0.016)
Number of observations	9562	8006	109,928	29,992	58,676	52,888	84,850	41,758	75,568

Notes: Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls all the level variables used in all the interactions (father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin), the number of matriculation credits achieved in 11th grade, average score in 11th grade. The quartiles of ability are derived from the distribution of the average blind scores in exams taken in 10th and 11th grades.

the credit-weighted average (blind) score in all previous matriculation exams and divided the distribution of this variable into quartiles. I then estimated the basic model for each of these quartiles separately. The estimates of the interaction of the non-blind score and the male indicators are shown in Table 7 for all nine subjects and all four quartiles. The pattern that emerges from the table is that the negative male effect on the non-blind score is reflected at all levels of student ability. In some subjects (biology, chemistry, computer science, history), the effect was strongest at the lower quartile; in others (bible studies, English, literature), it was strongest at the third or fourth quartiles. In all subjects except chemistry, however, the sign of the potential anti-male bias in all quartiles was negative and significantly different from zero. On the basis of this evidence, we may safely conclude that the potential negative male bias in the non-blind score is evident for students at all levels of ability as measured by their previous performance on matriculation exams.

5.4. Effects of other interactions with the non-blind score

The estimates presented above may reflect the effect of interactions between the non-blind test variable and other variables that correlate with gender. To assess this possibility, Eq. (2) was augmented by adding the interaction terms of the non-blind score indicator with the following dichotomous indicators of students' socio-demographic characteristics: recent immigrant, father with more years of schooling than the school mean, mother with more years of schooling than the school mean, more siblings than the school mean, and ethnic origin. Table 8 presents the results after all these additional interactions were included in Eq. (2). First, it should be noted that the inclusion of the nine additional interactions along with the non-blind score indicator did not change the estimated coefficient for the interaction between gender and non-blind test. The coefficient in the math equation, for example,

Table 8
Estimated gender bias in the non-blind scores versus other biases

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Non-blind score									
x Male	−0.069	−0.123	−0.087	−0.127	−0.081	−0.130	−0.176	−0.071	−0.053
	(0.032)	(0.042)	(0.012)	(0.024)	(0.013)	(0.023)	(0.013)	(0.016)	(0.014)
x Recent immigrant	0.440	−0.080	0.146	0.163	−0.089	0.202	0.377	0.051	−0.032
	(0.165)	(0.031)	(0.047)	(0.099)	(0.072)	(0.052)	(0.044)	(0.096)	(0.059)
x Father's schooling	0.004	−0.018	−0.024	0.015	−0.040	0.026	−0.040	−0.046	−0.012
	(0.031)	(0.034)	(0.014)	(0.026)	(0.016)	(0.022)	(0.015)	(0.018)	(0.014)
x Mother's schooling	0.026	0.011	−0.003	−0.054	−0.028	0.002	−0.027	0.010	−0.016
	(0.033)	(0.040)	(0.014)	(0.025)	(0.015)	(0.022)	(0.016)	(0.019)	(0.014)
x Number of siblings	0.061	0.011	−0.017	−0.050	0.036	0.013	0.012	−0.029	−0.001
	(0.048)	(0.040)	(0.013)	(0.032)	(0.014)	(0.020)	(0.014)	(0.021)	(0.012)
x Ethnic origin									
America–Europe	0.437	0.482	−0.189	0.980	0.209	−0.235	0.007	−0.465	0.494
	(0.345)	(0.064)	(0.281)	(0.105)	(0.186)	(0.046)	(0.327)	(0.207)	(0.244)
Asia–Africa	0.380	0.482	−0.193	0.980	0.272	−0.281	0.028	−0.334	−0.463
	(0.353)	(0.070)	(0.285)	(0.106)	(0.185)	(0.042)	(0.330)	(0.209)	(0.243)

Notes: Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls all the level variables used in all the interactions (father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin), the number of matriculation credits achieved in 11th grade, average score in 11th grade.

changed from − 0.086 to − 0.087; in English the change was from − 0.180 to − 0.176. Second, one of the newly added interactions, that between the non-blind test and immigration status, was positive and significant. The other interactions elicited no systematic pattern in terms of sign or significance level.

5.5. Does the potential negative male bias reflect statistical discrimination against boys?

The estimated gender difference may reflect statistical discrimination that may be motivated by the average superior performance of girls in the state exams. The blind and the non-blind scores are two sources of information about students' cognitive ability. If teachers are influenced by the expected higher performance of girls on the state exams (as shown in these data), male and female students will receive different scores on their school exams even if they perform at the same level. In this case, then, the estimated difference may reflect simple statistical discrimination (Cain, 1986).

Statistical discrimination against male students may also occur even if there are no real differences in cognitive performance between boys and girls. This will happen if teachers believe that the non-blind score is a less reliable signal of knowledge for boys than for girls (Aigner and Cain, 1977). Such beliefs may arise in a school context if, for example, male students are known to cheat, or are perceived of as cheating, more often than girls on school exams. The question of whether such perceptions are based on real evidence or are unfounded is irrelevant to the outcome of statistical discrimination.

Some of the evidence presented above does not support the interpretation of statistical discrimination. In English, in particular, boys outperformed girls on the state exams by a wide margin but possibly encountered bias in their school scores. Similar contrasting comparisons were found in some tests in other subjects as well. In advanced math, for example, boys had a higher blind-score average than girls but potentially sustained a bias in the school score. It is possible, however, that teachers form their expectations about gender differentials in true cognitive ability on the basis of overall performance in all subjects and not only in the subject that they teach. To assess this possibility, I estimated the models separately in two distinct environments: a sample of schools where boys outperform girls on average and a second sample of schools where girls do better on average than boys. First, I computed average performance on the basis of matriculation exams taken by the 2001 cohort while in tenth and eleventh grade. Table 9a presents the results of Estimation Eq. (2) for the two samples of schools based on this measure of average performance by gender. I also used school-average performance by gender on the basis of all matriculation exams taken by members of the 2000 graduating class. These results are presented in Table 9b.

Focusing first on Table 9a, we see clearly that in the sample of schools where girls outperform boys on average (the upper panel of Table 9a), the coefficients of the interaction between the male and the non-blind test indicator are negative in all subjects, namely the estimated gender difference is potentially biased against boys. More interesting results, however, are seen in the lower panel of Table 9a. In this sample, boys had a higher average external score on 11th grade matriculation exams in most subjects. The most noteworthy result, however, is that in eight of the nine subjects the potential bias is against male students, being negative and significantly different from zero. In five of the nine subjects, the estimates are even higher in this sample than in the sample of schools when girls dominate boys in average performance.

The outcomes presented in Table 9b generally confirm those in Table 9a. Overall, they do not support the hypotheses that the potential bias in the non-blind score reflects statistical discrimination against male students.

Table 9a
Estimated gender bias in the non-blind scores by gender average performance in 10th–11th grade matriculation exams

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Schools where girls are better than boys									
Mean score in 10–11th grades									
Girls	85.47	86.58	77.50	88.19	77.87	82.24	72.91	79.18	77.97
Boys	81.96	82.16	74.34	84.49	75.50	78.02	69.40	76.72	76.04
Male x (non-blind score)	−0.084	−0.131	−0.084	−0.141	−0.091	−0.128	−0.179	−0.065	−0.052
	(0.044)	(0.047)	(0.013)	(0.024)	(0.016)	(0.028)	(0.016)	(0.018)	(0.018)
Number of observations	6122	5450	77,952	22,570	37,436	39,864	58,236	28,844	46,276
Number of schools	107	121	214	150	163	126	212	141	175
Schools where boys are better than girls									
Mean score in 10–11th grades									
Girls	82.73	78.30	73.67	82.41	73.60	79.14	69.34	75.25	73.97
Boys	85.38	83.53	76.05	85.23	75.87	82.08	72.51	77.17	76.28
Male x (non-blind score)	−0.014	−0.128	−0.117	−0.172	−0.079	−0.123	−0.197	−0.090	−0.060
	(0.049)	(0.060)	(0.022)	(0.054)	(0.021)	(0.042)	(0.022)	(0.034)	(0.024)
Number of observations	3286	1958	30,826	6452	20,644	12,340	25,422	12,776	29,018
Number of schools	70	54	119	47	125	53	118	78	134

Notes: The performance is measured using 11th grade scores of the same students that are included in the sample. Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls students' lagged outcomes and background characteristics: father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin (Asia/Africa, America/Europe, Israel, Soviet Union Republics, Ethiopia and a category for students with missing data on ethnic origin).

Table 9b
Estimated gender bias in the non-blind scores by gender average performance of the 2000 cohort

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Schools where girls are better than boys									
Male	−0.047	0.018	−0.065	0.085	−0.248	−0.080	0.124	−0.102	−0.503
	(0.034)	(0.043)	(0.015)	(0.029)	(0.016)	(0.020)	(0.015)	(0.023)	(0.019)
Male x (non-blind score)	−0.092	−0.206	−0.097	−0.137	−0.078	−0.112	−0.191	−0.062	−0.049
	(0.037)	(0.047)	(0.012)	(0.028)	(0.014)	(0.024)	(0.014)	(0.022)	(0.016)
Number of observations	6860	4788	88,676	22,152	43,988	45,130	67,836	23,928	53,518
Number of schools	117	107	244	143	194	149	248	114	199
Schools where boys are better than girls									
Male	−0.087	0.013	−0.067	0.086	−0.192	0.058	0.120	−0.083	−0.423
	(0.077)	(0.086)	(0.027)	(0.064)	(0.038)	(0.059)	(0.034)	(0.043)	(0.026)
Male x (non-blind score)	0.028	−0.066	−0.079	−0.109	−0.131	−0.186	−0.148	−0.087	−0.085
	(0.080)	(0.084)	(0.029)	(0.051)	(0.027)	(0.075)	(0.032)	(0.035)	(0.029)
Number of observations	2162	1788	17,948	6074	12,226	6004	13,374	8224	19,754
Number of schools	49	41	78	48	81	27	72	51	96

Notes: The performance is measured using 12th grade scores of students who graduated from the same school as those included in the sample in 2000. Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls students' lagged outcomes and background characteristics: father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin (Asia/Africa, America/Europe, Israel, Soviet Union Republics, Ethiopia and a category for students with missing data on ethnic origin).

Table 9c presents estimates of the estimated gender difference when schools were divided into two samples in each subject according to the score on twelve graders' state exams in each subject in 2000. The results resemble those in Tables 9a and b: again the potential bias estimates were negative and significant in all subjects, even though the estimate of the male indicator was positive in each of the nine subjects.

5.6. Does the potential negative male bias reflect effect of different timing of exams?

The state-level exam is always after the school-level exam. Therefore the non-blind test is always first and the blind test is always second. This different timing could in theory account for some of the gap in performance between boys and girls in the school and the state-level exams. For example, if boys tend to study for the exam later than girls, perhaps because they tend to procrastinate or because they have a higher discount rate than girls, than they might be better prepared for the state exam than for the school-level exam and they may do better in the second than in the first. There are no cases where the school-level exams are taken before the state exams. However, the time difference between these two exams varies across subjects, typically being between one to three weeks. The fact that the potential bias is against boys in all subjects regardless of the time lag between the

Table 9c
Estimated gender bias in the non-blind scores by gender average performance in 12th grade blind tests in 2000

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Schools where girls are better than boys									
Male	−0.027	0.003	−0.083	0.047	−0.231	−0.067	0.131	−0.093	−0.481
	(0.037)	(0.053)	(0.021)	(0.031)	(0.016)	(0.021)	(0.032)	(0.028)	(0.016)
Male x (non-blind score)	−0.082	−0.169	−0.080	−0.130	−0.093	−0.122	−0.148	−0.060	−0.059
	(0.044)	(0.055)	(0.017)	(0.032)	(0.014)	(0.026)	(0.028)	(0.023)	(0.015)
Number of observations	5766	4024	47,732	15,040	46,928	38,496	15,156	19,540	71,848
Number of schools	92	83	162	99	203	121	79	93	272
Schools where boys are better than girls									
Male	−0.113	0.067	−0.056	0.147	−0.255	−0.064	0.124	−0.104	−0.507
	(0.062)	(0.062)	(0.018)	(0.047)	(0.040)	(0.047)	(0.015)	(0.030)	(0.094)
Male x (non-blind score)	−0.032	−0.167	−0.105	−0.135	−0.074	−0.116	−0.196	−0.080	−0.081
	(0.054)	(0.066)	(0.015)	(0.038)	(0.031)	(0.049)	(0.015)	(0.030)	(0.093)
Number of observations	3256	2566	58,892	13,186	9286	12,638	66,054	12,612	1424
Number of schools	74	66	160	92	72	55	241	72	23

Notes: The performance is measured using 12th grade blind-test scores of students who graduated the same school as those included in the sample in 2000. Dependent variables are standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The regressions include as controls students' lagged outcomes and background characteristics: father and mother schooling, number of siblings and 6 dummies as indicators of ethnic origin (Asia/Africa, America/Europe, Israel, Soviet Union Republics, Ethiopia and a category for students with missing data on ethnic origin).

two exams may be an indication that timing of the exams is not central to our results. However, there is one exceptional case where the time elapsed between the two exams is 3–4 months: students are allowed to take a second chance state-level exam in English and math, typically scheduled towards the end of the summer school break at the end of August. The relevant school-level score for these students is still the score in the school-level exams that was taken sometime between mid May to mid June. Factors related to different timing should most likely not be relevant when comparing the end August state-level score to the school-level score of 3–4 months ago.

Almost a fourth of students took the second chance exam in English at the end of August 2001, therefore the sample includes 20,236 observations versus 84,850 in the English sample used in Table 2.[13] I used this sample to estimate Eq. (1) while using the end of August score as the blind score instead of the respective June score. The estimated coefficient of the potential bias in this sample is − 0.145 (s.e. 0.023) while that reported in Table 2 is − 0.180 (s.e. 0.013). The two estimates have the same sign and are of similar size even though the time gaps between the two exams are very different. This similarity is another indication that the specific pattern in the timing of the state and school-level exams is not the cause of the pattern in our results.

5.7. When the external score is also non-blind: the case of thesis writing

Students enrolled in the advanced program in any subject (five credit units) are allowed to write a thesis as a substitute for the matriculation exams in the subject at hand. The grading protocol of the thesis includes a school score given by the subject teacher and a score given by an external examiner who grades the thesis. As part of this protocol, the thesis writer must meet the external examiner and present and discuss the thesis with him or her. In this case, then, the external state score is also non-blind. Although grading a paper may be different from grading an exam, insights may be gained by estimating the difference using the sample of thesis writers. I created such a sample by isolating subjects in which a reasonable number of students wrote theses. Since the sample in each subject was relatively small, I pooled all the thesis writers' together and introduced controls for each subject in the regression.

Table 10 presents the results of Estimation Eq. (1) based on the sample of thesis writers. The dependent variable was measured in terms of raw scores in the first column and in terms of standardized scores in the second column. Similar to the results based on the nine subjects, girls' theses received much higher external evaluations than boys'—about 0.16 of a standard deviation in the external evaluation distribution. However, unlike the results in the other seven subjects, the estimated coefficients of the male and (own teacher) non-blind score interaction were basically zero when raw scores were used and also when standardized scores were used instead. Since these outcomes were based on a much smaller sample than the samples used in the other nine subjects, zero difference between males and females in the grading of theses may have been a reflection of sampling variance. Therefore, I used the sample of thesis writers to estimate the difference in the nine subjects. Because math and English are compulsory, only for them could I find a reasonably large sample of thesis writers. Columns 3–4 in Table 10 present the basic model of Eq. (2) using the pooled sample of thesis writers who took math and English exams. The coefficient estimates were negative when both raw scores and standardized scores were used. The estimates were less precise than those in Table 3 but the size of the estimate in Column 4 of Table 10 (− 0.107) fell into the range between math (− 0.086) and English (− 0.180) shown in Table 3. This additional evidence, based on a quasi-natural experiment among thesis writers, strongly suggests that the estimated score difference originates in teachers' behavior. We address this issue next.

6. How can we explain the potential negative male bias? Teachers' versus students' behavior

Two broad sets of explanations may account for the estimated differences: one relating to students' behavior and another to teachers' behavior. If the latter is the source of the difference, one may conclude that the difference represents a form of discrimination. In this section, I present evidence suggesting that the test score difference indeed originates in teachers' behavior and that students' behavior is not a contributing factor.

6.1. Does the potential male bias in the school score reflect students' behavior?

The basic outcomes presented above about the negative male “bias” in school-level non-blind scores may be the results of behavioral differences between girls and boys. For example, students may not be at the same relative intrinsic ability level when performing in the state and school exams. There are several reasons why one could expect a time-varying gender difference in knowledge. First, female students may on average prepare early. On the other hand, boys may delay more of their preparation for the exam. Such differential patterns in preparation could reflect well-established differences in risk-aversion between males and females. Alternatively, even for initial identical level of preparation, male students may be better at improving their preparation between the first (school) and the second (state) exams. For example, males may be better at inferring from the school exams the type of questions that they should expect from the state exams or males may respond more vigorously than girls to a negative signal and make a relatively greater improvement as a consequence. The latter “wake up call” argument of this type implies that boys who receive an unexpectedly low school score may invest more effort and preparation and, for this reason, improve their performance on the state exam to a greater degree than girls would. Such a pattern may also emerge if boys consider the state exam more serious or challenging and, therefore, prepare for it more vigorously. All these various behavioral hypotheses suggest gender

[13] The sample based on the August second chance math exam of 7746 is much smaller than the respective sample that was used in Table 2 which included 109,928 observations. Nevertheless, the estimate from this sample is also negative though it is much less precise than the estimate reported in column 3 of Table 2.

Table 10
Estimated gender bias in the non-blind scores based on thesis writers: when the external score is also non-blind

	Grade in thesis		Grade in math and English exams	
	Raw	Standardized	Raw	Standardized
Intercept	88.238	0.051	77.218	0.043
	(1.083)	(0.069)	(1.368)	(0.090)
Male	−1.713	−0.156	0.834	0.032
	(0.829)	(0.084)	(1.172)	(0.062)
Non-blind score	3.684	0.008	4.234	0.082
	(0.860)	(0.103)	(0.682)	(0.047)
Male x non-blind score	0.803	−0.024	1.492	−0.107
	(1.017)	(0.123)	(0.976)	(0.060)
Number of students	1118		2442	
Number of schools	76		75	

differences in mean reversion patterns reflected in the "different difference" among boys and among girls between the two scores. However, the data does not allow distinguishing between these hypotheses but their joint effect can be tested by estimating the following non-linear model of mean reversion in the relationship between the school score and the state score. We define the following variables based on the non-blind score:

$$R_{1i} = \text{Maximum}\left[\left(S_i - S\right), 0\right]$$
$$R_{2i} = \text{Minimum}\left[\left(S_i - S\right), 0\right]$$

where S_i is the student non-blind score and $\hat{S}$ is the mean score of a reference group on the same exam.

Three reference groups may be considered: class peers, own-gender class peers, and each student's own scores on previous exams as well. For the first two groups, I used average respective scores on the same exam. In the third case, I used the own mean score in all other exams taken in tenth and eleventh grade. $R_{1i}=0$ for students who scored below the reference-group mean; $R_{2i}=0$ for students who scored above the reference-group mean. The estimation equation is:

$$A_i = \alpha + \beta R_{1i} + \gamma R_{2i} + \varepsilon_i \tag{4}$$

where A_i is the blind score on the state test of student i in a given subject. Eq. (4) allows the state score to respond to the school score non-symmetrically. Eq. (4) was estimated for each subject, separately for boys and girls, and the results are shown in Table 11. The table presents point estimates and not standard-error estimates because all coefficients were highly statistically significant.

Table 11
Estimated mean-reversion pattern differences by gender, while using three different reference groups

	Science subjects				Humanities subjects				
	Chemistry	Computer science	Math	Physics	Bible studies	Biology	English	History	Literature
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reference: class mean									
Girls below mean	0.780	0.716	0.850	0.726	0.799	0.589	0.729	0.871	0.804
	[0.406]	[0.395]	[0.396]	[0.390]	[0.423]	[0.404]	[0.450]	[0.422]	[0.356]
Boys below mean	0.739	0.765	0.799	0.604	0.790	0.579	0.655	0.832	0.866
	[0.497]	[0.455]	[0.464]	[0.442]	[0.556]	[0.552]	[0.484]	[0.508]	[0.597]
Girls above mean	0.564	0.450	0.366	0.447	0.577	0.410	0.641	0.448	0.490
Boys above mean	0.678	0.311	0.314	0.556	0.451	0.412	0.576	0.335	0.384
Reference: class mean by gender									
Girls below mean	0.785	0.731	0.840	0.744	0.813	0.583	0.736	0.863	0.774
Boys below mean	0.787	0.788	0.806	0.613	0.803	0.598	0.662	0.845	0.916
Girls above mean	0.549	0.456	0.342	0.449	0.512	0.406	0.633	0.434	0.468
Boys above mean	0.636	0.304	0.364	0.541	0.534	0.439	0.585	0.378	0.480
Reference: own mean									
Girls below mean	0.671	0.602	0.660	0.565	0.636	0.442	0.644	0.623	0.648
	[0.498]	[0.474]	[0.498]	[0.563]	[0.440]	[0.439]	[0.495]	[0.492]	[0.386]
Boys below mean	0.605	0.713	0.613	0.511	0.584	0.404	0.544	0.557	0.605
	[0.509]	[0.522]	[0.530]	[0.518]	[0.575]	[0.450]	[0.503]	[0.543]	[0.625]
Girls above mean	−0.346	−0.130	−0.143	−0.334	−0.289	−0.109	0.034	−0.398	−0.255
Boys above mean	−0.265	−0.162	−0.123	−0.257	−0.340	−0.026*	0.039	−0.328	−0.343

Notes: Standard errors are corrected for school-level clustering but are not presented in the table, since all the estimates are significantly different from zero (except those which are marked by *). The proportion of students in each group is presented in squared brackets. The class mean is based on all students in own class, who tested in the same exam. The own mean is based on all blind scores of exams taken in 10th and 11th grades.

The results are shown in three panels—the first from regressions based on class mean as the reference group, the second based on gender-specific within-class means, and the third based on the mean of the student's own lagged scores.

The first interesting comparison is the relative responses of girls and boys to a negative signal as reflected by a score lower than the mean of the reference group. The first two rows in each panel should be used for this comparison. When the class mean is used as the reference, girls' response to a negative signal exceeds that of boys in five subjects and in two subjects the estimated response coefficients are equal for boys and girls. The elasticities based on these estimates and the means of R_{1i} and R_{2i} for boys and girls also follow this pattern. The proportions of male and female students who are below their class mean are presented in squared brackets. On average, a higher proportion of the male students are below their class mean in comparison to the same statistic for women although the difference is not very large.

The comparison elicits a similar pattern when the own-gender mean in class is used as the referenced group (the first two rows of middle panel in Table 11). However, the response of girls to a score below their own previous mean (based on the mean of blind scores on previous matriculation exams) is much larger than that of boys in all subjects except computer science. This suggests that the hypothesis that boys have a higher mean reversion or a more vigorous response to a negative shock than girls is not supported by the data in any of the three definitions of reference group. Also note that the proportion of students who are below their own mean is very similar for girls and boys and in physics it is even higher among girls.

6.2. More on student behavior: female anxiety and stereotype threat

Another relevant form of differential behavior between female and male students is related to the level of anxiety under pressure. Recent studies suggest that women may be less effective than men in competitive environments, even if they are able to perform similarly in non-competitive environments (e.g., Gneezy et al. 2003). One may argue that state exams generate more stress and anxiety than school exams. If teachers expect these factors to have a larger effect on girls than on boys, they may compensate female students for their expected underachievement on the external test by giving them higher school scores. First, it should be noted that the scores on the school exam and the state exam are of identical importance in determining the final matriculation score, since both are equally weighted.[14] Therefore, they should cause similar levels of stress and anxiety among students, if they cause any stress at all. Second, in the foregoing results, girls outperformed boys on state exams in almost every subject and at almost every level of the curriculum. Third, the evidence that girls under-perform in pressure situations or competitive environments refers only, or mainly, to situations where stereotyping is a real threat or when the task is against the opposite gender.[15] Experiments have shown that girls' performance is not affected even in high-stake tests as long as there is no threat of gender stereotyping (e.g., Stangor and Sechrist, 1998). Gneezy et al. (2003) show that the increase in women's performance in a competitive environment is limited only when the rivals are men. The state-exam environment, however, is free of any obvious stereotype threat or overt competition against the other gender because the scoring is blind. Therefore, there is no reason to expect these factors, even if they are real, to have any gender-differentiated effect.

However, a dynamic stereotype threat may be at work in the context of the matriculation exams even if the current exam environment contains no active contemporaneous threat. For example, if girls react to previously experienced stereotypical situations, we would expect to find a strong correlation between blind versus non-blind score differences in exams in previous grades and the respective differences in exams in twelfth grade. Such a correlation may also be present if girls do suffer more than boys from high-stake exam anxieties. Some insight about these possible relationships may be gained by estimating Eq. (2) using two samples based on potential bias experienced by the students in the past—the sample of students who exhibited a positive difference between blind and non-blind scores on the eleventh-grade math matriculation exam, and a second sample composed of those whose difference between the scores was negative. Table 12 reports the results of the estimation of Eq. (2) on the basis of these two samples. Column 1 presents, for comparison, the results based on the sample of students who had math matriculation scores in both eleventh and twelfth grades. Column 2 presents results for students who had a positive difference between the two scores in eleventh grade, and Column 3 shows results for students who had a negative difference. The two respective estimated coefficients, − 0.103 and − 0.120, both precisely estimated, are not very far apart. Therefore, they do not support, in this limited

[14] Interestingly, in France, school exams are also an integral part of the matriculation system but the final scores are based solely on the national exams. The school exams are called by a name that has the same meaning as the Hebrew name of the corresponding exams in Israel.

[15] The "stereotype threat" theory (Steele, 1997) focuses on the consequences for individuals who contend with negative stereotypes related to their intellectual abilities. Stereotype threat is the psychological burden imposed by stereotype-based suspicions of inferiority in achievement. Stereotype threat has been shown to undermine academic achievement in two ways: by interfering with performance in mental tasks and, over time, by encouraging students to protect their self-esteem by disengaging from the threatened domain. Wanting not to perform badly, another possible result of stereotype threat, has been shown in experiments to impair performance in difficult cognitive tasks, either by simply distracting the performer or by eliciting a self-protective withholding of effort (Spencer et al., 1999). The social-psychology literature on stereotypes and discrimination suggests that students may react to a stereotyped situation in a variety of ways but that two ways are the most common. The first is avoidance of the stereotype situation. Female math majors, for example, may "dress down"—wear less markedly feminine clothing—in math classes than in humanities classes (Seymour and Hewitt, 1997). This suggests that in situations where women feel at risk of confirming a negative gender stereotype, they take steps to avoid projecting stereotypically feminine traits, thereby reducing the risk of being viewed through the lens of stereotype and being treated accordingly (Aronson et al., 1998) The second is confirmation of the stereotype, e.g., to perform below actual ability in a test that involves a stereotype threat. Such underachievement may result, for example, from anxiety and evaluation apprehension (Aronson et al., 1998). In the context of this paper, students' own perceptions of a stereotyped situation may be based on experience. For example, students who on a previous exam had a school score that was lower than the external blind score may blame some of the discrepancy on stereotype discrimination. According to the theories presented above, such students may respond in such a way as to refute the stereotype and make an effort to do better on the next exam. Alternately, they may succumb to the stereotype and do worse on the second external exam than on the first, thus ostensibly confirming the stereotype.

Table 12
Estimated 12th grade math gender bias in the non-blind score: dividing the sample according to the 11th grade difference in the blind and non-blind score

	All students	Students whom their 11th grade	
		Blind score> non-blind score	Blind score< non-blind score
Male	−0.093	−0.128	−0.058
	(0.022)	(0.030)	(0.025)
Non-blind score	0.346	−0.476	1.131
	(0.557)	(0.050)	(0.047)
Male x non-blind score	−0.118	−0.103	−0.120
	(0.021)	(0.025)	(0.031)
Number of students	23,170	13,008	10,162
Number of schools	270	256	222

Note: Standard errors are corrected for school-level clustering and are presented in parenthesis. The regressions include as controls all the level variables used in all the interactions (father and mother schooling, number of siblings, 6 dummies as indicators of ethnic origin), the number of matriculation credit units achieved in 11th grade, average score in 11th grade matriculation exams.

sample and within the specific context, the "stereotype threat" or the "anxiety" hypotheses as an explanation of the potential anti-male bias.

6.3. Does the potential anti-male bias reflect a steeper learning curve among male students?

In Section 5.2, we considered the effect of students' ability on the difference between blind and the non-blind scores and found that it had very little effect. However, another concept of ability, perhaps more behavioral in nature that may be relevant for differential outcomes is the ability of students to improve their performance between their junior and senior years in high school. Male students may have a higher positive slope in their learning curve than female students. If the school non-blind scores on senior-year matriculation exams are somewhat influenced by student performance in the junior year, a difference for male students between the senior-year blind and non-blind scores will emerge. To test this hypothesis, I computed two measures of average improvement between the junior and senior years, one based on blind scores and the other based on non-blind scores. I then re-estimated the basic model while adding these controls of average improvement one at a time. Table 13 presents the results. Column 1 shows the estimates without these controls for the sample for which these controls could be computed. The estimates in Column 1 are very similar to those reported in Tables 2 and 3. The estimates in Columns 2 and 3 are not very different from each other and from those in Column 1. Therefore, they do not support the interpretation that the potential anti-male bias occurs because male students have a steeper learning curve that is discounted somewhat by their teachers.

6.4. Is the gender difference the result of teachers' behavior?

If it is teachers' behavior that accounts mostly for the difference between male and female students, then the extent of this discrimination may vary among teachers or in accordance with teachers' characteristics, e.g., gender, age, years on the job, and,

Table 13
Estimated gender bias in the non-blind scores, while controlling for learning curve differences by gender

	The basic model	Controlling for average improvement	
		Using blind score	Using non-blind score
Chemistry	−0.055	−0.069	−0.054
n = 4780	(0.033)	(0.032)	(0.033)
Computer Science	−0.126	−0.137	−0.107
n = 3949	(0.042)	(0.042)	(0.042)
Math	−0.085	−0.087	−0.072
n = 54,286	(0.012)	(0.012)	(0.012)
Physics	−0.131	−0.134	−0.117
n = 14,965	(0.023)	(0.023)	(0.023)
Bible studies	−0.087	−0.100	−0.076
n = 29,195	(0.013)	(0.013)	(0.013)
Biology	−0.122	−0.123	−0.122
n = 26,362	(0.025)	(0.025)	(0.025)
English	−0.183	−0.183	−0.180
n = 41,450	(0.013)	(0.013)	(0.013)
History	−0.071	−0.077	−0.067
n = 20,821	(0.016)	(0.016)	(0.016)
Literature	−0.054	−0.060	−0.045
n = 37,605	(0.014)	(0.015)	(0.014)

Notes: Standard errors are corrected for school-level clustering and are presented in parentheses.

perhaps, marital status and number of children. Evidence of such variability may also be viewed as refuting the possibility that the estimated difference results from students' behavior because there is no reason to expect students' behavior to vary in a pattern consistent with the relationship between teachers' characteristics and the extent of the potential gender bias. Below I report results that allow the difference estimates to vary commensurate with the aforementioned characteristics of their teachers. Data that link students to their classroom teachers are available only for a sub-sample of schools and students and only in some subjects. Therefore, I re-estimate and report below the results of an estimation of the basic model (without interactions with teachers' characteristics) with the sub-samples for five subjects: English, math, physics, biology, and chemistry. In view of the size of the sample, the samples in biology and chemistry were pooled.

Table 14 presents the estimates of gender difference coefficients by teachers' characteristics. Four characteristics were considered: gender, age, years of teaching experience, and number of children. The effect of each characteristic on the estimated coefficient was examined sequentially and independently. For each characteristic the sample was divided into two: for gender by male and female and for the other characteristics by classroom teacher's being above or below the mean in the sample. The mean age was forty-six years, mean experience was twenty years, and mean number of children was one.

Column 1 in Table 14 presents the estimates from the sample of schools for which information on teachers' characteristics was available. Note that these estimates are not very different from those in Table 3, which were based on the full sample. For example, the two respective math estimates are identical (− 0.086), in English they are − 0.169 and − 0.180, in biology and chemistry − 0. 097 and − 0.090, and in physics − 0.054 and − 0.130.

Table 14 shows large variation by teachers' characteristics and a pattern that varies from subject to subject. Starting with teacher's gender, in math only one-third of teachers in the sample were male but they accounted for all of the differences in math; no differences observed among the sample of female math teachers. In physics, two-thirds of teachers were male and the differences were twice as large among female teachers as among male teachers. In biology and chemistry, most teachers were women and the difference, again, was due mostly to the behavior of female teachers. Large and significant variations in difference estimates were also observed when the samples were stratified by teacher age. In English and physics, the difference was evident mainly among young teachers while in math it originated mainly from older teachers. The variation in gender difference estimates by teachers' years of experience resembled the results by teacher's age. Finally, the potential negative male bias did not vary commensurate with the number of children a teacher had except in physics, where the difference was non-zero only among teachers with children. These patterns suggest that there is something about teachers that elicits a bias against male students; no sensible explanation links students' behavior to this pattern of variation in the bias against male students on the basis of teachers' characteristics.

It might be suspected that teachers are not assigned randomly to their classrooms and that certain types of teachers are assigned to certain types of students in a systematic way. This is not the case however. Teachers are assigned to subjects and not to certain classes. They are also pre-assigned to students in the tenth grade (receiving a fresh batch of 10th graders each year), and when students reach the 11th and 12th grades (when the tests are taken) teachers are not re-assigned.

6.5. Does the potential male bias result from higher achievement variance among boys?

Boys have, on average, a higher variance and a lower or equal mean score than girls in all subjects. This is reflected in the distribution of their matriculation test scores, in both the blind and the non-blind versions. The gender disparity in variance is highest in literature and chemistry and lowest in computer science and physics. Only in English and history do girls have higher variance than boys, and even here the difference is marginal. Table 15 presents the evidence. This pattern of higher performance

Table 14
Estimates of the gender bias in the non-blind scores, by teacher characteristics

Subject		Gender		Age (years)		Experience (years)		Children	
	The whole sample	Male	Female	<46	>46	<20	>20	0	>1
Biology and chemistry	−0.097	−0.008	−0.104	−0.080	−0.111	−0.165	−0.061	−0.109	−0.089
	(0.033)	(0.099)	(0.035)	(0.054)	(0.041)	(0.045)	(0.043)	(0.045)	(0.045)
Obs. schools	20,254	2130	18,124	8364	11,890	7268	12,986	7800	12,454
	120	20	108	57	80	40	93	56	76
Math	−0.086	−0.186	−0.037	−0.038	−0.135	−0.008	−0.107	−0.070	−0.104
	(0.030)	(0.048)	(0.034)	(0.045)	(0.041)	(0.067)	(0.034)	(0.036)	(0.046)
Obs. schools	9174	3006	6168	4698	4476	2730	6444	3856	5318
	40	28	38	37	36	28	38	32	36
Physics	−0.054	−0.061	−0.136	−0.145	0.025	−0.143	−0.040	−0.008	−0.104
	(0.049)	(0.065)	(0.074)	(0.057)	(0.074)	(0.064)	(0.064)	(0.074)	(0.060)
Obs. schools	7338	5216	2122	3162	4176	2054	5284	3764	3574
	84	56	28	32	52	20	64	44	40
English	−0.169	−0.136	−0.167	−0.223	−0.116	−0.190	−0.158	−0.170	−0.169
	(0.038)	(0.166)	(0.035)	(0.034)	(0.060)	(0.087)	(0.042)	(0.049)	(0.049)
Obs. schools	6834	558	6276	3478	3356	2136	4698	2540	4294
	38	11	38	31	33	25	36	28	34

Table 15
Standard deviation of the blind and non-blind score distribution, by gender blind score non-blind score

	Blind score		Non-blind score	
	Girls	Boys	Girls	Boys
Chemistry	14.6	16.0	10.9	12.4
	(78.8)	(76.8)	(86.4)	(84.2)
Computer science	18.90	18.9	13.8	14.9
	(72.7)	(73.0)	(85.0)	(83.0)
Math	23.9	25.0	17.8	19.2
	(79.5)	(77.3)	(82.1)	(79.1)
Physics	18.0	18.4	11.8	13.2
	(81.0)	(81.2)	(86.9)	(85.2)
Bible studies	20.2	21.5	14.7	15.7
	(66.8)	(61.4)	(77.7)	(72.5)
Biology	13.8	14.6	10.7	11.8
	(80.8)	(79.7)	(84.8)	(81.6)
English	17.3	16.9	12.9	12.8
	(70.3)	(72.2)	(77.5)	(76.9)
History	17.7	17.5	14.9	15.8
	(69.7)	(67.5)	(78.4)	(75.5)
Literature	13.4	16.7.	12.7	14.4
	(73.9)	(66.0)	(80.2)	(72.2)

Notes: The average score is presented in parentheses.

variance among boys may shed some light on the origin of the anti-male bias estimated above, through its interaction with the Ministry of Education rules–described in Section 3 of this paper–that "sanction" schools and teachers for large differences between blind and non-blind scores. One possible interactive mechanism is that teachers, who are probably aware of the gender-related variance disparities, may intentionally downward-bias the non-blind scores of boys in order to reduce the likelihood of overstepping the permissible range of difference between the scores and invoking the Ministry of Education sanctions. This mechanism may be more "operative" in an environment of large variance disparities.

Table 16 presents evidence when the sample was divided according to whether the performance variance of boys was higher or lower than that of girls. The data restrictions used to measure performance variance led to a smaller sample than that used in

Table 16
Estimated Gender bias in the non-blind scores using stratified samples according to the gender differences in the variance in test scores

	The basic model	School-level variance in all the subjects		School-level variance in the relevant subject	
		Var(boys)>Var(girls)	Var(boys)<Var(girls)	Var(boys)>Var(girls)	Var(boys)<Var(girls)
Chemistry	−0.093	−0.078	−0.190	−0.045	−0.172
	(0.040)	(0.043)	(0.083)	(0.043)	(0.068)
	2947	*2569*	*378*	*1813*	*1134*
Computer science	−0.183	−0.177	−0.261	−0.211	−0.161
	(0.053)	(0.056)	(0.137)	(0.072)	(0.058)
	1558	*1436*	*122*	*1021*	*537*
Math	−0.096	−0.087	−0.121	−0.089	−0.104
	(0.011)	(0.013)	(0.023)	(0.015)	(0.017)
	53,461	*39,905*	*13,556*	*32,492*	*20,969*
Physics	−0.158	−0.136	−0.257	−0.135	−0.186
	(0.024)	(0.027)	(0.049)	(0.036)	(0.031)
	11,561	*9507*	*2054*	*6769*	*4792*
Bible studies	−0.093	−0.088	−0.103	−0.069	−0.128
	(0.013)	(0.016)	(0.020)	(0.016)	(0.019)
	28,061	*20,489*	*7572*	*16,874*	*11,187*
Biology	−0.113	−0.128	−0.052	−0.118	−0.111
	(0.025)	(0.029)	(0.045)	(0.033)	(0.037)
	24,980	*20,426*	*4554*	*14,582*	*10,398*
English	−0.185	−0.194	−0.163	−0.146	−0.207
	(0.013)	(0.016)	(0.022)	(0.020)	(0.017)
	41,357	*29,741*	*11,616*	*14,766*	*26,591*
History	−0.076	−0.073	−0.089	−0.025	−0.122
	(0.016)	(0.018)	(0.033)	(0.023)	(0.019)
	19,923	*15,946*	*3977*	*9355*	*10,568*
Literature	−0.067	−0.077	−0.042	−0.059	−0.113
	(0.014)	(0.017)	(0.026)	(0.016)	(0.028)
	36,556	*27,006*	*9550*	*30,870*	*5686*

Notes: The number of students is in italics. School-level variance is based on all blind scores in a specific school. The sample is limited to those schools which have more than 10 students (of the same gender) with a blind score in the relevant subject.

Tables 2 and 3; therefore, Column 1 in the table presents the estimated coefficients of the gender difference as obtained from the full sample before stratification by performance variance. The estimates in Column 1 are not very different from those in Tables 2 and 3, confirming that there were no sampling differences in the estimation.

The first stratification was based on a variance computed from the distribution of scores in all subjects; the results are reported in Columns 2–3 of the table. This stratification divided schools on the basis of the overall variance of boys versus girls. Comparison of the estimates in Columns 2 and 3 does not elicit a clear pattern. In some subjects (e.g., biology and literature), the estimates were indeed much higher in the sample of schools that had higher variance among boys but in the other subjects the estimates were lower than those obtained from the sample in which girls had the higher variance.

I also stratified the sample uniquely for each subject according to the performance variance of boys and girls in that subject. The outcomes based on these stratifications are presented in Columns 5–6. When I compared the estimates in Column 4 with those in Column 5, again I found no systematic differences in the gender difference estimates obtained from samples of schools in which the performance variance of boys was lower or higher than that of girls. Actually, in the sample in subjects where male variance was much higher than female variance (chemistry and English), the bias against boys was smallest. Thus, the evidence does not seem to support higher variance among boys as an explanation for the downward bias in the non-blind scores of male students relative to their blind scores, a form of statistical discrimination.

7. Conclusions

Recent convincing evidence suggests that women face discrimination in the labor market in terms of both employment opportunities and wages (e.g., Goldin and Rouse, 2000). However, the question of whether discrimination against women is also partly responsible for the observed gender differences in human-capital investment and occupational choice has not been directly addressed.[16] This paper confronted one important alleged form of discrimination against women that may cause some of the skewed gender pattern in productivity-enhancing investments: discrimination against girls in school by their teachers. This discrimination, it is argued, emanates from a stereotyping of cognitive ability that causes female students to under-perform in, and to shy away from, math and science subjects in secondary and post-secondary schooling—a state of affairs that also affects occupational choice, of course. Women continue to avoid college majors and occupations that entail moderate amounts of coursework in mathematics, even though they match or surpass men's performance in high-school math and science courses (AAUW, 1999).[17] Gender-related differences in career choices also persist, especially in the fields of physical science, mathematics, and engineering, where women hold only about 10% of jobs.[18] In 1995, women constituted about 46% of the U.S. labor force but only 22% of scientists and engineers in the labor force.[19] In the U.K., only 6% of engineers in the labor force are women.[20] In Israel in 2005 only 20% of civil engineers and architects are women despite making up 52% of the labor force.[21]

The evidence presented in this study confirms that the previous belief that schoolteachers have a grading bias against female students may indeed be incorrect. On the contrary: on the basis of a natural experiment that compared two evaluations of student performance–a blind score and a non-blind score–the difference estimated strongly suggests a bias against boys. The direction of the bias was replicated in all nine subjects of study, in humanities and science subjects alike, at various level of curriculum of study, among underperforming and best-performing students, in schools where girls outperform boys on average, and in schools where boys outperform girls on average. Lindahl (2006) adopts a similar methodology and finds supporting evidence from national tests in Sweden that boys seem to be discriminated against in final tests, compared to girls. The bias against boys persisted when other interactions with the non-blind score were allowed, e.g., interactions with parental schooling and other student characteristics. When students' ability was added as a control or as an interaction with the bias against male students, the average difference did not disappear and was revealed at all levels of cognitive ability. The anti-male bias among teachers widened the difference between male and female students because the latter, on average, outperformed the former in almost all subjects. The size of the bias varied from − 0.053 to − 0.180 in the first estimates (Table 2), to − 0.062 to − 0.272 when controlling for the identical distribution of exams, with the approximate average value hovering at around − 0.10 in most of the different specifications. It is difficult to quantify exactly the effect of this gap on matriculation rates, but given that boys already have lower average marks than girls, this gap will lower their marks further, and could decrease their chances of acceptance into various programs after leaving school.

The anti-male bias may reflect differential behavior between male and female students, e.g., gender differences in the intrinsic relative ability level when performing in the school and state exams or gender differences in the ability to improve upon a bad signal. It may also reflect statistical discrimination against male students that emanates from the superior performance of female

[16] The empirical literature on gender discrimination has focused mainly on disparities in labor-market outcomes between males and females, foremost in earnings and hiring, given differences in observable productivity-altering characteristics. However, gender differences in productivity-altering characteristics, such as the form of human capital and occupational choice, have not been studied in this context of gender-biased practices (England, 1982). This lack of attention is surprising in view of the persistence of substantial male–female differences in educational and occupational choices and the expression of considerable concern about the low proportion of female scientists and engineers in the United States and Europe.

[17] In their initial college years, women are 2.5 times more likely than men to leave fields of mathematics, engineering, and physical sciences (Crocker et al., 1998). Girls in the U.S. also have lower enrollment rates than boys in advanced high-school science and math courses (Madigan, 1997).

[18] See, for example, Eccles (1994).

[19] National Science Foundation, 1998.

[20] For additional related evidence, see the 2003 report of the Equal Opportunity Commission. In U.K higher-education institutions in 2000/01, for example, the male/female student ratio was 1.6 in physical and mathematical sciences, 4.1 in computer science, and 5.4 in engineering and technology.

[21] Israeli Central Bureau of Statistics.

students in almost every subject or from the higher variance of performance of boys in every subject. The data do not support these two potential sources of the bias. On the other hand, variation of the estimated bias against male students in accordance with teachers' characteristics such as gender, age, and teaching experience strengthens the interpretation of the male bias as a form of discrimination resulting from teachers' behavior.

The magnitude of the grading bias against male students is relatively large and may harm students firstly by causing their final scores on some matriculation exams to fall below the passing mark and, thereby, by disqualifying them for a matriculation certificate. This may bar students from post-secondary schooling, at least temporarily, implying lower schooling attainment. Secondly, since admission to most university departments in Israel is based solely on the average score on the matriculation exams, the bias against boys lowers their average matriculation score and, by so doing, may reduce boys' prospects of admission to their preferred fields of study. Both effects have negative future labor-market consequences.

Finally, I should caution that the findings reported above that teachers are systematically giving lower grades to boys for the same quality of academic performance is not enough evidence to conclude that teaches are not engaging in other behavior that might reduce girls' desire to further human capital. In fact, it can be argued that the evidence of teachers male bias in grading tests could be consistent with teachers being tougher on boys because they have higher expectations for them. The tougher grades could also be viewed as a way to give male students stronger incentives to study harder in the future. Teachers may not use such an incentives device for girls if they do not have as high expectations for them. Such an alternative interpretation of the findings of this paper cannot be disproved since the evidence does not allow distinguishing between the hypotheses that teachers are lenient or merciful toward girls versus being tough or over demanding on boys.

References

American Association of University Women (AAUW), 1992. How Schools Short-Change Girls. AAUW Educational Foundation, USA.

American Association of University Women (AAUW), 1999. Gender Differences: Where Schools Still Fail our Children. Marlowe & Company, New York.

Aigner, Dennis J., Cain, Glen C., 1977. Statistical theories of discrimination in the labor. Industrial and Labor Relations Review 30 (2), 175–187 January.

Angrist, Joshua D., Lavy, Victor, 2004. The effect of high stakes high school achievement awards: evidence from a school-centered randomized trial. IZA Working Paper, vol. 1146. May.

Aronson, J., Quinn, D., Spencer, S.J., 1998. Stereotyping threat and the academic under-performance of minorities and women. In: Swim, J.K., Stanger, C. (Eds.), Prejudice. Academic Press.

Ben-Zvi Mayer, R., Hertz-Lazarovitz, Sefer, M., 1995. Teachers and pre-service teachers: on classifying boys and girls as distinguished students. In: Segen, N. (Ed.), Finding Gender Equality. Ministry of Education, Culture and Sports, Jerusalem, pp. 96–105 (Hebrew).

Bernard, M.E., 1979. Does sex role behavior influence the way teachers evaluate students? Journal of Educational Psychology 71 (4), 553–562.

Blank, Rebecca, 1991. The effects of double-blind versus single-blind refereeing: experimental evidence from the American economic review. American Economic Review 81 (5), 1041–1067 December.

Cain, Glen C., 1986. The economic analysis of labor market discrimination: a survey. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics. Amsterdam, North Holland, pp. 693–786.

Carr, M., Jessup, D., Fuller, D., 1999. Gender differences in first-grade mathematics strategy use: parent and teacher contributions. Journal for Research in Mathematics Education 30 (1), 20–46.

Crocker, J., Major, B., Steele, C., 1998. Social stigma, In: Gilbert, D., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, 4th ed. McGraw Hill, Boston, pp. 504–553.

Deaux, Kay, LaFrance, Marianne, 1998. Gender. The Handbook of Social Psychology, vol. II. McGraw-Hill. chap. 17.

Dovidio, J.F., Brigham, J.C., Johnson, B.T., Gaertner, S.L., 1996. Stereotyping, prejudice, and discrimination: another look. In: Macrae, N., Stangor, C., Hewstone, M. (Eds.), Stereotypes and Stereotyping. Guilford, New York, pp. 276–319.

Dusek, J.B., Joseph, G., 1983. The bases of teacher expectancies: a meta-analysis. Journal of Educational Psychology 75 (3), 327–346.

Eccles, J., 1994. Understanding women's educational and occupational choices: applying the Eccles et al. model of achievement-related choices. Psychology of Women Quarterly 18, 585–609.

Eccles, J.S., Jacobs, J.E., Harold, R.E., 1990. Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. Journal of Social Issues 46, 183–201.

England, Paula, 1982. The failure of human capital theory to explain occupational sex segregation. Journal of Human Resources 17 (3), 358–370 Summer.

Equal Opportunity Commission, 2003. Facts about women and men in Great Britain 2003. January. www.eoc.org.uk.

Fennema, E., Hart, L.E., 1994. Gender and the JRME. Journal for Research in Mathematics Education 25 (6), 648–659.

Fennema, E., Peterson, Penelope L., Carpenter, Thomas P., Lubinski, Cheryl A., 1990. Teachers' attributions and beliefs about girls, boys, and mathematics. Educational Studies in Mathematics 21 (1), 55–69.

Fiske, T. Susan, 1998. Stereotyping, prejudice, and discrimination. The Handbook of Social Psychology, vol. II. McGraw-Hill. chap. 25.

Glick, P., Wilk, K., Perrault, M., 1995. Images of occupations: components of gender and status in occupational stereotypes. Sex Roles 25 (5/6), 351–378.

Gneezy, Y., Niederle, M., Rustichini, A., 2003. Performance in competitive environments: gender differences. Quarterly Journal of Economics 1049–1074 August.

Goldin, Claudia, Rouse, Cecilia, 2000. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. American Economic Review 90 (4), 715–742 September.

Hallinan, M.T., Sorensen, A.B., 1987. Ability grouping and sex differences in mathematics achievement. Sociology of Education 60, 63–72 (April).

Hildebrand, G.M., 1996. "Redefining achievement," Murphy, P.F., Gipps, C.V., (Eds.), Equity in the Classroom (London and Washington, D.C. Falmer Press and UNESCO), pp. 149–171.

Hyde, J., S. Jaffee. "Perspectives from social and feminist psychology," Educational Researcher, 27 (5), pp. 14–16.

Jacobs, J.E., Eccles, J.E., 1992. The influence of parent stereotypes on parent and child ability beliefs in three domains. Journal of Personality and Social Psychology 63, 932–944.

Jacob, A. Brian, Levitt, Steven D., 2003. Rotten apples: an investigation of the prevalence and predictors of teachers cheating. Quarterly Journal of Economics 843–877 August.

Lavy, V., 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. Journal of Political Economy 10 (6), 1286–1318.

Lavy, Victor, 2004. Performance pay and teachers' effort, productivity and grading ethics. NBER Working Paper No. 10622.

Liang, Kung-Yee, Zeger, Scott L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Lindahl, E., 2006. Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden. Uppsala University. October 2006.

Madigan, T., 1997. Science proficiency and course taking in high school: the relationship of science course-taking patterns to increases in science proficiency between 8th and 12th grades. National Center for Education Statistics, NCES 97-838. U.S. Department of Education, Washington, D.C.

Madon, S., Jussin, L., Keiper, S., Eccles, J., Smith, A., Palumbo, P., 1998. The accuracy and power of sex, social class, and ethnic stereotypes: a naturalistic study in person perception. Personality and Social Psychology Bulletin 24 (12), 1304–1318 December.

Middleton, J., Spanias, P., 1999. Motivation for achievement in mathematics: findings, generalizations, and criticisms of the research. Journal of Research in Mathematics Education 30 (1), 65–88.
National Center for Educational Statistics, 1997. The Condition of Education 1997 (NCES 97-388). Author, Washington, D.C.
National Science Foundation, 1998. Women, Minorities, and Persons with Disabilities in Science and Engineering: 1994. VA, Arlington. (available on-line at www.nsf.gov/sbe/srs/nsf99338).
Rabin, Matthew, Schrag, Joel L., 1999. First impressions matter: a model of confirmatory bias. The Quarterly Journal of Economics 37–82 February.
Rebhorn, Leslie S., Miles, D.D., 1999. High-stakes testing: barrier to gifted girls in mathematics and science? School Science and Mathematics 99 (6), 313–318.
Rowsey, R.E., 1997. The effects of teachers and schooling on the vocational choice of university research scientists. School Science and Mathematics 72, 20–26.
Seymour, E., Hewitt, N., 1997. Talking about Leaving: Why Undergraduates Leave the Sciences. Westview Press, Boulder, CO.
Spencer, S.J., Steele, C.M., Quinn, D.M., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35, 4–28.
Stangor, C., Sechrist, G.B., 1998. Conceptualizing the determinants of academic choice and task performance across social groups. In: Swim, J.K., Stanger, C. (Eds.), Prejudice. Academic Press.
Steele, C.M., 1997. A threat in the air: how stereotypes shape intellectual identity and performance. American Psychologist 52, 613–629.
Tiedemann, J., 2000. Parents' gender stereotypes and teachers' beliefs as predictors of children' concept of their mathematical ability in elementary school. Journal of Educational Psychology 92 (1), 144–151.