NBER WORKING PAPER SERIES

ON THE ORIGINS OF GENDER-BIASED BEHAVIOR: THE ROLE OF EXPLICIT AND IMPLICIT STEREOTYPES

Eliana Avitzour
Adi Choen
Daphna Joel
Victor Lavy

Working Paper 27818
http://www.nber.org/papers/w27818

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2020

We thank B. Hameiri, U. Bram, Y. Bar-Anan, E. Duchini, E. Sand, and H. R. Trachtman for their useful comments. We also thank the Department of Education in Tel-Aviv municipality and the participating teachers for their cooperation in this study. Funding: None. Competing interests: None. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes
Eliana Avitzour, Adi Choen, Daphna Joel, and Victor Lavy
NBER Working Paper No. 27818
September 2020
JEL No. J16

ABSTRACT

In recent years, explicit bias against women in Science, Technology, Engineering and Math (STEM) is disappearing but gender discrimination is still prevalent. We assessed the gender-biased behavior and related explicit and implicit stereotypes of 93 math teachers to identify the psychological origins of such discrimination. We asked the teachers to grade math exam papers and assess the students' capabilities while manipulating the perceived gender of the students to capture gender-biased grading and assessment behavior. We also measured the teachers' implicit and explicit stereotypes regarding math, gender, and talent. We found that implicit, but not explicit, gender stereotypes correlated with grading and assessment behavior. We also found that participants who underestimated their own implicit stereotypes engaged in more pro-male discrimination compared to those who overestimated or accurately estimated them. Reducing implicit gender stereotypes and exposing individuals to their own implicit biases may be beneficial in promoting gender equality in STEM fields.

Eliana Avitzour
School of Psychological Sciences
Tel Aviv University
Tel Aviv
Israel
elianaa@mail.tau.ac.il

Adi Choen
School of Psychological Sciences
Tel Aviv University
Te Aviv
Israel
adi6@mail.tau.ac.il

Daphna Joel
School of Psychological Sciences
and Sagol School of Neuroscience
Tel Aviv University
Tel Aviv
Israel
djoel@tauex.tau.ac.il

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

1. Background

Why does gender discrimination in Science, Technology, Engineering, and Math (STEM) fields persist despite ongoing advancements in gender equality discourse and policies? In recent years, explicit bias against women in STEM has been rapidly disappearing (*GSS Data Explorer Key Trends*, 2019) and leading companies in STEM fields have been taking efforts to include more women within their ranks (*Women in the Workplace*, 2019). Yet despite these shifts in opinions and hiring policies, women still feel that their gender is a barrier to advancement (*Women in the Workplace*, 2019) and scholarly reviews suggest that gender discrimination is indeed still prevalent in the workplace, especially in STEM fields (Charlesworth & Banaji, 2019). For example, identical job applicants receive different offers based on their perceived gender: in a study concerning hiring behavior of STEM faculty, the same application was offered lower salaries and received more negative assessments of the candidate's competency and suitability, if the applicant name was female rather than male (Moss-Racusin et al., 2012). Gender discrimination is also well documented in math and science education. Multiple observational studies reveal that math teachers are susceptible to gender bias when grading examination papers (Breda & Hillion, 2016; Lavy, 2008; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018), and that this bias is consistent over their teaching careers (Lavy & Megalokonomou, 2019).

Charlesworth and Banaji suggest that implicit and explicit stereotypes are partially responsible for gender discrimination in STEM fields (Charlesworth & Banaji, 2019). Existing evidence points to a connection between lower representation of women in STEM fields and gender stereotypes regarding math and brilliance (Bian et al., 2018; Leslie et al., 2015; Nosek et al., 2009; Storage et al., 2016) and some observational studies demonstrate a possible relationship between teachers' implicit stereotypes and the achievements of students from stigmatized groups (Carlana, 2019; van den Bergh et al., 2010). However, only a handful of studies have investigated the relationship between stereotypes and discrimination experimentally, and those that did, did not contrast the importance of implicit versus explicit stereotypes (Moss-Racusin et al., 2012; Reuben et al., 2014). To more fully investigate the origins of gender-based discrimination, we designed a study that assessed participants' gender discrimination and related explicit and implicit stereotypes, in an ecological yet controlled experiment. We found that implicit, but not explicit, stereotypes correlated with discriminatory behavior. We also found that participants who underestimated their own implicit stereotypes engaged in more pro-male discrimination compared to those who overestimated or accurately estimated their own implicit stereotypes, a finding that could suggest that exposing individuals to their own implicit biases may be useful in promoting egalitarian behavior.

2. Method

Our experiment was conducted on primary school math teachers, who graded math examination papers of actual Fifth Grade students, assessed the students' capabilities, answered several questionnaires, and then performed the Gender-Science Implicit Association Test (IAT). Most studies about gender-biased grading use observational data in natural settings.[1] These studies employ various control methods in an attempt to rule out the possibility that differential grading is the result of the different exam-taking behaviors of girls and boys rather than the grading behavior of the teachers (Breda & Hillion, 2016; Lavy, 2008; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). We used an experimental design that rules out this confounding behavior completely by randomly assigning the gender of the student - every exam paper had a "female version" and a "male version" (that is, it appeared as though it was solved by a female or a male student) and so the exact same papers are graded by different teachers under different gender conditions. Several aspects of our design allowed us to boost the ecological validity of the study while still maintaining the methodological accuracy of a controlled experiment.

[1] An example of an exception is Hanna & Linden (2012), who randomly assigned student demographics to cover sheets of exam papers to capture gender and cast discrimination in teachers' grading in India.

First, our participants were professional math teachers and their task was to grade examination papers that were solved by actual Fifth Grade students. Second, most of our participants graded these papers from home, where, according to their own report, they usually grade papers. Lastly, we measured implicit gender stereotypes not only using the IAT but also in teachers' descriptions of their own students, in which we estimated and quantified their tendency to associate girls with hard work and boys with brilliance. Thus, our experiment comes close to the conditions of a field experiment in that we measured behaviors that are natural to our participants rather than a made-up task set in a lab room.

Participants: Ninety-three elementary school math teachers (88 women and 5 men) from numerous towns in central Israel participated in the experiment. Two additional teachers were dropped from the analyses because they had only taught in all-boy schools. Our sample size was based on the expectation that the effect of stereotypes on behavior would be medium (Cohen f^2=0.15). Detecting such an effect with detection power of 0.9 would require a sample size of around approximately 100 participants.

Procedure: Participants completed the study online at a time and place of their own convenience. The study was comprised of three parts. In Part I, participants graded the papers and answered questions about each student whose exam they had graded. In Part II, participants completed a questionnaire that included closed and open-ended questions. In Part III they took the Gender-Science IAT. Finally, they were thanked and awarded vouchers for a bookstore chain.

Materials:

Part I: Grading Exams and Assessing Students

Each exam paper was manipulated to appear as though it had been solved by a boy or a girl in accordance with its assigned condition. We graphically manipulated the text of the students' written answers - adding or removing affixes to words they wrote - to change the grammatical gender of the writer from male to female or vice versa (Fig. S1). We created four sets of exam papers. Each set contained the same twelve papers, solved by the same students, but with different papers manipulated and presented in the male and female conditions in each set. Teachers were randomly assigned one of the four sets at the beginning of the study.[2] Teachers were asked to grade the papers, write notes to the students where they thought appropriate, and write a final note for each student with advice or a summary of what the student should focus on[3]. After grading each paper, participants were also asked to fill out an assessment form about the student. Teachers were informed that the students would not see this form and were encouraged to answer honestly. The assessment form included questions about the student's (a) mathematical talent, (b) mathematical capability, (c) effort, and (d) chances of doing well in a top-level math class. The last question (e) asked teachers to advise on which math class level the student should be placed in the following school year (Table S1). The answers to all these questions were on 5-point Likert scales ranging from positive to negative assessments of the student (e.g. "Very Talented" to "Very Untalented").

Measuring gender-biased grading and assessment behavior

We used the grades and assessments that teachers gave the students to measure gender bias in their behavior. The average grade or assessment that an exam paper received across all conditions and teacher was used as a baseline for computing deviations for or against a specific "girl" or "boy".[4] If a teacher, for instance, awarded exam paper #137 two points more than the average points awarded to this exam paper, she demonstrated a two-point grading preference in favor of the student. If she indicated

[2] For further discussion see Supplementary Information

[3] We are currently analyzing teachers' notes to students, and these will be discussed in a separate article

[4] For further discussion see Supplementary Information

that student #137 was of high mathematical abilities (4 Likert scale points) and the rest of the sample indicated on average that student of exam paper #137 displayed medium mathematical abilities (3 Likert scale points), she demonstrated an assessment bias of one Likert scale point in favor of student #137. We calculated for each teacher a gender-biased grading behavior score using the following formula:

$$\frac{Teacher's\ average\ deviation\ for\ boys\ - Teacher's\ average\ deviation\ for\ girls}{SD\ of\ deviations\ across\ all\ teachers\ and\ papers}$$

Where positive values signify that a teacher favored boys and negative values signify that she favored girls. By subtracting a teacher's deviations in the Girl condition from her deviations in the Boy condition, we controlled for her personal tendency to give higher or lower grades.

Each teacher initially received six gender-biased behavior scores: one gender-biased grading score and five gender-biased assessment scores - of talent, capability, diligence, chances of doing well in a top-level class, and streaming advice. We tested for internal reliability of the five biases in assessment. Diligence is stereotypically associated with girls while the rest of the assessment items are stereotypically associated with boys (Bian et al., 2018; Leslie et al., 2015; Nosek et al., 2009; Storage et al., 2016). We therefore reversed the values of the diligence bias scores and tested the reliability of the five assessment biases. The analysis revealed questionable internal reliability (Cronbach's α = 0.63) due to the inclusion of the reversed diligence bias score and we therefore dropped the diligence score from the analysis. The remaining four biases (talent, capability, chances, and streaming advice) demonstrated excellent internal reliability (α = 0.90) and we therefore used their average as a single gender-biased assessment score. Finally, we created a combined gender-biased grading and assessment score by averaging the gender-biased assessment score and the gender-biased grading score (Cronbach's α = 0.73). This combined measure was our dependent variable in all further analyses. To simplify the term, we refer to this measure as *Gender-Biased Grading Behavior* throughout the text.[5]

Part II: The Survey

After grading and assessing the papers, teachers were asked to answer a survey. For all survey items see Table S2. The following paragraphs contain information regarding variables that were calculated using more than one survey item.

Implicit Gender-Brilliance Association: Teachers were asked to describe four students whom they had taught or are currently teaching: one of high potential who had succeeded, one of high potential who had failed, one of medium or low potential who had succeeded and one of medium or low potential who had failed. Ninety-three teachers wrote four descriptions of students. In 16 out of the 372 descriptions the gender of the students could not be inferred from the teachers' grammar and these were dropped from the analysis. We analyzed the content of the remaining 356 descriptions and tested for gender differences in the characteristics that emerged. The following paragraphs describe the group-level content analysis of the teachers' descriptions and the calculation of individual Implicit Gender-Brilliance Association scores in detail.

Content analysis: We removed all gender-identifying signs from the text, rendering all nouns, adjectives and verbs gender-neutral by using the male/female grammatical form, and asked two research assistants to identify prevalent student characteristics that appeared in the teachers' descriptions and that are relevant to success and failure. The research assistants were blind to the gender of the students in

[5] When using grading or assessment without averaging them, the results are qualitatively similar though the correlations and regression coefficient estimates are generally more precise when using grading than when using assessment.

the description texts. The first author of this article developed a coding scheme of eight characteristics based on the observations made by the research assistants and herself. Two other research assistants – also blind to the conditions and to the sex of the students – coded the recurrence of these characteristics in the description texts. The coding was done by each research assistant separately. Inter-coder agreement ranges from Krippendorf's Cuα = 0.76 to Cuα = 0.93 for all codes, indicating acceptable to excellent reliability (Table S3). After demonstrating inter-coder agreement, the two research assistants resolved the remaining differences of their coding by conversation. We then tested whether each of these eight characteristics was correlated with gender using chi-square test for independence (Fig. 2B; Table S3).

Implicit Gender-Brilliance Association scores: After demonstrating an association of boys with brilliance and girls with hard work at the group level (Fig. 2B) we calculated individual *Implicit Gender-Brilliance Association* scores by counting the number of stereotypical and counter-stereotypical statements in the descriptions of each teacher and subtracting the latter from the former. We defined a statement as stereotypical or counter-stereotypical based on the results of the group-level content analysis (Fig. 2B; Table S3). A stereotypical statement is therefore defined as describing a boy as (1) naturally talented, (2) lazy or unmotivated, or (3) messy. It is also defined as describing a girl as (4) lacking natural talent, (5) diligent or motivated, or (6) getting help from adults. A counter-stereotypical statement is the opposite (e.g. describing a girl as naturally talented or a boy as lacking natural talent). Statements referring to (7) having emotional problems and (8) lacking help from adults were not used in this calculation because they were not significantly associated with gender (Fig. 2B; Table S3).

Field-specific ability beliefs: Field specific ability beliefs are the view that brilliance is more crucial than effort for success in a certain field (Leslie et al., 2015). We asked three questions related to this belief. The first two were indirect, asking the teachers to indicate their levels of agreement with two opposing views: (1) "Most children have the necessary talent for math and the main reason for differences in performance is due to the effort that they invest in the subject" and (2) "Most children make an effort to succeed in math and the main reason for differences in performance is due to natural talent". The third item asked teachers to compare the importance of talent and effort for success in math directly, reading "What influences success in math more: innate talent or effort?". The first two items were meant to represent two mutually exclusive views and therefore be negatively correlated. However, there was no significant correlation between the two items [correlation coefficient r(92) = -0.13, P = 0.21]. We therefore dropped them from the analyses and used the direct comparison item alone.

Boy-Math stereotype: Two items were used to assess the extent to which teachers viewed boys as better in math. The two were averaged and were used as a single score (Cronbach's α = 0.71).

Gender Essentialism: four items were used to assess whether teachers saw men and women as inherently different from each other. Three of them represented essentialist views regarding gender (e.g. "Men and women are naturally different from each other in their ability, preferences and character") and one represented social constructionist views ("The differences in men and women's preferences and abilities are mostly the result of social circumstances: education, how they are treated, etc."). The three essentialist items demonstrated acceptable internal reliability (Cronbach's α = 0.77) but when adding the reversed social constructionist item, alpha was reduced to an unacceptable figure (Cronbach's α = 0.42). It was therefore dropped from the analysis and essentialism was measured through the average of the three essentialist items.

Feminism: Two items were used to assess support and identification with feminism. The two were averaged and used as a single score (Cronbach's α = 0.84).

Awareness of own implicit gender-science stereotype: To assess teachers' awareness of their own implicit stereotypes, we described the Gender-Science IAT prior to the test and asked them to predict their own IAT score on the 7-point Likert scale used by the Project Implicit website as feedback to participants. We then transformed their actual IAT results to the same 7-point scale as follows: *D*-scores >.65 were coded as +3, *D*-scores between.65 to.35 were coded as +2, *D*-scores between.35 and.15 were coded as +1, *D*-scores between.15 and −.15 were coded as 0, *D*-scores between -.15 and -.35 were coded as -1, *D*-scores between -.35 to -.65 were coded as -2 and *D*-scores <-.65 were coded as -3. These cut-offs were made according to conventions used on the Project Implicit website. We subtracted participants' IAT scores from their self-predicted scores and used the delta to assess participants' awareness of own implicit gender-science stereotype. A negative score signifies that a teacher predicted less stereotypical association than later demonstrated (underestimating one's own implicit Gender-Science stereotype) and a positive score signifies overestimating it. Participants with a score of zero have accurately predicted their own IAT results.

Part III: completing the Gender-Science Implicit Association Test

After answering the survey, participants completed a Hebrew version of the Gender-Science Implicit Association Test and their results were recorded in our database. We computed D-scores according to the scoring algorithm recommended in (Greenwald et al., 2003). Three of the participants did not complete the IAT due to technical problems. We used their data in all our analyses except those involving the IAT scores.

3. Analyses and Results:

The data relating to this study will be available at our websites upon publication of this manuscript. No code is available since we did not use it for any of the analyses.

Gender-biased behavior

We calculated a *Gender-Biased Grading Behavior* score for each teacher by comparing her grading and assessment behaviors in the "boy" and the "girl" papers. Consistent with previous findings (Breda & Hillion, 2016; Lavy, 2008; Lavy &

Megalokonomou, 2019; Lavy & Sand, 2018), some teachers favored boys when grading exams while others favored girls. The mean was not significantly different from zero (Fig. 1A).

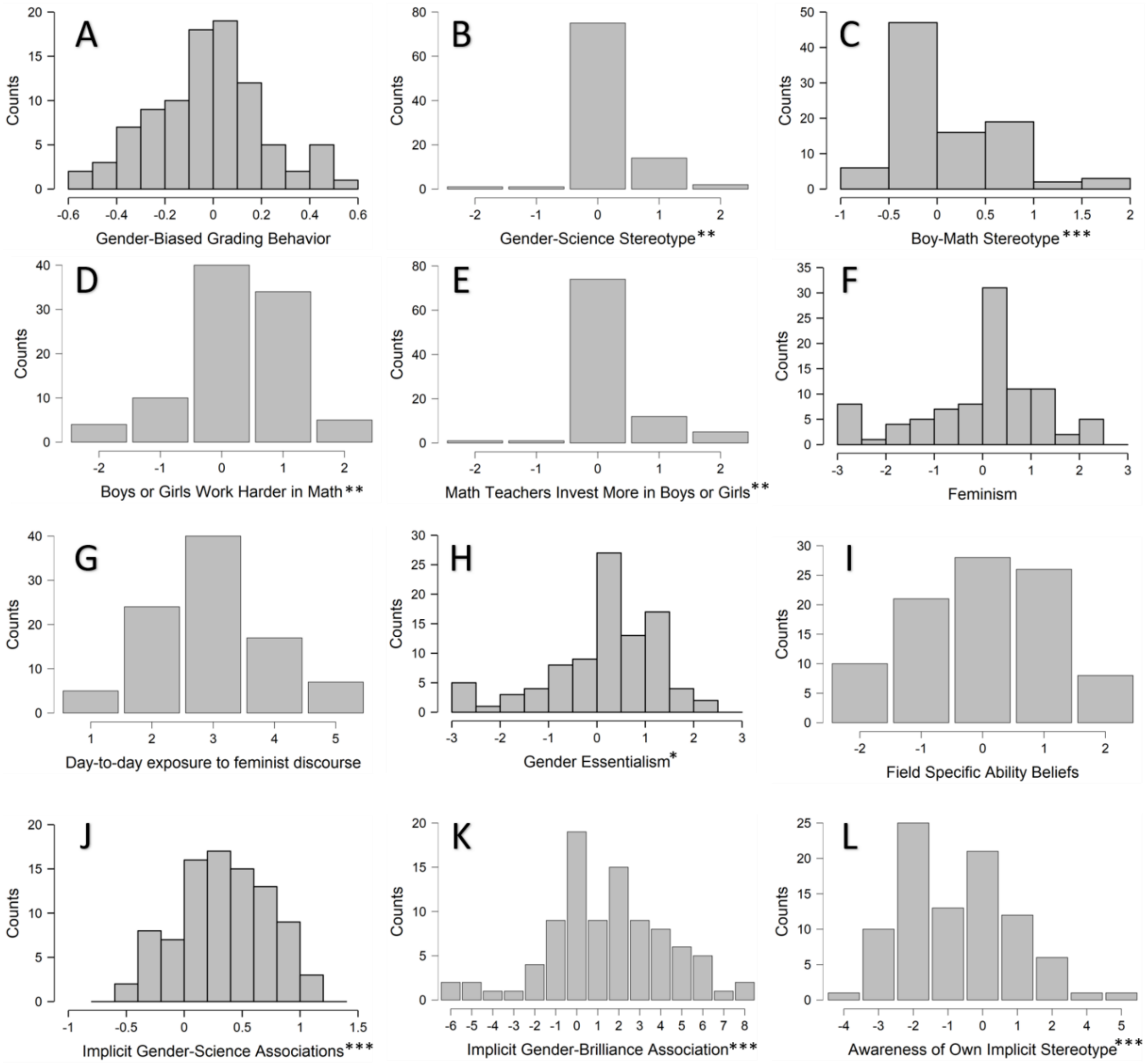


Fig. 1. Distributions of main variables in the study. *Notes.* Positive/higher values represent: (A) boy-favoring grading behavior; (B) male-science stereotype; (C) boy-math stereotype; (D) girls work harder; (E) invest more in boys; (F) feminist views; (G) more exposure; (H) essentialist views; (I) valuing talent over effort; (J) implicit men-science association; (K) implicit boy-brilliance association; (L) overestimating own stereotype. Distributions B-E and I theoretically range from -3 to +3 but the graphs only show values that were obtained in practice. Significance values refer to two-tailed one-sample t-tests against zero (**P*<0.05. ***P*<0.01. ****P*<0.001.). For descriptive statistics see table S4.

Implicit and explicit stereotypes and opinions

Teachers' explicit stereotypes and opinions were assessed with a series of questions regarding gender, STEM and brilliance (Table S2). On average, teachers expressed stereotypical and gender-essentialist views (Fig. 1B-D and Fig. 1H, respectively) and stated that teachers invest more in boys than girls (Fig. 1E). However, some of these results are driven by a small minority of teachers. For example, most teachers stated that men and women are equally suitable for science and humanities (Fig. 1B), that boys and girls are equally talented and successful in math (Fig. 1C), and that teachers invest their efforts in boys and girls equally (Fig. 1E). Yet because the non-egalitarian answers were overwhelmingly in the direction of boys, the averages are significantly higher than zero.

We measured two aspects of teachers' implicit gender stereotypes: (1) Implicit Gender-Science Association and (2) Implicit Gender-Brilliance Association. Teachers' *Implicit Gender-Science Association* was assessed using the Gender-Science Implicit Association Test (IAT). Consistent with the general population (Miller et al., 2015; Nosek et al., 2009), most teachers exhibited the stereotypical implicit association of males with science and females with the humanities (Fig. 1J). Teachers' *Implicit Gender-Brilliance Association* was assessed in their descriptions of their own students. Each teacher described four of her students: one of high potential who had succeeded, one of high potential who had failed, one of medium or low potential who had succeeded and one of medium or low potential who had failed. We found that teachers most frequently categorized male students as failing despite having high potential and female students as succeeding despite having medium or low potential (Fig. 2A). A logistic regression model with Potential (High/Medium or Low) and Outcome (Success/Failure) predicting the gender of the mentioned student shows significant main effects for both Potential ('high potential' predicting boy) [$\beta = 0.53$, Wald $\chi2 = 5.3$, $P = 0.02$, OR = 1.7] and Outcome ('success' predicting girl [$\beta = -0.83$, Wald $\chi2 = 14.1$, $P < 0.001$, OR = 0.418] (Table S5). We also analyzed the content of the descriptions and found that teachers tended to describe their male students as talented, messy, and lazy or unmotivated, and their female students as untalented, highly diligent or motivated, and as receiving help from adults (Fig. 2B; Table S3). It seems that the teachers remember their male students as messy geniuses who sometimes fail despite innate brilliance and their female students as mediocre students who sometimes succeed by working hard and getting help. In other words, they implicitly associate boys with brilliance and girls with hard work. For each teacher, we counted the number of stereotypical and counter-stereotypical statements in the descriptions they wrote of their students and subtracted the latter from the former to create individual *Implicit Gender-Brilliance Association* scores for further analyses (Fig. 1K).

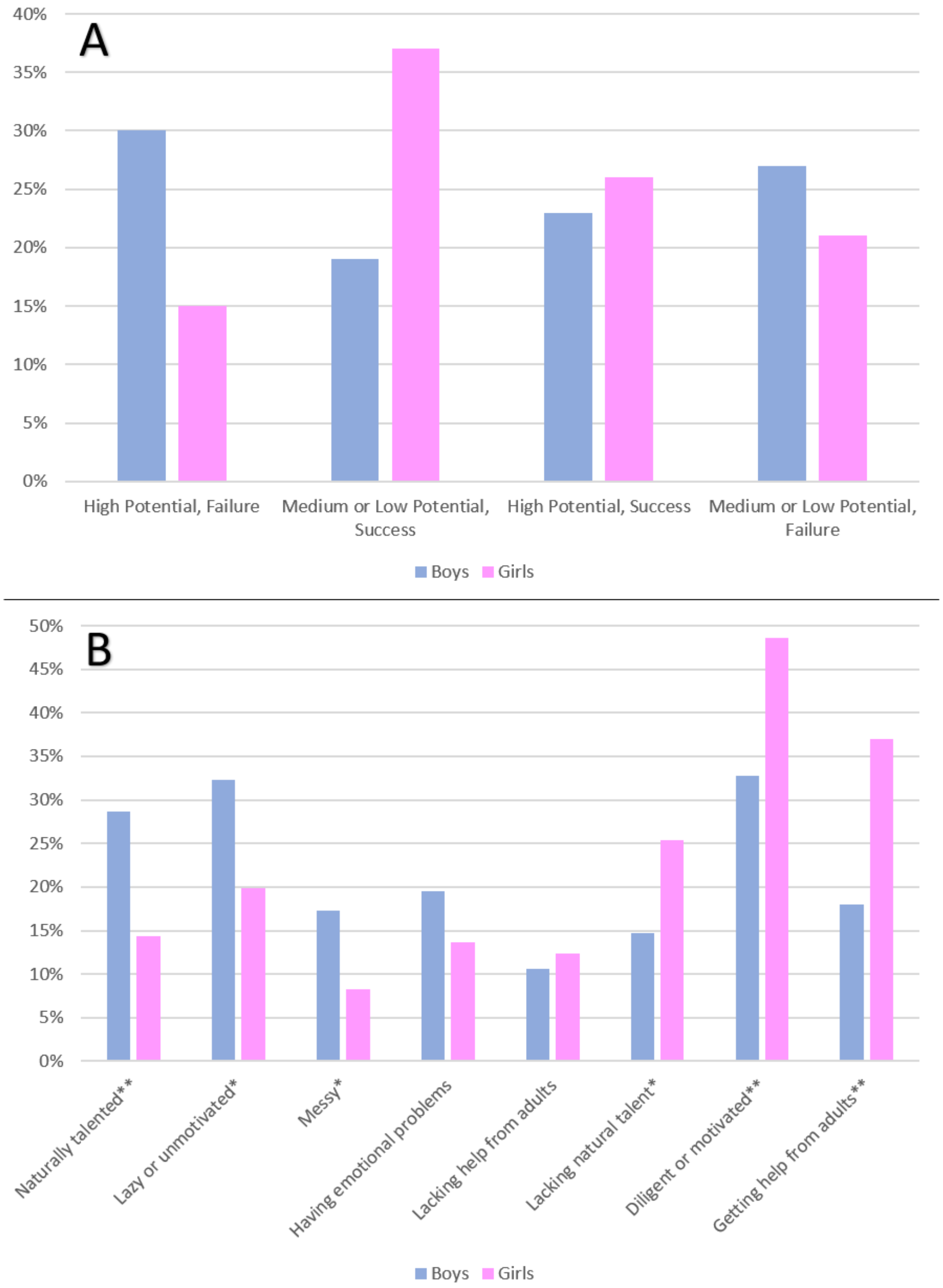


Fig. 2. Percentage of mentions of boys and girls (out of the total number of mentions of boys and girls, respectively) who were described in each category (A) and with each characteristic (B). Significance values refer to chi-square tests for independence (**P*<0.01. ***P* <0.001.).

The relations between gender-biased behavior and implicit and explicit stereotypes and opinions

To assess the relative contribution of explicit and implicit cognitions to teachers' grading behavior, we estimated the hierarchical regression model reported in Table 1. We started with a block of demographic control variables (model 1) and then added a block of explicit variables (model 2) and a block of implicit variables (model 3). None of the demographic or explicit opinion variables explained any variance in biased grading behavior (all p-values > 0.21) except one: believing that natural talent is more important for success in math than hard work. This belief, called *field-specific ability belief* (**9**) was positively correlated with boy-favoring grading behavior [correlation coefficient r(93) = 0.20, P = 0.049] and the coefficient remained significant when controlling for demographics [pr(93) = 0.24, P = 0.024]. While this field-specific ability belief was explicit, its correlation with boy-favoring grading behavior may reflect an implicit association between boys and brilliance. A recent study found that fields characterized by the belief that natural talent is more important for success have a lower representation of women (Leslie et al., 2015), and the authors hypothesized that this connection is mediated by the stereotype that women are less brilliant than men (The Field-Specific Ability Hypothesis; Leslie et al., 2015). Adding the belief that natural talent is more important for success in math than hard work to the block of implicit variables in the hierarchical regression model, we found that the coefficient remained significant (Table 1). Implicit gender-science associations (IAT scores) were also positively correlated with boy-favoring grading behavior [correlation coefficient r(90) = 0.23, P = 0.021]. This coefficient remained significant when controlling for demographics [pr(90) = 0.23, P = 0.028] and when added to the implicit block of the hierarchical regression model (Table 1). Last, Implicit gender-brilliance associations (stereotypical minus counter-stereotypical statements in descriptions of students) were also associated with boy-favoring grading behavior. The Pearson correlation coefficient approached significance [correlation coefficient r(93) = 0.19, P = 0.066], and became significant when controlling for demographics [pr(93) = 0.24, P = 0.023]. The coefficient was also significant when added to the implicit block of the hierarchical regression model (Table 1). Overall, the block of implicit variables (model 3) explained 15% of the variance in gender-biased grading behavior, while explicit variables (models 2) and demographics (model 1) explained none (Table 1).

Table 1. Hierarchical regression models predicting Gender-Biased Grading Behavior. N= 93 teachers except when stated. Significant statistics are bold. R^2 comparisons are always with the preceding model (to the left).

		Model 1			**Model 2**			**Model 3**		
Group	**Predictor**	**β**	**T**	**P**	**β**	**T**	**P**	**β**	**T**	**P**
Demographics	Year of birth	0.01	0.07	0.948	0.00	-0.02	0.987	0.02	0.13	0.899
	Education	-0.09	-0.79	0.430	-0.07	-0.55	0.582	-0.15	-1.27	0.208
	Experience	0.13	0.91	0.364	0.08	0.51	0.612	0.22	1.50	0.139
	Religiosity	0.05	0.49	0.628	0.07	0.55	0.587	0.09	0.80	0.424
Explicit Variables	Gender-Science Stereotype				0.00	-0.03	0.977	-0.03	-0.29	0.772
	Boy-Math Stereotype				0.08	0.64	0.521	-0.01	-0.07	0.943
	Boys or Girls Work Harder in Math				-0.07	-0.55	0.586	-0.12	-0.94	0.351
	Math Teachers Invest More in Boys or Girls				0.10	0.76	0.449	-0.04	-0.29	0.772
	Feminism				-0.07	-0.54	0.591	-0.09	-0.77	0.446
	Day-to-day Exposure to Feminist Discourse				0.04	0.33	0.745	-0.05	-0.38	0.704
	Gender Essentialism				-0.04	-0.32	0.752	-0.10	-0.84	0.406
Implicit Variables	Field-Specific Ability Beliefs							**0.25***	**2.23**	**0.029**
	Implicit Gender-Science Associations†							**0.26***	**2.24**	**0.028**
	Implicit Gender-Brilliance Association							**0.27***	**2.11**	**0.038**
Model Statistics	R^2	0.02			0.05			**0.20**		
	F for change in R^2	0.45			0.32			**4.78****		
	P for change in R^2	0.767			0.943			**0.004**		

*P<0.05.   **P<0.01.

† N = 90

Note. to ease the reading only key parameters are presented here. A fuller table, including unstandardized coefficients and 95% confidence intervals, are presented in Table S6.

Possible remedies for biased behavior: awareness of own implicit stereotypes

To assess teachers' awareness of their own implicit gender-science stereotype, we described the Gender-Science IAT prior to the test and asked the teachers to predict their scores. We then subtracted their actual score from their prediction to see if they underestimated their implicit stereotypes. On average, teachers in the sample underestimated their own stereotypes, predicting that their implicit associations will be significantly less stereotypical than they were (Fig. 1L). Further, the more they underestimated their stereotypical Gender-Science associations the more boy-favoring their grading behavior was [correlation coefficient r(90) = -0.212, P = 0.045]. The coefficient remained significant when controlling for demographics [pr(90) = -0.22, P = 0.039].

4. Discussion:

Overall, implicit measures of stereotypes were correlated with discriminatory behavior while explicit measures were not. Similarly, feminist views did not correlate with gender-biased grading behavior and neither did exposure to feminist discourses. In addition, the relations between implicit stereotypes and gender-biased grading behavior remained significant after controlling for explicit stereotypes, feminist views and exposure to feminist discourse (Table 1). Together, these findings suggest that the mechanisms that underlie biased behavior are not only implicit, they may also be independent from explicit opinions and identifications. Clearly, it is impossible to know whether scores on the different explicit measures reflect the true stereotypes and opinions of participants or only the greater ability to disguise these when explicit measures are used. Indeed, implicit associations may be driving behavior more than explicit stereotypes and opinions exactly because the former cannot be counteracted. This suggests, however, that reducing implicit stereotypes would help lessen discriminatory behavior and increase gender equality in STEM fields.

Reducing implicit stereotypes may contribute to gender equality in STEM fields not only because of their association with discriminatory behavior but also because it would create a more encouraging environment for girls and women. Teachers who discriminated in favor of girls typically exhibited less implicit stereotypes than those who discriminated in favor of boys, as reflected in the significant correlations between implicit gender stereotypes and gender-biased grading behavior. However, even within this subgroup of girl-favoring teachers, implicit stereotypes were significantly pro-male on average. Teachers with girl-favoring grading behavior demonstrated stereotypical gender-science (M=0.27, SD=0.38, t(46) = 4.92, P < 0.001) and stereotypical gender-brilliance (M=1.04, SD=2.91, t(48) = 2.49, P = 0.016) associations. If we consider the effects of self-fulfilling prophecies (Rosenthal & Jacobson, 1968), this could explain previous findings that boys are less negatively affected than girls by gender-biased grading (Lavy & Megalokonomou, 2019). Perhaps when boys are taught by a teacher who gives them lower grades and poorer assessments than they deserve, this teacher still associates male students with brilliance in math and interprets their failures as due to lack of effort. The student may internalize this teacher's view and conclude that his grades may improve if only he worked harder. By the same token, when girls are taught by a teacher who gives them lower grades and assessments, the teacher is likely to also have a stereotypical disassociation of female students with STEM and a tendency to interpret their success as due to effort, and failures as due to lack of talent. These girls may therefore feel that their low achievements are an accurate reflection of their abilities and consequently lose motivation to improve. A similar mechanism may be responsible for the observation that despite the conscious efforts made to increase the number of women in STEM companies, women still feel that their gender is a barrier to advancement and many do not maintain their jobs, a phenomenon referred to as the *leaky pipeline* of women in STEM (*Women in the Workplace*, 2019).

Another finding that demonstrates that general knowledge about stereotypes and biases may not be enough, is our observation of the different relations of discriminatory behavior with awareness of one's own bias versus awareness of other people's biases. While underestimating own implicit bias correlated with boy-favoring behavior, beliefs about gender-bias of other teachers did not. Responses to the question "Do most math teachers invest more effort in advancing and encouraging girls, or boys?" did not predict gender grading and assessment bias [correlation coefficient r(93) = 0.13, P = 0.21]. In other words, the general belief in the existence of biased behavior did not facilitate the rectification of one's own biases when grading and assessing students. This could be due to the low variability of teachers' answers to this question - 73% indicated that teachers are egalitarian in their encouragement of boys and girls - or because recognizing biased behavior in others allows one to feel protected from bias, whereas acknowledging one's own implicit bias promotes correction of biased behavior. Another beneficial tactic to promote gender equality may therefore be to increase individuals' awareness of their own implicit stereotypes, for example, by exposing them to their own IAT scores. Recent studies demonstrated the effectiveness of such interventions in the contexts of teachers' bias against immigrant children in Italy (Alesina et al., 2018) and of STEM faculty bias against women (Devine et al., 2017). We therefore expect that focusing on exposing individuals in positions of power, such as teachers and employers, to their own implicit biases and training them to overcome these biases will promote gender equality in STEM fields.

5. References

Alesina, A. F., Carlana, M., La Ferrara, E., & Pinotti, P. (2018). Revealing Stereotypes: Evidence from Immigrants in Schools. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3295948

Bian, L., Leslie, S. J., Murphy, M. C., & Cimpian, A. (2018). Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2017.11.006

Breda, T., & Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science*. https://doi.org/10.1126/science.aaf4372

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias*. *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjz008

Charlesworth, T. E. S., & Banaji, M. R. (2019). Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions. In *The Journal of neuroscience : the official journal of the Society for Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0475-18.2019

Devine, P. G., Forscher, P. S., Cox, W. T. L., Kaatz, A., Sheridan, J., & Carnes, M. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2017.07.002

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.85.2.197

GSS data explorer key trends. (2019). General Social Survey. https://gssdataexplorer.norc.org/trends

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*. https://doi.org/10.1257/pol.4.4.146

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

Lavy, V., & Megalokonomou, R. (2019). Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study. *Nber Working Paper Series*. https://doi.org/10.3386/w26021

Lavy, V., & Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*. https://doi.org/10.1016/j.jpubeco.2018.09.007

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*. https://doi.org/10.1126/science.1261375

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000005

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1211286109

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0809921106

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1314788111

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*. https://doi.org/10.1007/BF02322211

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0150194

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*. https://doi.org/10.3102/0002831209353594

Women in the Workplace. (2019). McKinsey & Company Report. https://www.mckinsey.com/featured-insights/gender-equality/women-in-the-workplace-2019

Supplementary Information for

On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes

Supplementary Text

Pilot Study

Gender-biased grading is often measured with the double-difference method (Botelho et al., 2015; Breda & Hillion, 2016; Burgess & Greaves, 2013; Lavy, 2008; Lavy & Megalokonomou, 2019; Lavy & Sand, 2018). This method compares the achievements of boys and girls in two separate examinations – one which is graded by teachers who know them (e.g. internal school exams), and one which is graded by blind examiners (e.g. national exams). One critique of this method is that the difference in the achievements of boys and girls in these two separate conditions might reflect the behavior of students rather than the behavior or biases of teachers (Lavy & Megalokonomou, 2019). To bypass this issue, we presented the same set of twelve papers to all participating teachers. Each paper was presented to half the teachers as belonging to a male student, and to the other half as belonging to a female student.

The exam included ten questions that resembled those of the national standardized math exams of recent years. Twenty-eight Fifth Grade students solved it. We chose sixteen of these twenty-eight papers for our pilot based on the following criteria:

a) Opportunity for bias: We read all twenty-eight papers and selected those that provided more room for subjective interpretation and grading. For example, papers with a simple computational error or with a correct answer to a multiple-choice question followed by an inaccurate explanation. In both examples, different teachers may grade differently.

b) Manipulation strength: We looked for the papers in which students used verbs that revealed their gender (in Hebrew, sentences like "I don't know", or "I think that…" are grammatically different in the male and female form). These cases allowed a simple graphical manipulation of student gender. By adding or removing affixes, we changed the grammatical gender from male to female or vice versa (Fig. S1). We created two versions of each paper – one of a 'boy' and another of a 'girl'.

Ten primary-school math teachers participated in our pilot. We asked the teachers to grade and assess the sixteen exam papers and to answer our questionnaire. We then analyzed the results and chose twelve out of the sixteen exam papers based on the following criteria:

1. High between-teachers grading variance (increasing the opportunity for biased grading).
2. High between-exam papers grade heterogeneity (i.e. some exam papers with low final grades and some with high final grades) to increase the ecological validity of the study.

During the pilot stage we measured the time it took teachers to grade the exams and conducted post-participation interviews. We used the interviews to ensure participants did not suspect a gender manipulation of the exams and to improve the overall design of the study. Based on these interviews, we decided to include only twelve exam papers instead of sixteen and to shorten the papers from the ten original questions to five questions in order to decrease the total amount of time teachers spent grading and assessing exams to below ninety minutes. This was done to prevent frustration and fatigue among participants, which could compromise their effectiveness and cooperation.

Recruiting participants:

The stated theme of the study was math instruction in Israel. To recruit participants, we first contacted the education departments of local authorities in central Israel and asked them to send an e-mail to their elementary school math teachers with an invitation to participate in the study (Table S7). To register for the study, participants had to confirm that they were practicing math teachers, who have taught or are teaching Fifth or Sixth Grade math. These teachers were invited to an event at Tel-Aviv University, which included a lecture on an unrelated topic (children's attention), followed by an online study in designated computer rooms6. As we learned that participants did not require any assistance to complete the study, subsequent participants were recruited via e-mail invitations by local authorities and via social media posts on math teachers designated Facebook pages and completed the study remotely (Table S7). After they registered online, teachers received a telephone call from a research assistant who confirmed their identity by asking them to repeat their name and the name of the school where they work while searching for their names on their school's website. The research assistant then gave them instructions on how to participate in the study and sent them a link and a personal ID number for cross referencing the different parts of the study. Participants completed the study by opening the link at a time and place of their own convenience.

Gender manipulation of the exam papers:

The first page of each exam paper presented fictitious demographics of the student. To ensure that the teachers paid attention to these details, we forced them to click on them in the following manner. Before grading each paper, participants landed on a page that presented a student ID number and three drop-down menus, each containing one demographic detail: (1) the student's year in school (5th grade in all cases), (2) gender (Girl / Boy), and (3) annual math Grade Point Average (GPA) of the student (High / Low) . Each drop-down menu contained only the 'correct' item for the student. For example, if student #137 was presented in the Girl condition then the gender drop-down menu would include only the item 'girl'. In the Boy condition, it would include only the item 'boy' (Fig. S2). We told teachers that students' demographics had already been entered into the system and that therefore each menu only contained the correct details of the student whose exam they are about to grade – but that due to a technical issue, this data had

[6] The lecture was unrelated to the study and was included to make the event attractive for teachers. In the main

paper we reported results using the entire sample. The same pattern of results was found with this initial set of data.

to be re-entered by them. We explained that they should therefore select the item which is available under each menu and move on to grading the exam. In addition, each exam paper was manipulated to appear as though it had been solved by a boy or a girl in accordance with its assigned condition. We graphically manipulated the text of the students' written answers - adding or removing affixes to words they wrote - to change the grammatical gender of the writer from male to female or vice versa (Fig. S1). We created four sets of exam papers, each containing the same twelve papers. The order of appearance of the twelve exam papers was randomized. The four conditions (High-GPA Girl, Low-GPA Girl, High-GPA Boy, Low-GPA Boy) were counter-balanced across the four sets. Teachers were randomly assigned one of the four sets at the beginning of the study. The GPA conditions (High/Low) were included in the study in order to explore whether gender-biased grading behavior is different for high and low achieving students. A two-way ANOVA with student gender (Girl/Boy) and GPA condition (High/Low) predicting grading behavior found no significant interaction between gender and GPA ($P>0.05$). We therefore ignored the GPA conditions in future calculations.

Measuring gender-biased behavior

Intuitively, measuring gender-biased behavior should be based on a comparison of the grade given by a teacher to some objective grades, for example grades of teachers who mark the papers gender-blindly. However, teachers are likely to imagine a male or a female student when marking, even if they are given no information about the student's gender. In that sense, it is not clear that a true 'gender-blind' condition can ever be achieved. In addition, Hebrew is a gendered language and the students' gender can be detected in their written answers. The conclusion is that it is hard to come up with an 'objective' grade. We therefore used as a reference point the average grade or assessment given to an exam paper across all conditions.

Additional Results:

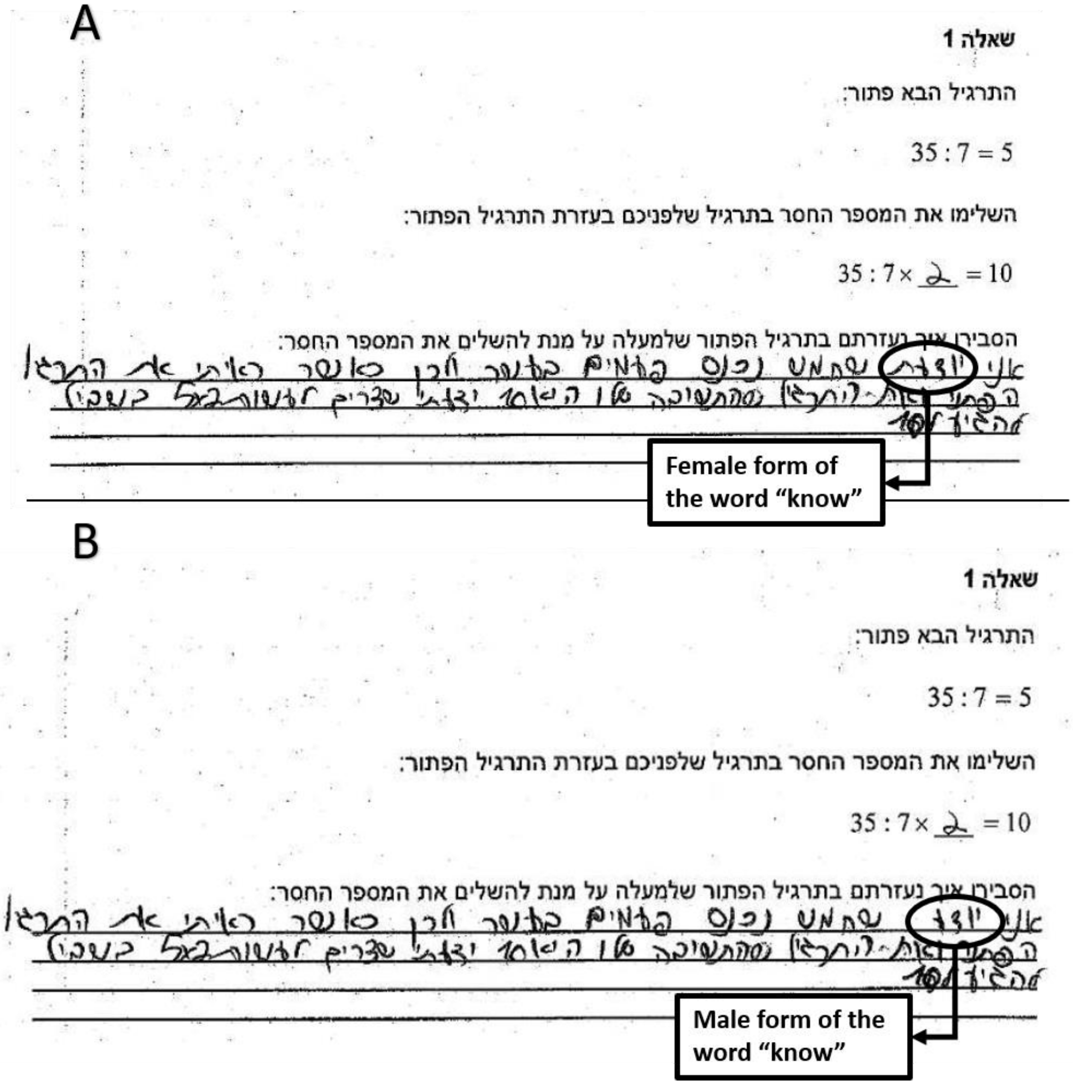
Validity of the gender-biased grading behavior variable: Teachers graded the exams of twelve students, each comprised of five questions. Each teacher's grading behavior was therefore recorded sixty times (12x5=60). Teachers also assessed the students' abilities over five questions, four of which were used in our analyses. Thus, assessment behavior was recorded forty-eight times for each teacher (12x4=48). The calculation of each teacher's gender-biased grading behavior is therefore based on one hundred and eight trials (60+48=108). This reduces the chances that the distribution of this measure is random noise. Furthermore, as reported in the main text, gender-biased grading behavior was correlated with all the implicit gender stereotypes that were measured in our study (together they explained 15% of the variability in gender-biased grading behavior, Table 1). That the distribution of gender-biased grading behavior is not random noise is further supported by the observation that a GPA-Biased Grading Behavior variable was not correlated with any of the variables in our study (all p-values > 0.153). GPA-Biased Grading Behavior was calculated using the following formula:

$$\frac{Teacher's\ average\ deviation\ for\ High\ GPA\ students - Teacher's\ average\ deviation\ for\ Low\ GPA\ students}{SD\ of\ deviations\ across\ all\ teachers\ and\ papers}$$

An alternative computation of gender-biased grading behavior variable: We used four sets of exam papers, each containing the same twelve papers. The four conditions (High-GPA Girl, Low-GPA Girl, High-GPA Boy, Low-GPA Boy) were counter-balanced across the four sets and teachers were randomly assigned one of the four sets when they started the study. Because the sets were randomized, the number of times an exam was presented in each condition was not identical across exams. To ensure that none of the conditions had a disproportionate effect on the exam average used for calculating deviations, we averaged each test in each of the four conditions separately and then used the average of these four averages as the baseline for calculating teachers' deviations. We repeated all our analyses with uncorrected averages and received the same pattern of results. We also repeated all our analyses while calculating each teachers' deviations from the average grade that was given to papers by all other 92 teachers, not including their own grade. Again, the same pattern of results was received.

Controlling for the number of Gender-Brilliance Association statements: A teacher's Implicit Gender-Brilliance Association score is defined as the difference between her stereotypical and counter-stereotypical statements. To ensure that teachers who wrote lengthier descriptions did not carry more weight in the results, we added the total number of statements (stereotypical and counter-stereotypical) as a controlling variable and repeated our analyses. The results remain unchanged, both in the simple Pearson correlation (the original coefficient is r(93) = 0.19 and when controlling for number of statements the partial coefficient is pr(93) = 0.19) and in the hierarchical regression model (Table S8). We therefore report the analyses without this control in the main text in order to simplify the interpretation of the parameter estimates.

An alternative computation of Awareness of Own Implicit Gender-Science Stereotype: To test the robustness of this variable we employed an alternative computation in addition to the one described in the Materials and Method section and repeated our analyses. We transformed both the teachers' predictions of their own IAT scores and their actual IAT scores to Z-Scores and subtracted Z-IATs from Z-Predictions. This alternative computation of awareness gives the same correlation between awareness and grading behavior which was reported in the main text but with a slightly stronger coefficient [r(90) = -0.23, P = 0.023]. This coefficient remains significant when controlling for demographics [pr(90) = -0.24, P = 0.023].



שאלה 1

התרגיל הבא פתור:

35 : 7 = 5

השלימו את המספר החסר בתרגיל שלפניכם בעזרת התרגיל הפתור:

35 : 7× 2 = 10

הסבירו איך נעזרתם בתרגיל הפתור שלמעלה על מנת להשלים את המספר החסר:

אני יודעת שחמש נכנס פעמים בעשר ולכן כאשר ראיתי את התרגיל
הפתור ואת התרגיל שהתשובה שלו היא 10 ידעתי שצריך לעשות 5x2 (כדי
להגיע ל10)

Female form of the word "know"

שאלה 1

התרגיל הבא פתור:

35 : 7 = 5

השלימו את המספר החסר בתרגיל שלפניכם בעזרת התרגיל הפתור:

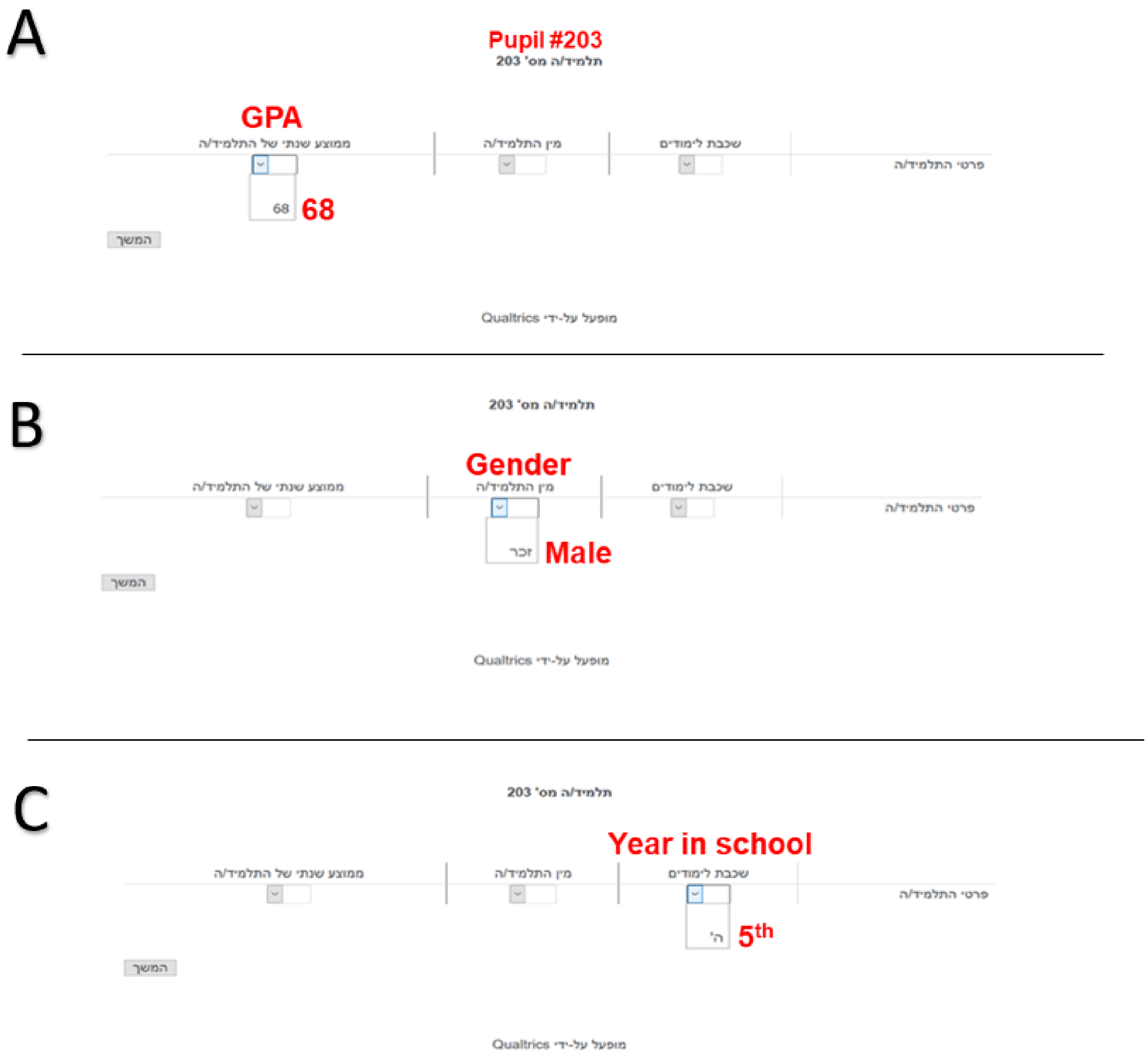
35 : 7× 2 = 10

הסבירו איך נעזרתם בתרגיל הפתור שלמעלה על מנת להשלים את המספר החסר:

אני יודע שחמש נכנס פעמים בעשר ולכן כאשר ראיתי את התרגיל
הפתור ואת התרגיל שהתשובה שלו היא 10 ידעתי שצריך לעשות 5x2 (כדי
להגיע ל10)

Male form of the word "know"

Fig. S1. Example of a student's written answer to question #1 of the exam paper, manipulated to appear in the female (A) and male (B) forms. The question reads: "Question 1 <break> The following exercise is solved: <break> 35:7=5 <break> complete the missing number in the following exercise with the solved exercise: <break> 35:7x___=10 <break> Explain how you used the solved exercise above to complete the missing number". The student answered: "I know that five can fit in ten twice and so when I saw the solved exercise and the exercise for which the answer is 10 I knew that I should do 5x2 to get to 10". The word "know" (circled in both sections) would take different grammatical forms for a male and female speaker. We graphically manipulated the answer to produce one male and one female version of the same answer, presenting it as a boy's paper to half the teachers and as a girl's paper to the other half.



תלמיד/ה מס' 203

פרטי התלמיד/ה	שכבת לימודים	מין התלמיד/ה	ממוצע שנתי של התלמיד/ה
			68

המשך

מופעל על-ידי Qualtrics

תלמיד/ה מס' 203

פרטי התלמיד/ה	שכבת לימודים	מין התלמיד/ה	ממוצע שנתי של התלמיד/ה
		זכר	

המשך

מופעל על-ידי Qualtrics

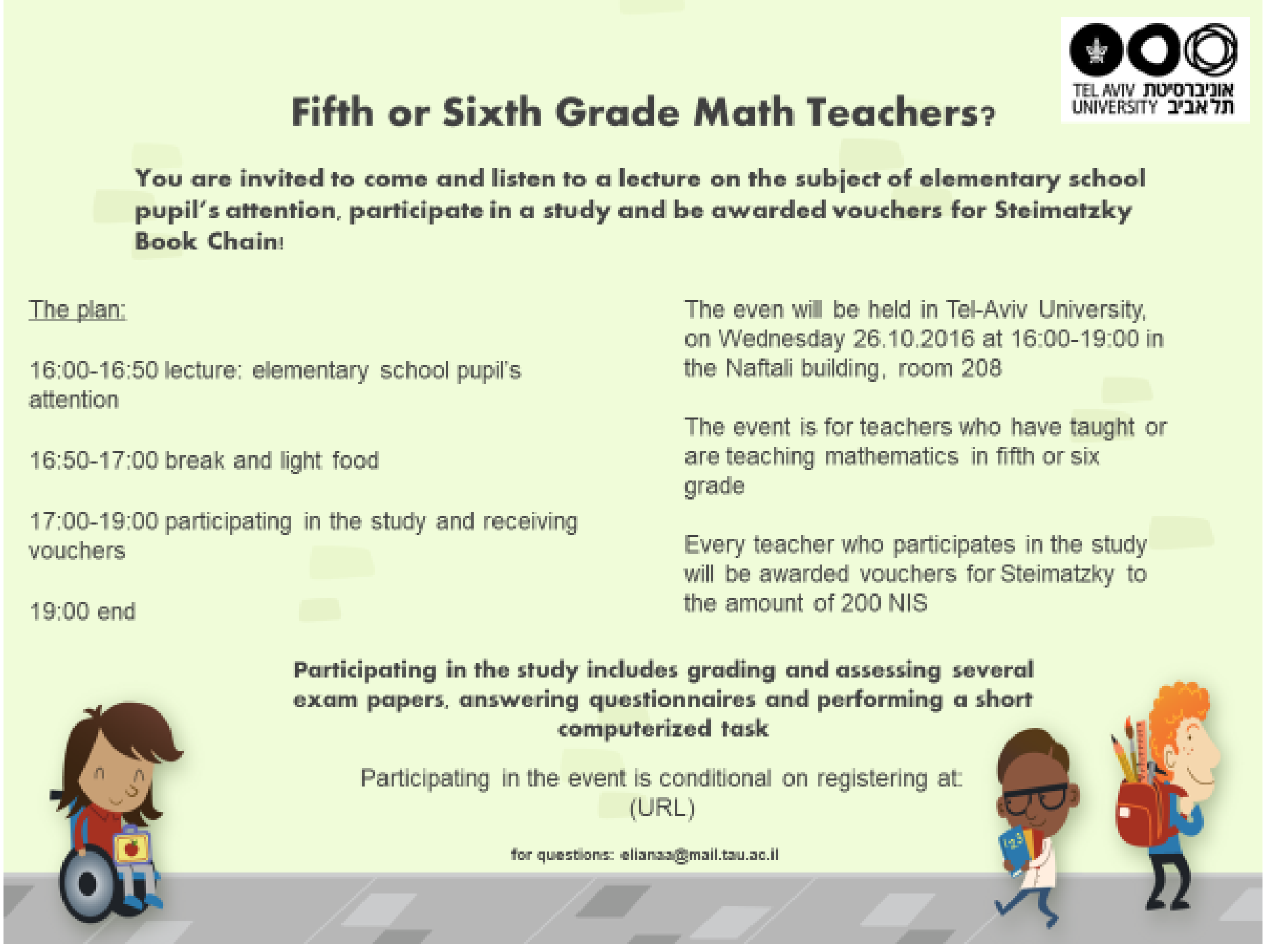
תלמיד/ה מס' 203

פרטי התלמיד/ה	שכבת לימודים	מין התלמיד/ה	ממוצע שנתי של התלמיד/ה
	ה'		

המשך

מופעל על-ידי Qualtrics

Fig. S2. Drop-down menus containing fictitious GPA (A), gender (B) and year in school (C). The teachers were forced to “re-enter” the demographics by selecting the only available option under each drop-down menu before moving on to grading the exam paper, thus ensuring they see the (fictitious) demographic details of each ‘student’.



TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Fifth or Sixth Grade Math Teachers?

You are invited to come and listen to a lecture on the subject of elementary school pupil's attention, participate in a study and be awarded vouchers for Steimatzky Book Chain!

The plan:

16:00-16:50 lecture: elementary school pupil's attention

16:50-17:00 break and light food

17:00-19:00 participating in the study and receiving vouchers

19:00 end

The even will be held in Tel-Aviv University, on Wednesday 26.10.2016 at 16:00-19:00 in the Naftali building, room 208

The event is for teachers who have taught or are teaching mathematics in fifth or six grade

Every teacher who participates in the study will be awarded vouchers for Steimatzky to the amount of 200 NIS

Participating in the study includes grading and assessing several exam papers, answering questionnaires and performing a short computerized task

Participating in the event is conditional on registering at: (URL)

for questions: elianaa@mail.tau.ac.il

Fig. S3. The (translated) invitation to participate in the study on campus. This flyer was sent to local authorities with a request to distribute to schools.

Table S1: Student Assessment Form

Please fill out the following regarding the student[a] whose exam sheet you have just marked.
Please note: the student will not see this assessment form. Please answer the questions honestly and as well as you can (you can go back and look at the exam again).

According to your assessment, how talented is the student in math?[b]

According to your assessment, what are the student's mathematical abilities?[c]

According to your assessment, to what extent is the student diligent? (D) (R)[b]

According to your assessment, what are the student's chances of being successful in a top-level class?[d]

Based on the exam that you have marked, in which class would you place this student?[e]

Notes. (D) indicates items that were dropped from the analysis due to low reliability. (R) indicates items that were reverse-scored.

[a] The word "student" is gendered in Hebrew, so teachers are reminded here again of the gender of the student whose exam paper they had marked.

[b] Responses to these items were given on a 5-point Likert scale (1=very little 5=very much).

[c] Responses to these items were given on a 5-point Likert scale (1=very low ability 5=very high ability).
Responses to these items were given on a 5-point Likert scale (1=very low chances 5=very high chance).

[e] Responses to these items were given on a 5-point Likert scale (1=Low achieving class 2=Unsure as to whether regular class or low achieving class 3=regular class 4=Unsure as to whether top-level or regular class 5=top level class).

Table S2: The survey items

Teachers' descriptions of own students

You will now be asked to think about a few students whom you have taught or are teaching and to tell us about them in two to three sentences (without using their names). Please describe the students as you see fit. If you're not sure what to write, you can describe the background of each student, their character and the reasons for their success and failure.[a]

Please briefly describe (in 2-3 sentences) a student of high potential in math who has succeeded:

Please briefly describe (in 2-3 sentences) a student of high potential in math who has not succeeded:

Please briefly describe (in 2-3 sentences) a student of medium or low potential in math who has succeeded:

Please briefly describe (in 2-3 sentences) a student of medium or low potential in math who has not succeeded:

Field-specific Ability Beliefs:

What influences success in math more: innate talent or effort?[b]

Most children have the necessary talent for math and the main reason for differences in performance is due to the effort that they invest in the subject[c] (D) (R)

Most children make an effort to succeed in math and the main reason for differences in performance is due to natural talent[c] (D)

Notes for this part of the table. (D) indicates items that were dropped from the analysis due to low reliability. (R) indicates items that were reversed.

[a] Since Hebrew is a gendered language, we wrote the word student in the conventional gender-neutral form of using a slash (/) to include both the male and female affix of the word.

[b]Responses to these items were given on a 5-point Likert scale (1=effort a lot more influential than talent 5=talent a lot more influential than effort).

[c]Responses to these items were given on a 6-point Likert scale (0=highly disagree 5=highly agree).

Table S2: The survey items – continued

Boy-Math Stereotype:

Who is better in math: boys or girls?[a]

Who receives higher grades in math: boys or girls?[a]

Gender-Science Stereotype:

You will now be asked to complete a computerized task. Before we start, please read the following paragraphs and answer the questions below. Different people have different opinions regarding the suitability of men and women for work in science. Our opinions can be comprised of a variety of factors and aspects: our life experiences, ideas that we know, our opinions about other matters and more. Examples of people's opinions regarding the suitability of men and women for work in science are:

- Women are more suitable for science and men are more suitable for the humanities
- Men are more suitable for science and women are more suitable for the humanities
- There is no relation between gender and the suitability for science

What is your opinion regarding the suitability of men and women for science and the humanities?[b]

Notes for this part of the table.

[a] Responses to these items were given on a 5-point Likert scale (1=girls a lot more than boys 5=boys a lot more than girls).

[b] Responses to these items were given on a 7-point Likert scale ranging from "men are a lot more suitable for science and women are a lot more suitable for the humanities" to "women are a lot more suitable for science and men are a lot more suitable for the humanities".

Table S2: The survey items – continued

Self-Forecast of IAT score

In addition to our conscious opinions, we all have automatic thoughts about men, women and science. Our automatic thoughts are triggered without our control, sometimes also without our knowledge. At times, our automatic thoughts are very different from our controlled thoughts. Because automatic thoughts can operate without our knowledge, they can influence our behavior and the way in which we judge and assess other people, unbeknownst to us.

The following task measures automatic thoughts. The task is called the "implicit association test". An association is the extent to which one term is connected or associated with another term. For example, a person can associate science with men more than science with women because of his belief about the different abilities of the two genders due to a social stereotype, or because of the different gender ratios of people who work in science.

Sometimes our associations are very different from our conscious thoughts. For example, a person might associate science with men more than science with women because of the different gender ratios of people who work in science, and yet hold the opinion that there is no connection between gender and the suitability for science.

The task will measure the extent to which you tend to associate science with men or women. It is possible to get the following results in this task:

- You have an automatic association between **men** and the exact sciences and between **women** and the humanities. This association may be strong, medium or weak.

- You have an automatic association between **women** and the exact sciences and between **men** and the humanities. This association may be strong, medium or weak.

- You have no automatic association between men and women and the exact sciences and the humanities.

Please answer the following question: What kind of automatic association do you think you will be shown to have between men and women and the exact sciences and the humanities?[a]

Notes for this part of the table.

[a] Responses to this item were given on the 7-point Likert scale in which IAT results are displayed in Project Implicit website, ranging from "strong association between men and the exact sciences and between women and the humanities" to "strong association between women and the exact sciences and between men and the humanities"

Table S2: The survey items - continued

Boys or Girls Work Harder in Math

Who invests more effort in math: boys or girls?[a] (R)

Math Teachers Invest More in Boys or Girls:

Do most math teachers invest more effort in advancing and encouraging girls or boys?[a]

Gender Essentialism

Men and women are naturally different from each other in their ability, preferences and character[b]

Men and women tend to think about solving mathematical problems differently[b]

The differences between men and women's preferences and abilities are mostly the result of biological differences between the sexes[b]

The differences in men and women's preferences and abilities are mostly the result of social circumstances: education, how they are treated, etc.[b] (R) (D)

Feminism:

I see myself as a feminist[a]

I support the feminist movement and its goals[a]

Day-to-Day Exposure to Feminist Discourse

To what extent are you exposed to discourses about feminism and the empowerment of women in your daily life?[b]

Demographics:

Age

Year of birth

Country of birth

Year of immigration (if relevant)

Religious stream

Town of residence

In what town is the school where you teach?

Do you have a teaching certificate?

What is your level of education?

What were your fields of study at university or college?

Notes for this part of the table. (D) indicates items that were dropped from the analysis due to low reliability. (R) indicates items that were reverse scored

[a]Responses to these items were given on a 5-point Likert scale (1=girls a lot more than boys 5=boys a lot more than girls).

[b]Responses to these items were given on a 5-point Likert scale (1=highly disagree 5=highly agree).

Table S2: The survey items – continued

Demographics - continued:

To what extent have you studied about feminism and the empowerment of women in high school?[a]

To what extent have you studied about feminism and the empowerment of women in higher education?[a]

In which of your higher education degrees have you studied about feminism and the empowerment of women?

How many years have you taught math?

Where do you usually grade papers?

Where did you grade papers in the context of this study?

Are you a homeroom teacher or have been one in the last three years?

Do you have other roles except for math teacher at your current school? (for example, coordinator)? If so, please describe.

How many students have you taught in the last year?

Of this number, how many were boys?

What is your family status?

Do you have children?

(If yes) How many sons do you have?

(If yes) How many daughters do you have?

What is the stream of education of your school?

Is the head of the school where you teach male or female?

How many math teachers are there in the entire school?

Of this number, how many are men?

Notes for this part of the table.

[a] Responses to these items were given on a 5-point Likert scale (1=very little 5=a lot).

Table S2: The survey items - continued

Filler items (interspersed throughout the questionnaire to mask the gender theme of the study):

Children from higher socio-economic backgrounds gain higher achievements in math than children from lower socio-economic backgrounds[a]

Some children are a lost cause when it comes to math[a]

Children who do not get homework support from their parents will find it hard to succeed in math[a]

Succeeding in math increases one's earning capacity[a]

It is important to encourage children to study math because succeeding in this field increases one's earning capacity[a]

Children do better academically when studying in homogenous groups with small gaps between children in the group[a]

In lower grades the gaps between children are smaller, and they grow larger in higher grades[a]

Children succeed or fail in math largely due to external circumstances such as family, socio-economic status, etc.[a]

It's possible to succeed in math without getting support from one's parents[a]

Most children like math[a]

Many children who like math are embarrassed to admit it[a]

Children who are good at math get positive reinforcement from their peer group[a]

In what should more time and resources be invested: promoting weak students or encouraging excellence in math?[b]

Do most teachers invest more effort in promoting weak students or in encouraging excellence in math?[b]

Notes for this part of the table.

[a] Responses to these items were given on a 5-point Likert scale (1=highly disagree 5=highly agree).

[b] Responses to these items were given on a 5-point Likert scale (1=invest in promoting weak students a lot more than in encouraging excellence 5=invest in encouraging excellence a lot more than in promoting weak students).

Table S3: Student characteristics in teachers' descriptions, by student gender.

Characteristic	# boys	# girls	% boys	% girls	χ2	P	Cu-α
Naturally talented	78	21	29%	14%	**12.06****	<0.001	0.90
Lazy or unmotivated	88	29	32%	20%	**8.57***	0.003	0.84
Messy	47	12	17%	8%	**7.01***	0.008	0.80
Having emotional problems	53	20	19%	14%	2.57	0.110	0.78
Lacking help from adults	29	18	11%	12%	0.20	0.700	0.86
Lacking natural talent	40	37	15%	25%	**6.89***	0.008	0.76
Diligent or motivated	89	71	33%	49%	**10.25***	0.001	0.91
Getting help from adults	49	54	18%	37%	**18.39****	<0.001	0.93

Notes. Significant statistics are bolded (**P*<0.01 ***P*<0.001). We present both the number (#) and the percentage (%) of boys and girls who were described with each characteristic. Each characteristic tested for independence of student gender with chi-squared test for independence.

Table S4. Descriptive statistics and t-tests for main variables in the study.

Variable	M	SD	T	DF	P
1. Gender-Biased Grading Behavior	-0.02	0.23	-0.86	92	0.389
2. Gender-Science Stereotype	__0.16__	__0.52__	__3.00**__	__92__	__0.003__
3. Boy-Math Stereotype	__0.34__	__0.59__	__5.67***__	__92__	__< .001__
4. Boys or Girls Work Harder in Math	__0.28__	__0.89__	__3.03**__	__92__	__0.003__
5. Math Teachers Invest More in Boys or Girls	__0.20__	__0.60__	__3.28**__	__92__	__0.001__
6. Feminism	0.25	1.26	1.88	92	0.062
7. Day-to-day Exposure to Feminist Discourse†	2.97	0.98	N/A		
8. Gender Essentialism	__0.25__	__1.13__	__2.13*__	__92__	__0.036__
9. Field-Specific Ability Beliefs	0.01	1.14	0.09	92	0.928
10. Implicit Gender-Science Associations	__0.35__	__0.39__	__8.54***__	__89__	__< .001__
11. Implicit Gender-Brilliance Associations	__1.45__	__2.88__	__4.86***__	__92__	__< .001__
12. Awareness of Own Implicit Gender-Science Stereotype	__-0.71__	__1.68__	__-4.02***__	__89__	__< .001__

Notes. Bolded statistics are significant. Significance values refer to two-tailed one-sample t-tests against zero (*P<0.05 **P<0.01 ***P<0.001). Positive = stereotypical views or behavior in all except in the following variables: 5 (positive=boys) 6 (positive=feminist views) 7 (1=very little 5=a lot) 8 (positive=gender essentialist views) 9 (positive=valuing talent over effort) and 12 (positive=overestimating own stereotype).

†t-test not applicable as exposure ranges from very little (1) to a lot (5).

Table S5. Logistic Regression Coefficients Indicating the Effects of Potential and Outcome Categories on Gender of Mentioned Student

					Wald Test		
Predictor	**Estimates**	**SE**	**OR**	**Z**	χ^2	**df**	**p**
(Intercept)	0.87	0.20	2.40	4.35	18.90	1	< .001
Potential	0.53	0.23	1.71	2.31	5.34	1	0.021
Outcome	-0.87	0.23	0.42	-3.76	14.11	1	< .001
Model Summary					19.74	353	< .001

Note. Potential: High=1 Medium or Low=0. Outcome: Success=1 Failure=0. Gender: Boy=1 Girl=0.

Table S6. Hierarchical regression models predicting Grading and Assessment Gender Bias. N= 93 teachers except when stated. Significant statistics are bold. R^2 comparisons are always with the preceding model (to the left).

		Model 1			Model 2			Model 3			95.0% Confidence Interval for B (Model 3)		Unstandardized Coefficients (Model 3)	
Group	Predictor	β	T	P	β	T	P	β	T	P	Lower Bound	Upper Bound	B	Std. Error
Demographics	Year of birth	0.01	0.07	0.948	0.00	-0.02	0.987	0.02	0.13	0.899	-0.01	0.01	0.00	0.00
	Education	-0.09	-0.79	0.430	-0.07	-0.55	0.582	-0.15	-1.27	0.208	-0.17	0.04	-0.07	0.05
	Experience	0.13	0.91	0.364	0.08	0.51	0.612	0.22	1.50	0.139	0.00	0.01	0.01	0.00
	Religiosity	0.05	0.49	0.628	0.07	0.55	0.587	0.09	0.80	0.424	-0.04	0.09	0.03	0.03
Explicit Variables	Gender-Science Stereotype				0.00	-0.03	0.977	-0.03	-0.29	0.772	-0.13	0.09	-0.02	0.06
	Boy-Math Stereotype				0.08	0.64	0.521	-0.01	-0.07	0.943	-0.10	0.09	0.00	0.05
	Boys or Girls Work Harder in Math				-0.07	-0.55	0.586	-0.12	-0.94	0.351	-0.09	0.03	-0.03	0.03
	Math Teachers Invest More in Boys or Girls				0.10	0.76	0.449	-0.04	-0.29	0.772	-0.11	0.08	-0.01	0.05
	Feminism				-0.07	-0.54	0.591	-0.09	-0.77	0.446	-0.06	0.03	-0.02	0.02
	Day-to-day Exposure to Feminist Discourse				0.04	0.33	0.745	-0.05	-0.38	0.704	-0.07	0.05	-0.01	0.03
	Gender Essentialism				-0.04	-0.32	0.752	-0.10	-0.84	0.406	-0.07	0.03	-0.02	0.02
Implicit Variables	Field-Specific Ability Beliefs							**0.25**	**2.23**	**0.029***	**0.01**	**0.10**	**0.05**	**0.02**
	Implicit Gender-Science Associations†							**0.26**	**2.24**	**0.028***	**0.02**	**0.30**	**0.16**	**0.07**
	Implicit Gender-Brilliance Associations							**0.27**	**2.11**	**0.038***	**0.00**	**0.04**	**0.02**	**0.01**
Model Statistics	R^2	0.02			0.05			**0.20****						
	F for change in R^2	0.45			0.32			**4.78**						
	P for change in R^2	0.767			0.943			**0.004**						

*P<0.05 **P<0.01

† N = 90

Table S7: Invitations to participate in the study

Text of e-mail sent to elementary school management teams inviting their teachers to participate in the study on campus (translated from Hebrew)

The following text was sent to local authorities with the request to forward it to primary school principals and math coordinators.

Tel-Aviv University is happy to invite current and former **teachers of 5th and 6th grade math** to attend a lecture on the topic of "Attention in Children of Primary School Age", to participate in a study about the instruction of math, and to receive a gift certificate to Steimatzky bookchain!

The event will be held in Tel-Aviv University, on Wednesday, 26.10.2016, at 16:00-19:00, in Sharet Building, room 110, and is designated for teachers who are currently teaching or have taught in the past 5th and 6th grade math. We will be grateful if you could forward this invitation to the relevant teachers in your schools.

For full details of the event, please the attached flyer [Fig. S3]. In addition, in accordance with the Chief Scientist of the Ministry of Education, I’m attaching a permit for conducting the study from the chief scientist and an official letter for the school teachers.

For questions you're welcome to reach out to Eliana in the address [e-mail address].

Sincerely,
The research team

Table S7: Invitations to participate in the study – continued

Text of e-mail sent to elementary school management teams inviting their teachers to participate in the study online (translated from Hebrew)

Hi [name of staff member],

I'm happy to invite the math teachers of [name of school] to participate in a study conducted by Tel-Aviv University. Participation will occur in the time and place which will be convenient for each teacher (it's online). Every participant will receive a modest gift.

Participating in the study will help us expand the knowledge about math studies is Israel. We hope it will add to the existing knowledge about the pedagogy of math instruction and therein lies its importance. Participation is open for teachers who are currently teaching or have taught math in 5th or 6th grade.

In accordance with the Chief Scientist of the Ministry of Education I'm attaching two documents: (1) a permit for conducting the study from the chief scientist and (2) an official letter for the school teachers. According to the Chief Scientist's regulations the permit is for the school management and the letter is to be distributed to the teachers.

I'll be grateful if you could distribute the letter to teachers and of course you are also welcome to participate.

For further details and for registration you can write to me at this e-mail address.

Best wishes,
Eliana, PhD student

Table S7: Invitations to participate in the study – continued

Official letter to teachers (attached to the e-mail that was sent to the school management)

Dear Teachers,

Subject: Participation in “Math Studies in Elementary Schools”

I have asked the Inspector General of Math in elementary education to distribute this letter among the math teachers for grades five and six.

Lately, the importance of math studies has come to the forefront of public discussion. The purpose of this research study is to examine different characteristics of teacher feedback and evaluation in the field of math, as well as the teachers’ standpoints about different educational and social subjects. We hope that this study will add to the existing knowledge in the scientific area of pedagogy of math, and therein lies its importance. This study is being performed in the context of my doctoral studies in the psychology department of Tel Aviv University, under the supervision of Prof. Daphna Joel.

For this study, we invite participation of teachers who teach or have taught math in grades five and six.

The data collection from the teachers will include these actions:

1. The teachers will assess and give feedback to a number of anonymous math exam papers.
2. The teachers will answer a questionnaire about their views on math pedagogy and about social phenomena and movements and perform a short, computerized task of social categorization of words.

The data collection is projected to last 1.5-2 hours. As compensation for their participation in the study, the subjects will receive a modest gift. Except for their email addresses (optional), participants will be instructed to maintain anonymity and not reveal identifying information. The study administrators will not record any identifying details. I would like to emphasize that:

-This study has been permitted by the Office of the Chief Scientist in the Ministry of Education, under the conditions of its permit (a copy of the permit has been delivered to the Inspector General of Math and is attached to this file for your perusal).

-The study results will be published in a way that will conceal the subjects’ identities.

-Other than the distribution of this letter, the Inspector General will not be involved in the data collection process so that, among other things, she will be unable to know which teachers agreed to participate in the study. Teachers who are interested in participating in the study are requested to respond directly to me at elianaa@mail.tau.ac.il to schedule a meeting in which they can receive a detailed explanation of the plan for the study.

Additionally, there are also plans for a continuing study on this subject. Teachers who are interested in receiving details about it are asked to write their email in the study questionnaire. The emails will be erased forever shortly after details about the study are sent, or by December 31, 2018, whichever comes first. When I email you the details, I will also attach the permit from the Chief Scientist’s Office for the continuing study.

Sincerely,
Eliana Avitzour
Primary Researcher

Table S7: Invitations to participate in the study – continued

Text of invitation posts in math teachers’ designated Facebook pages (translated from Hebrew)
Hi, as part of my PhD, I’m conducting a study about the instruction of math and am looking for participants. It’s a pleasant study that takes about two hours (online) and at the end you get a gift card for Steimatzky [a bookstore chain] of NIS 200 and the knowledge that you have contributed to the improvement of the quality of teaching in Israel ⍰ You can register at the following link: (URL)

Table S8. Hierarchical regression models predicting Gender-Biased Grading Behavior controlling for number of Gender-Brilliance statements. N= 93 teachers except when stated. Significant statistics are bold. R^2 comparisons are always with the preceding model (to the left).

		Model 1			**Model 2**			**Model 3**		
Group	**Predictor**	**β**	**T**	**P**	**β**	**T**	**P**	**β**	**T**	**P**
Demographics	Year of birth	0.01	0.07	0.948	0.00	-0.02	0.987	0.01	0.08	0.939
	Education	-0.09	-0.79	0.430	-0.07	-0.55	0.582	-0.16	-1.28	0.206
	Experience	0.13	0.91	0.364	0.08	0.51	0.612	0.22	1.43	0.158
	Religiosity	0.05	0.49	0.628	0.07	0.55	0.587	0.09	0.79	0.432
Explicit Variables	Gender-Science Stereotype				0.00	-0.03	0.977	-0.03	-0.26	0.797
	Boy-Math Stereotype				0.08	0.64	0.521	-0.01	-0.09	0.930
	Boys or Girls Work Harder in Math				-0.07	-0.55	0.586	-0.12	-0.96	0.341
	Math Teachers Invest More in Boys or Girls				0.10	0.76	0.449	-0.04	-0.33	0.740
	Feminism				-0.07	-0.54	0.591	-0.09	-0.77	0.446
	Day-to-day Exposure to Feminist Discourse				0.04	0.33	0.745	-0.04	-0.33	0.740
	Gender Essentialism				-0.04	-0.32	0.752	-0.10	-0.86	0.394
Implicit Variables	Field-Specific Ability Beliefs							**0.25***	**2.22**	**0.029**
	Implicit Gender-Science Associations†							**0.26***	**2.21**	**0.031**
	Implicit Gender-Brilliance Association							**0.28***	**2.06**	**0.043**
	Total number of stereotypical and counter-stereotypical Gender-Brilliance Association Statements (control variable)							-0.03	-0.22	0.828
Model Statistics	R^2		0.02			0.05			**0.20**	
	F for change in R^2		0.45			0.32			**3.55***	
	P for change in R^2		0.767			0.943			**0.011**	

*$P<0.05$

† N = 90