

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, Revised February 2021

We thank B. Hameiri, U. Bram, Y. Bar-Anan, E. Duchini, E. Sand, H. R. Trachtman, and participants in seminars and conferences 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2020 by Eliana Avitzour, Adi Choen, Daphna Joel, and Victor Lavy. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

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, Revised February 2021
JEL No. J16

ABSTRACT

This study examines the role of implicit and explicit stereotypes behind gender-discriminatory behavior. The empirical context is the grading discriminatory behavior of math teachers in experimental settings. Previous observational studies demonstrated that math teachers show gender bias when grading papers. The mechanisms behind this behavior are mostly unexplored. We asked teachers to grade gender-manipulated exam papers and measured their grading behavior and implicit and explicit gender stereotypes. We found that implicit gender stereotypes and underestimating own implicit stereotypes were associated with boy-favoring grading behavior. Reducing implicit gender stereotypes and exposing teachers to their implicit biases may promote gender equality in schools.

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. Introduction

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). Leading companies in STEM fields have been making 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 and 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 received lower wage offers and more negative assessments of the candidate’s competency and suitability if the applicant's name was female rather than male (Moss-Racusin et al. 2012; for similar results see Eaton et al., 2020). 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 (Lavy and Sand 2018; Lavy 2008; Breda and Hillion 2016; Lavy and Megalokonomou 2019). This bias is consistent over their teaching careers (Lavy and Megalokonomou 2019).

Charlesworth and Banaji suggest that implicit and explicit stereotypes are partially responsible for gender discrimination in STEM fields (Charlesworth and Banaji 2019). Existing evidence points to a connection between gender stereotypes regarding math and brilliance and lower representation of women in STEM fields (Leslie et al. 2015; Nosek et al. 2009; Bian et al. 2018; 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). Other observational studies demonstrate a similar association between implicit stereotypes and race discrimination (Rooth 2010; Glover, Pallais, and Pariente 2017).

However, only a handful of studies have investigated the relationship between stereotypes and discrimination experimentally, and those that have, do not contrast the importance of implicit versus explicit stereotypes (Moss-Racusin et al. 2012; Reuben, Sapienza, and Zingales 2014). This paper presents the first study that assessed participants' gender discrimination and related explicit and implicit stereotypes in an ecological, yet controlled experiment. We find that implicit, but not explicit, stereotypes are correlated with gender-discriminatory behavior. We also find that participants who underestimated their implicit stereotypes engaged in more pro-male discrimination compared to those who overestimated or accurately estimated their implicit stereotypes. This finding could suggest that exposing individuals to their own implicit biases may be useful in promoting egalitarian behaviors.

Our contribution relates to three strands of literature. First, studies on prejudice and discrimination often measure discriminatory behavior, explicit stereotypes, or implicit stereotypes. We extend this line of research by measuring all three in the same study and contrasting the relationships of the two types of stereotypes with gender-discriminatory behavior. Moreover, we measure teachers' implicit stereotypes not only by the Gender-Science Implicit Association Test, as has previously been done (for example Carlana, 2019), but also by content analysis of teachers' descriptions of their students.

Second, our study relates to recent studies on teachers' gender-biased behavior: Lavy (2008), Lavy and Sand (2019), and others¹ measured teachers' gender-based discriminatory behavior by calculating the difference between teachers' non-blind scores given to their students and blind scores of the same students on a different test (the "double-difference" method). We used a closed set of exam papers, each presented to teachers as either written by a girl or by a boy. We assessed teachers' gender-based discriminatory behavior by comparing a teacher's "assessment deltas" in the "boy" versus "girl" papers, where an "assessment delta" is defined as the difference between the grade

¹ For example, Cornwell, Mustard, and Van Parys (2013), Burgess and Greaves (2013), Diamond and Persson (2016), (Botelho, Madeira, and Rangel 2015), and (Terrier 2020) also use the systematic difference between non-blind and blind assessment across groups as a measure of such discrimination by teachers.

given to an exam paper by a specific teacher and the paper's average grade across the two gender conditions over the entire sample of teachers. The results of our study support the validity of this new measure as a proxy for discriminatory behavior. This new measure offers several advantages over the traditional double-difference method. For one, it is entirely experimental (because the same papers are presented as "boy" and "girl" to different teachers) yet ecologically valid (because we use papers solved by real students), which rules out any possibility of measuring student characteristics instead of teachers' behavior. In addition, it relies on a within-subject calculation, thus canceling out any differences in grading practices between the teachers except that of gender-discriminatory behavior. Lastly, it is flexible and scalable. For example, in the present study, teachers graded the exam papers of twelve students, each comprised of five questions. Each teacher's grading behavior was, therefore, recorded sixty times ($12 \times 5 = 60$). 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 ($12 \times 4 = 48$). Therefore, the calculation of each teacher's gender-discriminatory behavior is based on one hundred and eight trials ($60 + 48 = 108$), making it highly robust. However, the number of trials can easily be scaled up or down according to the desired robustness and the researchers' available resources. The "objective grade" of each exam paper was based on 93 separate gradings (as each teacher graded each exam), giving it high robustness, but this too could theoretically be calculated based on a smaller or larger sample in accordance with the researchers' needs.

The third strand to which our study contributes essential evidence is field-based literature on policies and interventions that aim to mitigate gender discrimination.² Our findings that implicit, more than explicit, stereotypes correlate with gender-

² See for example (Beaman et al. 2009) Bertrand & Duflo, 2016; Carnes et al., 2015; Devine et al., 2017; Kawakami et al., 2008; Kawakami, Phillips, et al., 2007.

discriminatory behavior suggest what should be the focus of corrective training to mitigate or reduce gender-based discrimination.

The rest of the paper is structured as follows. Section 2 presents the experiment. Section 3 describes the results and analysis. Section 4 offers extensions and robustness checks. Section 5 concludes with a discussion of our main findings.

2. The Experiment

Our experiment included a sample of primary school math teachers. We asked them to grade math examination papers of actual Fifth Grade students (who were unknown to them) and assess the students' capabilities. They also had to answer several questionnaires and then perform the Gender-Science Implicit Association Test (IAT). Most studies about gender-biased grading use observational data in natural settings.³ 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 (Lavy and Sand 2018; Lavy 2008; Breda and Hillion 2016; Lavy and Megalokonomou 2019). We used an experimental design that rules out this confounding behavior entirely 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 a female or a male student solved it). So the same papers are graded by different teachers under different gender conditions. Several aspects of our design allowed us to boost the study's ecological validity while still maintaining the methodological accuracy of controlled experiments. 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 reporting, they usually grade papers. Lastly, we measured implicit gender stereotypes not only using the IAT but also the teachers'

³ Two exceptions are Copur-Gencturk et al. (2020) and Hanna & Linden (2012), who manipulated the perceived gender of pupils experimentally.

descriptions of their own students – 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 laboratory room.

Participants: Ninety-three elementary school math teachers (88 women and five 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 impact with a detection power of 0.9 would require a sample size of approximately 100 participants.

Procedure: Participants completed the study online at a time and place of their 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.

Part I: Grading Exams and Assessing Students

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 were presented in the Girl condition, then the gender drop-down menu would include only the item 'girl'. The Boy condition would consist of only the item 'boy' (Fig. S1). We told teachers that students' demographics had already been entered into the system. Therefore, each menu contained only the correct details of the student whose exam they are about to grade – but that due to a technical issue, this data had to be re-entered by them. Therefore, we explained that they should select the available item under each menu before they can proceed to grade the exam.

Each exam paper was also manipulated to appear as though a boy or a girl had solved it 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 writer's grammatical gender from male to female or vice versa (Fig. S2). 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 counterbalanced 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 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. We, therefore, ignored the GPA conditions in future calculations.

Teachers were asked to grade the papers, write notes to the students where they thought appropriate, and write a final message for each student with advice or a summary of what the student should focus on.⁴ 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

⁴ We are currently analyzing teachers' notes to students, and these will be discussed in a separate article

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").

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 teachers was used as a baseline for computing deviations for or against a specific "girl" or "boy".⁵ 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 that student #137 was of high mathematical capabilities (4 Likert scale points) and the rest of the sample indicated on average that student of exam paper #137 displayed medium mathematical capabilities (3 Likert scale points), she demonstrated an assessment bias of one Likert scale point in favor of student #137. We calculated a gender-biased grading behavior score for each teacher using the following formula:

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

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 her 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,

⁵ For further discussion see Supplementary Information

diligence, chances of doing well in a top-level class, and streaming advice. We tested for internal reliability of the five biases in the assessment. Diligence is stereotypically associated with girls, while the rest of the assessment items are stereotypically associated with boys (Leslie et al. 2015; Nosek et al. 2009; Bian et al. 2018; Storage et al. 2016). Therefore, we 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 $\alpha = 0.63$) 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 ($\alpha = 0.90$), 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 $\alpha = 0.73$). 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.⁶

Part II: The Survey

After grading and assessing the papers, teachers were asked to answer a survey with items regarding beliefs about gender, STEM, and brilliance. 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 they had taught or are currently teaching: one of high potential who had

⁶ 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.

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. 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. 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 the description texts. We developed a coding scheme of eight characteristics based on the research assistants' and our own observations. Two other research assistants – also blind to the conditions and 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 Krippendorff's $C_{\alpha} = 0.76$ to $C_{\alpha} = 0.93$ for all codes, indicating acceptable to excellent reliability (Table 1). After demonstrating inter-coder agreement, the two research assistants resolved the remaining differences in their coding by conversation. We then tested whether each of these eight characteristics was correlated with gender using a chi-square test for independence (Fig. 1B ; Table 1). We then calculated *Implicit Gender-Brilliance Association* scores by counting the number of stereotypical and counter-stereotypical statements in each teacher's descriptions and subtracting the latter from the former. We defined a statement as stereotypical or counter-stereotypical based on the group-level content analysis (Fig. 1B ; Table 1). A stereotypical statement is therefore defined as describing a boy as (1) naturally talented, (2) lazy or unmotivated, or (3) messy, or 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. 1B ; Table 1).

Field-specific ability belief: Field-specific ability belief is the belief 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 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 $\alpha = 0.71$).

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”). 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 $\alpha = 0.77$), but when adding the reversed social constructionist item, alpha was reduced to an unacceptable figure (Cronbach’s $\alpha = 0.42$). Therefore, it was 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 $\alpha = 0.84$).

Awareness of own implicit gender-science stereotype: To assess teachers' awareness of their implicit stereotypes, we described the Gender-Science IAT before the test and asked them to predict their own IAT score on a 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 according to conventions used on the Project Implicit website (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). We subtracted participants' IAT scores from their self-predicted scores and used the delta to assess participants' awareness of their 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. Results and 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 (Lavy and Megalokonomou 2019; Lavy 2008; Lavy and

Sand 2018; Breda and Hillion 2016), some teachers favored boys when grading exams while others favored girls. The mean was not significantly different from zero (Fig. 2A).

Explicit stereotypes and opinions

Teachers' explicit stereotypes and opinions were assessed with various questions regarding gender, STEM, and brilliance (Table S2). On average, teachers expressed stereotypical and gender-essentialist views (Fig. 2B-D and Fig. 2H, respectively) and stated that teachers invest more in boys than girls (Fig. 2E). 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. 2B), that boys and girls are equally talented and successful in math (Fig. 2C), and that teachers invest their efforts in boys and girls equally (Fig. 2E). Yet because non-egalitarian answers were overwhelmingly in the boys' direction, the averages are significantly higher than zero.

Implicit stereotypes

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 (Nosek et al. 2009; Miller, Eagly, and Linn 2015), most teachers exhibited the stereotypical implicit association of males with science and females with the humanities (Fig. 2J).

Teachers' *Implicit Gender-Brilliance Association* was assessed through their descriptions of four students they had taught or are currently teaching who have high or medium/low potential and who had succeeded or failed (for more details, see Section 2 and Table S2). 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. 1A). A logistic regression model with Potential (High/Medium or Low) and Outcome (Success/Failure) predicting the gender of the

mentioned student showed significant main effects for both Potential ('high potential' predicting boy) [$\beta = 0.53$, Wald $\chi^2 = 5.3$, $P = 0.02$, OR = 1.7] and Outcome ('success' predicting girl [$\beta = -0.83$, Wald $\chi^2 = 14.1$, $P < 0.001$, OR = 0.418] (Table S3). We also analyzed the descriptions' content. We found that teachers tended to describe their male students as talented, messy, lazy, or unmotivated, and their female students as untalented, highly diligent or motivated, and receiving help from adults (Fig. 1B; Table 1). Thus, 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.

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

To assess the relative contributions of explicit and implicit stereotypes to teachers' grading behavior, we estimated the regression models reported in Table 2. We started with a specification that includes demographic control variables (model 1) and then added a set of explicit stereotypes and opinions measures (model 2) and a set of implicit stereotypes measures (model 3). None of the demographic or explicit stereotypes and opinions variables explained any variance in biased grading behavior (all p -values > 0.21) except one: believing that natural talent is more important for math than hard work. This belief, called *field-specific ability belief*, 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 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). 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 2). 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 2). Lastly, 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 2). 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 2).

Possible remedies for biased behavior: awareness of own implicit stereotypes

On average, teachers in the sample underestimated their stereotypes, predicting that their implicit associations will be significantly less stereotypical than they were (Fig. 2L). 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. Robustness Checks and Additional Results:

The 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 ($12 \times 5 = 60$). 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 ($12 \times 4 = 48$). Therefore, the calculation of each teacher's gender-biased grading behavior is 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 above, gender-biased grading behavior was correlated with all the implicit gender stereotypes measured in our study (together, they explained 15% of the variability in gender-biased grading behavior, Table 2). Further validation of the gender-biased grading behavior variable comes from observing that a GPA-Biased Grading Behavior variable was not correlated with any of our study variables (all p-values > 0.153). GPA-Biased Grading Behavior was calculated using the following formula:

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

An alternative computation of gender-biased grading behavior variable: Because teachers were randomly assigned to one of four sets of exam papers, 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. We then used the average of these four averages as the baseline for calculating teachers' deviations. We repeated all our analyses without this procedure, using the simple average of each test across the 93 teachers who graded it as the baseline for deviations, and received the same pattern of results. We also repeated all our analyses while calculating

each teachers' deviations from the average grade given to papers by all other 92 teachers, not including their grades. 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 the number of statements, the partial coefficient is $pr(93) = 0.19$) and in the regression model (Table S4). Therefore, we report the results without this control in the main text 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 Section 2 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 yields the same correlation between awareness and grading behavior 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 [$pr(90) = -0.24$, $P = 0.023$].

5. Discussion

Overall, we find that implicit measures of stereotypes are correlated with discriminatory behavior, while explicit stereotypes are not. Similarly, feminist views do not correlate with gender-biased grading behavior, and neither does exposure to feminist discourses. In addition, the relations between implicit stereotypes and gender-biased grading behavior remain significant after controlling for explicit stereotypes, feminist views, and exposure to feminist discourse (Table 2). Together, these findings suggest that the mechanisms that underlie discriminatory behavior are not only implicit, they may also be independent from explicit opinions and identifications. It is impossible to know whether scores on the different explicit measures reflect the true stereotypes and views of participants or a greater ability to conceal them when explicit measures are used. Indeed, implicit associations may be driving behavior more than explicit stereotypes and opinions precisely because the former cannot be counteracted. Whereas changing explicit stereotypes is not expected to affect discriminatory behavior, reducing implicit stereotypes might help lessen discriminatory behavior and increase gender equality in STEM fields.

Reducing implicit stereotypes may contribute to gender equality in STEM fields because of their association with discriminatory behavior and 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 sub-group of girl-favoring teachers, implicit stereotypes are 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 and Jacobson 1968; Glover, Pallais, and Pariente 2017), this could explain previous findings that boys are less negatively affected than girls by gender-biased grading (Lavy and 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 male 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 have a stereotypical disassociation of female students with STEM and a tendency to interpret their failures 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 the awareness of one's own bias versus awareness of other people's prejudices. 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 discriminatory behavior did not facilitate the rectification of one's own biases when grading and assessing students. Perhaps recognizing discriminatory behavior in others allows one to feel protected from discrimination, whereas acknowledging one's own implicit bias promotes correction of discriminatory conduct. Another beneficial tactic to encourage gender equality may therefore be to increase individuals' awareness of their 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). Therefore, we expect that 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.

6. References

- Alesina, Alberto F., Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2018. "Revealing Stereotypes: Evidence from Immigrants in Schools." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3295948>.
- Beaman, Lori, Raghendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics*. <https://doi.org/10.1162/qjec.2009.124.4.1497>.
- Bergh, Linda van den, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W. Holland. 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>.
- Bertrand, Marianne, and Esther Duflo. 2016. "Field Experiments on Discrimination." *National Bureau of Economic Research*. <https://doi.org/10.3386/w22014>.
- Bian, Lin, Sarah Jane Leslie, Mary C. Murphy, and Andrei Cimpian. 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>.
- Botelho, Fernando, Ricardo A. Madeira, and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics*. <https://doi.org/10.1257/app.20140352>.
- Breda, Thomas, and Mélina Hillion. 2016. "Teaching Accreditation Exams Reveal Grading Biases Favor Women in Male-Dominated Disciplines in France." *Science*. <https://doi.org/10.1126/science.aaf4372>.
- Burgess, Simon, and Ellen Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics*. <https://doi.org/10.1086/669340>.
- Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias*." *The Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjz008>.

- Carnes, Molly, Patricia G. Devine, Linda Baier Manwell, Angela Byars-Winston, Eve Fine, Cecilia E. Ford, Patrick Forscher, et al. 2015. "The Effect of an Intervention to Break the Gender Bias Habit for Faculty at One Institution: A Cluster Randomized, Controlled Trial." *Academic Medicine*.
<https://doi.org/10.1097/ACM.0000000000000552>.
- Charlesworth, Tessa E.S., and Mahzarin R. Banaji. 2019. "Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*.
<https://doi.org/10.1523/JNEUROSCI.0475-18.2019>.
- Copur-Gencturk, Yasemin, Joseph R. Cimpian, Sarah Theule Lubienski, and Ian Thacker. 2020. "Teachers' Bias Against the Mathematical Ability of Female, Black, and Hispanic Students." *Educational Researcher*.
<https://doi.org/10.3102/0013189X19890577>.
- Cornwell, Christopher, David B. Mustard, and Jessica Van Parys. 2013. "Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School." *Journal of Human Resources*.
<https://doi.org/10.3368/jhr.48.1.236>.
- Devine, Patricia G., Patrick S. Forscher, William T.L. Cox, Anna Kaatz, Jennifer Sheridan, and Molly Carnes. 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>.
- Diamond, Rebecca, and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." *National Bureau of Economic Research Working Paper Series*.
- Eaton, Asia A., Jessica F. Saunders, Ryan K. Jacobson, and Keon West. 2020. "How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors' Biased Evaluations of Physics and Biology Post-Doctoral Candidates." *Sex Roles*. <https://doi.org/10.1007/s11199-019-01052-w>.
- Glover, Dylan, Amanda Pallais, and William Pariente. 2017. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Store." *Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjx006>.
- Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 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. 2019.
<https://gssdataexplorer.norc.org/trends>.

- Hanna, Rema N., and Leigh L. Linden. 2012. "Discrimination in Grading." *American Economic Journal: Economic Policy*. <https://doi.org/10.1257/pol.4.4.146>.
- Kawakami, Kerry, John F. Dovidio, and Simone van Kamp. 2007. "The Impact of Counterstereotypic Training and Related Correction Processes on the Application of Stereotypes." *Group Processes and Intergroup Relations*. <https://doi.org/10.1177/1368430207074725>.
- Kawakami, Kerry, Jasper Moll, Sander Hermesen, John F. Dovidio, and Abby Russin. 2000. "Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation." *Journal of Personality and Social Psychology*. <https://doi.org/10.1037/0022-3514.78.5.871>.
- Kawakami, Kerry, Curtis E. Phillips, Jennifer R. Steele, and John F. Dovidio. 2007. "(Close) Distance Makes the Heart Grow Fonder: Improving Implicit Racial Attitudes and Interracial Interactions Through Approach Behaviors." *Journal of Personality and Social Psychology*. <https://doi.org/10.1037/0022-3514.92.6.957>.
- Kawakami, Kerry, Jennifer R. Steele, Claudia Cifa, Curtis E. Phillips, and John F. Dovidio. 2008. "Approaching Math Increases Math = Me and Math = Pleasant." *Journal of Experimental Social Psychology*. <https://doi.org/10.1016/j.jesp.2007.07.009>.
- Lavy, Victor. 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, Victor, and Rigissa Megalokonomou. 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, Victor, and Edith Sand. 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, Sarah Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. "Expectations of Brilliance Underlie Gender Distributions across Academic Disciplines." *Science*. <https://doi.org/10.1126/science.1261375>.
- Miller, David I., Alice H. Eagly, and Marcia C. Linn. 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, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 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, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, et al. 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, Ernesto, Paola Sapienza, and Luigi Zingales. 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>.
- Rooth, Dan Olof. 2010. "Automatic Associations and Discrimination in Hiring: Real World Evidence." *Labour Economics*. <https://doi.org/10.1016/j.labeco.2009.04.005>.
- Rosenthal, Robert, and Lenore Jacobson. 1968. "Pygmalion in the Classroom." *The Urban Review*. <https://doi.org/10.1007/BF02322211>.
- Storage, Daniel, Zachary Horne, Andrei Cimpian, and Sarah Jane Leslie. 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>.
- Terrier, Camille. 2020. "Boys Lag behind: How Teachers' Gender Biases Affect Student Achievement." *Economics of Education Review*. <https://doi.org/10.1016/j.econedurev.2020.101981>.
- "Women in the Workplace." 2019. McKinsey & Company Report. 2019. <https://www.mckinsey.com/featured-insights/gender-equality/women-in-the-workplace-2019>.

7. Tables and Figures

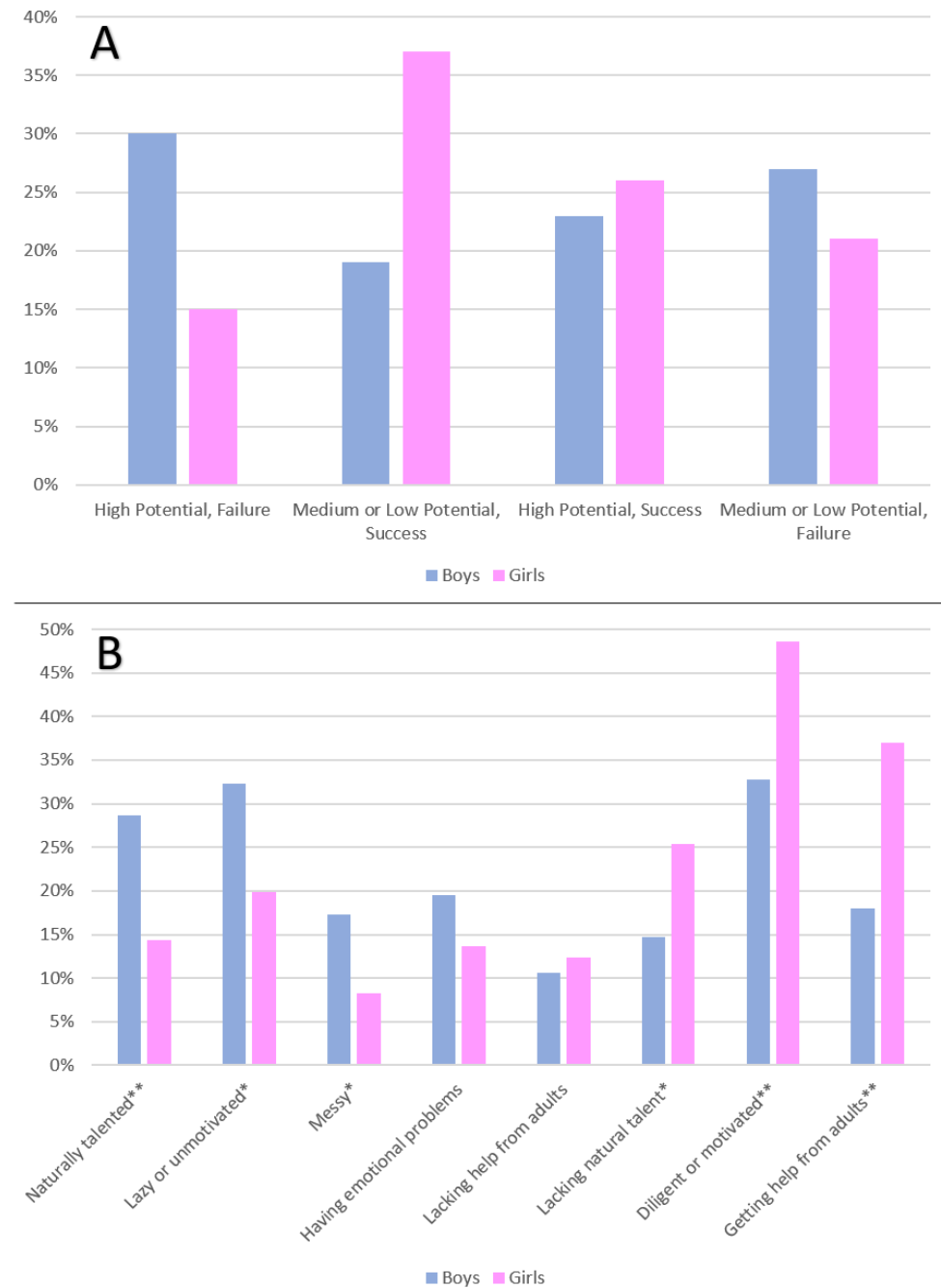


Fig. 1. Percentage of mentions of boys and girls (out of the total number of mentions of boys and girls, respectively) who were described in each category (A) and with each characteristic (B). Significance values refer to chi-square tests for independence (* $P < 0.01$, ** $P < 0.001$).

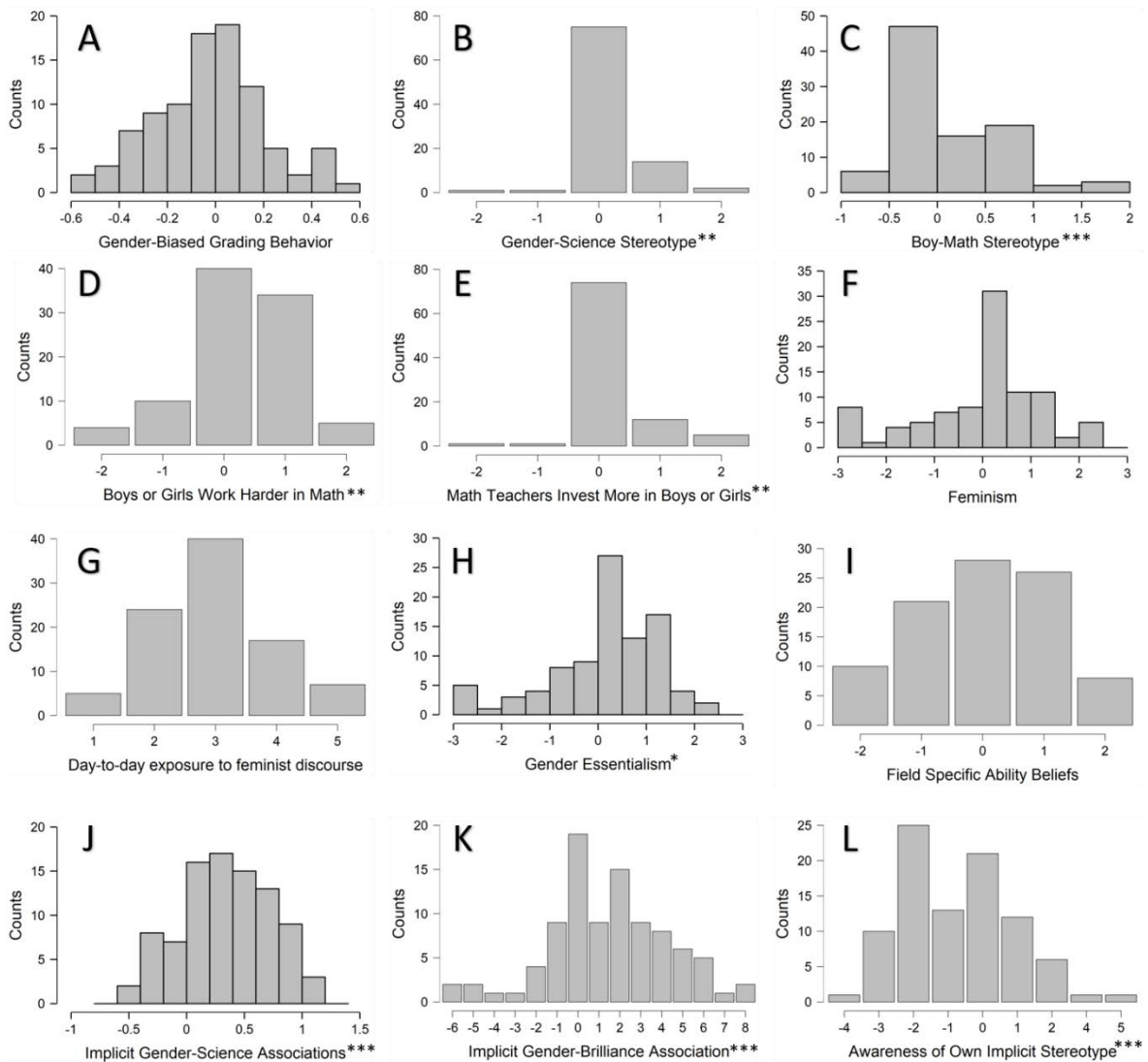


Fig. 2. Distributions of the main variables in the study. *Notes.* Positive/higher values represent: (A) boy-favoring grading behavior; (B) male-science stereotype; (C) boy-math stereotype; (D) girls work harder; (E) invest more in boys; (F) feminist views; (G) more exposure to feminism; (H) essentialist views; (I) valuing talent over effort; (J) implicit men-science association; (K) implicit boy-brilliance association; (L) overestimating one's own stereotype. Distributions B-E and I theoretically range from -3 to +3, but the graphs only show values 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 3.

Table 1: Student Characteristics in Teachers' Descriptions by Student Gender.

| Characteristic | # boys | # girls | % boys | % girls | χ^2 | 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 in bold. We present both the absolute number (#) and the percentage (%) of boys and girls described with each characteristic out of the total number of boys and girls, respectively. Each characteristic was tested for the independence of student gender with a chi-squared test for independence.

Table 2. Regression Models of Grading and Assessment Gender Bias.

| Group | Predictor | Model 1 | | | Model 2 | | | Model 3 | | | |
|---------------------------|--|-------------|-------|-------|-------------|-------|-------|--------------------|-------------|--------------|--------------------|
| | | B | T | P | β | T | P | β | T | P | B |
| Demographics | Year of birth | 0.01 | 0.07 | 0.948 | 0.00 | -0.02 | 0.987 | 0.02 | 0.13 | 0.899 | 0.00 (0.00) |
| | Education | -0.09 | -0.79 | 0.430 | -0.07 | -0.55 | 0.582 | -0.15 | -1.27 | 0.208 | -0.07 (0.05) |
| | Experience | 0.13 | 0.91 | 0.364 | 0.08 | 0.51 | 0.612 | 0.22 | 1.50 | 0.139 | 0.01 (0.00) |
| | Religiosity | 0.05 | 0.49 | 0.628 | 0.07 | 0.55 | 0.587 | 0.09 | 0.80 | 0.424 | 0.03 (0.03) |
| Explicit Variables | Gender-Science Stereotype | | | | 0.00 | -0.03 | 0.977 | -0.03 | -0.29 | 0.772 | -0.02 (0.06) |
| | Boy-Math Stereotype | | | | 0.08 | 0.64 | 0.521 | -0.01 | -0.07 | 0.943 | 0.00 (0.05) |
| | Boys or Girls Work Harder in Math | | | | -0.07 | -0.55 | 0.586 | -0.12 | -0.94 | 0.351 | -0.03 (0.03) |
| | Math Teachers Invest More in Boys or Girls | | | | 0.10 | 0.76 | 0.449 | -0.04 | -0.29 | 0.772 | -0.01 (0.05) |
| | Feminism | | | | -0.07 | -0.54 | 0.591 | -0.09 | -0.77 | 0.446 | -0.02 (0.02) |
| | Day-to-day exposure to Feminist Discourse | | | | 0.04 | 0.33 | 0.745 | -0.05 | -0.38 | 0.704 | -0.01 (0.03) |
| | Gender Essentialism | | | | -0.04 | -0.32 | 0.752 | -0.10 | -0.84 | 0.406 | -0.02 (0.02) |
| Implicit Variables | Field-Specific Ability Beliefs | | | | | | | 0.25 | 2.23 | 0.029 | 0.05 (0.02) |
| | Implicit Gender-Science Associations† | | | | | | | 0.26 | 2.24 | 0.028 | 0.16 (0.07) |
| | Implicit Gender-Brilliance Associations | | | | | | | 0.27 | 2.11 | 0.038 | 0.02 (0.01) |
| Model Statistics | R ² | 0.02 (0.23) | | | 0.05 (0.24) | | | 0.20 (0.22) | | | |
| | F for change in R ² | 0.45 | | | 0.32 | | | 4.78 | | | |
| | P for change in R ² | 0.767 | | | 0.943 | | | 0.004 | | | |

† N = 90

Notes: N= 93 teachers except when stated otherwise. Significant statistics are bold. R² comparisons are always with the preceding model (to the left). Standard errors are in brackets. B indicates the parameter estimate, T the t-value, and P the respective level of significance.

Table 3. 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 |

Online appendix

Not for publication

Supplementary Information for

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

This PDF file includes:

Supplementary text
Figures S1 to S3
Tables S1 to S5

Supplementary Information 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 boys and girls' achievements in two separate examinations – one graded by teachers who know the students (e.g., internal school exams), and another exam graded by blind examiners (e.g., national exams). One critique of this method is that the difference in boys and girls' achievements in these two different conditions might reflect students' behavior 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 study 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 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 the student gender. By adding or removing affixes, we changed the grammatical gender from male to female or vice versa (Fig. S2). 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 study. 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 study's ecological validity.

We measured the time it took teachers to grade the exams during the pilot stage and conducted post-participation interviews. We used the interviews to ensure participants did not suspect a gender manipulation of the exams and improve the study's overall design. Based on these interviews, we decided to include only twelve exam papers instead of sixteen and shorten the papers from the ten original questions to five questions 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 to the main study:

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. We asked them to send an e-mail to their elementary school math teachers with an invitation to participate in the study (Table S5). 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 rooms. 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 S5). 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.

Tables and Figures

A

Pupil #203
תלמידה מס' 203

GPA

ממוצע שנתי של התלמיד/ה

68

68

מסך המשך

מספד על-ידי Qualtrics

B

תלמידה מס' 203

Gender

ממוצע שנתי של התלמיד/ה

זכר

Male

מסך המשך

מספד על-ידי Qualtrics

C

תלמידה מס' 203

Year in school

ממוצע שנתי של התלמיד/ה

ה'

5th

מסך המשך

מספד על-ידי Qualtrics

Fig. S1. 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 that they see the (fictitious) demographic details of each ‘student’. The English translation of the Hebrew text is given in red.

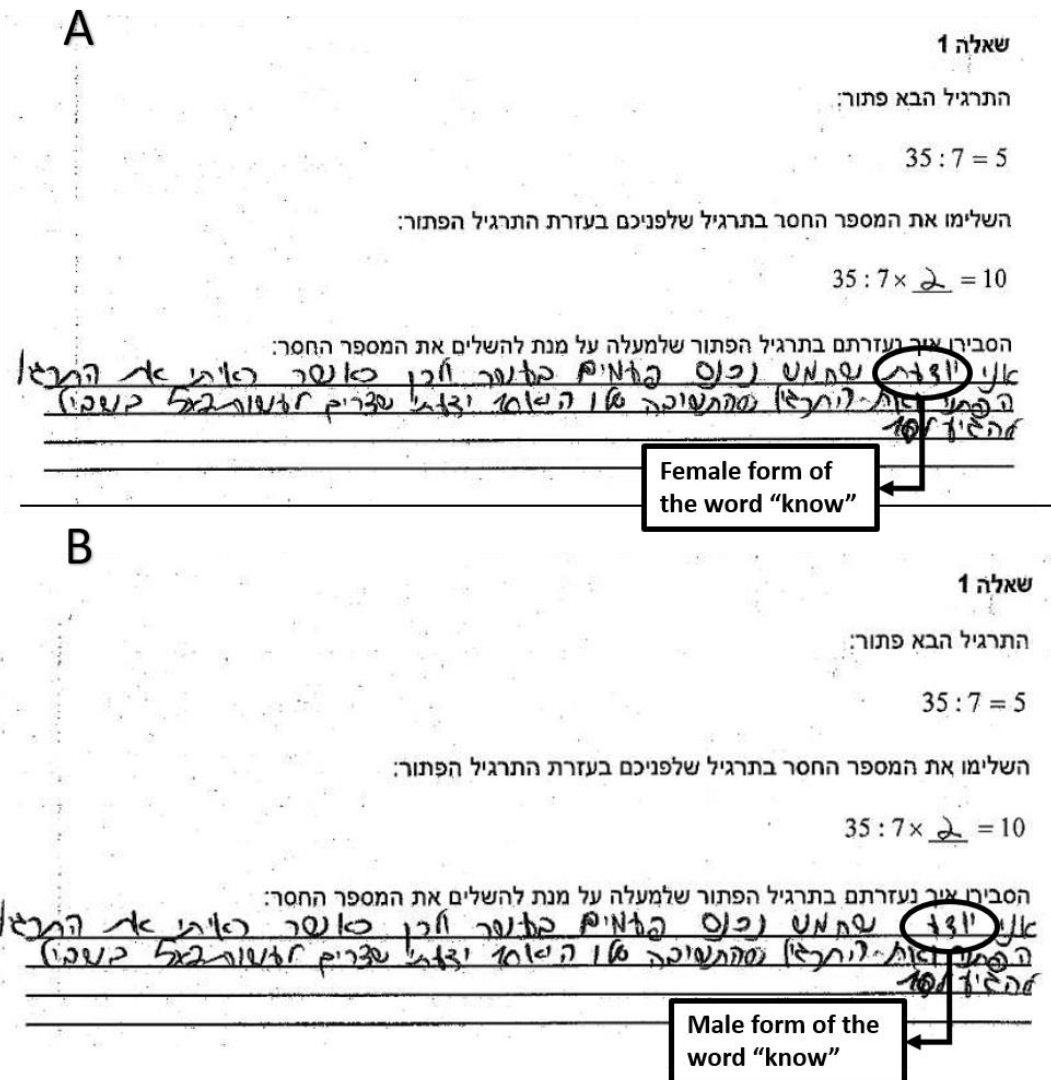


Fig. S2. Example of a student's written answer to question #1 of the exam paper, manipulated to appear in the female (A) and male (B) forms. The question reads: "Question 1 < break > The following exercise is solved: < break > $35:7=5$ < break > complete the missing number in the following exercise with the solved exercise: < break > $35:7 \times \underline{\quad} = 10$ < 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 5×2 to get to 10". The word "know" (circled in both sections) takes different grammatical forms for a male or 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 of the teachers and as a girl's paper to the other half.

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 event 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 sixth 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: eliana@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? (1=very little ... 5=very much)
- According to your assessment, what are the student's mathematical capabilities? (1=very low ability ... 5=very high ability)
- According to your assessment, to what extent is the student diligent? (1=very little ... 5=very much) (D) (R)
- According to your assessment, what are the student's chances of being successful in a top-level class? (1=very low chances ... 5=very high chance)
- Based on the exam that you have marked, in which class would you place this student? (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)

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

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

Table S2: The survey items

Teachers' descriptions of their 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 or 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? (1=effort does a lot more influential than talent ... 5=talent does a lot more influential than effort)
- 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 (0=highly disagree ... 5=highly agree) (D) (R)
- Most children make an effort to succeed in math and the main reason for differences in performance is due to natural talent (0=highly disagree ... 5=highly agree) (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.

^aSince 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.

Table S2: The survey items – continued

Boy-Math Stereotype:

Who is better at math: boys or girls? (1=girls are a lot more than boys ... 5=boys are a lot more than girls)

Who receives higher grades in math: boys or girls? (1=girls do a lot more than boys ... 5=boys do a lot more than girls)

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 to 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 appropriate for the humanities
- Men are more suitable for science, and women are more appropriate for the humanities
- There is no relationship between gender and the suitability for science

What is your opinion regarding the suitability of men and women for science and the humanities? (1=men are a lot more suitable for science and women are a lot more suitable for the humanities ... 7=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 how 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 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. Yet, they 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? (1=strong association between men and the exact sciences and between women and the humanities” ... 7=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? (1=girls do a lot more than boys ... 5=boys do a lot more than girls) (R)

Math Teachers Invest More in Boys or Girls

Do most math teachers invest more effort in advancing and encouraging girls or boys? (1=girls a lot more than boys ... 5=boys a lot more than girls)

Gender Essentialism

- Men and women are naturally different from each other in their ability, preferences, and character (1=highly disagree ... 5=highly agree)
- Men and women tend to think about solving mathematical problems differently (1=highly disagree ... 5=highly agree)
- The differences between men and women's preferences and abilities are mostly the result of biological differences between the sexes (1=highly disagree ... 5=highly agree)
- The differences in men and women's preferences and abilities are mostly the result of social circumstances: education, how they are treated, etc. (1=highly disagree ... 5=highly agree) (R) (D)

Feminism

- I see myself as a feminist. (1=highly disagree ... 5=highly agree)
- I support the feminist movement and its goals (1=highly disagree ... 5=highly agree)

Day-to-Day Exposure to Feminist Discourse

To what extent are you exposed to discourses about feminism and women's empowerment in your daily life? (1=very little ... 5=a lot)

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?

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

Table S2: The survey items – continued

Demographics - continued:

- What is your level of education?
- What were your fields of study at university or college?
- 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 you 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 at 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?

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 (1=highly disagree ... 5=highly agree)
- Some children are a lost cause when it comes to math (1=highly disagree ... 5=highly agree)
- Children who do not get homework support from their parents will find it hard to succeed in math (1=highly disagree ... 5=highly agree)
- Succeeding in math increases one's earning capacity (1=highly disagree ... 5=highly agree)
- It is important to encourage children to study math because succeeding in this field increases one's earning capacity (1=highly disagree ... 5=highly agree)
- Children do better academically when studying in homogenous groups with small gaps between children in the group (1=highly disagree ... 5=highly agree)
- In lower grades, the gaps between children are smaller, and they grow larger in higher grades (1=highly disagree ... 5=highly agree)
- Children succeed or fail in math primarily due to external circumstances such as family, socio-economic status, etc. (1=highly disagree ... 5=highly agree)
- It's possible to succeed in math without getting support from one's parents (1=highly disagree ... 5=highly agree)
- Most children like math (1=highly disagree ... 5=highly agree)
- Many children who like math are embarrassed to admit it (1=highly disagree ... 5=highly agree)
- Children who are good at math get positive reinforcement from their peer group (1=highly disagree ... 5=highly agree)
- In which of the following should more time and resources be invested: promoting weak students or encouraging excellence in math? (1=invest in promoting weak students a lot more than encouraging excellence ... 5=invest in encouraging excellence a lot more than in promoting weak students)
- Do most teachers invest more effort in promoting weak students or in encouraging excellence in math? (1=invest in promoting weak students a lot more than encouraging excellence ... 5=invest in encouraging excellence a lot more than in promoting weak students)

Table S3. Logistic Regression Coefficients Indicating the Effects of Potential and Outcome Categories on Gender of Mentioned Student

| Predictor | Estimates | SE | OR | Z | Wald Test | | |
|---------------|-----------|------|------|-------|-----------|-----|--------|
| | | | | | 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 S4. 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. R² comparisons are always with the preceding model (to the left).

| Group | Predictor | Model 1 | | | Model 2 | | | Model 3 | | |
|---------------------------|---|---------|-------|-------|---------|-------|-------|--------------|-------------|--------------|
| | | 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 |
| Implicit Variables | Gender Essentialism | | | | -0.04 | -0.32 | 0.752 | -0.10 | -0.86 | 0.394 |
| | 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 |
| Model Statistics | Total number of stereotypical and counter-stereotypical Gender-Brilliance Association Statements (control variable) | | | | | | | -0.03 | -0.22 | 0.828 |
| | 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

Table S5: 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 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 "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 book chain!

The event will be held at 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 see 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 S5: 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, convenient for each teacher (it's online). Every participant will receive a modest gift.

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

Following 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 would 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, Ph.D. student

Table S5: 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 of 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 and the teachers’ standpoints about various educational and social subjects. We hope that this study will add to the existing knowledge in the scientific area of math pedagogy, 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 teachers who teach or have taught math in grades five and six. The data collected from the teachers will include these actions:

1. The teachers will assess and give feedback to a number of anonymous maths exam papers.
2. The teachers will answer a questionnaire about their views on math pedagogy and 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:

-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 license has been delivered to the Inspector General of Math and attached to this file 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 where they can receive a detailed explanation of the study's plan.

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 continuing study.

Sincerely,

Eliana Avitzour

Primary Researcher

Table S5: 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)
