AMERICAN PSYCHOLOGICAL ASSOCIATION

Developmental Psychology

ISSN: 0012-1649

2021, Vol. 57, No. 6, 940–950
https://doi.org/10.1037/dev0001186

Stereotype Threat Effects on Italian Girls' Mathematics Performance: A Failure to Replicate

Franca Agnoli, Francesca Melchiorre, Claudio Zandonella Callegher, and Gianmarco Altoè
Department of Developmental Psychology and Socialization, University of Padova

Many studies have found that males, on average, perform better than females in mathematics, although the size of this gender gap is small and varies considerably across countries. Stereotype threat has been proposed as a principal cause of this gender gap. From this perspective, females' performance is affected by fear of confirming a negative stereotype about females' mathematical ability and this stereotype can be activated by an experimental manipulation that reminds females of the stereotype. Yet, evidence of a stereotype threat effect on mathematics performance in childhood and adolescence has been mixed. The present study replicated a highly cited study of stereotype threat among Italian adolescents with a much larger sample of Italian ninth grade (89 male, 75 female, mean age = 14.2) and eleventh grade (84 male, 80 female, mean age = 16.2) public high school students. Performance in tests administered both before and after the experimental manipulations were analyzed with a series of logistic mixed-effects models. Model comparisons confirmed that males performed better than females, but the probability of a stereotype threat effect was infinitesimal. We conclude that Italian adolescent gender differences in mathematics may not be explained by stereotype threat effects.

Keywords: gender differences, mathematics performance, model comparison, replication, stereotype threat

Stereotyping mathematics as a male domain is very common in many countries, and mathematics-gender stereotypes are well known by elementary school children. Cvencek et al. (2011) found that the cultural stereotype of boys being better than girls in mathematics is present in the United States as early as second grade, a result found with both explicit and implicit measures of stereotyping. In Italy, using implicit measures of stereotyping, Passolunghi et al. (2014) found that third, fifth, and eighth grade girls associate mathematics with the male gender. Furthermore, Nosek et al. (2009) conducted a large cross-national study with more than 500,000 participants from 34 countries and found that more than 70% of men and women associated male with science and female with liberal arts more easily than the reverse as measured by the gender-science Implicit Association Test (IAT; see Greenwald & Banaji, 1995). In addition, the strength of nation-level stereotypes (measured implicitly) predicted nation-level gender differences in eighth grade mathematics achievement (Nosek et al., 2009). This finding of a relationship between stereotypes and performance raises the question of whether and how a negative stereotype about women in mathematics impairs their performance.

Stereotype threat theory postulates a situational decrement in a person's performance owing to the awareness that his or her own ingroup is considered to be less skillful in the domain in which he or she is going to be tested (Spencer et al., 2016; Steele et al., 2002). It provides an explanation for situational underperformance in the presence of a negative stereotype. Stereotype threat has been proposed as an explanation for the gender gap in mathematics tests (Spencer et al., 1999). Spencer et al. showed that women from a highly selected sample (i.e., college students very good in mathematics, aware of their own good mathematics ability, and who considered good mathematics performance important) underperformed on difficult mathematics tests when the negative gender stereotype was activated. They showed that the gender difference in performance could be eliminated when stereotype threat was lowered by describing the test as not producing gender differences.

Other evidence that stereotype threat can impair mathematics performance was reported by Shih et al. (1999), who manipulated the salience of different identities in a sample of female Asian American college students by activating either the person's gender or ethnic identity prior to a mathematical test. Performance declined when the female identity had been made salient and was enhanced when the Asian identity had been made salient, in comparison to the performance of a control group for whom no particular identity was salient. For additional evidence of stereotype threat effects on adult mathematics performance, see Davies et al. (2002).

Studying the emergence of stereotype threat effects on mathematics performance among children and adolescents is important

Franca Agnoli https://orcid.org/0000-0001-7940-4080
Claudio Zandonella Callegher https://orcid.org/0000-0001-7721-6318
Gianmarco Altoè https://orcid.org/0000-0003-1154-9528

The data set, supplemental materials (including detailed analyses and analysis code), and test booklets (including instructions, photographs, and mathematics tests) are available via the Open Science Framework and can be accessed at https://doi.org/10.17605/OSF.IO/HZ5G7.

Correspondence concerning this article should be addressed to Franca Agnoli, Department of Developmental Psychology and Socialization, University of Padova, Via Venezia 8, Padova, 35131, Italy. Email: franca.agnoli@unipd.it

for at least three reasons: (a) to determine when, developmentally, the stereotype threat effect emerges; (b) to study the effect in a natural setting such as a classroom (adult studies are often performed in a lab); and (c) to identify potential moderators of the gender stereotype threat effect. A meta-analysis based on 47 comparisons of mathematical performance of girls (ranging in age from 6 to 17.5 years) in either a stereotype threat condition or a control condition found a small average standardized mean difference equal to −.22 in the expected direction (Flore & Wicherts, 2015). Because most of the studies included in this meta-analysis were published, however, the authors considered whether this small effect was inflated by a publication bias. They used several methods to find evidence for publication bias (Rothstein et al., 2005) and used funnel plot asymmetry to correct for the effect of this bias. The revised estimated stereotype threat effect size was only −.07, a very weak effect.

Flore and Wicherts (2015) also noted in their meta-analysis that small imprecise study samples showed larger negative effects and studies with large samples found small effects. Most published studies testing the stereotype threat model with children and adolescents have very small sample sizes, and because of the publication bias for statistically significant results, these published studies may be exaggerating a weak effect. This may explain why studies with larger samples have failed to replicate reported studies with imprecise samples (Maxwell et al., 2015; Szucs & Ioannidis, 2017). Indeed, concerns have also been raised about the robustness and replicability of adult research on stereotype threat, which is often based on small samples of participants (Stoet & Geary, 2012).

Consistent with the conclusions of the meta-analysis by Flore and Wicherts (2015), which raised doubts about whether this stereotype threat effect occurs in childhood and adolescence, developmental studies with large sample sizes have found little or no evidence of the effect. Ganley et al. (2013) conducted three large studies in the United States with children and adolescents (ranging in age from 9–18; $N = 931$) using a range of implicit and explicit stereotype activation methods. They found better mathematics performance for boys than girls, but they did not find any evidence of stereotype threat influence on the mathematics performance of school-age girls. Ganley et al. (2013) concluded that claims of a stereotype threat affecting the performance of girls on mathematics tests may reflect a publication bias, in accordance with the conclusion of Flore and Wicherts (2015).

Flore et al. (2018) conducted a large study ($N = 2{,}064$) of the influence of gender stereotype threat on the mathematics performance of Dutch second-year high school students, typically 13 to 14 years of age. In addition to its large sample size, this study had several notable strengths. The experimental paradigm included an explicit stereotype threat manipulation (Spencer et al., 1999) and a control condition in which the negative stereotype was nullified. The explicit manipulation was a statement that boys and girls do not perform equally well on the test, and the control manipulation was a statement that they do perform equally well. Other strengths included the following:

1. The study was preregistered. Preregistration avoids some of the methodological problems that may be responsible for inconsistent results in this area (for discussion, see Simmons et al., 2011).
2. An a priori power analysis was conducted. As Button and Munafò (2017; p. 23) explain, low-powered studies decrease the frequency of true statistically significant findings and consequently increase the proportion of findings that are false positives.
3. A multilevel statistical analysis avoided the use of covariates. Studies of adults have used some measure of mathematics performance as a covariate when assessing the effect of a stereotype threat manipulation, but this analytic approach is inconsistent with the assumptions of covariate analyses (Wicherts, 2005; Wicherts et al., 2005).

Flore et al. (2018) attempted to maximize the stereotype threat effect by selecting high achieving students, including boys, and employing a difficult mathematical test. The data showed no evidence of a stereotype threat effect on the mathematical test performance of girls in Dutch high schools, a conclusion reinforced by Bayesian analyses that showed strong evidence in favor of the null hypothesis of no stereotype threat effect.

The meta-analysis of Flore and Wicherts (2015) and the results of both Ganley et al. (2013) and Flore et al. (2018) provide reason to doubt the existence of stereotype threat effects on the mathematical performance of young and adolescent girls. If there is an effect, these data suggest that it is very small. It is, of course, possible that stronger stereotype threat effects would be found in some cultural settings. Indeed, Flore et al. (2018) suggest that stereotype threat may not influence Dutch adolescents. Because gender stereotypes and their effects on young women may depend on local cultural factors and may change over time, it is extremely important to replicate studies that obtained statistically significant effects. The discrepancy of these results with previous results obtained with 12- to 16-year-old students in different countries (Delgado & Prieto, 2008; Huguet & Regner, 2007; Huguet et al., 2009; Keller & Dauenheimer, 2003; Muzzatti & Agnoli, 2007) suggests the need for replications in different social/cultural settings where the sensitivity to the negative stereotype could be stronger compared with the Dutch society. The latest Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development [OECD, 2019] found no significant gender difference in the mathematics performance of 15-year-old Dutch adolescents, and it is feasible that the gender stereotype has weakened and does not affect their mathematical performance anymore.

In contrast to the results of Ganley et al. (2013), Muzzatti and Agnoli (2007) found that the performance of 13-year-old Italian girls was consistent with a stereotype threat effect, but the performance of younger girls showed no evidence of this effect. In two experiments they investigated stereotype threat susceptibility in Italian children from ages 7 to 13 years using an implicit manipulation. In the experimental condition, children viewed information about the frequencies of male and female famous mathematicians (i.e., men have been overrepresented in the realm of extraordinary achievement in mathematics). A vignette presented to the children portrayed 10 famous mathematicians, including nine men and only one woman. This implicit manipulation reminded participants that mathematics is stereotyped as a male domain, thereby making salient both the mathematics

domain and female underrepresentation in mathematics. The striking male overrepresentation in extraordinary mathematics achievement was expected to make the mathematics gender stereotype salient and thereby depress girls' mathematics performance. These experiments found no stereotype threat effect from second through fifth grade (from 7 to 10 years of age). Only the eighth grade (mean age = 13.0) girls showed a decrement in performance due to the stereotype threat manipulation.

We decided that a replication of Muzzatti and Agnoli (2007) was warranted. Their article has been cited more than 200 times according to Google Scholar, often as evidence for the negative effect of stereotype threat on girls' mathematical performance. A sizable gender gap in the mathematics performance of Italian adolescents has persisted over decades, which could, in part, be attributable to effects of stereotype threat. Italian 15-year-old boys' performance in the PISA mathematics tests was 17 points higher than girls in 2006 (OECD, 2007; p. 232), 20 points higher in 2015 (OECD, 2016; p. 199), and 16 points higher in 2018 (OECD, 2019; p. 146).

In replicating Muzzatti and Agnoli (2007), we sought to take advantage of the strengths of their methodology while correcting some weaknesses. Their experiments had four characteristics of a well-controlled study of stereotype threat:

1. Both boys and girls participated. Stereotype threat manipulations are expected to affect girls' performance and boys serve as a control group.

2. Participants were randomly assigned to an experimental condition or a control condition.

3. The mathematics problems were difficult, especially in the second experiment in which the mean performance of experimental groups ranged from 16 to 41% correct. More difficult problems are considered more likely to result in a larger stereotype threat effect (Delgado & Prieto, 2008; O'Brien & Crandall, 2003; Steele, 2010).

4. The mathematics test was administered in class, avoiding the limitations of a lab study.

The experiments reported by Muzzatti and Agnoli (2007) share weaknesses with many other studies of the stereotype threat effect in children and adolescents; these experiments had small sample sizes and used analysis of variance with covariates in the statistical analyses. Muzzatti and Agnoli (2007) found a significant stereotype threat effect only for their oldest participants (eighth grade), and this group had the smallest sample size (60 students divided into four groups).

Another problem is the statistical analysis chosen to test the stereotype threat effect. Muzzatti and Agnoli (2007) used performance on a pretest as a covariate to assess stereotype threat effects, and almost all studies of stereotype threat effects, both with adults and children, have used analysis of variance and covariates (for an exception, see Flore et al., 2018). Both ANOVA and ANCOVA are inappropriate statistical methods for analyzing stereotype threat effects. One fundamental problem is that the data are categorical (correct, incorrect, or missing responses to mathematics problems), and these methods are not appropriate for categorical data analysis (Agresti, 2002) despite their widespread use in psychological research. Mathematical performance is typically measured as proportion correct, and its variance is a function of proportion correct, violating the equal-variance assumption of these analysis methods. As Jaeger (2008) demonstrates, these analysis methods can yield statistically significant interactions (the primary evidence of a stereotype threat effect) when there is no actual interaction between the factors (see also Gelman, 2015; Flore, 2018; pp. 132–135).

There are additional problems with the way ANCOVA has been used to analyze stereotype threat effects. Measures of mathematical ability have often been used as covariates when analyzing stereotype threat effects, despite the problems with covariate corrections that Wicherts (2005) and others have highlighted. A fundamental assumption of ANCOVA is that the covariate is independent of the experimental effect, but group differences, such as a difference in the mathematical ability of boys and girls, violate that assumption. The variance explained by gender and the variance explained by mathematics ability cannot be separated, and spurious effects can arise. As Wicherts (2005) observed, "stereotype threat theory *explicitly* predicts violations of practically all assumptions underlying ANCOVA."

The present study[1] replicates the method, materials, and procedure of Experiment 2 in Muzzatti and Agnoli (2007), with modifications to the participants, the mathematical problems, and the data analysis methods.

Participants in the two experiments of Muzzatti and Agnoli ranged in age from 7 to 13 years, and a significant stereotype threat effect was found only for the oldest children, suggesting that stereotype threat originates during early adolescence. To be confident that participants were old enough to exhibit the effect, we selected 14- and 16-year-old adolescents. In the Italian school system, adolescents of this age are enrolled in high school. After finishing middle school (where the curriculum is standardized nationwide), Italian students select a secondary level school that offers the educational track they want to pursue. We sampled students enrolled in *Liceo Scientifico*, which is among the primary college-preparatory high schools. In 2015, the year this research was conducted, 22% of all Italian students entering high school selected a *Liceo Scientifico*, and 42% of these students were female (Ministero dell'Istruzione, dell'Università e della Ricerca, 2015). These girls should be more likely than girls in other types of high school to exhibit the effect because theory predicts that stereotype threat will only undermine mathematics test performance for girls who consider the subject of mathematics to be important to them (Keller, 2007; Spencer et al., 1999).

Because these participants were high school students studying mathematics, we could not use the same mathematics problems solved by the 13-year-old students who had shown a stereotype threat effect in Muzzatti and Agnoli (2007). As in Muzzatti and Agnoli, we selected challenging mathematics problems appropriate for each age group because difficult items have been found to produce stronger stereotype threat effects (Campbell & Collaer, 2009; O'Brien & Crandall, 2003; Spencer et al., 1999; Wicherts et al., 2005). Difficult problems were selected from study books for the first and third years of *Liceo Scientifico* and from the INVALSI

[1] This study was not preregistered.

standardized tests[2] for these two age groups (INVALSI, 2011). With the exception of the participants and the mathematics problems, this replication study followed the same method as in the second experiment of Muzzatti and Agnoli (2007).

We analyzed the data using mixed-effects logistic models. Logistic models describe the odds of each response instead of the proportion, and unlike proportions, odds are not bounded between 0 and 1. Mixed-effects logistic models allow including subjects and items as random effects. Using these models, we can take into account the variability attributable to subjects in both the pretest and the posttest and the variability attributable to the mathematics problems. Accounting for the variance owing to these random effects should increase the sensitivity of the critical test for stereotype threat effects, which is the interaction between gender, time of test (pre/post), and condition.

Method

Participants

Three hundred twenty-eight high school students (155 females and 173 males) from northeastern Italy participated in the study. Participants attended either the ninth or eleventh grade of two public college-preparatory, science-oriented high schools (first and third year of *Liceo Scientifico* in Italy). The first group (ninth graders) included 164 students (75 females and 89 males) with a mean age of 14.2 years. The second group (eleventh graders) included 164 students (80 females and 84 males) with a mean age of 16.2 years. Our sampling strategy was to obtain at least double the number of 13-year-old participants in Experiment 2 of Muzzatti and Agnoli (60 students), and we sampled all available students at these two high schools.

Design and Manipulation

Photographs

The design and manipulation were the same as in Experiment 2 of Muzzatti and Agnoli (2007). Participants were randomly assigned to the experimental (stereotype threat) or control condition. In the experimental condition, the mathematics gender stereotype was made salient by a task designed to remind participants of the female underrepresentation in mathematics. Students viewed an array of photographs of ten famous mathematicians, including nine males and only one female. Participants were asked to count the number of male and female mathematicians, calculate the proportion of each gender, and construct a graphical representation of this numerical relationship. This simple task was intended to ensure that participants were actually aware of the difference in frequencies of male and female mathematicians represented in the photographs. Participants in the experimental condition were 39 females and 45 males in the ninth grade and 42 females and 41 males in the eleventh grade.

In the control condition the task portrayed an array of ten neutral photographs including nine flowers and one fruit. Participants performed exactly the same task, but this time they were counting and calculating the proportion of flowers and fruit. Participants in the control condition were 36 females and 44 males in the ninth grade and 38 females and 43 males in the eleventh grade.

Mathematical Tests

Two mathematical tests, a pretest and a posttest, were constructed for each grade. Each test was composed of 18 problems adapted from the INVALSI standardized tests (INVALSI, 2011) and taken from study books for these two grades. Mathematical problems were selected concerning properties of numbers, arithmetic, and algebra from tests appropriate for each grade. There were no problems concerned with geometry, trigonometry, or other advanced topics. Each problem had multiple-choice answers (one correct of four alternatives). The pretests and posttests were constructed as parallel versions, with corresponding problems that required the same skills and knowledge. Two versions of the pretest and posttest were constructed for each grade that differed only in the order of the problems. The problem order in one version was the opposite of the order in the other version. These two problem orders were randomly assigned to participants.

During the posttest, the salience of the experimental manipulation was maintained with a message at the bottom of each page. For the control group this message said, "What was portrayed in the page you read? (a) 1 flower and 9 fruits, (b) 9 flowers and 1 fruit, or (c) some vegetables." For the experimental group this message said, "Who was portrayed in the page you read? (a) 1 male mathematician and 9 female mathematicians, (b) 9 male mathematicians and 1 female mathematician, or (c) famous writers." As a final check that participants remained aware of the manipulation, all participants were asked after the posttest to describe what the 10 images they had seen represented and in what proportions. All responses to this question were correct.

After completing both the pretest and posttest, participants were required to evaluate their own performance (*How well do you think you performed in this test?*) on a Likert scale from 1 (*Not well at all*) to 5 (*Very well*) and judge the difficulty of the tests (*How difficult do you rate this test?*) on a 5-point scale from 1 (*Not difficult at all*) to 5 (*Very difficult*).

All the materials, including instructions, photographs, and mathematics tests, are available on the Open Science Framework (see Agnoli et al., 2021) at https://doi.org/10.17605/OSF.IO/HZ5G7.

Procedure

There were four steps in recruiting participants and obtaining informed consent. First, a letter sent to the principals of the two high schools described the research plan and requested participation from their schools. Second, the principals presented the research plan to their school boards, who approved the plan. Third, the parents of all potential participants were asked to provide written informed consent in a letter that described the research and explained that the results would all be confidential in accordance with the Italian privacy law. Fourth, students were assured that all results were collected anonymously and their participation would remain confidential. Students were free to interrupt their participation at any time without any consequences. Approval was not requested from the University of Padova Institutional Review

[2] INVALSI is the National Institute of Educational System Evaluation in Italy. All Italian students must take grade-appropriate written tests designed to evaluate their learning levels in Italian language and mathematics.

Board because we followed the laws and regulations governing research conducted in Italian high schools as described above.

The study was conducted in classrooms by two female experimenters who instructed the participants and timed the tests. Booklets were prepared in advance containing all the materials for either the control or stereotype threat condition, and the order of the booklets was randomized. Students were seated in their usual assigned seats in the same configuration that is used during in-class exams, in which copying or looking at other students' exams is forbidden. The two experimenters distributed the randomly ordered booklets to the seated students and collected the booklets at the end of the experiment.

Total time for the experiment was 45 minutes. First, participants were allowed 20 minutes to complete the 18 problems of the pretest. Second, during the five-minute experimental manipulation, participants looked at the pictures with 9-to-1 ratios of flowers and fruit (control condition) or male and female mathematicians (experimental condition), and they answered questions about the number and proportion of these pictures. Third, participants were allowed 20 minutes to complete the 18 problems of the posttest.

Results

The raw data and supplemental materials (including details of all statistical analyses with analysis code and graphical representation of results) can be accessed at https://doi.org/10.17605/OSF.IO/HZ5G7.

Missing Responses in Pretest Performance

Each participant's response to each problem could be correct, wrong, or simply missing, and analyzing only the correct responses may overlook differences in wrong or missing responses (to our knowledge, stereotype threat theory makes no predictions about whether performance differences will be due to wrong or missing responses). To examine gender differences in the likelihood of responding to all problems, we analyzed missing responses in the pretest. Overall, 54% of ninth graders and 37% of eleventh graders responded to all problems on the pretest and consequently had 0 missing answers. Figure 1 in the OSF supplemental materials shows the distributions of missing responses for males and females in both grades.

A logistic regression model was computed to test the role of gender and grade in predicting the likelihood of responding to all problems in the pretests. Males were more likely than females to respond to all problems in the ninth grade (64% vs. 41%) and in the eleventh grade (43% vs. 32%), as confirmed by the model (B = .93, Z = 2.88, p = .004, 95% CI [.30, 1.57]). Ninth graders responded to all problems more than the eleventh graders (B = .86, Z = 2.77, p = .006, 95% CI [.26, 1.48]). The interaction between grade and gender was not significant. Furthermore, the odds of missing one or more responses were more than twice as large for females than for males in both grades (*OR* for gender = 2.53, 95% CI [1.35, 4.81]). Details of the logistic regression model and the odds ratios can be found in Tables 4 and 5, respectively, of the OSF supplemental materials.

In the analyses that follow, we consider missing responses to be incorrect, assuming that missing responses occur when a participant does not know how to solve the problem.

Gender Differences in Mathematics Performance in the Pretest

Table 1 presents mean accuracy and standard deviations for mathematical performance by gender and grade for just the pretests.

To take into account variability attributable to subjects and mathematical problems, we used a mixed-effects logistic regression model (see Pinheiro & Bates, 2000) in which both gender and grade are treated as fixed effects. This modeling method can also weigh the contributions of multiple random effects to the overall variance, and both participants (164 ninth graders + 164 eleventh graders = 328 levels) and mathematical problems (18 + 18 = 36 levels) are treated as random effects. Accuracy in the pretest was the dependent dichotomous variable with two levels: 0 (incorrect or missing response) and 1 (correct response), producing 5904 observations in total (328 participants × 18 problems). Data analyses were performed using free software R and the lme4 package (Bates et al., 2014; R Core Team, 2018). Table 2 reports the results of the full model, with predictors for fixed and random effects.

Males solved a higher proportion of problems compared with females in both the ninth and eleventh grade (.72 vs .66 and .61 vs .53), as confirmed by the model (B = −.40, Z = −2.41, p = .016, 95% CI [−.72, − .07]). The mathematical problems were more difficult for the eleventh graders than the ninth graders (B = −.62, Z = −1.88, p = .06, 95% CI [−1.27, .04]). The Gender × Grade interaction was not significant.

Table 3 presents the odds ratios for accuracy in the pretests. Using the odds ratios as estimated effect sizes, the odds of a correct answer for females is only 2/3 of the odds for males (*OR* = .67, 95% CI [.49, .93]). See the OSF supplemental materials for figures showing the distributions of responses, model estimates, distributions of the random effects of participants and items, and Cook's distance.

Models Testing Gender Stereotype Threat Effects

Table 4 presents the mean accuracy and standard deviations in the pretest and posttest for each gender, grade and condition (Control or Stereotype Threat). See the OSF supplemental materials for boxplots of mean accuracy in each condition.

Stereotype threat theory predicts that women (as well as adolescents of the age tested in this study) would show a decrement in mathematical performance when gender stereotypes are activated through an experimental manipulation. To evaluate the presence of the gender stereotype threat effect, we compared seven different mixed-effects logistic regression models. First, we describe the

Table 1
Mean Accuracy and Standard Deviations in the Pretest ($N_{subjects}$ = 328; N_{items} = 36; $N_{observations}$ = 5,904)

	Male			Female		
Grade	*n*	*M*	*SD*	*n*	*M*	*SD*
9th	89	0.72	0.18	75	0.66	0.17
11th	84	0.61	0.20	80	0.53	0.18

Table 2
Estimated Parameters for Accuracy in the Pretest ($N_{subjects}$ = 328; N_{items} = 36; $N_{observations}$ = 5,904)

Parameter			95% CI			
Name	Estimate	*SE*	Lower	Upper	*Z*	*p* value
Main effects						
Intercept	1.21	0.23	0.75	1.67	5.21	<.001
Gender (Female)	−0.40	0.17	−0.72	−0.07	−2.41	.016
Grade (11th)	−0.62	0.33	−1.27	0.04	−1.88	.060
Interaction effect						
Gender × Grade	−0.06	0.23	−0.51	0.40	−0.24	.810
Random effects						
Subjects (Intercept)	0.88		0.78	0.99		
Item (Intercept)	0.85		0.68	1.11		

Note. Baseline category for gender is male. Baseline category for grade is ninth. Confidence intervals were computed using the profile likelihood because the likelihood function is not symmetric in the case of random effects parameters.

seven models shown in Table 5, and then we compare the models using information criteria (see McElreath, 2016; pp. 188–205). In all models, the dependent variable was the response given by a participant (0 for wrong answer or missing response, 1 for correct response). All models account for the variability of both the participants and the mathematical problems by including these variables as random effects.

The simplest model (m0 in Table 5) has no fixed effects; it considers only the random effect of participants ($n_{subjects}$ = 328) and the random effect of problems (n_{items} = 72). This is used as a reference model to evaluate the potential contribution to performance of the independent variables in the experimental design. The next three models (m1, m2 and m3) evaluate, respectively, the potential contributions to performance of gender, gender plus grade, and gender plus grade and their interaction.

The next three models (m4, m5 and m6) evaluate the potential effect of stereotype activation on mathematical performance. Model m4 adds the interaction of condition (control/stereotype threat) and time (pre/post) to the fixed effects of gender and grade. Models m5 and m6 are central to testing the main hypothesis of stereotype threat theory. Model m5 tests the three-way interaction of condition, time, and gender, which has been reported as evidence of stereotype threat in previous studies that included both males and a control group in the experimental design. Model m5 is consistent with the hypothesis that threat due to stereotype activation differentially affects boys and girls. Finally, model m6 evaluates the four-way interaction between condition, time, gender, and grade. This could be the best model if the experimental manipulation affects girls differently in the two different age groups, implying a developmental change in the stereotype threat effect. Note that models m1 and m5 are the most relevant in the analysis as they support, respectively, the presence of gender differences in mathematics and the presence of a stereotype threat that differentially affects girls and boys.

Table 3
Odds Ratios for Accuracy in the Pretest ($N_{subjects}$ = 328; N_{items} = 36; $N_{observations}$ = 5,904)

		95% CI	
Parameter	*OR*	Lower	Upper
Main effects			
Intercept	3.35	2.12	5.31
Gender (Female)	0.67	0.49	0.93
Grade (11th)	0.54	0.28	1.04
Interaction effect			
Gender × Grade	0.94	0.60	1.49

Note. *OR* = odds ratio. Baseline category for gender is male. Baseline category for grade is ninth.

Increasing the number of model parameters (as in the sequence of models shown in Table 5) improves the fit to the data but risks overfitting the data. Model comparison is a method for selecting the most plausible statistical model, given the data, among a set of candidate models (McElreath, 2016). Table 6 presents the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and weights for both criteria for the seven models. Information criteria estimate the average deviance (i.e., error) of a model's ability to predict new data, and thus lower criterion values are evidence of a better model (Wagenmakers & Farrell, 2004). The criterion weights can be interpreted as an estimated probability that each model will perform best on future data (assuming this experiment was exactly replicated). Evaluating models using information criteria permits a trade-off between parsimony and goodness-of-fit (Vandekerckhove et al., 2015).

As Table 6 shows, model m1, with gender as the only fixed effect, has the lowest information criteria and consequently is the most likely model given the data and the set of models considered. The weights indicate that model m1 has an estimated probability of 63% (AIC weight) or 78% (BIC weight) of performing better than all the other six models, reflecting the fact that male participants performed better than female participants on these mathematical problems. The goodness-of-fit of model m1 can be evaluated using Conditional R^2, which was .36 and estimates the variance explained by the gender fixed effect and two random factors of the model, as explained in the OSF supplemental materials. Table 7 shows the estimated parameters of the mixed-effects logistic regression for model m1, indicating that male participants performed better than female participants ($B = -.33$, $Z = -3.50$, $p < .001$, 95% CI [−.51, −.14]). The odds ratio for the gender effect of model m1 is .72, 95% CI [.60, .87].

Table 6 shows that model m2, with gender and grade as fixed effects, had the second lowest AIC, with an estimated probability

Table 4
Mean Accuracy and Standard Deviations in the Pretests and Posttests ($N_{subjects}$ = 328; N_{items} = 72; $N_{observations}$ = 11,808)

		Males					Females				
			Pretest		Posttest			Pretest		Posttest	
Condition	Grade	*n*	*M*	*SD*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*
Control	9th	44	0.72	0.18	0.62	0.14	36	0.71	0.15	0.58	0.14
	11th	43	0.62	0.21	0.67	0.16	38	0.49	0.18	0.61	0.15
ST	9th	45	0.73	0.19	0.59	0.12	39	0.61	0.17	0.54	0.14
	11th	41	0.59	0.20	0.67	0.17	42	0.56	0.18	0.65	0.15

Note. ST = stereotype threat.

of 24% of performing better than all other five models. Model m0, with no fixed effects but only random effects, had the second lowest BIC, with an estimated probability of 21% of performing better than the other models. According to stereotype threat theory, model m5 should have been the best model, but the information criteria were large for this model. The estimated probability that model m5 would perform better than the other models is infinitesimal for both information criteria.

In these analyses of pretest and posttest performance, missing responses were considered errors. In the OSF supplemental materials we consider how the results and conclusions would be affected if missing responses were excluded, again concluding that the stereotype threat model m5 is extremely unlikely to perform better than the other models.

Overall, the model comparisons presented in Table 6 give support to the hypothesis that there are gender differences in mathematics in this study favoring males, both in the first and third year of a science-oriented high school. There was no support, however, for the hypothesis that activation of a gender stereotype depresses mathematical performance of 14- and 16-year-old girls.

Evaluating the Full Model

Although model m6 was among the least likely models in the model comparison, it deserves additional consideration as a full model of the effects of condition, time (pre/post), gender, grade, and their interactions. The three-way interaction between time, condition, and gender represents the stereotype threat effect, which was captured in model m5, and model m6 adds the possibility that the effect might vary with age. An analysis of model m6 (presented in detail in the OSF supplemental materials) determined that the four-way interaction of time, condition, gender, and age was significant. Because this interaction was significant, we performed two planned comparisons to estimate the stereotype threat effects, one for ninth grade female participants and one for eleventh grade female participants, contrasting the difference between posttest and pretest performance in the stereotype threat condition with the difference in the control condition. Both estimated stereotype threat effects were nonsignificant (Z = 1.53, p = .25 for ninth grade female participants and Z = −.70, p = .97 for eleventh grade female participants).

Power Analysis

No power analysis was conducted before collecting and analyzing the data, but we can evaluate the inferential risk related to our study in a retrospective design analysis (Gelman & Carlin, 2014), which defines the plausible effect size in accordance with findings reported in the literature. To determine the sensitivity of our analyses to a stereotype threat effect, we conducted a series of simulations using the simr package (Green & MacLeod, 2016). These simulations (presented in detail in the OSF supplemental materials) indicate that with a small effect size of d = −.25, the statistical power associated with our study is about .72, and with the effect size reported by Flore and Wicherts (2015) it is about .60. For a moderate effect size with d = −.50, the power of our study is almost 1.

Discussion

One reason that stereotype threat theory deserves study is that it offers a potential explanation for gender differences in mathematics

Table 5
Specifications of Mixed-Effects Logistic Regression Models of Pretest and Posttest Performance

Model		Predictors	
Name	Type	Fixed effects	Random effects
m0	Baseline		Subject, Problem
m1	Additive	Gender	Subject, Problem
m2	Additive	Gender + Grade	Subject, Problem
m3	2-way interaction	Gender × Grade	Subject, Problem
m4	2-way interaction	Condition × Time + Gender + Grade	Subject, Problem
m5	3-way interaction	Condition × Time × Gender + Grade	Subject, Problem
m6	4-way interaction	Condition × Time × Gender × Grade	Subject, Problem

Note. $n_{subjects}$ = 328, n_{items} = 72, $n_{observations}$ = 11,808. Models that include interactions also include the main effects of the interaction terms.

Table 6
Model Comparison Using AIC and BIC

Model	*df*	AIC	AIC weights	BIC	BIC weights
m0	3	13,099.85	0.00	13,121.98	0.21
m1	4	13,089.84	0.63	13,119.34	0.78
m2	5	13,091.78	0.24	13,128.66	0.01
m3	6	13,093.71	0.09	13,137.97	0.00
m4	8	13,097.42	0.01	13,156.44	0.00
m5	11	13,099.78	0.00	13,180.92	0.00
m6	18	13,097.89	0.01	13,230.67	0.00

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion. $n_{subjects}$ = 328; n_{items} = 72; $n_{observations}$ = 11,808.

performance. Averaging across all participating countries in the OECD, PISA has consistently found that, among 15-year-olds, girls outperform boys in reading and, to a lesser extent, boys outperform girls in mathematics (OECD, 2019). There is, however, considerable variability in overall performance and in these gender gaps across countries. The gender gap in mathematics is found both in the United States, where boys outperformed girls by 9 points in the mathematics test of the latest PISA study, and in Italy where boys outperformed girls by 16 points. Only three countries had a larger mathematics gender gap than Italy in the 2018 PISA test. Furthermore, standardized tests of all Italian children in school years 2, 5, 6, 8, and 10 (ages 7 to 15) reveal that boys outperform girls at every age, and the mathematics gender gap increases with age (Contini et al., 2017).

Striking differences in the gender gap in mathematics across nations suggest that this gender gap emerges from the influence of cultural-societal norms, and the increase in gender gap with age may be due to a growing influence of these norms throughout development. Gender stereotypes are an element of cultural-societal norms that may contribute to gender gaps in performance. These include stereotypes regarding how a male or female person should behave and stereotypes regarding people with different roles in society. According to stereotype threat theory, activating a stereotype that defines boys as better than girls in mathematics causes a decrement in girls' performance, especially in the context of tests that are critical for academic and career advancement.

Muzzatti and Agnoli (2007) investigated the development of a stereotype threat effect in mathematics in children ranging in age from 7 to 13 years, finding no evidence of a gender stereotype threat effect for children ranging in age from 7 to 10 years but a statistically significant effect for 13-year-old children. Their results are often cited as evidence for the development of a gender stereotype threat effect. However, later studies with larger sample sizes in both the United States and the Netherlands found no evidence of a stereotype threat effect on the mathematics performance of children and adolescents ranging in age from 8 to 17.5 years (see Ganley et al., 2013; Flore et al., 2018). The authors of these studies suggested that reports of gender stereotype threat effects among children and adolescents are overrepresented in the literature because of a publication bias that favors studies with statistically significant findings over those that find no evidence of an effect.

There is a clear conflict between the finding of Muzzatti and Agnoli (2007) of a gender stereotype threat effect in the mathematics performance of 13-year-old adolescents and more recent findings of no effect in the United States and the Netherlands in girls of about the same age and with the same or more explicit stereotype activation methods. Possibly this conflict is explained by geographic differences in the gender gap and in gender stereotypes; Italy has a larger gender gap than the United States and the Netherlands. We concluded that a replication of Muzzatti and Agnoli (2007) was needed to investigate whether the effect they found persists and is replicable in Italy.

Although we used the same experimental design and methods as in Muzzatti and Agnoli (2007), we revised some aspects of the research. Our participants were adolescents 14 and 16 years old, and consequently more likely to be vulnerable to a stereotype threat. Further, they were from a college preparatory high school where performance in mathematics is highly valued. Instead of analyzing performance with an ANCOVA with pretest performance as a covariate, we computed and compared seven mixed-

Table 7
Estimated Parameters of Mixed-Effects Logistic Regression Model m1 ($N_{subjects}$ = 328; N_{items} = 72; $N_{observations}$ = 11,808)

Parameter			95% CI			
Name	Estimate	*SE*	Lower	Upper	*Z*	*p* value
Main effects						
Intercept	0.82	0.15	0.52	1.11	5.46	<.001
Gender (Female)	−0.33	0.09	−0.51	−0.14	−3.50	<.001
Random effects						
Subjects (Intercept)	0.74		0.67	0.83		
Item (Intercept)	1.14		0.97	1.37		

Note. Baseline category for gender is male.

effects logistic regression models. Most studies of effects on mathematical performance have analyzed each participant's mean performance across all problems, but our analyses took into consideration random variability due to both participants and items (i.e., mathematical problems). As Gelman and Carlin (2014) observed, "psychology research involves particular challenges because it is common to study effects whose magnitudes are unclear (e.g., consider the literature on priming and stereotype threat) in a context of large uncontrolled variation (especially in between-subjects designs) and small sample sizes. The combination of high variation and small sample sizes in the literature imply that published effect-size estimates may often be overestimated" (p. 684; see also Gelman, 2015).

We compared a series of seven plausible models (McElreath, 2016) of increasing complexity. In the simplest model, variability in mathematical performance is considered to be attributable only to variability in the random effects of participants and items. Subsequent models added the fixed effects of gender, grade, and their interaction. More complex models added the condition by time interaction. According to stereotype threat theory, this interaction should be a determinant of mathematical performance, because girls' performance in the posttest should be impaired only if they are in the stereotype threat condition, but the models with this interaction term were found to be very unlikely. The most plausible model attributes variability in mathematical performance to only the gender effect and variability in the random effects. The gender effect is congruent with the statistical difference found in the Italian PISA results for mathematics.

Our finding that gender stereotype threat played no role in the performance of adolescents aligns with the results of Ganley et al. (2013) and Flore et al. (2018). Similarly, Wille et al. (2018) found no stereotype threat effect in a large sample of German fifth-grade children.

What can explain our current finding of no gender stereotype threat effect given the finding by Muzzatti and Agnoli (2007) of a significant stereotype threat effect for Italian eighth graders using the same experimental methods? We consider some possible explanations.

Some researchers have suggested that people are not all equally affected by stereotype threat manipulations. For example, some have argued that female participants with a higher level of domain identification (i.e., a greater interest in mathematics) should find the manipulation more threatening and show the effect more strongly (Keller, 2007; Steele, 1997). The problem with this explanation is that the participants of this study were all enrolled in *Liceo Scientifico,* which Italian students select after middle school knowing that mathematics and science will be central to their studies for five years of high school. Students in the third year should be highly committed because they would have had opportunities to transfer to a different high school less focused on mathematics.

Some researchers have suggested that all mathematical problems are not equally effective at revealing the stereotype threat effect, and some (Beilock et al., 2007; Keller, 2007; Nguyen & Ryan, 2008; Spencer et al., 1999) claim that stereotype threat has a detrimental effect on mathematical performance only when the task is particularly difficult. The mathematical problems we used appear, however, to have been difficult for our participants. The mean percentage of correct answers ranged from 49% (eleventh grade females in the pretest control condition) to 73% (ninth grade males in the pretest stereotype threat condition). As noted by Flore et al. (2018), this range of performance reflects a realistic classroom testing situation.

Another possibility is that adolescents of the ages included in our research have changed in some ways that diminished the role of stereotype threat since the studies were conducted by Muzzatti and Agnoli (2007). The stereotype that boys are better than girls in mathematics may be less prominent or less threatening. The data indicate, however, that the boys outperformed the girls in our sample, and Italian boys outperform Italian girls in the PISA test of mathematics. Note that Flore et al. (2018) also found a significant gender gap, with Dutch boys outscoring Dutch girls, but found no evidence of stereotype threat effect.

If a stereotype threat effect exists, we should not expect to observe it unless the experimental manipulation successfully activates the stereotype. We used an implicit method to activate the stereotype that repeatedly reminded all participants in the experimental condition that the ratio of male to female famous mathematicians is 9 to 1. Ganley et al. (2013) used a similar manipulation as well as more direct manipulations and did not find evidence of a stereotype threat effect with any manipulation. The manipulation we used, however, was apparently effective for the oldest participants in the study by Muzzatti and Agnoli (2007).

Gender differences in mathematics continue to be larger in Italy than in most other OECD countries. Widely held stereotypes about gender differences in Italy may contribute in some manner to these differences, but a reminder in a classroom setting that mathematicians of extraordinary achievement are more often men than women did not cause a substantial decrement in performance. Apparently, being reminded that males outperform females in mathematics does not impair adolescent female performance in classrooms. We should seek other evidence of a causal relationship between the sociocultural environment, including stereotypes, and gender differences in mathematics performance.

Despite the clear importance of replication in science, replications have not been highly regarded in the social sciences. As Chambers (2017) noted "the academic culture in psychology places little emphasis in repeating the experimental methods of other psychologists." We believe that given the recent presence of null results for stereotype threat effects on females' mathematical performance, both in adults (e.g., Finnigan & Corker, 2016; Pennington et al., 2019) and adolescents (e.g., Flore et al., 2018), replications would be the best instrument to test the reliability of stereotype threat studies. The generalizability of stereotype threat effects could be tested in different settings (lab and/or school settings) and different cultures (different nations and, within a nation, different levels of status for women/girls).

References

Agnoli, F., Melchiorre, F., Zandonella Callegher, C., & Altoè, G. (2021). *Stereotype threat effects on Italian girls' mathematics performance: A failure to replicate* [Data set and supplemental materials]. https://doi.org/10.17605/OSF.IO/HZ5G7

Agresti, A. (2002). *Categorical data analysis*. Wiley.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. https://arxiv.org/abs/1406.5823

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal*

of Experimental Psychology: General, *136*(2), 256–276. https://doi.org/10.1037/0096-3445.136.2.256

Button, K. S., & Munafò, M. R. (2017). Powering reproducible research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 22–33). Wiley Blackwell.

Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly*, *33*(4), 437–444. https://doi.org/10.1111/j.1471-6402.2009.01521.x

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Contini, D., Di Tommaso, M. L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, *58*, 32–42. https://doi.org/10.1016/j.econedurev.2017.03.001

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, *28*(12), 1615–1628. https://doi.org/10.1177/014616702237644

Delgado, A. R., & Prieto, G. (2008). Stereotype threat as validity threat: The anxiety–sex– threat interaction. *Intelligence*, *36*(6), 635–640. https://doi.org/10.1016/j.intell.2008.01.008

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality*, *63*, 36–43. https://doi.org/10.1016/j.jrp.2016.05.009

Flore, P. C. (2018). *Stereotype threat and differential item functioning: A critical assessment* (Doctoral dissertation). Tilburg University. https://research.tilburguniversity.edu/files/23445144/Flore_Stereotype_7_3_2018.pdf

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, *49*(10), 1886–1897. https://doi.org/10.1037/a0031412

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, *41*(2), 632–643. https://doi.org/10.1177/0149206314525208

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295x.102.1.4

Huguet, P., Dumas, F., Marsh, H., Wheeler, L., Seaton, M., Nezlek, J., Suls, J., & Régner, I. (2009). Clarifying the role of social comparison in the big-fish–little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

Huguet, P., & Regner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology*, *99*(3), 545–560. https://doi.org/10.1037/0022-0663.99.3.545

INVALSI. (2011). *Le competenze in scienze, lettura e matematica degli studenti quindicenni: Rapporto nazionale PISA 2009* [Competencies in science, reading and mathematics of fifteen-year-old students: National report PISA 2009]. https://www.invalsi.it/invalsi/ri/Pisa2009/documenti/RAPPORTO_PISA_2009.pdf

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *The British Journal of Educational Psychology*, *77*(Pt. 2), 323–338. https://doi.org/10.1348/000709906XI

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, *29*(3), 371–381. https://doi.org/10.1177/0146167202250218

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Ministero dell'Istruzione, dell'Università e della Ricerca. (2015, May). https://www.istruzione.it/allegati/2015/Focus_iscrizioni_as2015_2016_pubblicazione.pdf

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology*, *43*(3), 747–759. https://doi.org/10.1037/0012-1649.43.3.747

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

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*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*(6), 782–789. https://doi.org/10.1177/0146167203029006010

OECD (2007). *PISA 2006: Science competencies for tomorrow's world* (Vol. 2: Data). PISA, OECD Publishing, Paris. http://www.oecd.org/pisa/pisaproducts/39703566.pdf

OECD (2016). *PISA 2015 results: Excellence and equity in education* (Vol. I). PISA, OECD Publishing, Paris. https://doi.org/10.1787/9789264266490-en

OECD (2019). *PISA 2018 results (Volume II): Where all students can succeed*. PISA, OECD Publishing, Paris. https://doi.org/10.1787/19963777

Passolunghi, M. C., Ferreira, T. I. R., & Tomasetto, C. (2014). Math–gender stereotypes and math-related beliefs in childhood and early adolescence. *Learning and Individual Differences*, *34*, 70–76. https://doi.org/10.1016/j.lindif.2014.05.005

Pennington, C. R., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European*

Journal of Social Psychology, *49*(4), 717–734. https://doi.org/10.1002/ejsp.2540

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

R Core Team. (2018). R: A language and environment for statistical computing. https://www.R-project.org/

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). Publication bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 1–7). Wiley.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*(1), 80–83. https://doi.org/10.1111/1467-9280.00111

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steele, C. M. (2010). *Whistling Vivaldi: And other clues to how stereotypes affect us*. WW Norton & Company.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, *34*, 379–440.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102. https://doi.org/10.1037/a0026617

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, Z. Wang, J. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist*, *60*(3), 267–269. https://doi.org/10.1037/0003-066X.60.3.267

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. https://doi.org/10.1037/0022-3514.89.5.696

Wille, E., Gaspard, H., Trautwein, U., Oschatz, K., Scheiter, K., & Nagengast, B. (2018). Gender stereotypes in a children's television program: Effects on girls' and boys' stereotype endorsement, math performance, motivational dispositions, and attitudes. *Frontiers in Psychology*, *9*, 2435. https://doi.org/10.3389/fpsyg.2018.02435

Received October 19, 2019
Revision received February 16, 2021
Accepted March 5, 2021 ■