Review of General Psychology

Can Stereotype Threat Explain the Gender Gap in Mathematics Performance and Achievement?

Gijsbert Stoet and David C. Geary

Online First Publication, January 16, 2012. doi: 10.1037/a0026617

Review of General Psychology

1089-2680/12/$12.00 DOI: 10.1037/a0026617

Can Stereotype Threat Explain the Gender Gap in Mathematics Performance and Achievement?

Gijsbert Stoet
University of Leeds, United Kingdom

David C. Geary
University of Missouri

Men and women score similarly in most areas of mathematics, but a gap favoring men is consistently found at the high end of performance. One explanation for this gap, stereotype threat, was first proposed by Spencer, Steele, and Quinn (1999) and has received much attention. We discuss merits and shortcomings of this study and review replication attempts. Only 55% of the articles with experimental designs that could have replicated the original results did so. But half of these were confounded by statistical adjustment of preexisting mathematics exam scores. Of the unconfounded experiments, only 30% replicated the original. A meta-analysis of these effects confirmed that only the group of studies with adjusted mathematics scores displayed the stereotype threat effect. We conclude that although stereotype threat may affect some women, the existing state of knowledge does not support the current level of enthusiasm for this as a mechanism underlying the gender gap in mathematics. We argue there are many reasons to close this gap, and that too much weight on the stereotype explanation may hamper research and implementation of effective interventions.

Keywords: stereotype threat, gender gap, mathematics performance

The issue of gender differences in mathematics achievement and performance continues to capture much attention within and beyond academia, especially with respect to the greater number of men than women at the high end of the continuum (Ceci & Williams, 2010; Halpern et al., 2007; Hyde & Mertz, 2009). Much of the conjecture and debate in this area is because mathematics is a gateway to employment in well-paying and prestigious science, technology, engineering, and mathematics professions (STEM) (Ceci & Williams, 2007; Ceci, Williams, & Barnett, 2009; Halpern et al., 2007). One explanation for the underrepresentation of women at the high end of mathematics achievement and accomplishment, stereotype threat, has received considerable attention and is often described or implied as the cause of these differences in textbooks (e.g., Myers, 2008), in books for the general public (e.g., Fine, 2010), in professional periodicals (e.g., Goodall, 2010), and in other media (e.g., Erbé, 2008).

We review the strength of the associated evidence, specifically articles that have provided a potential replication of the original critical study that demonstrated women's performance on mathematics tests is lower in contexts that trigger the stereotype that "men are better in mathematics" compared with a context in which this stereotype is not explicitly or implicitly primed (Spencer et al., 1999). The critical test of this research is the interaction between gender and context (threat vs. nonthreat); that is, the performance difference between men and women is smaller or nonexistent in the nonthreat condition than in the threat condition, in which women underperform relative to men.

Before discussing these tests and the associated evidence for stereotype threat and women's performance in mathematics, we provide a short overview of recent empirical research on actual gender differences in mathematics achievement and theoretical accounts.

Gender Differences in Mathematics Performance

One cannot discuss explanations of gender differences in mathematics before having reviewed the actual empirical evidence for any such differences; do they really exist? There are numerous reviews and large-scale studies of gender differences in mathematics and its various subdomains (Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Lindberg, Hyde, Petersen, & Linn, 2010; Penner, 2003; Strand, Deary, & Smith, 2006). Despite this wealth of data, a consensus has yet to emerge with respect to mean differences and differences in variance. In a recent meta-analysis of studies published between 1990 and 2007, Lindberg et al. (2010) concluded there were no gender differences in mean performance and nearly equal variability within each gender (see also Hyde & Mertz, 2009). Their conclusions were based on averaging effect sizes over age, mathematical content, nationality, and other variables. Such averaging may be misleading, as the differences are smaller (0.07 *SD* advantage for boys and men) in lower ability than higher ability samples (0.40 *SD* difference), and across grade level (e.g., 0.15 *SD* female advantage in preschool and 0.23 male

Gijsbert Stoet, Institute of Psychological Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom; David C. Geary, Department of Psychological Sciences, University of Missouri, Columbia.

We thank Jeff Rouder, Ben Winegard, Jon Oxford, and Drew Bailey for insightful comments and discussion. We thank Kimmo Eriksson, Chad Forbes, Jason Lawrence, Torun Lindholm, Lauri O'Brien, and Toni Schmader for providing data necessary for the meta-analysis, and the anonymous reviewers for helpful comments.

Correspondence concerning this article should be addressed to Gijsbert Stoet, Institute of Psychological Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom; or David C. Geary, Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211-2500. E-mail: g.stoet@leeds.ac.uk

advantage in high school), as just two examples. Another recent publication, based on analyses of the kindergarten to 5th grade performance of 20,000 U.S. children, concluded that a roughly 0.20 *SD* advantage emerged by 3rd grade, favoring boys in every mathematical domain they assessed and across all socioeconomic groups (Fryer & Levitt, 2010). By the end of 5th grade, the ratio of boys to girls in the top 5% of performance was 3:1. Another recent meta-analysis of 30 years of SAT and ACT scores from national talent searches confirmed a very large 13:1 ratio of middle school boys to girls at the highest levels of performance (>700 on the SAT) in the early 1980s (Benbow & Stanley, 1980, 1983), which declined to roughly 4:1 by 1991 and has remained at that level since then (Wai, Cacchio, Putallaz, & Makel, 2010).

In summary, we have one new large-scale longitudinal study and two large meta-analyses on this issue published in the same year, each reaching different conclusions. The final word on the existence of gender differences in mean mathematics scores and in variability in these scores has yet to be uttered. At this point, our interpretation of this literature is that there are likely small mean differences, typically favoring boys but sometimes favoring girls depending on grade and mathematical content, and critical differences in the ratio of boys to girls and men to women at higher levels of performance (Geary, 1996).

Explanations of Gender Differences and the Stereotype Threat Hypothesis

Irrespective of mathematics performance, there is little debate about the existence of gender differences in a variety of cognitive domains (Chrisler & McCreary, 2010; Geary, 2010; Halpern, 2000). Debate continues, however, over their theoretical and practical implications (e.g., Hyde, 2005). Although psychologists routinely agree that most complex abilities, personality dispositions, or behaviors result from some combination of biological and social mechanisms, debate on the gender differences in mathematics has seemed to bifurcate into two extreme positions, or at least it is presented as such in the media and by some scientists. These positions are in regard to the origins of these gender differences, and at one end include proximate biological mechanisms, such as prenatal exposure to sex hormones, sometimes placed in a wider evolutionary context (Geary, 2010) and sometimes not (Hines, 2010), and at the other mechanisms that are largely or entirely socially focused (Hyde & Mertz, 2009). Stereotype threat is an example of the latter category.

Stereotype threat was first proposed by Steele and Aronson (1995) and refers to the conscious or unconscious belief that one belongs to a group stereotypically known for specific performance deficits. The hypothesis is that these beliefs lead to suboptimal performance on tasks specific to the stereotype. The mechanism has been applied to the study of performance gaps of various groups, in particular ethnic minorities and women. Following the publication of the influential article by Spencer et al. (1999) stereotype threat is a frequently used explanation for gender differences in mathematics performance. The article, from here on referred to as "the original", has received considerable attention. For example, it was cited 452 times, which is 10 times more than the average number of citations of the other 25 articles in the same volume of the *Journal of Experimental Social Psychology*.[1] But the attention reached far beyond academia. The stereotype threat explanation of the gender gap in mathematics is regularly mentioned in the press, and it is frequently assumed that this is the only factor that can explain it. In fact, the current review was directly triggered by an article in "*Times Higher Education*" (Goodall, 2010), a popular British magazine for and written by professionals working in higher education. In this it was stated that "The psychologist Claude Steele has shown that girls perform less well in math tests if they are told beforehand that on average males outperform females." This is a mischaracterization of Steele's original claims (later, we address issues regarding the degree to which academic researchers carry responsibility for the misrepresentation of stereotype threat research; see also Sackett, Hardison, and Cullen, 2004a).

The current article does not aim to review evidence for the stereotype threat hypothesis in general, especially because there are various reviews on this topic (Kit, Tuokka, & Mateer, 2008; Nguyen & Ryan, 2008; Sackett et al., 2004a, Smith, 2004; Wheeler & Petty, 2001). Our review focuses on whether stereotype threat is a viable explanation of the gender gap in mathematics. We believe that this explanation deserves critical review, because it is so often uncritically accepted as the explanation of the earlier noted gender differences in mathematics both inside and outside academia; the gender gap in mathematics achievement is of enormous theoretical and practical significance, and the debate that the president of Harvard University in 2005, Professor Lawrence Summers, ignited is a good and often cited example of that (Halpern et al., 2007).

We address two questions about the stereotype threat explanation of the gender gap in mathematics that we believe have not received sufficient attention. The first is whether the available empirical data really support the hypothesis. Our motivation for answering this question stems not only from the influence of the hypothesis in academia and in the popular press, but also from the fact that the hypothesis is quite extraordinary. There is the implication that stereotyped beliefs, even if implicitly primed by nonmathematical content (below) can (in some characterizations) have a large and pervasive effect on girl's and women's mathematical performance and long-term achievement. Furthermore, there is the proposal that this belief can be relatively easily overridden by simply stating that men and women perform equally or by presenting information on successful women before the mathematical task; these types of manipulations are often used for controls in these studies. Hence, beliefs not only provide a potential explanation for the gender difference but also a prospect for eliminating the difference. And with Carl Sagan's famous quote that "extraordinary claims require extraordinary evidence" in mind, we review the published data, methods, and inferential statistical tests to determine how extraordinary the evidence is: Do the data really support the claim?

Given the many citations and the important implications of the original article by Spencer et al. (1999), one would expect that the phenomenon has now been well studied and replicated. Therefore, we begin with the original claims as expressed by Spencer et al. (1999), and then review the research that has tested their critical gender by threat-condition interaction.

[1] In June 2011, as determined by the Web of Science, which is the world's largest database for scientific research articles.

The Original Article and Its Evidence

The original trio of studies that tested the hypothesis that stereotype threat contributes to the gender difference in mathematics performance was published in 1999 (Spencer et al., 1999). In the first experiment, 28 men and 28 women, all of whom had taken at least one calculus course, were administered easy (Quantitative section, Graduate Record Examination [GRE] items) and more difficult (Advanced GRE mathematics test) mathematics items. The results confirmed a common pattern in the literature; that is, no gender differences on comparatively easy mathematics tests but a male advantage on more difficult ones. The interpretation of this pattern is grounded in the assumption that the stereotype threat emanating from difficult problems is larger than that emanating from easy problems. It is important to note that the authors acknowledged there are other explanations of this pattern: A difficult test has more discriminative power than an easy one. Given that both explanations would have predicted exactly the same outcome, this experiment did not provide discriminative support for the stereotype threat hypothesis, but it was helpful in identifying a mathematics test that captures the gender difference.

The second experiment with 24 men and 30 women tested the effect of different threat conditions on mathematics performance. The test used was the difficult one of the first experiment, which resulted in a floor effect on the second of two halves of this experiment, and thus these items were not analyzed. One group of participants was told that the mathematics test showed gender differences in the past (without stating the direction of the effect), and the other group was told that the mathematics test did not show gender differences in the past. The critical interaction between threat condition and gender was significant. Specifically, among the four groups of participants (men and women in threat and no-threat conditions), the group of women who were told there were gender differences performed significantly worse than any of the other three groups. Women in the no-threat condition scored no differently than men.

The final experiment with 31 men and 36 women was similar to the second, but the threat was manipulated differently. In the no-threat condition, participants were told (as in Experiment 2) there was no gender difference in performance on the mathematics test they were about to take. In the threat condition, participants were not told anything about gender differences. The assumption was that women are by default experiencing a threat when solving difficult mathematics problems, as argued in the first experiment. As in the second experiment, the group of women in the threat condition performed worse than any of the other three groups.

The fascinating feature of this final experiment was that the group of women who were explicitly told that men and women perform equally on the mathematics test performed no differently than the men in the same condition, and significantly better than women who were told nothing (Experiment 3). In terms of statistical analysis, this effect was expressed by an interaction between gender and stereotype threat condition, whereby the difference between men and women in the threat condition was statistically significant, with no significant difference between men and women in the no-threat condition.

Few experiments are without confounds, and this was the case here: The research has several methodological shortcomings that need consideration given the weight these and follow-up studies have had in the scientific and popular literature.

First, the findings of the initial experiment could be largely explained by the higher discriminative power of the more difficult test. As the authors noted, the results of this experiment do not allow us to distinguish the stereotype threat hypothesis from the hypothesis that men have a real advantage on more difficult mathematics problems (Penner, 2003). Second, half of the data from Experiment 2 were discarded because of a floor effect; there was no gender difference because neither men nor women did well on these problems. A gender difference was found for the remaining half of the test, which suggests the test material had low split-half reliability (otherwise, no such dramatic differences between the two halves of the test could have been found). With low reliability, it is difficult to interpret results from the items that were analyzed. Finally, in Experiment 3, some participants were excluded for not making reasonable effort; that is, completing the 20 min test in 5 min. If it was reasonable to exclude these participants, then one could argue that more participants from the female group in the threat condition should have been excluded. Examination of their Figure 3, and assuming that the bars indicate *SD* (the article does not indicate), around four women in the threat condition group must have scored zero or lower; even when they appeared to make a reasonable effort, their score certainly suggested that they made no reasonable effort (i.e., why only take the duration but not the mathematics score as a criterion of reasonable effort?). Although we are not sure removing two participants out of 67 would alter the overall results, we wonder whether inclusion of up to four women with zero scores out of 17 or 18 in the threat group (again, the presented data make it difficult to determine cell size) would have changed the critical gender $\times$ threat condition interaction.

Despite some reservations, the findings were original, exciting, and potentially very important. The issue is whether the results provided extraordinary evidence for an extraordinary claim. Of course, it is at this point where scientific replication and extension become critical, as reported in the next section.

Support in Subsequent Articles

The Web of Science and Scopus databases were searched for replications of the effect reported by Spencer et al. (1999). It was assumed that if a study replicated the original, it would have cited it.[2] There was one minor problem with this assumption. Coincidentally, another research group (Brown & Josephs, 1999) had asked exactly the same question as the original and their paper was accepted for publication only 3 months after the original had been accepted. This second article has been cited far less than the original, and it appears that all except one study that replicate either one of these studies cites the original. Hence, we will include these two extra articles in our analyses.

The database Web of Science alone list 452 articles that cited the original at the time of writing this review (we used Scopus to check whether Web of Science missed any relevant articles). The original article has been cited for many different reasons. For example, it has been cited by studies generally related to stereotype

[2] The database interfaces both have a function to retrieve all the manuscripts that cite a selected manuscript.

threat or general studies about gender gaps in nonmathematical areas. Still, 141 articles were related to mathematics. We aimed to determine how many of these studies had a design that could be classified as replicating that of the original. We determined that there are five essential components of a replication of the original. The first is that both men and women were tested. Obviously, both genders are needed to replicate the original, and thus, the many experiments studying just one gender, typically female, cannot replicate the original gender by threat interaction (see below for a criticism of stereotype threat studies without a control group). Second, a mathematics test was used. Third, men and women were recruited irrespective of preexisting beliefs about the gender stereotypes.[3] Fourth, men and women were randomly assigned to two different stereotype threat conditions (as was the case in Experiment 2 and Experiment 3 of the original study).[4] Fifth, if the study would explicitly state that it cannot be considered a replication of the original, despite fulfilling the first four criteria, it would not be included either, which was the case for only one study (Schmader & Johns, 2003, p. 444).

A total of 23 studies met these criteria. Across the 20 studies with adult participants (3 were with participants under 18), there was little variation in experimental conditions (Table 1). Most studies ($n = 16$) used a mathematics test based on material typical of exams and admission tests (e.g., GRE or SAT), but some used alternative standardized mathematics tests or customized tests (e.g., testing how well participants could determine the correctness of equations). The participants in the different studies were also similar (typically undergraduate students, but sometimes neither age nor recruitment method were specified). None of the studies aimed to merely replicate the original, and many considered here are only one part of a larger study.

Table 1
Overview of Articles Replicating the Original (Spencer et al., 1999) in Adult Participants

Study	Effect	Adjusted
Josephs, Newman, Brown, & Beer (2003)	Yes	Yes
Brown & Josephs (1999)	Yes	Yes
Lesko & Corpus (2006)	Yes	Yes
Gonzales, Blanton, & Williams (2002)	Yes	Yes
Johns, Schmader, & Martens (2005)	Yes	Yes
Keller (2002)	Yes	Yes
Osborne (2007)	Yes	Yes
Marx & Roman (2002)	Yes	Yes
Wicherts, Dolan, & Hessen (2005)[a]	Yes	No
Smith & Postmes (2011)	Yes	No
Davies et al. (2002)	Yes	No
Inzlicht & Ben-Zeev (2000)	No	Yes
Schmader (2002)	No	Yes
Grand et al. (2011)	No	No
Eriksson & Lindholm (2007)	No	No
O'Brien & Crandall (2003)[b]	No	No
McIntyre, Paulson, & Lord (2003)	No	No
Good et al. (2008)	No	No
Forbes & Schmader (2010)	No	No
Lawrence & Charbonneau (2009)	No	No

Note. The column "Effect" indicates whether a significant ($p < .05$) interaction between stereotype threat and gender was found. The column "Adjusted" indicates whether scores were adjusted for a previous math score (e.g., SAT).
[a] This study only found an effect in the number series subtest. [b] The authors of this study concluded that there is a statistically significant effect, but based that on a one-sided *t* test only (O'Brien & Crandall, 2003, p. 786).

It was determined that a study replicated the original findings if the study (or any one of the experiments of the sequence of experiments) found a significant (alpha criterion = 5%) interaction between gender and stereotype threat, and when women performed significantly worse in the threat than in the no-threat condition compared with men. Based on this criterion, 11 of 20 articles (55%) replicated the original effect.

One difficulty in interpreting many of these analyses is that mathematics scores were adjusted in half the studies for previous mathematics performance, using a preexisting mathematics score (e.g., SAT) as a covariate.[5] The reasoning was that the adjustment creates a baseline score for each participant. As pointed out elsewhere (Halpern et al., 2007), it was hitherto unclear if the covariates were a problem:

> Sackett et al. (Sackett et al., 2004a; Sackett, Hardison, & Cullen, 2004b) also raised concerns about the use of covariates and other statistical procedures used to demonstrate stereotype threat. Steele and Aronson (2004) responded to the concerns raised by Sackett et al. by referring to the large number of studies that found evidence for stereotype threat and by pointing out that many of these studies do not rely on the use of covariates to demonstrate the effect. This exchange, which was published in one of psychology's leading journals, shows the disagreement over the concept of stereotype threat and its importance in real-life settings. No one has yet conducted a meta-analysis of these stereotype-threat studies, so the size of the effect is unknown, but some studies show large effects [. . .] It is also unknown how altering test scores by removing stereotype threat from the testing setting affects the validity of the scores in predicting grades or other indicators of success. (p. 34)

The essence of the problem is that the mathematics score is the outcome of interest, and adjusting for preexisting differences on this or a very similar outcome creates confounds (for an in depth treatment of this and other types of problems associated with covariates see Chapman & Miller, 2001). For example, if it is the case that only women suffer from stereotype threat in mathematics assessments, the adjustment of scores for preexisting mathematics assessment scores is different for the different experimental groups. That is a problem, because an important assumption of a covariate analysis is that the groups do not differ on the covariate. But that group difference is exactly what stereotype threat theory tries to explain! This is an irreconcilable difference between the theory and the statistical assumptions underlying covariate analysis.

[3] An example of a study that did not fulfill this criterion was one that specifically selected participants who had heard of a gender stereotype in regard to mathematics and women (Martens, Johns, Greenberg, & Schimel, 2006).

[4] Gender relevant stereotype threat manipulations can take many forms, such as introducing positive role models or varying the number of members of the opposite sex being present. We did not include a study by Grimm, Markman, Maddox, and Baldwin (2009), because we believe that giving points for good answers versus taking points for wrong answers in itself is not a gender-relevant stereotype-threat manipulation when compared to the manipulations in the other studies we selected.

[5] The original did not.

Given these problems, we believe that it would be better not to consider the studies using adjusted scores, especially because the original 1999 study did not use adjustment, probably exactly for these reasons. If such a strict but arguably reasonable and logical criterion is applied, however, only 3 out of 10 (30%) articles replicated the original results.

Finally, of the 23 articles that included studies with features that could replicate the original studies, three included 7- to 16-year-olds (Keller, 2007; Keller & Dauenheimer, 2003; Muzzatti & Agnoli, 2007). One study of 7- to 13-year-olds found a significant interaction between gender and stereotype threat, as predicted, but only for 13-year-old girls (Muzzatti & Agnoli, 2007). This is an interesting and potentially important result, because it might reflect the development of the stereotype and thus the potential for poor performance under threat conditions. But the remaining studies by the same research group sometimes found the predicted interaction in 16-year-olds (Keller & Dauenheimer, 2003), and sometimes not (Keller, 2007). Altogether, it appears there are not yet sufficient data to draw conclusions about developmental aspects of stereotype threat.

Meta-Analysis

A meta-analysis can, in principle, quantify some of the issues highlighted earlier. In analogy to the large meta-analysis on the effect of stereotype threat in stereotyped populations conducted by Nguyen and Ryan (2008), we decided to analyze the effect of stereotype threat on women in the aforementioned studies, and we used the standardized mean as effect size (as detailed in Viechtbauer, 2010, p. 7; Figure 1).[6] Our aim was not just to quantify the strength of the stereotype threat effect on women, but also to take into account the adjustment of mathematics scores as a moderator variable. This analysis was directly inspired by the aforementioned quote of the review of Halpern et al. (2007).

We calculated the model estimates using a random effects model ($k = 19$) with a restricted likelihood function (Viechtbauer, 2010). We found that for the adjusted data sets, there was a significant effect of stereotype threat on women's mathematics performance (estimated mean effect size ± 1 SEM; -0.61 ± 0.11, $p < .001$), but this was not the case for the unadjusted data sets (-0.17 ± 0.10, $p = .09$). In other words, the moderator variable "adjustment" played a role; the residual heterogeneity after including the moderator variable equals $\tau^2 = 0.038$ (± 0.035), $Q_{\text{residual}}(17) = 28.058$, $p = .04$, $Q_{\text{moderator}}(2) = 32.479$, $p < .001$ (compared to $\tau^2 = 0.075$ (± 0.047), $Q(18) = 43.095$, $p < .001$ without a moderator), which means that 49% of the residual heterogeneity can be explained by including this moderator.

Mischaracterization of the Role of Stereotype Threat in the Gender Gap in Mathematics Performance

The available evidence suggests some women's performance on mathematics tests can sometimes be negatively influenced by an implicit or explicit questioning of their mathematical competence, but the effect is not as robust as many seem to assume. This is in and of itself not a scientific problem, it simply means that we do not yet fully understand the intrapersonal (e.g., degree of identification with mathematics) and social mechanisms that produce the gender by threat interactions when they are found.

The issue is whether there has been a rush to judgment. It is possible that academic researchers understand that the phenomenon is unstable, with much left to be discovered, and it is reporters and other members of the popular press who have overinterpreted the scientific literature, and as a result mischaracterized the phenomenon as a well established cause of the gender differences in mathematics performance. No doubt this is indeed the case for many researchers but in many other cases there is undo optimism about the stability and generalizability of this phenomenon. An example of this idea can be found in a recent publication by Grand, Ryan, Schmitt, and Hmurovic (2011), which did not replicate the stereotype threat effect on women's mathematics performance and downplayed the importance of replication as follows: "However, Nguyen and Ryan (2008) emphasized that the important question for future research in addressing this issue is not whether the results of the theory can be replicated consistently, as meta-analytic evidence across multiple comparison groups have clearly demonstrated its robustnes (p. 22)." We do not share this enthusiasm, given the just described failures to replicate the original effect unambiguously.

We list just a few more examples of typical overconfident statements. Good, Aronson, and Harder (2008) state "It is well established that negative stereotypes can undermine women's performance on mathematics tests." Similarly, Eriksson and Lindholm (2007) note "It is well established that an emphasis on gender differences may have a negative effect on women's math performance in U.S.A., Germany and the Netherlands." Davies et al. (2002) wrote "Women in quantitative fields risk being personally reduced to negative stereotypes that allege a gender-based math inability. This situational predicament, termed stereotype threat, can undermine women's performance and aspirations in all quantitative domains." In summary, there is a mismatch between the strength of the stereotype threat explanation of the gender difference in some areas of mathematics and the way many researchers describe it in the abstracts of their scientific publications. This might well have contributed to the common misrepresentation of stereotype threat hypothesis in the popular media.

We think there are two possible reasons for the misrepresentation of the strength and robustness of the effect. On the one hand, we assume that there has simply been a cascading effect of researchers citing each other, rather than themselves critically reviewing evidence for the stereotype threat hypothesis. For example, if one influential author describes an effect as robust and stable, others might simply accept that as fact.

Another reason for the overenthusiastic support of the stereotype hypothesis is that many of the associated studies only assessed the presumably stigmatized group rather than including a control group. For example, Wraga, Helt, Jacobs, and Sullivan (2008) aimed to dispel the idea that women's underperformance in mathematics relative to men might be related to biological factors and did so using functional MRI (fMRI). Their study only included female participants and concludes that "By simply altering the context of a spatial reasoning task to create a more positive message, women's performance accuracy can be increased sub-

[6] We excluded Davies, Spencer, Quinn, and Gerhardstein (2002), because the article did not present sufficient data and the corresponding author could not provide us with the requested data.

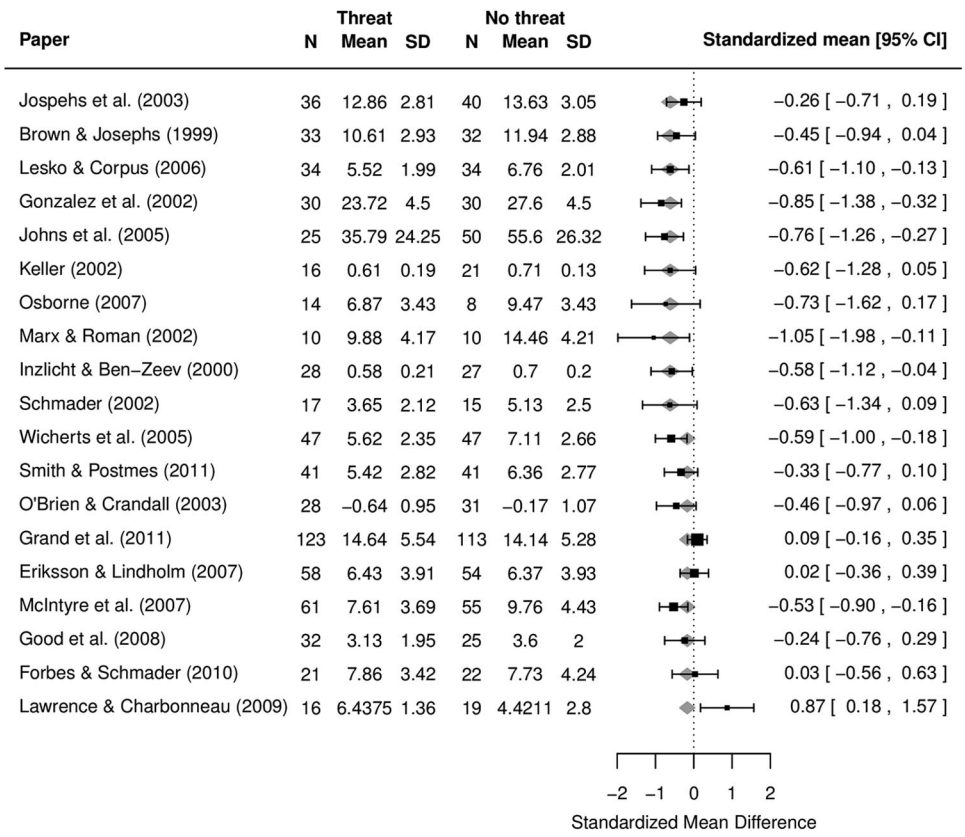

Figure 1. Forest plot of the 19 studies included in the meta-analysis. Of each study, the number of participants, the *M* and *SD* of women under threat and no threat are shown (*N* is often not reported for each cell of the design; in those cases, we have assumed equal numbers. Inzlicht & Ben-Zeev (2000) only report the total number of participants, but the number of male and female participants can be calculated based on a press release (Brown University News Service, 2000). The plot shows standardized means (negative values mean mathematics performance decrease because of threat), as well as confidence intervals (on the right). The size of each black square relates to the sample size of the study. Each gray rhombus (i.e., the ◇ symbol) indicates the standardized mean estimate for the group of adjusted (-0.61 ± 0.11) and nonadjusted studies (-0.17 ± 0.10). Altogether, only the group of studies with adjusted scores confirmed a statistically significant effect of stereotype threat.

stantially through greater neural efficiency." Krendl, Richeson, Kelley, and Heatherton (2008) carried out another fMRI study of stereotype threat in women only and concluded that "The present study sheds light on the specific neural mechanisms affected by stereotype threat, thereby providing considerable promise for developing methods to forestall its unintended consequences." The results of these studies are interesting and potentially important, but in the absence of a male control group it is difficult, if not logically impossible, to draw conclusions about gender differences in performance.[7] Some of these studies do not explicitly state that their findings tell us something about women (in comparison to men), but the focus on the gender of the female participants when describing the data, and the discussions within these studies about social policy imply that is exactly what the authors mean. At the very least, we cannot expect that the general public would understand this distinction.

It seems likely that the strong conclusions of these latter studies again lead to other studies claiming that there is strong support for women being disadvantaged by social stereotyping. For example, Derks, Inzlicht, and Kang (2008) cite the Krendl et al. (2008) study and amplify the stereotype threat hypothesis as follows: "The incremental benefit of the fMRI work here is in the ability to test a behavioral theory at the biological level of action, the brain. This can serve as a springboard for further theory-testing investigations. The fact that these results converge with the behavioral work of others provides consistency across different levels of analysis and organization, an important step toward the broad understanding of any complex phenomenon (p. 169)." And Stewart, Latu, Kawakami, and Myers (2010) cite the Krendl et al. (2008) article to support the following statement: "In 2005, Harvard President

[7] If we would accept that a study merely with female participants would reveal something unique about women, one could make the same argument for any other group category unique to all participants. The mistake of lacking a control group becomes clearer if one would conclude that a study with people who all wear cloths says something unique about people wearing clothes. We can only draw such conclusions by including a control group (i.e., men in studies that aim to draw conclusions about women in comparison to men, or naked people in studies that aims to draw conclusions about people wearing clothes).

Lawrence Summers publicly suggested that innate gender differences were probably the primary reason for women's underrepresentation in math and science domains. His remarks caused a stir in academic and nonacademic communities and are at odds with considerable research suggesting that women's underperformance in math and science is linked to situational factors (p. 221)." The problem here is that although the neuroscientific studies are interesting and important in and of themselves, they do not inform us about whether women are at a disadvantage *in comparison to men* on the tasks used in these studies. This is critical if we are to fully assess the stereotype threat explanation of the gender gap in mathematics performance.

Finally, we also felt that there were some potential problems with the presentation and interpretation of data. There was often an incomplete description of results (e.g., only figures with no reported *M*s or *SD*s), alpha values were relaxed when it matched the hypothesis, and the analyses of different representations of the same data, such as number correct and percent correct. Granted, it can be reasonable to explore different dependent measures, but it was sometimes the case that significant or marginally significant effects (e.g., percentage correct) were reported in the text and nonsignificant effects (of the same data) in a footnote. Moreover, there was no consistency in the dependent measure of choice, except that the significant one was highlighted in the text and abstract and the nonsignificant one placed in a footnote.

Conclusions and Outlook

We started our review with an overview of research on gender differences in mathematics performance and achievement. Based on the various large surveys on this topic, it seems reasonable to conclude that at least in the higher levels of performance, male mathematical achievers appear to outnumber female mathematical achievers. This is not only reflected in mathematics exams, but also in the number of jobs related to mathematics held by men and, for example, the prestigious Fields Medal for mathematical achievement, which has been won by men only since it was first awarded in 1936. While few researchers will deny that there are gender differences in mathematics achievement, the really interesting question is what factors contribute to these differences, especially given that it will be impossible to close the gender gap without understanding these factors.

This article reviewed evidence for the stereotype threat explanation of gender differences in performance, favoring boys and men, on difficult mathematics tests (gender differences are not typically found on comparatively easy tests; Penner, 2003). The question was whether the published research provides strong and stable evidence for the stereotype threat hypothesis as the primary causal explanation of this gender difference. Even when assuming that all failures to replicate have been reported, we can only conclude that evidence for the stereotype threat explanation of the gender difference in mathematics performance is weak at best, as less than half of the studies from which clear and unconfounded conclusions can be drawn did not show such an effect.

We also discussed the extent to which existing literature has amplified the stereotype threat hypothesis such that uncritical reading of the literature would lead one to conclude, as many have, that the hypothesis is strongly supported. Given the many enthusiastic statements about the stereotype threat effect, one of the most surprising findings of our review was that there were only 21 studies (including the original) that compared mathematics performance of men and women who were randomly assigned to threat conditions. This seems to be quite a contrast to larger reviews, such as by Nguyen and Ryan (2008). We identified three main reasons for the difference. First, their review was very general, whereas ours focused on the stereotype threat explanation of the gender gap in mathematics performance only. Thus, the many articles that were about other groups that might be affected by stereotype threat were not included. Second, we only included studies that had a male control group. Third, we only included published studies. We believe that this is reasonable, because it is difficult to determine the scientific credibility of unpublished data. Furthermore, we do not think that a possible file drawer effect, which is the likelihood of missing articles that have not been published, would change our conclusion. More likely than not, unpublished studies would have found no differences between experimental conditions, although we can only speculate about this.

We think that a more critical reading and interpretation of results is a necessity. After all, the aforementioned types of flaws in scientific work will ultimately damage the reputation of the whole field of psychology as a science, as was recently addressed in a *New York Times* article by Carey (2011) about forgery and unchallenged data massage:

> The scandal, involving about a decade of work, is the latest in a string of embarrassments in a field that critics and statisticians say badly needs to overhaul how it treats research results. In recent years, psychologists have reported a raft of findings on race biases, brain imaging and even ESP that have not stood up to scrutiny. Outright fraud may be rare, these experts say, but they contend that Dr. Stapel took advantage of a system that allows researchers to operate in near secrecy and massage data to find what they want to find, without much fear of being challenged.
>
> "The big problem is that the culture is such that researchers spin their work in a way that tells a prettier story than what they really found," said Jonathan Schooler, a psychologist at the University of California, Santa Barbara. "It's almost like everyone is on steroids, and to compete you have to take steroids as well." (p. A3)

We are hopeful that our review can help to counterbalance the mischaracterization of stereotype threat and the common belief that this threat has a strong causal effect on women's performance, which may influence policy making and popular beliefs in ways that, ironically, perpetuate the gender difference (Geary, 1996). When policymakers believe that achievement differences in mathematics can be overcome by simply reducing stereotypical beliefs (as the literature suggests), they might not be willing to invest in the study of other potential contributing factors and thus will not pursue solutions for these factors.[8]

But what if other factors, such as the gender difference in three-dimensional spatial cognition or interest in nonsocial domains, contribute to the gap (Baron-Cohen, Knickmeyer, & Belmonte, 2005; Ceci & Williams, 2010; Ceci et al., 2009; Farrell, 2005; Ferriman, Lubinski, & Benbow, 2009; Geary, 1996; Lubin-

[8] Educationalists indeed argue that interventions based on the belief of students rather than on teaching methods and materials should be treated cautiously (Yeager & Walton, 2011).

ski, Benbow, & Sanders, 1993)? In this situation, the focus on stereotype threat will ensure that other potential contributing factors are ignored or downplayed and thus efforts will not be invested in better understanding them or interventions to change them. For example, instructing women on how to use spatial diagrams to setup complex multistep mathematical word problems (which many men do without instruction) can reduce women's disadvantage on these problems (Johnson, 1984). Downplaying the relation between spatial abilities and performance in some areas of mathematics and corresponding gender differences (e.g., Casey, Nuttall, & Benbow, 1995) ensures that interventions to address these differences will not be developed, and absence of such interventions ensures the gap will remain.

Finally, given the importance and influence of the stereotype threat hypothesis, it is essential that researchers conduct studies aimed to replicate the original findings, and more fully address the relative contribution of threat to the overall magnitude of the gender differences in mathematics performance. In general, researchers should be more precise in their characterization of previous research and more circumspect about the meaning of their results.

Altogether, we hope that our review will encourage researchers to put more effort in testing the basic hypothesis that the gender gap in mathematics performance can be explained as a stereotype threat, put more effort in a clear characterization of data, and last, but not least, be more careful with characterizing the current state of knowledge in the scientific literature and with discussions with the popular press. Ultimately, both psychological science and the general public, in particular girls and women, will be the winners if such an attitude is taken.

References

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: Implications for explaining autism. *Science, 310,* 819–823. doi:10.1126/science.1115455

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210,* 1262–1264. doi:10.1126/science.7434028

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science, 222,* 1029–1031. doi:10.1126/science.6648516

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76,* 246–257. doi:10.1037/0022-3514.76.2.246

Brown University News Service. (2000). *Women perform better in math when tested without men, study says* [Press release]. Retrieved from http://brown.edu/Administration/News_Bureau/2000-01/00-023.html

Carey, B. (2011, November). Fraud case seen as a red flag for psychology research. *New York Times,* A3.

Casey, M. B., Nuttall, R., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test-scores across diverse samples. *Developmental Psychology, 31,* 697–705. doi:10.1037/0012-1649.31.4.697

Ceci, S. J., & Williams, W. M. (Eds.). (2007). *Why aren't more women in science? Top researchers debate the evidence.* Washington, DC: APA. doi:10.1037/11546-000

Ceci, S. J., & Williams, W. M. (2010). *The mathematics of sex: How biology and society conspire to limit talented women and girls.* New York, NY: Oxford University Press.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135,* 218–261. doi:10.1037/a0014412

Chapman, J. P., & Miller, G. A. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. doi:10.1037/0021-843X.110.1.40

Chrisler, J. C., & McCreary, D. R. (Eds.). (2010). *Handbook of gender research in psychology.* New York, NY: Springer.

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,* 1615–1628. doi:10.1177/014616702237644

Derks, B., Inzlicht, M., & Kang, S. (2008). The neuroscience of stigma and stereotype threat. *Group Processes and Intergroup Relations, 11,* 163–181. doi:10.1177/1368430207088036

Erbé, B. (2008, July). *To the contrary.* Public Broadcasting Service (PBS). Retrieved from http://www.youtube.com/watch?v=qVlyJi5hwhA

Eriksson, K., & Lindholm, T. (2007). Making gender matter: The role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. *Scandinavian Journal of Psychology, 48,* 329–338. doi:10.1111/j.1467-9450.2007.00588.x

Farrell, W. (2005). *Why men earn more: The startling truth behind the pay gap—And what women can do about it.* New York, NY: AMACOM.

Ferriman, K., Lubinski, D., & Benbow, C. P. (2009). Work preferences, life values, and personal views of top math/science graduate students and the profoundly gifted: Developmental changes and sex differences during emerging adulthood and parenthood. *Journal of Personality and Social Psychology, 97,* 517–532. doi:10.1037/a0016030

Fine, C. (2010). *Delusions of gender: The real science behind sex differences.* London, England: Icon Books Ltd.

Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of Personality and Social Psychology, 99,* 740–754. doi:10.1037/a0020971

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal-applied Economics, 2,* 210–240. doi:10.1257/app.2.2.210

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences, 19,* 229–247. doi:10.1017/S0140525X00042400

Geary, D. C. (2010). *Male, female: The evolution of human sex differences* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/12072-000

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28,* 659–670. doi:10.1177/0146167202288010

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology, 29,* 17–28. doi:10.1016/j.appdev.2007.10.004

Goodall, A. (2010). Sister's winning formula. *Times Higher Education, 1967,* 34–39.

Grand, J. A., Ryan, A. M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24,* 1–28. doi:10.1080/08959285.2010.518184

Grimm, L. R., Markman, A. B., Maddox, W. T., & Baldwin, G. C. (2009). Stereotype threat reinterpreted as a regulatory mismatch. *Journal of Personality and Social Psychology, 96,* 288–304. doi:10.1037/a0013463

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., &

Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science, 8,* 1–51.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45. doi:10.1126/science.7604277

Hines, M. (2010). Sex-related variation in human behavior and the brain. *Trends in Cognitive Sciences, 14,* 448–456. doi:10.1016/j.tics.2010.07.005

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581–592. doi:10.1037/0003-066X.60.6.581

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Science USA, 106,* 8801–8807. doi:10.1073/pnas.0901265106

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155. doi:10.1037/0033-2909.107.2.139

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11,* 365–371. doi:10.1111/1467-9280.00272

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle. *Psychological Science, 16,* 175–179. doi:10.1111/j.0956-7976.2005.00799.x

Johnson, E. S. (1984). Sex differences in problem solving. *Journal of Educational Psychology, 76,* 1359–1371. doi:10.1037/0022-0663.76.6.1359

Josephs, R. A., Newman, M. L., Brown, R. P., & Beer, J. M. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science, 14,* 158–163. doi:10.1111/1467-9280.t01–1-01435

Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles, 47,*193–198. doi:10.1023/A:1021003307511

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology, 77,* 323–338. doi:10.1348/000709906X113662

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,* 371–381. doi:10.1177/0146167202250218

Kit, K. A., Tuokka, H. A., & Mateer, C. A. (2008). A review of the stereotype threat literature and its application in a neurological population. *Neuropsychology Review, 18,* 132–148. doi:10.1007/s11065-008-9059-9

Krendl, A. C., Richeson, J. A., Kelley, W. M., & Heatherton, T. F. (2008). The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math. *Psychological Science, 19,* 168–175. doi:10.1111/j.1467-9280.2008.02063.x

Lawrence, J. S., & Charbonneau, J. (2009). The link between basing self-worth on academics and student performance depends on domain identification and academic setting. *Learning and Individual Differences, 19,* 615–620. doi:10.1016/j.lindif.2009.08.005

Lesko, A. C., & Corpus, J. H. (2006). Discounting the difficult: How high math-identified women respond to stereotype threat. *Sex Roles, 54,*113–125. doi:10.1007/s11199-005-8873-2

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123–1135. doi:10.1037/a0021276

Lubinski, D., Benbow, C. P., & Sanders, C. E. (1993). Reconceptualizing gender differences in achievement among the gifted. In K. A. Heller, F. J. Monks, & A. H. Passow (Eds.), *International handbook of research and development of giftedness and talent* (pp. 693–707). London, England: Pergamon Press.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42,* 236–243. doi:10.1016/j.jesp.2005.04.010

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin, 28,* 1183–1193. doi:10.1177/01461672022812004

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39,* 83–90. doi:10.1016/S0022-1031(02)00513-9

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

Myers, D. G. (2008). *Social psychology* (9th ed.). New York, NY: McGraw-Hill Companies.

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,* 1314–1334. doi:10.1037/a0012702

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

Osborne, J. W. (2007). Linking stereotype threat and anxiety. *Educational Psychology, 27,* 135–154. doi:10.1080/01443410601069929

Penner, A. M. (2003). International gender-by-item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology, 95,* 650–655. doi:10.1037/0022-0663.95.3.650

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004a). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist, 59,* 7–13. doi:10.1037/0003-066X.59.1.7

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004b). On the value of correcting mischaracterizations of stereotype threat. *American Psychologist, 59,* 48–49. doi:10.1037/0003-066X.59.1.48

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38,* 194–201. doi:10.1006/jesp.2001.1500

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452. doi:10.1037/0022-3514.85.3.440

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review, 16,* 177–206. doi:10.1023/B:EDPR.0000034020.20317.89

Smith, L. G. E., & Postmes, T. (2011). Shaping stereotypical behaviour through the discussion of social stereotypes. *British Journal Of Social Psychology, 50,* 74–98. doi:10.1348/014466610X500340

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811. doi:10.1037/0022-3514.69.5.797

Steele, C. M., & Aronson, J. A. (2004). Stereotype threat does not live by Steele and Aronson (1995) alone. *American Psychologist, 59,* 47–48. doi:10.1037/0003-066X.59.1.47

Stewart, T. L., Latu, I. M., Kawakami, K., & Myers, A. C. (2010). Consider the situation: Reducing automatic stereotyping through situational attribution training. *Journal of Experimental Social Psychology, 46,* 221–225. doi:10.1016/j.jesp.2009.09.004

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive

abilities test scores: A UK national picture. *British Journal of Educational Psychology, 76,* 463–480. doi:10.1348/000709905X50906

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48.

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence, 38,* 412–423. doi:10.1016/j.intell.2010.04.006

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797–826. doi:10.1037/0033-2909.127.6.797

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,* 696–716. doi:10.1037/0022-3514.89.5.696

Wraga, M., Helt, M., Jacobs, E., & Sullivan, K. (2008). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Scan, 2,* 12–19.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81,* 267–301. doi:10.3102/0034654311405999

Received August 14, 2011
Revision received October 14, 2011
Accepted November 14, 2011 ■