Journal of Educational Psychology
2013, Vol. 105, No. 3, 850–860

0022-0663/13/$12.00 DOI: 10.1037/a0032094

Order of Administration of Math and Verbal Tests: An Ecological Intervention to Reduce Stereotype Threat on Girls' Math Performance

Annique Smeding
University of Lausanne

Florence Dumas
Aix-Marseille Université and Centre National de la Recherche Scientifique

Florence Loose
Montpellier 2 University and MRM (Montpellier Research in Management), Montpellier, France

Isabelle Régner
Aix-Marseille Université and Centre National de la Recherche Scientifique

In 2 field experiments, we relied on the very features of real testing situations—where both math and verbal tests are administered—to examine whether order of test administration can, by itself, create vs. alleviate stereotype threat (ST) effects on girls' math performance. We predicted that taking the math test before the verbal test would be deleterious for girls' math performance (ST effect), whereas taking the verbal test before the math test would benefit their math performance. We also explored whether ST (if any) may spill over from the math test to the verbal test in a real-world testing situation. The studies were conducted among French middle-school students (*N*s = 1,127 and 498) during a regular class hour. In both studies, whereas girls underperformed on the math test relative to boys in the math-verbal order condition (ST effect), they performed as well as boys in the verbal-math order condition. Moreover, girls' math performance was higher in the verbal-math order condition than in the math-verbal order condition. Test order affected neither girls' verbal performance (no ST spillover) nor boys' verbal or math performance. In Study 2, additional measures pertaining to students' self-evaluations in and perceptions of the math and verbal domains provided complementary evidence that only girls who took the math test first experienced ST. Implications of order of test administration for women's experience in math, for ST effect and ST spillover research, and for educational practices are discussed.

Keywords: stereotype threat, spillover phenomenon, math performance, order of test administration, classroom intervention

At school, a widespread practice in student evaluation consists in administering standardized cognitive tests, usually comprising both math and verbal sections. In France, all second, third, fifth, and sixth graders complete standardized math and verbal tests during National Evaluations (Ministry of National Education, 2008).[1] Also, the Graduate Record Exam (GRE) and the SAT Reasoning Test, with their math and verbal sections, play a central role in admission decisions at most universities in the United States. At an international level, the Organisation for Economic Co-operation and Development Programme for International Student Assessment (OECD-PISA, 2010) assesses, every 3 years, math, verbal, and scientific knowledge among 15-year-olds in participating countries. Another feature of these testing situations is variation in the order of administration of math and verbal sections, with quite different practices in each setting. The order of verbal and math sections of the French National Evaluations is left up to teachers' decision. Math and verbal sections of the SAT and GRE tests may appear in any order. In PISA, a cluster rotation design is used to form different test booklets beginning with either math, science, or verbal items.

Because they are widely used in educational settings, it is essential that these standardized tests and their administration are as fair as possible. However, findings indicate that boys usually outperform girls at the math section of the French standardized

This article was published Online First March 18, 2013.

Annique Smeding, Laboratoire de psychologie sociale, Université de Lausanne, Lausanne, Switzerland; Florence Dumas, Laboratoire de Psychologie Cognitive (LPC), UMR CNRS 7290, Aix-Marseille Université, Marseille, France, and Centre National de la Recherche Scientifique, Lyon, France; Florence Loose, Institut Universitaire de Technologie, Département GEA–CC 411, Montpellier 2 University, and Montpellier Research in Management, Montpellier, France; Isabelle Régner, Laboratoire de Psychologie Cognitive (LPC), UMR CNRS 7290, Aix-Marseille Université, and Centre National de la Recherche Scientifique.

Florence Dumas is also at IFROSS Recherche, Jean Moulin Lyon 3 University, Lyon, France.

We thank the schools' teachers for their help with designing the standardized tests as well as Céline Darnon and Pascal Huguet for their comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Annique Smeding, Laboratoire de psychologie sociale, Institut des sciences sociales, Université de Lausanne, Bâtiment Géopolis, 1015 Lausanne, Switzerland. E-mail: Annique.Smeding@unil.ch

[1] French standardized National Evaluations of third and sixth graders have been deleted in 2009 (1 year after the present studies were conducted).

National Evaluations (Ministry of National Education, 2009). Likewise, men perform better than women on the math section of the SAT and GRE (College Entrance Examination Board, 1997). The OECD-PISA (2010) results show that, on average, 15-year-old boys obtain higher scores than girls in mathematics. Although the extent to which biology explains these gender differences in math is still highly controversial (e.g., Halpern et al., 2007), it is now well established that the threat of confirming a negative stereotype about women's math ability harms their performance on standardized math tests, a phenomenon known as *stereotype threat* (ST; e.g., Ben-Zeev, Duncan, & Forbes, 2005; Schmader, Johns, & Forbes, 2008; Spencer, Steele, & Quinn, 1999). Although less documented, laboratory studies also show that once ST has occurred on a math test, it may spillover and deteriorate women's performance in domains unrelated to math (e.g., Beilock, Rydell, & McConnell, 2007; Inzlicht & Kang, 2010). Both ST effects and ST spillover were examined here in two field experiments. We relied on the very features of real testing situations—where both math and verbal sections are administered—to examine whether their order of administration can, by itself, create *versus* alleviate ST in girls' math performance. Additionally, because the present studies were conducted in real-world testing settings, we could explore, for the first time, whether ST spillover findings generalize outside the laboratory on the verbal section of a standardized test.

Effects of ST on Performance

ST refers to a decrease in test performance in situations in which individuals feel threatened by the possibility that their performance will confirm—to others and/or themselves—a negative stereotype about their group ability in the performance domain (Steele, 1997). This situational threat increases concern about being stereotypically judged and mistreated, which depletes executive resources and leads to underperformance (Mazerolle, Régner, Morisset, Rigalleau, & Huguet, 2012; Régner et al., 2010; Schmader & Johns, 2003; Schmader et al., 2008). The deleterious effects of ST on women's math performance are well documented (Ambady, Paik, Steele, Owen-Smith, Mitchell, 2004; Cadinu, Maass, Rosabianca, & Kiesner, 2005; Inzlicht & Ben-Zeev, 2000; Schmader, 2002; Schmader & Johns, 2003; Spencer et al., 1999). In Spencer et al.'s (1999) research, for example, women typically underperform relative to equally qualified men on standardized math tests when simply told that the test measures math skills or that the test is gender-biased, but perform as well as men when told that the test is gender-fair.

Girls are also susceptible to ST in math quite early in their cognitive development (Ambady, Shih, Kim, & Pittinsky, 2001; Muzzatti & Agnoli, 2007). In the laboratory, Ambady et al. (2001) showed that Asian American girls from lower elementary and middle-school grades performed significantly worse on a math test when their gender identity was activated, compared with when their ethnic identity (associated with a positive stereotype in math) was activated or when no identity was activated. Likewise, Muzzatti and Agnoli (2007) found ST effects in 10-year-old Italian girls when they were reminded that "extraordinary achievement in math is typically a male phenomenon" (p. 747). Huguet and Régner (2007, 2009) provided evidence of ST outside the laboratory, among middle-school girls in their natural school environment. In these studies, French middle-school girls and boys had to learn a complex figure and then to reconstruct it from memory on paper. The task was presented as diagnostic of either geometry ability or drawing ability. Whereas girls underperformed (i.e., recalled fewer units) relative to boys in the geometry condition, they outperformed them in the drawing condition. In other words, ST occurred among young girls when they were simply (although erroneously) led to believe that they were taking a math (geometry) test.

Additionally, there is evidence that ST does not end on math performance but can have lingering effects in unrelated domains (Beilock et al., 2007; Inzlicht & Kang, 2010; Inzlicht, Tullett, Legault, & Kang, 2011). For example, women who experienced ST on a math test were found more likely to engage in aggressive (Study 1) or unhealthy food behaviors (Study 2) as compared with women taking the same math test but instructed how to cope with ST (Inzlicht & Kang, 2010). As an explanation, Inzlicht and Kang (2010) suggest that individuals who experience ST try to suppress the related negative thoughts, which leaves self-control resources depleted and impairs the capacity to efficiently monitor subsequent effortful behaviors. Of particular interest here, Beilock et al. (2007) showed that women who underperformed on a math test in an ST condition also underperformed on a subsequent verbal working memory task. According to these authors, spillover occurred because both tests heavily relied on the same type of working memory resources that ST also consumes.

Reducing ST

Given such negative effects on performances, researchers have looked at ways to take ST away. Several methods have proved efficient for women in the math domain. For example, one method consists in describing the math test as insensitive to gender differences (e.g., Spencer et al., 1999). This makes the gender-ability stereotype irrelevant to the performance at hand and neutralizes the fear of confirming it. Women also perform better on a difficult math test when they are told about another woman who excels in math (McIntyre et al., 2005; McIntyre, Paulson, & Lord, 2003), exactly as one would expect if women had engaged in upward-comparison assimilation (for this notion, see Huguet et al., 2009; Mussweiler & Strack, 2000). Another method consists in encouraging women to cope with ST by cognitively reappraising their emotions in order to suppress the stress and the negative thoughts associated with the gender-math stereotype (Inzlicht & Kang, 2010). Female students' math performance can also be restored when they are given the opportunity to affirm their self-concept in an unrelated domain before taking a math test (Croizet, Désert, Dutrévis, & Leyens, 2001; Martens, Johns, Greenberg, & Schimel, 2006). Researchers assume that making positive self-information accessible helps to override the impact of the imbalance between female students' need of self-worth and the expectation that their gender group should fail in math (Rydell, McConnell, & Beilock, 2009).

All these methods have undoubtedly important implications for women's math performance. However, because they necessitate implementation of instructions and cover stories that are not naturally present in real-world testing situations, these methods are at the core of a debate: Can ST naturally occur and be alleviated in the real world in the absence of such artificial interventions (Cullen, Waters, & Sackett, 2006; Sackett & Ryan, 2012; Stricker &

Ward, 2004)? A less controversial, but rarely used method to test and counteract ST in the real world consists in using existing features of actual testing situations. This is what Stricker and Ward (2004) did. As they were "working under the auspices of the Educational Testing Service" (Aronson & Dee, 2012, p. 271), Stricker and Ward took advantage of the fact that most standardized testing settings require participants to report demographic information. On this basis, they asked their participants to report gender either prior to or after taking a standardized math test. Although this research is often viewed as the most determining experimental study examining ST in operational testing situations, its results are highly debated (for an overview, see Sackett & Ryan, 2012). On one hand, Stricker and Ward concluded that varying the timing of collection of gender information had no significant effect on women's math performance. On the other hand, Danaher and Crandall (2008), who reanalyzed the data, concluded that moving the gender question to the end of the math test significantly increased women's math performance, a conclusion that Stricker and Ward (2008) viewed as unwarranted because of flawed estimates and extrapolations.

This debate typically illustrates the skepticism regarding ST generalizability to real-world high-stakes testing situations. Very recently, an amicus brief on ST has been cowritten by several ST researchers and a group of lawyers. This brief has been filed with the U.S. Supreme Court in order to acquaint the Court with ST and its relevance to affirmative action in public higher education (Brief of Experimental Psychologists, 2012). Given the ongoing debate regarding ST generalizability and the related educational policy implications, demonstrating that ST occurs and can be alleviated in the real world is certainly an important challenge for current and future ST research. Consequently, studies designed to examine ST and ways to reduce it by relying on naturally occurring features of the real-world testing situation are needed to strengthen confidence in ST generalizability. It therefore represented the first aim of the present studies. Additionally, because ST spillover has been scrutinized in laboratory studies only, examining this phenomenon in real testing situations would be a novel contribution to this growing field of research. This was the second, although more exploratory, aim of our studies.

The Present Research

We took advantage of the fact that in many high-stakes testing situations, a naturally occurring feature is variation in the order of administration of math and verbal standardized tests. This is interesting because women, although negatively stereotyped in the math domain, are positively stereotyped (as compared with men) in verbal domains (Hyde & Kling, 2001; Skaalvik & Rankin, 1990; Sommers, 2000). Given these negative and positive stereotypes, we hypothesized that the order of test administration— math-verbal order *versus* verbal-math order—is by itself likely to impact women's math performance, for the worst in the former case (ST effect) and for the best in the latter (ST reduction). Indeed, as simply describing a test as assessing math skills is sufficient to induce ST, women taking a math test before a verbal test would experience a classical ST effect on the math test. On the contrary, taking a verbal test before a math test would prevent women's performance deficit on the math test, as past research has shown that women can resist ST on a math test when they have the opportunity to self-affirm in an unrelated domain (Martens et al., 2006) or when positive stereotypes targeting them are activated (Ambady et al., 2001; Shih, Pittinsky, & Ambady, 1999). Because women are positively stereotyped in verbal domains, we reasoned that taking a verbal test first would activate the related positive stereotype and protect them from ST on the subsequent math test. Testing this hypothesis merely requires manipulating the order of tests administration, which has not yet been done in ST research. If our hypothesis were confirmed, it would indicate that an entirely ecological way (i.e., requiring neither specific test instructions nor cover stories) to reduce ST among women in math would be to fix the order of test administration: verbal tests before math tests.

Second, we explored whether ST (if any) may or may not spill over from the math section to the verbal section of a standardized test in real-world testing situations. Past ST spillover findings (Beilock et al., 2007; Inzlicht & Kang, 2010; Inzlicht et al., 2011) may lead one to expect that if girls underperform on the math test (when taking the math test first), they would also underperform on the subsequent verbal test. An alternative hypothesis, however, suggests that ST spillover is unlikely when the subsequent test is related to and has the potential to activate a positive stereotype. Indeed, whereas ST should increase attempts to suppress related negative thoughts and deplete self-control resources, the presence of positive (self-relevant) features in the test-taking environment is likely to counteract depletion by reassuring the self. In other words, the presence of a verbal test and the related positive stereotypic expectancies for women may represent a sufficiently strong incentive to counteract ST spillover. This alternative hypothesis would be in line with research demonstrating that self-affirmation (Schmeichel & Vohs, 2009) or the induction of positive mood (Tice, Baumeister, Shmueli, & Muraven, 2007) can mitigate the consequences of ego depletion (for a review, see Inzlicht & Schmeichel, 2012).

Regarding Beilock et al.'s (2007) research, we have two additional reasons to not necessarily expect a replication of their laboratory findings in the present field experiments. First, Beilock et al. outlined that spillover occurs as long as the subsequent ability test depends heavily on the same type of working memory resources that ST also consumes. Although the verbal tests we used are likely to rely on working memory to some degree, they were not pure working memory tasks (contrary to the task used by Beilock et al., 2007). In addition, these tests were made up of familiar verbal problems adapted to students' curriculum and thus likely to activate the gender stereotype favoring girls in that domain. Second, Beilock et al. used very explicit ST instructions stating that their research was aimed at better understanding why women consistently score lower than men on math tests. Consequently, it is quite likely that, even when performing the subsequent verbal working memory task, women in Beilock et al.'s study still had in mind the research's aim regarding women's inferiority in math. In our studies, in contrast, there was no mention of gender-math differences at any time, and our participants were clearly informed that they were going to take two different and separate tests: one test to evaluate their math abilities and another test to evaluate their verbal abilities. The salience of negative math thoughts when performing the verbal test should thus be less likely for our female participants than for those of Beilock et al. and, if any, should be counterbalanced by the positive verbal thoughts activated when taking the verbal test.

Finally, although the present research was designed to test the influence of order of test administration on girls' performance, it raised another question largely overlooked in the ST literature. If girls are positively stereotyped in the verbal domain as compared with boys, could it be that taking the verbal test first leads to poorer *verbal* performance for *boys* (compared with taking the math test first)? Probably not. Indeed, research on stereotype development among children indicates an important difference between negative stereotypes targeting women in the math domain and negative stereotypes targeting men in the verbal domain (Frome & Eccles, 1998; Meece, Parsons, Kaczala, & Goff, 1982; see also Kiefer & Shih, 2006). For children, stereotypes about gender differences in math performance reflect differences in innate ability between boys and girls (suggesting an inherent female inferiority in math ability), but gender differences in verbal performance are rather attributed to differences in effort, motivation, and interest (suggesting *no* inherent male inferiority in verbal ability). In other words, whereas women in the math domain are targeted by a negative-*ability* stereotype, this is not the case for men in the verbal domain. Consequently, it is unlikely that men spontaneously experience the fear of confirming a negative-*ability* stereotype in the verbal domain.

In support of this view, ST research has shown that nonchronically stigmatized individuals (such as White men) are unlikely to activate negative-ability stereotypes targeting them if no explicit ST instructions are provided (e.g., Seibt & Förster, 2004). The very few studies in which a performance decrement was observed among men on a verbal test did explicitly manipulate negative versus positive expectations regarding men's verbal abilities (Keller, 2007; Seibt & Foster, 2004). Because our test description mirrored real test instructions (i.e., merely describing the tests as being diagnostic of math or verbal ability) with no additional information focusing on gender-ability differences, there was no reason for boys to be aware that they could be viewed through the lens of a negative-ability stereotype in the verbal domain. Therefore, boys' verbal test performance should be unaffected by order of test administration.

To sum up, we expected that girls' math performance would be harmed in the math-verbal order condition, but would benefit from the verbal-math order condition, whereas girls' verbal performance, and boys' verbal as well as math performances, would be unaffected by test ordering. We tested for these hypotheses among French middle-school students in their regular classrooms (a population for which ST effects on girls' math performance have already been demonstrated; Huguet & Régner, 2007, 2009), with real-life test instructions stating that the math and verbal tests assessed math and verbal abilities, respectively. As such, the testing setting was similar to the high-stakes testing situations commonly experienced by students during their academic curriculum.

Study 1

Method

Participants and design. Participants were 1,127 eighth graders (586 girls; mean age = 14, SD = .67) from nine French middle schools (schools were located in urban areas in Southern France). All schools were very similar as they were selected from the same geographic (urban) area, were all public, socially and culturally mixed schools, and were not high-ranked or elite institutions. Furthermore, the math and French curricula were very similar across schools (and across classes as a matter of fact), because the French Ministry of National Education provides a very precise program about what must be taught in a given grade in all of the country's public schools. Across schools, classrooms comprised, on average, 50.62% of girls (SD = 12.62), which reflects the common practice in French middle schools to equalize classes in terms of gender ratios. The study was a 2 (gender: male vs. female) × 2 (test order: math-verbal vs. verbal-math) between-subjects design.

Procedure and measures. Tests were taken during a Trimester 2 regular class hour. In each school, half of the classrooms were randomly assigned to the math-verbal order condition (math test taken before the verbal test) and the other half to the verbal-math order condition (resulting in 291 girls in the math-verbal condition, 295 girls in the verbal-math condition, 258 boys in the math-verbal condition, and 283 boys in the verbal-math condition). As for the French national evaluations, schoolteachers administered the tests. All were trained so as to standardize test administration, and none of them were participants' math or French instructors. All participants were told that they would complete two separate tests: one assessing their math abilities and another one assessing their verbal abilities (or the reverse depending on test order condition).

Tests were modeled after those used for Grade 6's French national evaluations (which are standardized math and verbal tests completed by all French students). They were designed by math and French teachers to be equally difficult and adapted to Grade 8's academic curriculum. The math test (16 items) focused on algebra, geometry, and operations. The verbal test (30 items) focused on reading, vocabulary, and language comprehension.

For each test, students were given 20 min to work through the problems, with only a few minutes break between the tests (about 5 min). Math and verbal performances were measured by the percentage of correct responses. Finally, students' grades in math and French (at Trimester 1) were taken from official school records and were used to control for prior differences in math and verbal abilities, respectively. These grades corresponded to the mean grades students obtained during the trimester on continuous assessments.[2]

Results

Preliminary analyses. To examine whether there were systematic variations in covariates (grades) across test order conditions, we submitted participants' math and French grades to two 2 (gender) × 2 (test order) analyses of variance (ANOVAs). Only a main effect of gender appeared for French grades, with girls (M = 12, SE = .14) outperforming boys (M = 10.66, SE = .14), $F(1, 1123) = 45.76$, $p < .001$, d = .40. Because there were no systematic variations in grades across test order conditions, math

[2] In Study 1, mean grades in French for each group were as follows: $M_{Girls/Math\text{-}Verbal}$ = 11.94 (SD = 3.33); $M_{Girls/Verbal\text{-}Math}$ = 12.04 (SD = 3.03); $M_{Boys/Math\text{-}Verbal}$ = 10.80 (SD = 3.37); $M_{Boys/Verbal\text{-}Math}$ = 10.51 (SD = 3.48). As for math, mean grades were as follows: $M_{Girls/Math\text{-}Verbal}$ = 10.81 (SD = 3.85); $M_{Girls/Verbal\text{-}Math}$ = 10.51 (SD = 4.13); $M_{Boys/Math\text{-}Verbal}$ = 10.47 (SD = 4.05); $M_{Boys/Verbal\text{-}Math}$ = 10.55 (SD = 4.13). In Study 2, mean verbal scores on the national evaluations were as follows: $M_{Girls/Math\text{-}Verbal}$ = 62.00 (SD = 16.60); $M_{Girls/Verbal\text{-}Math}$ = 64.87 (SD = 16.44); $M_{Boys/Math\text{-}Verbal}$ = 56.04 (SD = 18.19); $M_{Boys/Verbal\text{-}Math}$ = 55.13 (SD = 17.81). For math, mean scores were as follows: $M_{Girls/Math\text{-}Verbal}$ = 67.61 (SD = 15.94); $M_{Girls/Verbal\text{-}Math}$ = 70.15 (SD = 14.87); $M_{Boys/Math\text{-}Verbal}$ = 73.69 (SD = 15.11); $M_{Boys/Verbal\text{-}Math}$ = 72.37 (SD = 15.31).

and French grades were used as covariates in our main analyses. In addition, to further ensure that the effects of gender and test order would be estimated without bias, we also entered the Grades × Test Order interactions in our models. This adjustment is recommended when the covariate (e.g., test grades) is related to the measured independent variable (e.g., gender; see Yzerbyt, Muller, & Judd, 2004).

Order of test administration. Math and verbal test performances were submitted to two 2 (gender) × 2 (test order) between-subjects analyses of covariance (ANCOVAs), controlling for math grades, $F(1, 1121) = 131.34$, $p < .001$, and French grades, $F(1, 1121) = 79.42$, $p < .001$, respectively, and the Grades × Test Order interactions, $F(1, 1121) = 2.69$, *ns*, for math performance; and, $F(1, 1121) = 0.90$, *ns*, for verbal performance.

Math performance. For math performance, only the predicted Gender × Test Order interaction was significant, $F(1, 1121) = 7.37$, $p < .01$ (see Figure 1). As expected, simple main effect analyses showed that whereas girls ($M = 51.01$, $SE = 1.11$) underperformed relative to boys ($M = 54.69$, $SE = 1.18$) when the math test was taken first, $F(1, 1121) = 5.16$, $p < .03$, $d = .20$, girls ($M = 54.67$, $SE = 1.10$) performed as well as boys ($M = 52.21$, $SE = 1.12$) when the verbal test was taken first, $F(1, 1121) = 2.43$, *ns*. Girls taking the verbal test first performed better than those taking the math test first, $F(1, 1121) = 5.46$, $p < .03$, $d = .19$. Thus, and as predicted, girls who took the math test before the verbal test experienced a classical ST effect on the math test, whereas those who took the verbal test first resisted ST on the math test. Figure 1 also seems to illustrate a stereotype lift effect on boys' math performance (i.e., a performance boost caused by a downward comparison with a negatively stereotyped outgroup; Walton & Cohen, 2003): Boys' math performance tended to be slightly higher in the math-verbal order condition than in the verbal-math order condition. However, this difference did not reach significance, $F(1, 1121) = 2.31$, *ns*.

Verbal performance. For verbal performance, results revealed a main effect of gender, $F(1, 1121) = 5.35$, $p < .03$, $d = .14$, with girls ($M = 56.56$, $SE = .53$) performing better than boys ($M = 54.77$, $SE = .55$). No other effect was significant, indicating that boys' (as well as girls') verbal performance was unaffected by the order of test administration.[3]

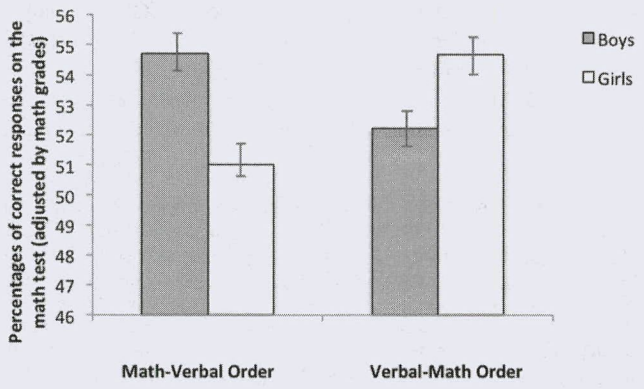

Figure 1. Math test performance as a function of gender and test order in Study 1. Error bars represent the standard error of the mean.

Discussion

Study 1 demonstrated that, as expected, girls underperformed on the math test relative to boys in the math-verbal order condition, but performed as well as boys in the verbal-math order condition. Moreover, girls' math performance was higher in the verbal-math order condition than in the math-verbal order condition. This suggests that girls suffered from ST when the math test was completed first, but not when this test was preceded by a verbal test. Importantly, neither girls' verbal performance nor boys' math or verbal performances were affected by test ordering. We did not find evidence of ST spillover from the math test onto the verbal test among girls in the math-verbal order condition either. Given the important educational implications of these results, and because effect sizes were in the small range, we conducted a second study to test their replicability and generalizability to a younger sample. Study 2 was also designed to provide complementary evidence, besides performance, that only girls who took the math test first experienced ST. We thus examined whether measures other than performance were affected by order of test administration. We relied on past research showing that ST effects are not confined to performance but impact students' perceptions and self-evaluations in the threatening and/or nonthreatening domains.

For example, some research has shown that girls rate their relative standing in math less positively than do boys (Huguet & Régner, 2007) and that ST increases negative math-related thoughts in women (Cadinu et al., 2005). We assume that taking the verbal test first, a domain in which girls and women are positively stereotyped, may be self-reassuring and help them restore higher self-evaluations in the math domain. If this is the case, then girls would report lower self-evaluations in math compared with boys in the math-verbal order condition, but not in the reverse order condition. Another consequence of repeated exposure to ST is disidentification from the stigmatized domain, resulting in a decrease of the importance or relevance of the threatening domain for the self (Steele, 1997). So, if taking the math test first leads girls to further experience ST, they would attach less interest or importance to the math domain.

An alternative hypothesis can be made, however, because it has proved difficult to devalue a domain that is highly valued in society (Crocker & Major, 1989; Steele, Spencer, & Aronson, 2002). Because math is such an important domain in the academic curriculum, girls may cope with the threat of taking the math test first not by rejecting the importance of math altogether, but by enhancing their temporary interest in the nonthreatening verbal domain. In line with this, Davies, Spencer, Quinn, and Gerhard-

[3] In both studies, some students declared that, "besides French, they spoke another language at home" (the item was stated as such). This was the case for 19.9% of students in the math-verbal condition and 20.6% in the verbal-math condition in Study 1, and for 11.5% of students in the math-verbal condition and 11.9% in the verbal-math condition in Study 2. However, entering the dichotomous variable "Other language spoken at home besides French" in the analytic model for verbal performance did not change our findings in any of the studies. Analyses only revealed a main effect of this variable in Study 1 (with students reporting speaking another language at home having lower verbal scores than those who do not). This main effect was not replicated in Study 2, and no interaction effects occurred with any of the other variables in any of the studies. Given that no interaction effects were found with gender or test order condition, we do not discuss this finding further.

stein (2002) showed that exposure to ST during test taking led women to avoid math items in favor of verbal items (Study 2) and to report more interest in verbal domains (Study 3). Although it may seem puzzling to predict that test ordering will influence girls' interest in the verbal domain while we expected and found no effect on verbal performance (Study 1, no ST spillover effect), both are completely compatible. If ST in math may indeed be unlikely to spill over onto girls' subsequent performance in a positively stereotyped domain, it is nevertheless likely to increase their need of self-worth. Such a need can be easily met by valuing domains in which their gender group fares well (Crocker & Major, 1989).

In sum, Study 2 was not a mere replication of Study 1. It was designed to provide converging evidence from both self-reports and test performance that ST was operating among girls taking the math test first. Another improvement compared with Study 1 was the use of standardized test scores as covariates, which are less biased indicators of prior individual performances than are teachers' academic grades.

Study 2

Method

Participants and design. Participants were 498 seventh graders (267 girls; mean age = 13, *SD* = .48) from four French public middle schools located in the same geographical area as in Study 1. These schools held comparable characteristics as those of Study 1 (i.e., socially and culturally mixed schools, not high-ranked or elite institutions, similar math and French curricula). Likewise, as in Study 1, gender ratios were quite homogeneous across classrooms, with the average percentage of girls being 53% (*SD* = 11.47). The study was again a 2 (gender: male vs. female) × 2 (test order: math-verbal vs. verbal-math) between-subjects design.

Procedure and measures. Procedure and measures closely paralleled those from Study 1. In each school, half of the classrooms were randomly assigned to the math-verbal order condition and the other half to the verbal-math order condition (131 girls in the math-verbal condition; 136 girls in the verbal-math condition; 129 boys in the math-verbal condition; 102 boys in the verbal-math condition). Tests were modeled after those used for Grade 6's French national evaluations, but adapted to Grade 7's academic curriculum. The math test comprised 16 items and the verbal test 21. Percentages of correct responses were computed as performance indicators. Upon test completion, participants indicated, on 5-point scales, how important it was for them to be good in the math and verbal domains (1 = *Not at all important*, 5 = *Very important*), how interesting they thought it was to work on a particular exam in the math/verbal domain (1 = *Not at all interesting*, 5 = *Very interesting*), and their standing in comparison with their classmates in both domains (1 = *Among the worst*, 5 = *Among the best*). All these self-reports were collected after (not before) test completion, as the experience of ST has proved to increase gradually during test taking due to the cumulative negative effects of thought intrusions (Cadinu et al., 2005; Spencer, Steele, & Quinn, 2002). Importantly, this timing was also compatible with the ecological testing conditions we wanted to preserve in the present research (i.e., no implementation of unusual features before tests completion). We obtained from school records participants' test scores on the French national evaluations they took the preceding year when they were sixth graders. This allowed us to control for prior individual differences in abilities using standardized (rather than teacher-graded) measures in the math and verbal domains.

Results

Preliminary analyses. To examine whether there were systematic variations in covariates across test order conditions, we submitted participants' math and verbal test scores on the national evaluations (NE) to two 2 (gender) × 2 (test order) ANOVAs. For the NE-math score, only a main effect of gender was found, with boys (M = 73.03, SE = 1.04) outperforming girls (M = 68.88, SE = .94), $F(1, 494) = 9.04$, $p < .01$, $d = .27$. This difference replicates the typical gender gap in math observed on standardized tests. For the NE-verbal score, a main effect of gender was found, with girls (M = 63.43, SE = 1.06) outperforming boys (M = 55.58, SE = 1.14), $F(1, 493) = 25.46$, $p < .001$, $d = .46$. No other effects were significant. Thus, as in Study 1, there were no systematic variations on the covariates as a function of test order conditions. These NE-test scores and their interaction with test order condition were entered in the models.

Order of test administration. Math and verbal performances were submitted to two 2 (gender) × 2 (test order) between-subjects ANCOVAs, controlling for NE-math scores, $F(1, 492) = 101.64$, $p < .001$, and NE-verbal scores, $F(1, 491) = 76.266$, $p < .001$, respectively, and their interaction with test order, $F(1, 492) = .23$, *ns*, for math performance and, $F(1, 491) = 1.06$, *ns*, for verbal performance. The same analyses were run on the self-report measures, with the covariates adapted to the domains (math and verbal). Degrees of freedom varied depending on the outcome under investigation, as there were some missing data.

Math performance. As in Study 1, only the predicted Gender × Test Order interaction was significant, $F(1, 492) = 4.76$, $p < .04$. Replicating Study 1's findings, simple main effect analyses showed that girls (M = 49.79, SE = 1.36) underperformed relative to boys (M = 53.71, SE = 1.38) when the math test was taken first, $F(1, 492) = 4.01$, $p < .05$, $d = .26$ (see Figure 2). In addition, girls taking the verbal test first (M = 54.06, SE = 1.32) not only performed as well as boys (M = 51.87, SE = 1.52) in the same condition, $F(1,$

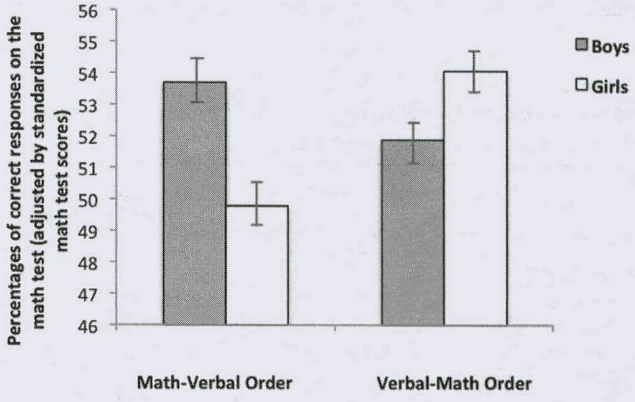

Figure 2. Math test performance as a function of gender and test order in Study 2. Error bars represent the standard error of the mean.

492) = 0.28, *ns*, but they also performed better than girls taking the math test first, $F(1, 492) = 5.10$, $p < .03$, $d = .28$.

Verbal performance. For verbal performance, no main or interaction effects were found (all $Fs < 2$), indicating that, as in Study 1, test order did not influence performance on the verbal test ($M = 79.52$, $SE = .55$).

Interest in and importance of the domains. Descriptive statistics for all self-report measures are displayed in Table 1. Regarding interest in math, results revealed only a main effect of test order, with participants reporting higher interest in the math domain when the verbal test was taken first ($M = 3.83$, $SE = .07$) than when the math test was completed first ($M = 3.51$, $SE = .06$), $F(1, 488) = 12.95$, $p < .001$, $d = .32$. In the verbal domain, a main effect of gender was found, with girls ($M = 3.51$, $SE = .06$) reporting higher interest than boys ($M = 3.31$, $SE = .07$), $F(1, 487) = 4.75$, $p < .04$, $d = .20$. This effect was qualified by a Gender × Test Order interaction effect, $F(1, 487) = 4.67$, $p < .04$: Girls reported higher interest in the verbal domain ($M = 3.66$, $SE = .09$) than did boys ($M = 3.24$, $SE = .09$) when the math test was taken first, $F(1, 487) = 10.22$, $p < .01$, $d = .41$, but not when the verbal test was completed first ($F < 1$, *ns*).

Regarding perceived importance of the math and verbal domains, results revealed no main or interaction effects (all $Fs < 2$), both domains being rated as equally highly important by all students ($Ms = 4.37$ and $SEs = .04$).

Self-evaluations. Participants' perceived standing as compared with their classmates was examined as a function of gender and test order. In the math domain, results revealed only a marginally significant interaction effect between gender and test order, $F(1, 489) = 3.42$, $p < .07$. As expected, girls' self-evaluations in the math domain ($M = 3.29$, $SE = .08$) were lower than those of boys ($M = 3.58$, $SE = .08$) when the math test was taken first, $F(1, 489) = 6.0$, $p < .02$, $d = .32$, but not when the verbal test was taken first ($F < 1$, *ns*). Regarding self-evaluations in the verbal domain, no main or interaction effects were found ($M = 3.25$, $SE = .04$).

Discussion

Findings from Study 2 replicated and extended those from Study 1. First, they demonstrated, as in Study 1, that completing a math test before a verbal test undermined girls' math performance, whereas taking the verbal test before the math test protected their math performance from ST. As in Study 1, test order did not affect boys' math performance, or the verbal performance of any of the gender groups. There was no sign of ST spillover either. Importantly, these results were obtained with standardized covariates, namely, students' test scores on the standardized national evaluations. In addition, Study 2's findings indicated that completing the math test first did not undermine girls' identification with the math domain as a whole but rather increased their interest in the verbal domain in which they are positively stereotyped. Such an increase in interest in nonthreatening domains in reaction to negative stereotypes targeting another domain has already been reported in the literature (Davies et al., 2002). This is consistent with the idea that girls try to cope with ST by favoring nonthreatening domains rather than by reducing their identification with and interest in such an important domain as math. Interestingly, all participants—girls and boys—reported higher interest in math when the verbal test was completed first, indicating that both gender groups took some benefits from the verbal-math order condition.

Eventually, findings relating to the self-evaluation measures suggested that taking the math test first led girls to rate their relative standing in math as less favorably than did boys, whereas no differences were found when the verbal test was completed first. Order of test administration did not affect boys' self-evaluations in math, or any of the gender groups' ratings in the verbal domain. Importantly, the very fact that girls reported lower self-evaluations in math compared with boys in the math-verbal order condition but not in the reverse order provides further evidence that ST was operating when the math test was taken first but not when the verbal test was taken first. These findings also suggest that self-presentational concerns were probably not of major influence here. Indeed, if girls in the math-verbal order condition had biased their responses in a socially desirable manner, they would have rated their relative standing in math as high, possibly as a means to react to the fear of confirming the negative stereotype (see Kray, Tompson, & Galinsky, 2001, for research on stereotype reactance). This is not what we found. Together, findings therefore indicate that, in accordance with the ST hypothesis, the math-verbal order condition had negative effects for girls only and in the math domain only.

Table 1
Means (and Standard Errors) for the Self-Report Measures in Study 2 as a Function of Gender and Test Order

	Math-verbal order		Verbal-math order	
Self-report measure	Girls	Boys	Girls	Boys
Math domain				
Interest	3.52 (.09)	3.49 (.09)	3.74 (.09)	3.92 (.10)
Importance	4.38 (.07)	4.37 (.07)	4.37 (.07)	4.37 (.08)
Self-evaluations	3.29 (.08)	3.58 (.08)	3.47 (.08)	3.45 (.09)
Verbal domain				
Interest	3.66 (.09)	3.24 (.09)	3.37 (.09)	3.37 (.10)
Importance	4.49 (.07)	4.31 (.08)	4.34 (.07)	4.36 (.09)
Self-evaluations	3.29 (.08)	3.35 (.08)	3.21 (.08)	3.17 (.09)

General Discussion

Given that ST effects on women's math performance is a pervasive phenomenon that can seriously impact them in a variety of domains, research has evidenced several promising remediation strategies (e.g., exposing female students to positive role models, telling them that the test is gender-fair, or asking them to write about their positive values just before taking a math test). However, all these methods require implementing specific test instructions or cover stories that are not naturally present in real testing situations. The aim of the present field experiments was to test the efficiency of a simple and ecological intervention inspired from the real-world testing situation itself. As taking standardized tests comprising both math- and verbal-related sections is common practice in educational settings, we reasoned that we could rely on this testing situation to reduce ST effects among girls in math. We argued that the order of administration of these math and verbal tests would be crucial. On the basis of past ST research, we hypothesized that if the math test were completed before the verbal test, we would observe a classical ST effect on girls' math perfor-

mance. On the contrary, we expected that taking the verbal test before the math test would protect girls' subsequent math performance from ST.

Results of both studies were consistent with our expectations. Taking a verbal test (a domain in which women are targeted by a positive ability stereotype) before a math test (a domain in which women are targeted by a negative ability stereotype) seems to produce similar effects (ST reduction) as lab interventions. Our ecological intervention indeed benefited girls' math performance, suggesting that taking the verbal test first contributed to limiting the negative consequences of ST for girls in the math domain. Findings relating to a series of self-report measures in Study 2 further supported the ST hypothesis. Girls who completed the math test first reported more interest, than any of the other groups, in working on a nonthreatening verbal exam, and rated their standing in math less favorably than did their male peers. Together, results from the two studies provided convergent support for the detrimental effect of the math-verbal test order and the beneficial effect of the reverse order for girls in math. These findings clearly illustrate that the very features that are naturally present in real-world testing settings have the potential to *both elicit and alleviate* ST, which is an important contribution to the issue of ST generalizability to the real world. Because taking standardized tests comprising verbal- and math-related sections is a frequent practice in educational settings, ensuring that the verbal sections are completed before the math sections represents a realistic means to create a virtuous cycle for women in math. This easy-to-implement, though theory-driven, intervention would be one way to support girls' and women's educational and professional aspirations through a reduction of bias in test scores and self-evaluations in math.

This ecological method seems even more interesting as it benefited girls' math performance without significant costs for boys' math or verbal performances. First, although boys' math performance tended to be slightly lower in the verbal-math order condition than in the math-verbal order condition (stereotype lift tendency), this difference was not significant. This finding is not surprising as it is clearly consistent with the existing literature showing that stereotype lift effects are much lower in magnitude than ST effects (Walton & Cohen, 2003). In addition, according to Walton and Cohen (2003), nontargets of a negative-ability stereotype (here, boys in math) can underperform when the stereotype is *explicitly* refuted (e.g., describing the math test as gender-fair) because they lose the benefit of downward comparison to a negatively stereotyped outgroup (girls in math). Here, the test order manipulation did not explicitly invalidate the negative stereotype targeting women in the math domain, explaining why the stereotype lift effect among boys did not reach significance. Our findings are therefore all the more encouraging from an applied educational perspective: They suggest that taking a verbal test before a math test is a successful method to improve women's math performance without significantly decreasing that of men.

Second, boys' verbal performance was unaffected by the order of test administration. This finding suggests that boys did not experience ST on the verbal test. This is consistent with the idea that men in the verbal domain are not targeted by a negative-*ability* stereotype, their poor performances (when any) in this domain being rather attributed to a lack of effort and interest (Frome & Eccles, 1998; Meece et al., 1982). The fact that our male participants obtained lower verbal scores than girls regardless of test order may thus reflect less interest in that domain. Results regarding the self-report measures tend to support this interest explanation, as boys, regardless of test order condition, were less interested in working on a verbal exam than girls. However, because most ST research has focused on math stereotypes and their effects on girls' and women's performances, further research is needed to better understand boys' underperformance as compared with girls in the verbal domain.

Another aim of the present studies was to explore whether ST spillover may generalize outside the laboratory on the verbal section of a standardized test in real test-taking settings. It seems not. The present findings indicate that there was no ST spillover from the math test onto the verbal test in any of our studies and, more generally, that girls' verbal performance was unaffected by order of test administration. These results suggest that ST spillover is not systematic and provide support for a boundary condition that has been previously proposed by Beilock et al. (2007). According to them, one condition for spillover is that the subsequent test depends on the same type of working memory resources that ST also consumes. If this condition was indeed fulfilled in their research (their verbal task was a direct measure of working memory), this was not the case in the present studies: Although the verbal tests we used certainly relied on working memory to some degree, they were not pure working memory tasks.

Furthermore, contrary to Beilock et al. (2007), we never mentioned that our studies were aimed at examining gender differences in math, which strongly reduced the possibility that girls were thinking about the negative math stereotype throughout the whole study. Precisely because the verbal tests consisted in familiar verbal problems and because no explicit threat instructions were given, girls were likely to activate the gender stereotype favoring women in the verbal domain. This was probably not the case in the Beilock et al. study because the two-back working memory task was rather novel for participants. Our findings therefore suggest, in accordance with research on ego depletion (Inzlicht & Schmeichel, 2012; Schmeichel & Vohs, 2009), that if positive cues are present in the test-taking situation and activated, lingering effects of ST can be mitigated.

The absence of an ST spillover effect in our research is, again, encouraging. It suggests that spillover on the verbal test is likely to occur only when study or test instructions explicitly state that the aim is to explore gender differences and/or when no positive cues are present in the situation. It also suggests that, pending ecological interventions like the ones tested in the present studies are implemented, ST spillover from the math to the verbal domain is not observed in real-life testing situations. Nevertheless, to strengthen our conclusions, future research should more systematically examine this issue by comparing the influence of order of test administration when gender differences are explicitly mentioned versus when no specific instructions are provided, or in the presence versus absence of positive cues. This would contribute to highlight all the possible boundary conditions for ST spillover.

A potential limitation of our research is that we used standardized tests modeled after those used for Grade 6's French national evaluations and adapted to Grades 7 and 8's academic curriculum (there is no standardized national evaluation for seventh and eighth graders). It would be interesting in future research to investigate whether our results would be replicated on French second, fifth. or

sixth graders for whom the standardized national evaluations are used to assess math and verbal abilities. Likewise, future research could test the generalizability of our findings to the GRE and SAT tests. The fact that the effect of test ordering on girls' math performance was found here in two studies and among two different age groups of students is quite encouraging, to say the least. Another issue that could be addressed is the delay between the verbal and the math tests. In the present studies, we fixed a few-minutes delay between the tests. It thus would be interesting to determine which delay would be optimal for female students' subsequent math performance. In addition, it would be worthwhile in future studies to investigate what would be the optimal test order condition for individuals who are targeted by a negative gender-math stereotype, but simultaneously by a positive ethnic-math stereotype, as is the case for Asian women (see Shih et al., 1999).

Another avenue for future research would be to explore the precise mechanisms through which the verbal-math order reduces ST effects. Because women are positively stereotyped in the verbal domain, it might be that taking the verbal test first increases their feelings of self-efficacy and/or their performance expectancies (compared with taking the math test first), resulting in higher math performance. Such a mechanism would be in line with research showing that performance expectancy is a partial mediator of ST effects (Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003). Also in line with this idea is research using the ST-reducing technique of self-affirmation (Legault, Al-Khindi, & Inzlicht, 2012), which postulates that when threatened individuals have the possibility to self-affirm in a valued, nonthreatening domain, their sense of competence is restored, although empirical evidence is lacking.

Finally, although the present findings were small in magnitude, it is important not to underestimate their practical significance. Small effect sizes are quite common when predicting a multiply determined outcome like academic achievement (Ahadi & Diener, 1989; Noftle & Robins, 2007). Consistent with this idea, ST is one of the ways by which negative gender stereotypes affect achievement, whereas other literatures have focused, for instance, on how these stereotypes influence girls' self-perceptions of ability and their educational choices (e.g., Eccles, 1987; Eccles, Jacobs, & Harold, 1990; Jacobs, 1991; Tenenbaum & Leaper, 2003; Tiedemann, 2000; Wigfield et al., 1997). Furthermore, small effect sizes can have a major impact on outcomes over time (Abelson, 1985; Rosenthal & Rubin, 1982).

To conclude, the present findings show that the ordering of math and verbal sections of standardized tests can affect girls' math performance, for the worst when the math test is completed first (i.e., ST effect), or for the best when the verbal test is taken first (i.e., ST reduction). These results emphasize that great importance should be attached to tests' ordering. However, to date, this issue is addressed quite differently in the real-world testing situations highlighted in the present article (i.e., random administration, cluster rotation design, teachers' decision). Our findings, rather, encourage fixing the order of math and verbal sections of standardized tests: verbal tests before math tests. This would help female students performing at optimal levels in the math domain without significantly impairing boys' math performance or any of the gender groups' verbal performance. One may think that randomization would be a better solution, based on the assumption that the verbal-math order might compensate for the negative effect of the math-verbal order on girls' math performance. Such compensation is unlikely, however. The present studies showed that girls in the verbal-math order condition performed equally well as boys on the math test but not better than them. Randomization may thus, at best, reduce the female disadvantage on the math test but is unlikely to eliminate it. Consistent with this, the gender gap is present on the math sections of both SAT and GRE despite randomization. Perhaps more importantly, the math-verbal order would still be unfair to those unfortunate female students who would have been randomly assigned to this condition, with potential negative consequences for their math-related careers. Consequently, making students complete verbal tests before math tests is much more advantageous than randomization. Our findings demonstrate that it is possible to work efficiently with the very features of the existing test-taking procedures to improve girls' and women's prospects in math and to make educational settings fairer places.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133. doi:10.1037/0033-2909.97.1.129

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56,* 398–406. doi:10.1037/0022-3514.56.3.398

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchel, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology, 40,* 401–408. doi:10.1016/j.jesp.2003.08.003

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12,* 385–390. doi:10.1111/1467-9280.00371

Aronson, J., & Dee, T. (2012). Stereotype threat in the real world. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 264–279). New York, NY: Oxford University Press.

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,* 256–276. doi:10.1037/0096-3445.136.2.256

Ben-Zeev, T., Duncan, S., & Forbes, C. (2005). Stereotypes and math performance. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 235–249). New York, NY: Psychology Press.

Brief of Experimental Psychologists. et al. as Amici Curiae Supporting Respondents, Fisher v. University of Texas, August 13, 2012 (No. 01–1015).

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology, 33,* 267–285. doi:10.1002/ejsp.145

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16,* 572–578. doi:10.1111/j.0956-7976.2005.01577.x

College Entrance Examination Board. (1997). *National report on college-bound seniors, various years.* New York, NY: Author.

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96,* 608–630. doi:10.1037/0033-295X.96.4.608

Croizet, J.-C., Désert, M., Dutrévis, M., & Leyens, J.-P. (2001). Stereotype threat, social class, gender, and academic under-achievement: When our reputation catches up to us and takes over. *Social Psychology of Education, 4,* 295–310. doi:10.1023/A:1011336821053

Cullen, M. J., Waters, C. M., & Sackett, P. R. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance, 19,* 421–440. doi:10.1207/s15327043hup1904_6

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology, 38,* 1639–1655. doi:10.1111/j.1559-1816.2008.00362.x

Davies, P. G., Spencer, S. J., Quinn, D., & 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

Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly, 11,* 135–172. doi:10.1111/j.1471-6402.1987.tb00781.x

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues, 46,* 183–201. doi:10.1111/j.1540-4560.1990.tb01929.x

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology, 74,* 435–452. doi:10.1037/0022-3514.74.2.435

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 in the Public Interest, 8,* 1–51. doi:10.1111/j.1529-1006.2007.00032.x

Huguet, P., Dumas, F., Marsh, H. W., Régner, I., Wheeler, L., Suls, J., . . . Nezlek, J. (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,* 156–170. doi:10.1037/a0015558

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

Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology, 45,* 1024–1027. doi:10.1016/j.jesp.2009.04.029

Hyde, J. S., & Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly, 25,* 364–378. doi:10.1111/1471-6402.00035

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

Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology, 99,* 467–481. doi:10.1037/a0018951

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science, 7,* 450–463. doi:10.1177/1745691612454134

Inzlicht, M., Tullett, A. M., Legault, L., & Kang, S. K. (2011). Lingering effects: Stereotype threat hurts more than you think. *Social Issues and Policy Review, 5,* 227–256. doi:10.1111/j.1751-2409.2011.01031.x

Jacobs, J. E. (1991). Influence of gender stereotypes on parent and child mathematics attitudes. *Journal of Educational Psychology, 83,* 518–527. doi:10.1037/0022-0663.83.4.518

Keller, J. (2007). When negative stereotypic expectancies turn into challenge or threat: The moderating role of regulatory focus. *Swiss Journal of Psychology, 66,* 163–168. doi:10.1024/1421-0185.66.3.163

Kiefer, A., & Shih, M. (2006). Gender differences in persistence and attributions in stereotype relevant contexts. *Sex Roles, 54,* 859–868. doi:10.1007/s11199-006-9051-x

Kray, L. J., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology, 80,* 942–958. doi:10.1037/0022-3514.80.6.942

Legault, L., Al-Khindi, T., & Inzlicht, M. (2012). Preserving integrity in the face of performance threat: Self-affirmation enhances neurophysiological responsiveness to task errors. *Psychological Science, 23,* 1455–1460. doi:10.1177/0956797612448483

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

Mazerolle, M., Régner, I., Morisset, P., Rigalleau, F., & Huguet, P. (2012). Stereotype threat strengthens automatic recall and undermines controlled processes in the older adults. *Psychological Science, 23,* 723–727. doi:10.1177/0956797612437607

McIntyre, R. B., Lord, C. G., Gresky, D. M., Ten Eyck, L. L., Frye, G. D. J., & Bond, C. F. (2005). A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. *Current Research in Social Psychology, 10,* 116–136.

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

Meece, J. L., Parsons, J. E., Kaczala, C. M., & Goff, S. B. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin, 91,* 324–348. doi:10.1037/0033-2909.91.2.324

Ministry of National Education. (2008). *L'évaluation des élèves de 6ème* [The evaluation of sixth graders]. Retrieved from http://evace26.education.gouv.fr/

Ministry of National Education. (2009). *Filles et garçons sur le chemin de l'égalité de l'école à l'enseignement supérieur* [Girls and boys on the road of equality from primary school to higher education]. Retrieved from http://media.eduscol.education.fr/file/2009/33/6/F_&_G_sur_le_chemin_de_l_egalite_2009_web_45336.pdf

Mussweiler, T., & Strack, F. (2000). The "relative self": Informational and judgmental consequences of comparative self-evaluation. *Journal of Personality and Social Psychology, 79,* 23–38. doi:10.1037/0022-3514.79.1.23

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

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93,* 116–130. doi:10.1037/0022-3514.93.1.116

Organisation for Economic Co-operation and Development. (2010). *OECD Programme for International Student Assessment (PISA).* Retrieved from http://www.pisa.oecd.org/pages/0,3417,en_32252351_32235907_1_1_1_1_1,00.html

Régner, I., Smeding, A., Gimmig, D., Thinus-Blanc, C., Monteil, J. M., & Huguet, P. (2010). Individual differences in working memory moderate stereotype-threat effects. *Psychological Science, 21,* 1646–1648. doi:10.1177/0956797610386619

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169. doi:10.1037/0022-0663.74.2.166

Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology, 96,* 949–966. doi:10.1037/a0014846

Sackett, P. R., & Ryan, A. M. (2012). Concerns about generalizing ste-

reotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 249–263). New York, NY: Oxford University Press.

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

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115,* 336–356. doi:10.1037/0033-295X.115.2.336

Schmeichel, B. J., & Vohs, K. D. (2009). Self-affirmation and self-control: Affirming core values counteracts ego depletion. *Journal of Personality and Social Psychology, 96,* 770–782. doi:10.1037/a0014635

Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology, 87,* 38–56. doi:10.1037/0022-3514.87.1.38

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

Skaalvik, E. M., & Rankin, R. J. (1990). Math, verbal, and general academic self-concept: The internal/external frame of reference model and gender differences in self-concept structure. *Journal of Educational Psychology, 82,* 546–554. doi:10.1037/0022-0663.82.3.546

Sommers, C. H. (2000, May). The war against boys. *Atlantic Monthly, 285,* 59–70.

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

Spencer, S. J., Steele, C. M., & Quinn, D. M. (2002). Stereotype threat and women's math performance. In A. E. Hunter & C. Forden (Eds.), *Readings in the psychology of gender: Exploring our differences and commonalities* (pp. 54–68). Needham Heights, MA: Allyn & Bacon.

Steele, C. M. (1997). A threat in the air. *American Psychologist, 52,* 613–629. doi:10.1037/0003-066X.52.6.613

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press. doi:10.1016/S0065-2601(02)80009-0

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693. doi:10.1111/j.1559-1816.2004.tb02564.x

Stricker, L. J., & Ward, W. C. (2008). Stereotype threat in applied settings re-examined: A reply. *Journal of Applied Social Psychology, 38,* 1656–1663. doi:10.1111/j.1559-1816.2008.00363.x

Tenenbaum, H. R., & Leaper, C. (2003). Parent–child conversations about science: The socialization of gender inequities? *Developmental Psychology, 39,* 34–47. doi:10.1037/0012-1649.39.1.34

Tice, D. M., Baumeister, R. F., Shmueli, D., & Muraven, M. (2007). Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology, 43,* 379–384. doi:10.1016/j.jesp.2006.05.007

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92,* 144–151. doi:10.1037/0022-0663.92.1.144

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39,* 456–467. doi:10.1016/S0022-1031(03)00019-2

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89,* 451–469. doi:10.1037/0022-0663.89.3.451

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40,* 424–431. doi:10.1016/j.jesp.2003.10.001

Received June 13, 2012
Revision received January 9, 2013
Accepted January 25, 2013 ■