Journal of Educational Psychology
2007, Vol. 99, No. 3, 545–560

0022-0663/07/$12.00 DOI: 10.1037/0022-0663.99.3.545

Stereotype Threat Among Schoolgirls in Quasi-Ordinary Classroom Circumstances

Pascal Huguet
Centre National de la Recherche Scientifique and Aix-Marseille University

Isabelle Régner
Centre National de la Recherche Scientifique and University Toulouse-Le-Mirail

There is ample evidence today in the stereotype threat literature that women and girls are influenced by gender-stereotyped expectations on standardized math tests. Despite its high relevance to education, this phenomenon has not received much attention in school settings. The present studies offer the 1st evidence to date indicating that middle school girls exhibit a performance deficit in quasi-ordinary classroom circumstances when they are simply led to believe that the task at hand measures mathematical skills. This deficit occurred in girls working alone or in mixed-gender groups (i.e., presence of regular classmates) but not in same-gender groups (i.e., presence of only same-gender classmates). Compared with the mixed-gender groups, the same-gender groups were also associated for girls in the stereotype threat condition with greater accessibility of positive role models (i.e., female classmates who excel in math), at the expense of both stereotypic in-group and out-group members (i.e., low-math-achievement girls and high-math-achievement boys). Finally, the greater accessibility of positive role models mediated the impact of the activated stereotype on girls' performance, exactly as one would expect from C. M. Steele's (1997) stereotype threat theory. Taken together, these findings clearly show that reducing stereotype threat in the classroom is a crucial challenge for both scientists and teachers.

Keywords: stereotype threat, schoolgirls, classroom, visual memory, ROCF task

Women all over the world have participated (as mathematicians, physicists, astronomers, and so on) in unraveling the secrets of nature. Yet they remain underrepresented in mathematically intensive disciplines and careers such as the natural and physical sciences and engineering. In Western Europe in 2003, for example, 35.7% of university graduates in mathematics, natural and physical sciences, and computer science were women (averaged over the 15 member states hereafter referred to as *EU-15*; see European Research Council, 2003). Similarly, among engineering graduates, women were in the minority (20.2%) across the EU-15. Women outnumbered men only in the field of education and were more or less equal in number of graduates to men in fields such as the humanities and the arts, and health and social services. Women were underrepresented in science-related activities and careers as well, whether among senior university staff (34%), as members of scientific boards (27%), and within the business sector (15%). Finally, women applicants from EU-15 were slightly but consistently less successful than men in receiving funding for their scientific projects. The fate of women scientists in the United States is roughly the same (see National Science Foundation, 2004). In 2001, they accounted for more than half of all graduate students in science fields like psychology (74%), biology (54%), and the social sciences (52%), but they were in the minority in mathematics (35%), physical sciences (30%), computer science (30%), and engineering (20%). Women made up about 26% of employed science or engineering doctorate holders in 2001. Thus, men still predominate in scientific institutions both in the United States and abroad.

Mathematics has been identified as the critical filter that prevents women from gaining access to the hard sciences and related occupations (Sells, 1973). In line with this, the existence of gender differences favoring men on standardized math tests (especially word problems and geometry items from the math portion of the SAT) has frequently been a topic of debate.

GENDER DIFFERENCES ON STANDARDIZED MATH TESTS: FROM BIOLOGY TO SOCIALIZATION

Relying on very large samples (thousands) of intellectually talented adolescents (12- to 14-year-olds), Benbow and Stanley (1980) reported that boys outperformed girls by about half a standard deviation (0.40) on the math portion of the SAT. Although they attributed this difference to "superior male mathematical ability, which may in turn be related to greater male ability in spatial tasks" (p. 1264), they also noted that this superiority was "probably an expression of a combination of both endogenous and exogenous variables" (ibid). Benbow and Stanley (1983) later concluded that boys predominate in the highest ranges of mathematical reasoning ability before they enter adolescence (with a male-to-female ratio of 13:1 in the group that scored above 700).

Pascal Huguet, Cognitive Psychology Laboratory, UMR CNRS 6146, Centre National de la Recherche Scientifique; and Aix-Marseille University, Marseille, France. Isabelle Régner, Social Psychology Laboratory, UMR CNRS 5551, Centre National de la Recherche Scientifique; and University Toulouse-Le-Mirail, Toulouse, France.

This research was funded by Grant JC6082 from the French Ministry of Higher Education to both authors.

Correspondence concerning this article should be addressed to Pascal Huguet, Université Aix-Marseille 1, 3 Place Victor Hugo, Case D, 13331 Marseille Cedex 3, France. E-mail: huguet@up.univ-mrs.fr

Although they again acknowledged that the "reasons for this sex difference are unclear" (p. 1031), Benbow and Stanley (1983) suggested sexual differentiation of the brain as a possible candidate (relying on Guy & McEwen, 1980; see also Benbow, 1988; Benbow & Benbow, 1984).[1]

Eccles and Jacobs (1986) challenged this conclusion. Gender differences on standardized math tests, they claimed, are mostly rooted in stereotypic gender-role beliefs that parents (and teachers) communicate to their children (and students) on a daily basis, resulting in differential expectations, confidence, and attitudes toward math in boys and girls (see also Eccles, Adler, & Meece, 1984; Eccles, Jacobs, & Harold, 1990; Jacobs & Eccles, 1992). Consistent with this, Eccles et al. (1990) showed that parents' beliefs about their daughter's versus son's math abilities were related subsequently to their child's math self-efficacy, identification with math, and math performance.

Hyde, Fennema, and Lamon (1990) found that gender differences favoring boys on standardized math tests did not emerge until the high school years. Their meta-analytic findings (100 studies, 254 independent effect sizes) also helped clarify the overall picture. For example, girls were superior to boys in arithmetic, and there were no significant gender differences in the understanding of mathematical concepts per se. Likewise, using meta-analytic findings from the 1970s as a baseline, Hyde et al. showed that the magnitude of gender differences varied over time (see also Feingold, 1988), which is hard to accommodate with any biological account (for a similar argument regarding spatial math ability, see Halpern, 1992). As reported by Hyde et al., however, gender differences favoring men grew larger with increasingly selective samples and were the largest for highly selective samples and for samples of gifted persons (no significant gender differences were found among adults or children taken from the general population).

Hedges and Nowell (1995) showed that the male advantage among the higher ability samples was not an artifact caused by biased sample selection and reported a higher proportion of boys at both extremes of the distribution.[2] More evident among the better high school students, the male advantage is in fact minimal on easy test items but increases as the items become more difficult, even when gender differences in variability are controlled (Penner, 2003). Combined with the fact that spatial aptitude (i.e., mental rotation of three-dimensional objects) mediates gender differences in math (Casey, Nuttall, & Pezaris, 1997; see also Benbow, 1988), this observation suggests that the gender differences in math ability, if any, only show up on difficult material.

STEREOTYPE THREAT AS AN ALTERNATIVE EXPLANATION TO BIOLOGY AND SOCIALIZATION

However, it is precisely in this situation (i.e., when they are faced with difficult math tests) that women who are good at math may feel threatened by the possibility that their performance will confirm—to others, to themselves, or both—the negative stereotype about their gender's math abilities (Spencer, Steele, & Quinn, 1999; Steele, 1997). This threat unfortunately leads to poorer performance and thus produces the expected negative outcome.

In Spencer et al.'s (1999) studies, for example, women with high math ability performed less well (than equally qualified men) on difficult math tests both when they were told that the test produced gender differences and when that information was not given, but performed as well as men when told that no gender differences had been found (for similar findings, see Johns, Schmader, & Martens, 2005; Keller & Dauenheimer, 2003; O'Brien & Crandall, 2003; Quinn & Spencer, 2001; Sekaquaptewa & Thompson, 2003; for a review, see Ben-Zeev, Duncan, & Forbes, 2005). The very fact that falsifying the gender stereotype about math not only reduced the male advantage but eliminated it altogether runs counter to any biological account of gender differences in this domain. In Cadinu, Maass, Rosabianca, and Kiesner's (2005) study, women who were told that the math test they were about to take produced gender differences engaged in negative math-related thoughts (e.g., "These exercises are too difficult for me") that were associated with poorer performance (compared to women who were told that no gender differences had been found). The negative thoughts mediated the impact of the activated stereotype on math performance, exactly as one would expect from Steele's (1997) stereotype threat theory (STT; see also Steele & Aronson, 1995).

According to STT, women and minority-group members are expected to experience additional tension—over and above that associated, for most people, with taking difficult tests—because they are preoccupied by fears of confirming a negative stereotype. From a purely cognitive point of view, such a preoccupation might reduce working memory capacity, which is critical to performing well on complex intellectual tasks (Kane et al., 2004). Schmader and Johns (2003) showed that stereotype threat was indeed associated with a lower working memory capacity (compared with a no-threat condition), which in turn led to lower math performance in women.

The fact that gender differences favoring men on standardized math tests are confined to the higher ability samples is also clearly consistent with STT. As noted by Steele (1997), susceptibility to stereotype threat derives not from internal doubts about one's ability based on one's history of failure and/or the internalization of the stereotype under the influence of socialization, but from one's identification with the critical domain and the resulting concern about being stereotyped in that domain. To the extent that women who excel in math identify strongly with this domain—in the sense that they perceive math as self-relevant—stereotype threat is expected to be especially prominent in women from higher ability samples. One may argue that the persistent presence of women in these selective samples is inconsistent with STT (as well as with biological accounts). However, if STT is supported, then women who excel in math would do even better in stereotype-free environments. Having to face stereotype threat might, over

[1] According to Benbow (1988), at least three physiological factors are relevant to understanding gender differences in mathematical reasoning ability, namely (a) left-handedness and (b) symptomatic atopic disease (allergies), which may be related to bihemispheric representation of cognitive functions or to the impact of prenatal testosterone exposure; and (c) myopia. As noted by Benbow, "these physiological factors, especially prenatal testosterone exposure, lend credence to the view that sex differences in extremely high mathematical reasoning ability may be, in part, physiologically determined" (p. 180).

[2] The National Science Foundation (2006) recently released a report essentially declaring that the gender gap on standardized math tests has disappeared, but the related statistics did not focus specifically on highly selective samples.

time, lead women to disidentify with math as an important domain, that is, "avoid or drop the domain as an identity or basis of self-esteem" (Spencer et al., 1999, p. 6). This additional facet of the theory helps explain why women remain underrepresented in mathematically intensive disciplines and careers.

Thus, there are ample reasons to believe today that gender differences favoring men on standardized math tests are due to the temporary intervention of a negative stereotype, as suggested by STT.

STEREOTYPE THREAT IN CHILDREN

Interestingly, there is also evidence that this negative stereotype operates in children. Not only do children become aware of their gender identity at a very young age (Aboud, 1988; Hirschfeld, 1996; Huston, 1987), they also believe—as early as first grade—that boys are better than girls in math-related areas (Lummis & Stevenson, 1990; see also Eccles, Wigfield, Harold, & Blumenfeld, 1993; Wigfield et al., 1997). This belief is not necessarily erroneous, as differences favoring boys on standardized math tests have sometimes been found, especially among math-talented children (Mills, Ablard, & Stumpf, 1993; Robinson, Abbott, Berninger, & Busse, 1996). Again, however, the fact that the boys' advantage expresses itself more clearly when highly selective samples are retained is consistent with STT.

Direct evidence of stereotype threat in girls facing a math test was found by Ambady, Shih, Kim, and Pittinsky (2001). In the laboratory, Ambady et al. assigned Asian American girls from the lower (kindergarten to Grade 2, aged 5–7) and upper (Grades 3–5, aged 8–10) elementary school grades and from middle school (Grades 6–8, aged 11–13) to one of three conditions: Asian identity activated, female identity activated, and no identity activated (control). In the ethnic-identity condition, the younger children (kindergarten to Grade 2) were simply asked to color a picture depicting two Asian children eating with chopsticks out of rice bowls. In the gender-identity condition, the picture showed a girl holding a doll. In the control condition, it was a landscape. For the two groups of older children, the identity manipulation was made by asking them to answer questions related either to their ethnicity (e.g., whether they had any non-Caucasian/non-White classmates) or their gender (e.g., whether most of their friends were boys or girls). Those in the control group answered neutral questions (e.g., favorite season). All children were then given math-related questions from a standardized test designed for their particular grade. Both lower elementary school and middle school girls performed significantly worse on the 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. Although the upper elementary school girls performed best in the gender-identity condition than in the other two conditions, Ambady et al.'s findings indicate that girls can be influenced by the negative gender stereotype quite early in their cognitive development.

THE PRESENT STUDIES

Despite its high relevance to education (particularly educational equality), the stereotype threat phenomenon has not received much attention in real classrooms, especially among children. Like most stereotype threat research (for reviews, see Maass & Cadinu, 2003; Steele, Spencer, & Aronson, 2002), past studies on gender and math performance has tested participants (generally young adults) in the laboratory, individually or in small groups (see Ben-Zeev et al., 2005). Some studies (Good, Aronson, & Inzlicht, 2003; Keller, 2002; Keller & Dauenheimer, 2003; Massey, Charles, Lundy, & Fischer, 2003) were exceptions. None of them, however, can be taken as direct evidence that girls are influenced by stereotype threat in ordinary classroom circumstances.

In Good et al.'s (2003) study, adolescent girls who were encouraged by college students to view intelligence as malleable and/or to attribute academic difficulties to the novelty of the educational setting earned significantly higher math standardized test scores than girls in the control condition (where these encouragements were not made). This interesting finding suggests that stereotype threat is the default mindset for female students in math classes and shows how to change this mindset. But it cannot be taken as direct evidence that girls (especially before adolescence) facing math in ordinary classroom circumstances are affected by stereotype threat.

Likewise, Massey et al.'s (2003) ambitious survey offered evidence that stereotype threat is a significant component of the academic underachievement of low-status group members. This survey focused on African Americans and Latinos, however, and did not specifically address the question of stereotype threat among girls, a fortiori children in math classes.

Studies by Keller (2002) and Keller and Dauenheimer (2003) indicated that young women (*M* age = 18.7, for the 2002 study) and adolescent girls (*M* age = 15.7, for the 2003 study) were influenced by stereotype threat in the classroom setting. However, in the stereotype threat (vs. control) condition, participants were told that the math test had (vs. had not) been shown to produce gender differences, which is not what teachers typically tell their students. In Keller's study, not only was this gender difference information given, but participants in the stereotype threat condition also learned explicitly that men had outperformed women in past research. So one can still wonder whether the negative gender stereotype operates in the school context under more ordinary circumstances (both when the other classmates are physically present and when the stereotype is not explicitly activated), especially with children or girls younger than those tested by Keller and Dauenheimer.

There is little doubt that this stereotype can be subtly or implicitly activated (see Smith, 2002), even in children (Ambady et al., 2001). In Ambady et al.'s laboratory study, however, the negative stereotype was activated directly via tasks that made it salient, rather than indirectly as is probably the case in the school context. McKown and Weinstein (2003) offered evidence, in this context, of indirectly induced stereotype threat (via task characterization) among children from stigmatized ethnic groups, but this finding does not tell us about the ultimate fate of schoolgirls in similar circumstances.

Clearly, there is today no direct evidence that stereotype threat occurs among schoolgirls in ordinary classroom circumstances. Not only is direct evidence clearly needed, but there are some reasons to believe that stereotype threat does not necessarily generalize to test settings outside the laboratory. As noted by Wicherts, Dolan, and Hessen (2005), only a few studies have looked into the debilitating effects of stereotype threat on test performance in test settings high in ecological validity and/or settings with consequential test outcomes. And stereotype threat effects were generally small or nonexistent in these studies (despite the large and representative samples used). Cullen, Hardison, and Sackett (2004), for example, relied upon the differential predictions procedure that is the gold standard for examining test bias and found no support for stereotype threat effects in educational and military job settings. Likewise, Stricker and Ward (2004) conducted two field studies within an actual high-stakes test situation but were unable to replicate the strong negative effects of asking for biographical information prior to taking a test (i.e., group prime; see Steele & Aronson, 1995) on minority and female test performance. In addition, three recent lab experiments addressed the effects of stereotype threat on Blacks' test performance in a job selection context (McFarland, Lev Arey, & Ziegert, 2003; Nguyen, O'Neal, & Ryan, 2003; Ployhart, Ziegert, & McFarland, 2003). In these studies, test-taking motivation was enhanced by the promise of financial rewards for high test scores. Despite the use of manipulations with well-established effects (i.e., race prime and test diagnosticity), the debilitating effects of stereotype threat on minority test performance were generally absent. Sackett (2003) suggested that these results imply that the generality of stereotype threat effects to motivational job selection contexts is limited. Likewise, Stricker and Ward suggested that their studies indicate that high test stakes appear to be capable of overriding the negative effects of stereotype threat on test performance.

The two studies reported in the present article tested whether girls (10–12 years old) were influenced by stereotype threat in quasi-ordinary classroom circumstances (i.e., as close as possible to normal classroom conditions). More specifically, Study 1 looked at whether stereotype threat occurred among middle school girls in their natural school environment when they were simply led to believe (erroneously) that the task they would perform measured mathematical skills. As such, this study also tested the possibility that stereotype threat among schoolgirls encompasses a broader set of intellectual tasks rather than just math tests, which would make it even more problematic than previously assumed. We used a modified version (see the Appendix) of the Rey-Osterrieth Complex Figure (ROCF) recall memory task (Osterrieth, 1944; Rey, 1941) that is still in widespread use today for neuropsychological assessment (of adults as well as children), especially when perceptual organization, visuospatial constructional ability, planning, and visual memory are being evaluated (e.g., Akshoomoff & Stiles, 1995). The complex figure, which Rey developed for administration to individuals with brain damage, is a two-dimensional line drawing that has no particular meaning and is difficult to encode verbally (Waber, Bernstein, & Merola, 1989). The ROCF task was particularly useful for our purposes here because it can be presented as either a geometry test, a memory game, or a drawing test (Huguet, Brunot, & Monteil, 2001), at least with children.[3] It is therefore well suited for subtly or indirectly activating the negative gender stereotype. This task is also thought to tap a variety of skills that are important or even essential components of academic performance (Kirkwood, Weiler, Bernstein, Forbes, & Waber, 2001). As noted by Kirkwood et al., successful encoding of the complex figure depends on the intactness of fundamental visual–perceptual and visual–spatial skills, as well as on cognitive and metacognitive processes such as attention, organization, and strategy use. Here, we hypothesized that at least some of the fundamental skills and processes involved in the ROCF task would be temporarily disrupted under stereotype threat, resulting in girls' poorer performance when the task was characterized as a geometry test rather than in a stereotypically irrelevant way (e.g., as a memory game). Whereas Study 1 tested this hypothesis with a selective sample of students working alone on the ROCF task, Study 2 tested it under conditions close to ordinary classroom tests (i.e., classmates present), with students from the general population.

Not only did Study 2 test whether schoolgirls were affected by stereotype threat in these more ordinary circumstances, but it also varied the gender composition of the classroom (mixed-gender vs. same-gender setting). Several laboratory findings have suggested that gender composition of the performance setting indeed matters for stereotype threat to occur. Inzlicht and Ben-Zeev (2000, Experiment 1) found that college women's math performance was significantly lower when they were asked to take the test in a mixed-gender setting with a majority of men (2 male participants and 1 female participant) than when they worked in a same-gender setting (groups of 3 female participants; for conceptually similar results, see Sekaquaptewa & Thompson, 2003). Being outnumbered by men, Inzlicht and Ben-Zeev (2000) reasoned, probably led women to increased awareness of their gender group and related stereotypes (which is clearly consistent with numerous findings in the self-concept and stereotype literature). Also consistent with this account, Inzlicht and Ben-Zeev (2000, Experiment 1) showed that, in this same condition (being outnumbered by men), women's performance in a nonstereotyped domain (a verbal test) was unaffected. Similarly, when men took a math test while being outnumbered by women (Inzlicht & Ben-Zeev, 2000, Experiment 2), their test performance was the same as when they took the test in the presence of other men. More recently, Inzlicht and Ben-Zeev (2003) showed that women's math-performance deficit in mixed-gender minority conditions occurred in public as well as private environments, indicating just how far-reaching the effects of stereotype threat can be. Gender composition therefore deserves special attention when the impact of the negative gender stereotype is being tested in math classes.

The negative gender stereotype, we reasoned, should be more salient for girls when both male and female classmates are physically present than when only same-gender classmates are involved, resulting in a smaller (or even null) stereotype threat effect in the latter setting. Note here that the effects possibly related to the gender composition of the classroom were neglected in past stereotype threat research within the school context (see previous section). Furthermore, the fact that there are numerous lab findings in favor of our gender composition hypothesis does not make it trivial. As noted earlier in this article, there are some reasons to believe that what has been discovered about stereotype threat in

[3] In their study, Huguet, Brunot, and Monteil (2001) modified Rey's complex figure to increase its complexity and remove certain components (i.e., two Christian crosses and one circle including three dots evoking a human face). Not only could these components be encoded more easily, but they were not appropriate for characterizing the task as a geometry test. Only schoolboys participated in Huguet, Brunot, and Monteil's (2001) study, which offered preliminary evidence that the ROCF task characterization (geometry vs. control) makes a difference in students' recall performance (it was found to depend on their academic history in math).

the context of the laboratory does not necessarily generalize to test settings high in ecological validity and/or to settings with consequential test outcomes, such as the classroom. In addition, it should be pointed out that the use of the ROCF task does not undermine the authenticity of the present experiments. In France, children aged 10 to 12 years old are frequently faced with this type of task in their classroom both when they are faced with the national standardized evaluation (the ROCF task comes close to many subtests of this evaluation, which takes place in the first year of secondary school) and in the course of regular classroom exercises (the ROCF task is quite similar to several exercises from the geometry section of students' math books).

Finally, the psychological correlates of gender composition were explored in light of Marx, Stapel, and Muller's (2005) recent findings on stereotype threat and the salience of successful role models. In their study, describing a math test as being diagnostic of math ability (vs. reasoning ability) made women's membership in the stereotyped group (women) especially salient. Of particular interest here is the fact that for women in the stereotype threat condition, learning about another woman who excelled in math neutralized the negative effects of the threatening gender stereotype (see also Marx & Roman, 2002), exactly as one would expect if female participants had engaged in upward-comparison assimilation (see Mussweiler & Strack, 2000; Stapel & Suls, 2004).

Based on these and the gender-composition findings, we assumed that increased salience of positive role models would be precisely what would protect women and girls against the negative gender stereotype when working on math tests in same-gender settings rather than in mixed-gender ones. In the stereotype threat condition (geometry test), we predicted that the presence of only same-gender classmates would lead girls to focus more exclusively on positive role models (i.e., female classmates who excel in math), possibly at the expense of both stereotypic in-group and out-group members (i.e., low-math-ability girls and high-math-ability boys).

Study 1

Method

Participants

Two weeks before the start of the study, two math teachers from two French public middle schools selected 40 (among 172) of their students (20 girls and 20 boys) in sixth or seventh grade (ages 11–13). The selection was made on the basis of how students evaluated their actual standing in math relative to their classmates (from 1 = *much worse* to 5 = *much better,* including 3 = *the same*), hereafter called *comparative evaluation*. Only those who chose Points 4 and 5 on the comparative evaluation scale (M = 4.47, SD = 0.51) were retained for the study, resulting in a very selective sample. Consistent with their comparative evaluations, these students' average math grade (at Trimester 2) was equal to 16.25 (SD = 2.43) on a scale ranging from 0 to 20, which in France is a very high grade. Students were also asked to indicate the importance they personally attributed to math (on a scale ranging from 1 = *not important* to 5 = *extremely important*). The stereotype threat literature typically assumes that high performance in a domain is equivalent to being strongly identified with the domain. Consistent with this, the present high-math-ability participants rated math as very important for them personally (M = 4.75, SD = 0.44).[4] No gender differences were found on these measures. All parents allowed their children to participate in a study focusing on "children's behavior at school."

Procedure

The children were met individually by a same-gender experimenter in one of their regular classrooms. They were informed that (a) they had been chosen at random among a larger sample; (b) the session would include a "geometry test" (for half of them) versus a "memory game" (for the other half) on which they might do well or poorly, followed by a brief questionnaire; and (c) their data would be kept confidential (they were told not to write their names on the answer sheet or questionnaire). The participants were told that the task they would be given might help develop a new geometry test for a textbook (geometry condition) versus a new game for a fun magazine (memory game condition). Although they were given the opportunity to not participate if they so desired, all students chose to continue with the session. Within each gender, students were assigned at random to one of the two conditions (geometry vs. memory game). They were systematically and explicitly asked to take the test (or game) very seriously and to put in as much effort as they could. After answering any questions, the experimenter asked the student to get ready for the first part of the test or game (figure encoding), put the figure on the student's desk, and left the room for exactly 50 s. He or she then returned to the room, took the figure, and gave the student a pencil and a sheet of paper for the second part (figure reproduction from memory). Then the experimenter left the room again and allowed 5 min for reproducing the figure (which was sufficient, as a pretest had indicated).

In all conditions, the students were then asked to rate the extent to which they agreed with the idea that what they did was a scholastic test (on a scale ranging from 1 = *completely disagree* to 5 = *completely agree*), hereafter called *manipulation check*. All students also rated the degree to which the task was difficult and interesting (on scales ranging from 1 = *not at all* to 5 = *extremely*), hereafter called *self-reports*. The children were debriefed and thanked afterwards.

As suggested by Rey and Osterrieth themselves, recall performance was measured in terms of both the number and quality of the units reproduced from the complex figure (there were 22 units in the present version). The scoring was done by two independent judges blind to the experimental conditions and the students'

[4] However, we agree with an anonymous reviewer that high performance in a domain is not necessarily equivalent to being strongly identified with the domain. High math achievers may sometimes perceive math as important in their parents' or teachers' eyes, more than for themselves personally (Goethals & Darley, 1987). As shown by Huguet, Brunot, and Monteil (2001), low math achievers sometimes remain strongly identified with math, at least temporarily, perhaps for similar reasons (parents' and teachers' influences). Whatever the underlying process, however, the importance ratings indicated whether the domain mattered in some ways. And the more a domain was rated as important, the more self-related threats due to one's personal history of failure or one's group's bad reputation in this domain might have been salient for the students.

gender ($\alpha = .98$). Two points were given if the unit was correct and properly positioned, 1 if it was either altered but correctly placed or not altered but incorrectly placed, 0.5 if it was altered and in the wrong place, and 0 if it was missing or unrecognizable. The best possible score was 44 (22 units × 2 points).

Results

The assumptions of all statistical analyses presented in this article were systematically tested. Only significant deviations to these assumptions (when any) are reported.

Manipulation Check

As indicated by a 2 (student gender) × 2 (task characterization) analysis of variance (ANOVA), the students perceived the task as a scholastic test to a greater extent in the geometry condition ($M = 3.10$, $SD = 1.07$) than in the memory game condition ($M = 2.15$, $SD = 1.09$), $F(1, 36) = 7.37$, $p < .01$, $\eta^2 = .07$. No other effects were found.

Recall Performance

The students' overall recall score was only 21.66 points ($SD = 5.22$) out of a maximum of 44, indicating that the task was difficult for them. As expected, the Student Gender × Task Characterization interaction was significant, $F(1, 36) = 8.89$, $p < .006$, and was associated with a large effect size ($\eta^2 = .20$). The present pattern of means (see Figure 1) was beyond our expectations. As revealed by simple contrasts, girls performed less well (i.e., recalled fewer units) than boys in the geometry condition, $F(1, 36) = 3.92$, $p < .05$, $\eta^2 = .10$, whereas they outperformed the boys in the memory game condition, $F(1, 36) = 4.99$, $p < .03$, $\eta^2 = .12$. In addition, whereas boys performed better in the geometry condition than in the memory game condition, $F(1, 36) = 6.19$, $p < .02$, $\eta^2 = .15$, the opposite effect—though marginally significant—was obtained for girls, $F(1, 36) = 2.99$, $p < .09$, $\eta^2 = .08$.

Self-Reports

The 2 × 2 ANOVA mentioned above yielded no effects. Regardless of their gender and assigned condition, students perceived the task as both difficult ($M = 3.87$, $SD = 0.76$) and interesting ($M = 3.42$, $SD = 0.71$). Single-sample t tests showed that these ratings differed significantly ($ps < .001$) from the midpoint (3) of the rating scale.

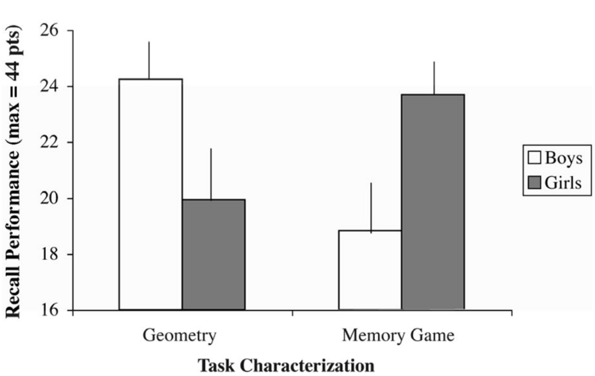

Figure 1. Mean recall performance as a function of student gender and task characterization. Error bars represent standard error.

Discussion

Although very simple, Study 1 gives rise to some exciting results. First of all, it provides initial evidence of stereotype threat in schoolgirls (at least in those who excel in math) who were simply (but erroneously) led to believe that the task being performed measured mathematical skills. The negative gender stereotype was activated here both implicitly and indirectly by means of task characterization, rather than explicitly via gender-related information or directly via preliminary tasks that made it salient. This implicit as well as indirect activation was sufficient for a performance deficit to occur in the stereotype threat condition.

Whereas girls underperformed relative to boys in the geometry condition, they outperformed them in the memory game condition, which was unexpected. Although completely speculative, one explanation for this finding is that boys were influenced by a stereotype favoring girls in the memory domain (for evidence that men can be affected by stereotypes favoring women, see Leyens, Désert, Croizet, & Darcis, 2000). However, this stereotype, if there is any, is not as strong or as salient in our society as the one favoring men in math, which runs against the strength of boys' underperformance effect in the memory game condition. Perhaps boys were simply more reluctant to expend effort on the ROCF task when characterized as a memory game (compared with when characterized as a geometry test), because this characterization is not (or is less) relevant in the school context. As shown by the self-report findings, however, participants in both conditions rated the task as interesting. Stereotype lift is another explanation. According to Walton and Cohen (2003), *stereotype lift* is a performance boost that occurs in members of nonstereotyped groups when they are made aware that an outgroup is negatively stereotyped, or simply (as in the geometry condition here) when the task performed triggers the negative stereotype. Walton and Cohen (2003) suggested that negative out-group stereotypes improve performance by encouraging downward comparisons with a denigrated outgroup. By comparing themselves with a socially discredited group, nonstereotyped individuals may experience increased feelings of self-efficacy (Bandura, 1986), which may in turn improve their performance, especially on difficult, self-relevant tests requiring persistence in the face of frustration. In line with this, the present data indicate that the task was perceived as both difficult (as shown by the difficulty ratings) and self-relevant (as shown by the importance ratings). Although this study was not specifically designed to test stereotype lift, the boys' performance in the two characterized conditions was entirely consistent with this effect.

These preliminary findings leave a crucial question unanswered, however: Can stereotype threat be found in schoolgirls under more ordinary classroom circumstances? The fact that girls exhibited a performance deficit merely because they believed that the task measured mathematical skills is one thing. Whether this deficit occurs under conditions close to those found in an ordinary classroom is another, especially if the gender composition of the classroom is considered. As suggested earlier, there are good reasons to predict that, compared with a mixed-gender setting, the presence of only same-gender classmates will be associated with a smaller if not null stereotype threat effect. Likewise, in the threatening (geometry) condition, the salience of successful role models should be higher in a same-gender setting than in a mixed-gender setting. This possibility was explored in Study 2 by means of what is hereafter referred to as the *nomination task*.

Study 2

Method

Participants

The participants were 454 students (223 girls and 231 boys, aged 11–13) from nine French public middle schools. These schools were selected because of their representativeness (in terms of social class and math achievement-test scores) of the student population of this age. Not surprisingly, the participants' math grades ($M = 12.38$, $SD = 3.54$) as well as their comparative evaluations in this discipline ($M = 2.99$, $SD = 0.80$) were below those of the highly selective sample in Study 1. Examining these preliminary data further, we found a significant main effect of student gender on the comparative evaluations in math when math grades were controlled, $F(1, 437) = 15.74$, $p < .001$, $\eta^2 = .04$. Consistent with the negative math-related gender stereotype, girls ($M = 2.87$, $SE = 0.04$) rated their relative standing in this subject less positively than did boys ($M = 3.12$, $SE = 0.04$). This bias did not exist in the students' comparative evaluations in drawing, $F(1, 438) = 1.86$, $p = .17$, which were taken as a control and measured during the same time period (while controlling for drawing grades). No gender difference was found in the importance ratings students attributed to math. Overall, these ratings remained high ($M = 3.92$, $SD = 0.91$). As in Study 1, these two measures (i.e., comparative evaluations and importance) were taken before the study began. All parents allowed their children to participate under the same premise as used before.

Procedure

Students were met collectively by two experimenters (one man and one woman) in one of their regular classrooms. They were informed that (a) their class had been chosen at random from among a larger sample; (b) the session would include a geometry test (vs. a drawing test) on which they might do well or poorly, followed by a brief questionnaire; and (c) their data would be kept confidential. Although they were given the opportunity to not participate if they so desired, all students chose to continue with the session. Then the experimenters divided each class into two mixed-gender or two same-gender subgroups of 10 to 14 students. Half of the students in each class remained in the same room with one of the two experimenters, and the other half moved to an adjacent room with the other experimenter (in the mixed- as well as same-gender subgroups, the experimenter's gender did not necessarily match that of participants). The reason for the move was said that it would simply be easier to carry out the upcoming activity. In order to minimize suspicion possibly related to the way each class was divided, the students in the mixed-gender groups were told that group assignment had been done at random; those in the same-gender groups were told that it was simply a much faster way to divide the students than any other solution. Students were seated separately to prevent cheating and were told again, to make it very clear, that the task was designed to evaluate their "ability in geometry" (or their "ability in drawing"). Drawing was used in Study 2 rather than memory game to see whether the findings of Study 1 could be replicated when the stereotype-irrelevant condition implied an activity not only assumed to be nondiagnostic of math-related skills but also not explicitly related to students' memory ability. Furthermore, drawing is still part of the curriculum (in France, drawing is a mandatory subject in these grades), which may have helped prevent a motivation-loss phenomenon, if any. Because the task was administered collectively, students were given more time (90 s) than in Study 1 (50 s) to encode the complex figure. As before, however, they had 5 min to reconstruct it from memory. The experimenter remained in the room throughout the encoding and recall phases, during which he or she simply sat at the teacher's desk and watched the entire group silently.

Students then rated the extent to which they perceived the task as a valid measure of drawing ability versus geometry ability (manipulation check), on a single scale ranging from 1 (*valid test of drawing*) to 5 (*valid test of geometry*), including 3 (*valid test of both*). As before, they also rated the degree to which the task was difficult and interesting (self-reports). Then they estimated their task-specific effort (on a scale ranging from 1 = *none at all* to 5 = *a lot*). This additional item was included as a means of knowing whether the stereotype threat and stereotype-lift effects (if any) were associated with changes in self-reported effort.

Finally, on the last page of the questionnaire, and whichever way the task was characterized, students were asked to nominate the highest and lowest math achievers in their class, including themselves if they desired, listing up to three names in each category (nomination task). As noted earlier, the nomination task was used as a means of finding out whether girls' salience of positive role models (i.e., high-math-ability female classmates) changed as a function of task characterization and gender composition. Recall that, under stereotype threat, positive-role salience was expected to be higher in the same-gender context than in the mixed-gender context. If this were the case, then the former setting should be associated for girls with a higher tendency (compared to the mixed-gender setting) to nominate high-math-ability classmates from their own gender group. As we see later, this is exactly what happened.

Results

Manipulation Check

The 2 (student gender) × 2 (task characterization) × 2 (gender composition) ANOVA yielded only a significant main effect of task characterization, $F(1, 443) = 18.39$, $p < .001$, $\eta^2 = .04$. As expected, students perceived the task as a better measure of their geometry ability in the geometry condition ($M = 3.45$, $SD = 0.90$) than in the drawing condition ($M = 3.11$, $SE = 0.85$). Single-sample t tests showed that the students' ratings of the task as a geometry test differed significantly from the midpoint (3) of the scale only in the geometry condition ($p < .001$), although this difference was marginally significant in the drawing condition ($p < .06$).

Recall Performance

As in Study 1, task performance was scored by two independent judges ($\alpha = .97$). The overall score was 25.08 points ($SD = 6.34$)

versus 21.66 in Study 1, which was not surprising given that more time was allowed for encoding. Students' math grades correlated significantly with their recall performance, $r(454) = .29$, $p < .001$. Although this correlation did not occur in Study 1, $r(40) = .06$, perhaps due to low variance in grades, it was significant or marginally significant in each Study 2 condition (*rs* ranged from .19 to .42) but one: the mixed-gender setting among girls who were faced with the task characterized as a geometry test, $r = .05$, *ns*, which was the most self-threatening condition for these participants. All subsequent analyses were thus performed using math grades as a covariate (Aronson, Fried, & Good, 2002). Moreover, this technique made the present analysis more comparable to that conducted by Keller and Dauenheimer (2003) on a student sample taken from the general population. Using their participants' math grades as a covariate, these authors showed that stereotype threat was not necessarily limited to the highest math achievers, at least in a school setting. Note that testing this important aspect of Keller and Dauenheimer's findings on younger students was an integral part of the present study.

The 2 (student gender) × 2 (task characterization) × 2 (gender composition) analysis of covariance (ANCOVA)[5] revealed a main effect of gender composition, $F(1, 442) = 34.54$, $p < .001$, $\eta^2 = .07$, indicating that students performed better (i.e., recalled more units) in the same-gender condition ($M = 26.40$, $SE = 0.36$) than in the mixed-gender condition ($M = 23.09$, $SE = 0.43$). This effect, however, was qualified by a two-way interaction between student gender and task characterization, $F(1, 442) = 8.53$, $p < .004$, $\eta^2 = .02$, itself qualified by a three-way interaction between student gender, task characterization, and gender composition, $F(1, 442) = 4.14$, $p < .04$, $\eta^2 = .01$ (see Figure 2). Because stereotype threat was expected to occur to a greater extent, if not solely, in the mixed-gender setting, we broke this higher order interaction down by gender composition. As expected, whereas the Student Gender × Task Characterization interaction was clearly significant in the mixed-gender setting, $F(1, 176) = 11.62$, $p < .001$, $\eta^2 = .06$, it was not in the same-gender setting, $F(1, 266) = 0.46$, $p = .50$, $\eta^2 = .00$ (despite higher statistical power).[6] Simple main analyses on the mixed-gender setting revealed that girls underperformed relative to boys in the geometry condition, $F(1, 176) = 5.44$, $p < .02$, $\eta^2 = .03$, but outperformed them in the drawing condition, $F(1, 176) = 6.19$, $p < .01$, $\eta^2 = .03$. Moreover, girls performed significantly better in the drawing condition than in the geometry condition, $F(1, 176) = 11.41$, $p < .001$, $\eta^2 = .06$. Although the boys' performance in both conditions was consistent with a stereotype-lift effect (i.e., higher performance in geometry than in drawing), this effect in Study 2 was nonsignificant, $F(1, 176) = 1.96$, $p = .16$, $\eta^2 = .01$. Finally, Figure 2 suggests that boys in the geometry condition did much better in the same-gender setting compared with the mixed-gender setting, which was indeed true, $F(1, 442) = 10.53$, $p < .001$, $\eta^2 = .02$.

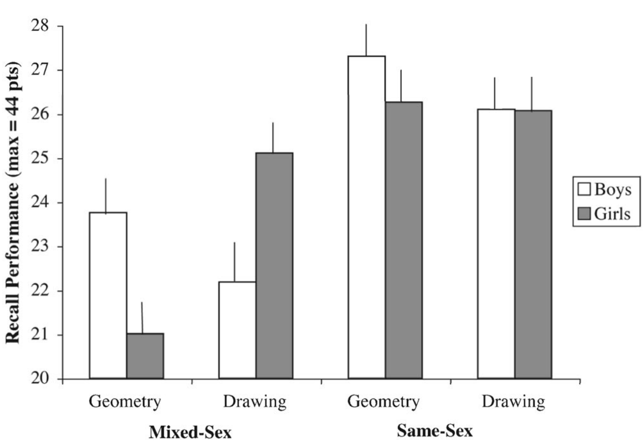

Figure 2. Mean recall performance as a function of student gender, task characterization, and gender composition. Error bars represent standard error.

Self-Reports

The 2 × 2 × 2 ANCOVA revealed a marginally significant main effect of task characterization on students' interest in the task, $F(1, 440) = 3.27$, $p < .07$, $\eta^2 = .01$, with more interest in the geometry condition ($M = 3.27$, $SE = 0.07$) than in the drawing condition ($M = 3.09$, $SE = 0.07$). A main effect of gender composition occurred on the difficulty ratings, $F(1, 437) = 5.76$, $p < .02$, $\eta^2 = .01$, which were higher in the mixed-gender setting ($M = 3.49$, $SE = 0.07$) than in the same-gender setting ($M = 3.28$, $SE = 0.05$). A main gender effect was also significant, $F(1, 437) = 5.96$, $p < .01$, $\eta^2 = .01$, indicating that girls ($M = 3.49$, $SE = 0.06$) perceived the task as more difficult than did boys ($M = 3.28$, $SE = 0.06$). This effect, however, was qualified by a marginally significant Student Gender × Task Characterization interaction, $F(1, 437) = 3.54$, $p < .06$, $\eta^2 = .01$. As the simple main effects showed, girls in the geometry condition ($M = 3.61$, $SE = 0.09$) rated the task as more difficult than did boys ($M = 3.23$, $SE = 0.08$), $F(1, 437) = 9.80$, $p < .002$, $\eta^2 = .02$. Girls in the geometry condition also thought the task was more difficult than did girls in the drawing condition ($M = 3.38$, $SE = 0.09$), $F(1, 437) = 3.43$, $p < .07$, $\eta^2 = .01$ (see Figure 3). The assumption of homogeneity of variances was not met here. The Student Gender × Task

[5] Following Yzerbyt, Muller, and Judd (2004), each interaction term between the covariate (students' math grades) and the two manipulated variables (task characterization and gender composition) were entered into the ANCOVA. As stated by these authors, whenever a covariate is potentially confounded with the measured independent variable (here, students' gender) and, additionally, participants have been randomly assigned to levels of the manipulated independent variables, then the predicted interaction between the measured and manipulated independent variables will be confounded with the interaction between the covariate and the manipulated independent variables. When this is the case, then the interaction between the manipulated and measured independent variables will be estimated with bias unless both the covariate and the interactions between the covariate and the manipulated variables are controlled.

[6] We deliberately retained a larger sample for the same-gender setting ($n = 272$), compared with the mixed-gender setting ($n = 182$). The resulting higher statistical power in the former setting reduced the probability of Type II error that could have been associated with the decision of accepting the null stereotype threat effect that was expected in this setting. Also, although the omnibus F corresponding to the Student Gender × Task Characterization interaction was not significant in the same-gender setting ($p < .50$), one may wonder whether girls ($M = 26.27$, $SE = 0.70$) still significantly underperformed relative to boys ($M = 27.31$, $SE = 0.71$) in the geometry condition. A simple main analysis showed that this gender difference was clearly not significant, $F(1, 266) = 1.68$, $p < .30$, $\eta^2 = .00$.

Characterization interaction, however, was clearly significant when using a nonparametric test (Kruskal–Wallis, $p < .01$). Finally, no effects were found on the students' self-report of effort, which, overall, was quite high ($M = 3.45$, $SE = 0.04$).

Nomination Task

The vast majority of students (96%, $n = 434$) nominated at least one classmate in each math ability category (high or low). Students nominated in the high-ability category were high math achievers ($M = 16.14$, $SD = 1.42$), and students nominated in the low-ability category were low math achievers ($M = 8.12$, $SD = 1.87$), indicating that participants understood the nomination task correctly. Single-sample t tests showed that the high and low math achievers scored respectively higher, $t(441) = 55.61$, $p < .0001$, and lower, $t(424) = -47.04$, $p < .0001$, than the "average" student ($M = 12.38$, $SD = 3.54$). The nominations in both categories (highest and lowest math achievers) were not restricted to the classmates who were physically present at the time of the nomination task. Overall, 40% of the nominated highest achievers and 46% of the nominated lowest achievers were not physically present in the immediate testing situation. This observation can be taken as further evidence that the participants understood the nomination task correctly (i.e., they did not believe that only physically present classmates could be nominated).[7]

On this basis, we computed two indexes: (a) the number of high-math-ability girls nominated minus the number of low-math-ability girls nominated (hereafter referred to as *Index 1*), and (b) the number of high-math-ability girls nominated minus the number of high-math-ability boys nominated (hereafter referred to as *Index 2*); a positive difference score indicating more high-math-ability girls nominated (for both indexes). These two indexes were then examined via a 2 (student gender) × 2 (task characterization) × 2 (gender composition) ANCOVA, using either Index 1 or Index 2 as the dependent variable. The covariate was the proportion of girls who were high math achievers in each class (i.e., those whose math grades were 1 standard deviation above the mean of their math class).

Index 1. This analysis revealed a main effect of gender composition, $F(1, 432) = 17.56$, $p < .001$, $\eta^2 = .04$: Students nominated a greater number of high-math-ability girls in the same-gender condition ($M = 0.65$, $SE = 0.08$) than in the mixed-gender condition ($M = 0.13$, $SE = 0.09$). This effect was qualified, however, by three interactions, namely: student gender by task characterization, $F(1, 432) = 10.97$, $p < .001$, $\eta^2 = .03$; task characterization by gender composition, $F(1, 432) = 21.85$, $p < .0001$, $\eta^2 = .05$; and the three-way interaction between student gender, task characterization, and gender composition, $F(1, 432) = 4.63$, $p < .03$, $\eta^2 = .01$ (see Table 1, Index 1). Because the salience of high-math-ability female classmates was expected to vary among girls alone, we broke this higher order interaction down by student gender. The Task Characterization × Gender Composition interaction was, in fact, significant for girls, $F(1, 212) = 20.37$, $p < .0001$, $\eta^2 = .09$, and for boys, $F(1, 217) = 4.30$, $p < .04$, $\eta^2 = .02$.

As expected, girls facing the geometry condition in the mixed-gender setting nominated more low- than high-math-ability girls ($M = -0.35$, $SE = 0.19$), whereas those in the same-gender setting did exactly the opposite, that is, they nominated essentially high-math-ability girls ($M = 0.95$, $SE = 0.15$), $F(1, 212) = 29.67$, $p < .0001$, $\eta^2 = .12$. In contrast, girls in the drawing condition nominated mostly high-math-ability girls, whatever the gender composition ($M = 0.91$, $SE = 0.20$, and $M = 0.48$, $SE = 0.19$, for mixed- and same-gender contexts, respectively), $F(1, 212) = 2.19$, $p < .14$, $\eta^2 = .01$.

Boys in the geometry condition in the mixed-gender setting nominated the same number of low- and high-math-ability girls ($M = 0.00$, $SE = 0.16$), whereas those in the same-gender setting nominated more high- than low-math ability girls ($M = 0.94$, $SE = 0.15$), $F(1, 217) = 20.17$, $p < .0001$, $\eta^2 = .09$. Boys in the drawing condition nominated the same number of low- and high-math-ability girls in both gender composition settings ($M = 0.00$, $SE = 0.19$, for the mixed-gender setting; $M = 0.23$, $SE = 0.18$, for the same-gender setting), $F(1, 217) = 0.66$, *ns*, $\eta^2 = .00$.

Index 2. Here, whereas the Task Characterization × Gender Composition interaction was clearly significant for girls, $F(1, 216) = 19.42$, $p < .0001$, $\eta^2 = .08$, it was nonsignificant for boys, $F(1, 224) = 2.33$, $p = .13$, $\eta^2 = .01$. Consistent with the previous findings, girls facing the geometry condition in the mixed-gender setting nominated more high-math-ability boys than high-math-ability girls ($M = -0.13$, $SE = 0.23$), whereas those in the same-gender setting nominated essentially high-math-ability girls ($M = 0.88$, $SE = 0.19$), $F(1, 216) = 11.89$, $p < .001$, $\eta^2 = .05$. Girls in the drawing condition also nominated mostly high-math-

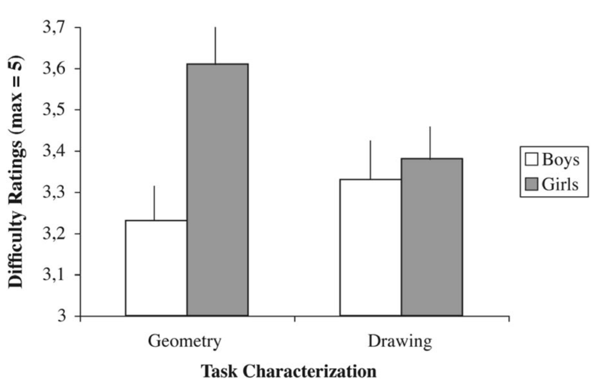

Figure 3. Mean difficulty ratings as a function of task characterization and student gender. Error bars represent standard error.

[7] One may wonder whether the same-gender condition did or did not bias the nominations toward same-gender classmates (i.e., more students nominated among those physically present in the same- than in the mixed-gender condition). It did not (quite the contrary), at least for the nominated highest achievers (62% of students nominated from among those physically present for the mixed-gender setting vs. 58% for the same-gender setting, $t(440) = 1.27$, $p < .21$). It did, however, for the nominated lowest achievers (49% vs. 58%, respectively, $t(423) = -2.71$, $p < .01$): Relative to the mixed-gender condition, participants' nominations in the same-gender condition relied more on the students physically present in the immediate testing situation. Whereas those percentages were identical for the nominated highest and lowest achievers in the same-gender condition, they were not in the mixed-gender condition (62% vs. 49%, $t(173) = 4.68$, $p < .0001$), in which participants relied more on those physically present for the nominated highest achievers than for the nominated lowest achievers.

Table 1
Mean (Standard Error) Scores of Nominations as a Function of Student Gender, Task Characterization, and Gender Composition

	Geometry		Drawing	
Nomination	Mixed gender	Same gender	Mixed gender	Same gender
For girls				
Index 1	−0.35 (.19)	0.95 (.15)	0.91 (.20)	0.48 (.19)
Index 2	−0.13 (.23)	0.88 (.19)	1.42 (.24)	0.36 (.23)
For boys				
Index 1	0.00 (.16)	0.94 (.15)	0.00 (.19)	0.23 (.18)
Index 2	−0.87 (.23)	−0.243 (.20)	−0.55 (.27)	−0.71 (.25)

Note. Index 1 refers to the number of high-math-ability girls nominated minus the number of low-math-ability girls nominated. Index 2 refers to the number of high-math-ability girls nominated minus the number of high-math-ability boys nominated. For both indexes, a positive difference score indicates more high-math-ability girls nominated.

ability girls, but did so to a greater extent when working in the mixed-gender setting (M = 1.42, SE = 0.24) than in the same-gender setting (M = 0.36, SE = 0.23), $F(1, 216) = 8.92$, $p < .003$, $\eta^2 = .04$. The corresponding three-way interaction between student gender, task characterization, and gender composition, however, was not significant, $F(1, 443) = 2.60$, $p = .11$, $\eta^2 = .01$ (see Table 1, Index 2).

Mediation findings. As indicated by the nomination data, the presence of only same-gender classmates led girls in the stereotype threat (geometry) condition to focus more on positive role models (i.e., high-math-ability girls) than on either stereotypic in-group or out-group members (i.e., low-math-ability girls and high-math-ability boys). Given that stereotype threat was eliminated in the same-gender setting, testing whether girls' nominations mediated the interaction between task characterization and gender composition on task performance was tempting. In order to test for mediated moderation, we followed the four steps outlined by Baron and Kenny (1986; see also Muller, Judd, & Yzerbyt, 2005). Step 1 consisted of showing that the Task Characterization × Gender Composition interaction was a significant predictor of girls' performance. In Step 2, the same interaction needed to predict girls' nominations. In Step 3, girls' nominations needed to predict performance when task characterization, gender composition, and their interaction were controlled. For Step 4, the direct effect of the Task Characterization × Gender Composition interaction on performance needed to be significantly reduced when girls' nominations were included in the analysis. Following this procedure, we conducted two separate mediated moderation analyses, one for each mediator (i.e., nomination Indexes 1 and 2), while controlling for the proportion of girls who were high math achievers in each class (and for biases due to possible interactions between the covariate and the independent variables, as suggested by Yzerbyt et al., 2004).

The regression analysis revealed that the Task Characterization × Gender Composition interaction predicted girls' performance (B = –1.21, SEB = .41, β = –.19, $t(218) = 2.92$, $p < .004$; Step 1). Then, the Task Characterization × Gender Composition interaction predicted girls' nominations, when retaining as dependent variables both Index 1 (B = –0.36, SEB = .09, β = –.23, $t(214) = –3.95$, $p < .0001$) and Index 2 (B = –0.48, SEB = .11, β = –.25, $t(218) = –4.47$, $p < .0001$, Step 2). In Step 3, the analyses revealed that only Index 1 comparing high- to low-math-ability girls nominated predicted girls' performance (B = 0.78, SEB = .34, β = .19, $t(209) = 2.31$, $p < .02$), when the experimental variables and their interaction were controlled. Finally, the analysis revealed that the Task Characterization × Gender Composition interaction no longer predicted girls' performance when the nomination data (Index 1) were controlled (B = –0.53, SEB = .50, β = –.08, $t(209) = –1.07$, $p < .29$). A Sobel test (Step 4) showed that the corresponding drop in unstandardized coefficients (from –1.21 to –0.53) was significant, Z = –1.99, p = .046 (see Figure 4).

Because Step 1 (the Task Characterization × Gender Composition interaction effect on task performance) was not significant for boys (B = 0.16, SEB = .42, β = .03, $t(226) = .38$, *ns*), their nominations (Index 1 as well as Index 2) could not be taken as possible mediators.

Finally, although participants' nominations were fairly accurate (overall, less than 6% of objectively high achievers were nominated as low achievers and vice versa), we also examined whether the mediation analyses led toward the same conclusion when removing the few nominations that were inadequate. These new mediation analyses (on Index 1 and Index 2 as before) led toward exactly the same conclusions.

Discussion

These new findings replicate and extend those of Study 1 in a number of important ways. Again, whereas girls performed less well than boys in the geometry condition, they outperformed them in the no-threat (drawing) condition, at least in the mixed-gender setting. This pattern offers direct evidence that in conditions close to ordinary classroom circumstances where both genders are typically represented, schoolgirls were influenced by the negative stereotype about their gender group's math ability. The fact that the girls' performance deficit was eliminated in the same-gender setting clearly shows that gender composition matters for this stereotype. This finding is consistent with past laboratory research (Inzlicht & Ben-Zeev, 2000, 2003; Sekaquaptewa & Thompson, 2003) and is especially interesting in relation to the girls' nominations. As expected, in the geometry condition, the presence of only same-gender classmates was associated for girls with a greater tendency (compared to the mixed-gender setting) to nominate high-math-ability girls, as opposed to both low-math-ability girls and high-math-ability boys.

In light of Marx et al.'s (2005) findings, which showed that the salience of positive role models counteracts stereotype threat, the girls' nomination tendency helps explain why their performance deficit was eliminated in the same-gender setting. To the extent that girls' nominations reflected the salience of positive role models, the mediation findings (at least those associated with Index 1) are clearly consistent with this explanation. In fact, the girls' tendency to nominate high-math-ability girls was found in every condition except the one where they underperformed (i.e., geometry condition in the mixed-gender setting) and showed up especially in the mixed-gender setting when the task was said to assess drawing ability. Although unexpected, this last finding can be

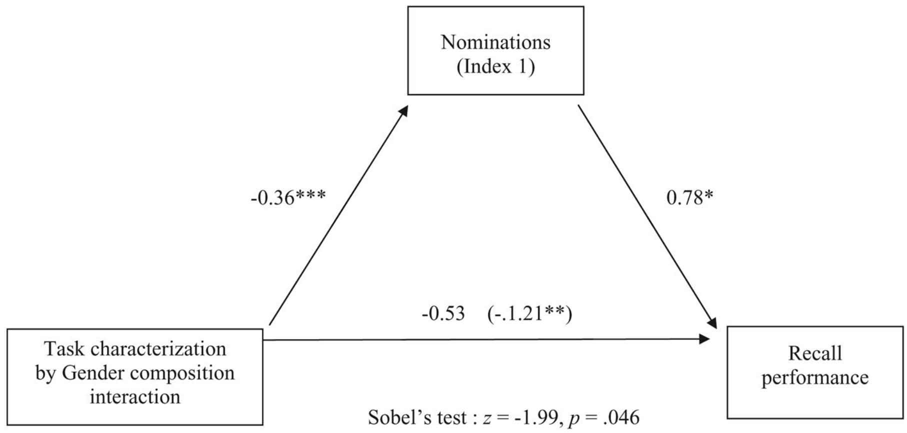

Figure 4. Girls' nominations (assessing the cognitive salience of high-ability girls, Index 1) as a mediator of the Task Characterization × Gender Composition interaction effect on recall performance. Unstandardized coefficients are reported. The number in parentheses indicates the direct effect of the Task Characterization × Gender Composition interaction on the recall performance prior to the inclusion of Index 1 in the regression equation. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

taken as evidence that the physical presence of the opposite gender does not necessarily undermine girls' salience of positive role models in the stereotyped domain, at least when the test at hand is not perceived as clearly diagnostic of math ability.

It was in this particular context (i.e., task characterized as a drawing test in a mixed-gender setting) that girls outperformed boys, which is very interesting for several reasons. As shown by the manipulation check, the way the task was characterized did not always lead to the expected effect. Although the task was clearly perceived as diagnosing geometry ability in the geometry condition, its perceived meaning in the drawing condition was more ambiguous, falling somewhere between geometry and drawing. Girls thus outperformed boys when the latter were present even though the change in the task's meaning was only slight. One may think that, when faced with a drawing task in a mixed-gender setting, the boys were affected by an implicitly activated stereotype favoring girls in the artistic domain, which was inhibited in the same-gender setting. Again, however, the task was not clearly perceived as an assessment of drawing in this condition, so this account should be considered with caution. Furthermore, the comparative evaluation findings in drawing run directly against the hypothesis of a stereotype favoring girls in this domain. Once more (see the *Participants* section for Study 2), whereas girls rated their relative standing in math less positively than did boys (while controlling for students' math grades), the reversed bias did not exist in the students' comparative evaluations in drawing. As before, the motivation-loss hypothesis for boys was not supported by the self-reports, which simply showed more interest toward the ROCF task in the geometry condition than in the drawing condition for both gender groups, whatever the gender composition setting. Likewise, the invocation of stereotype lift seems inadequate here. To begin with, in Study 2, the stereotype lift effect in the mixed-gender setting was not significant. From a stereotype lift approach, there is also no reason to expect the boys in the geometry condition to do better in the same-gender setting compared with the mixed-gender setting. Not only would the stereotype lift approach predict the reverse trend, but it would also lead toward the expectation of a lower performance in girls compared to boys in the same-gender setting when the task was characterized as drawing, provided that a stereotype favoring girls in drawing is operating (but we have seen that such a stereotype is not guaranteed). Instead, there was absolutely no gender difference on performance in the same-gender drawing context. On a larger scale, the performance findings indicate that whereas girls did equally well in each condition but the one in which they could indeed feel threatened, the boys did better in both task characterization conditions in same-gender environments compared to mixed-gender environments. As such, the present pattern suggests that same-gender environments can have two complementary effects: They can remove the stereotype threat obstacle in math for the girls, and they can apparently lead boys to pay more attention to the focal task.

Finally, the participants' self-reports also merit special attention. The girls perceived the task as especially difficult in the geometry condition, both compared to boys in the same condition and to other girls in the drawing condition, regardless of gender composition (and there is evidence that increased difficulty of the task at hand is associated with increased anxiety in female participants; see Stricker & Bejar, 2004). Stereotype threat was thus eliminated in the presence of same-gender classmates only, despite the fact that the task in this setting was perceived as equally difficult as in the mixed-gender setting. This finding indicates that schoolgirls can escape the negative gender stereotype even when the test is difficult and perceived as such. It is also worth noting that, in the mixed-gender context, participants exposed to the geometry condition did not report lower interest (quite to the contrary) or lower effort than those in the drawing condition. It is therefore unlikely that the girls' suboptimal performance in the geometry condition resulted solely from a conscious reduction in effort (e.g., giving up in the face of a difficult task). Of course, they may have been unwilling to report a low level of effort because this behavior is socially undesirable. However, the girls did not report greater effort (or interest; see also Study 1) in the drawing condition where they performed much better, despite the social desirability of this kind of effort. Because the participants' estimates of interest and effort did not vary as a function of gender composition either, an account in terms of conscious or strategic motivations also seems inadequate for explaining the elimination of stereotype threat in the same-gender setting.

General Discussion

There is ample evidence today from laboratory experiments that women and girls show a performance deficit on standardized math tests taken under stereotype threat. Regarding this important part of the stereotype threat literature, the present findings make at least four contributions.

First, they provide the first evidence to date of the impact of stereotype threat on schoolgirls in quasi-ordinary classroom circumstances, that is, when both genders are represented and the negative gender stereotype is both implicitly and indirectly (rather than explicitly and directly) activated. Past research on this topic has tested participants in the laboratory, so the generalizability of stereotype threat effects to girls in the natural context of school has remained unclear. Keller and Dauenheimer's (2003) study showed that girls (older than those tested here) can be influenced by stereotype threat in this context. Once again, however, the participants in their study who were faced with the threat (vs. control participants) were told that the math test they would take had (vs. had not) been shown to produce gender differences. Because this is not what teachers typically tell their students, the question remained as to whether the negative gender stereotype takes effect both in school when the stereotype is not explicitly activated, and for girls younger than the ones in the Keller and Dauenheimer's study. This is no longer an unanswered question. We do not argue that our experimental situations were strictly identical to ordinary classroom circumstances. One notable difference comes from the fact that the students were not exposed to their regular teachers. Instead, they were faced with a person (the experimenter) they did not know (as were the older students in Keller and Dauenheimer's study). This situation, however, comes close to what may happen at the beginning of the school year, especially in the first year of secondary school, when most students are faced with new teachers. Furthermore, the present data complement rather than compete with Keller and Dauenheimer's findings. Study 2 indeed showed that stereotype threat in schoolgirls can be observed with students taken from the general population, which is what Keller and Dauenheimer claimed on the basis of their own findings with older students. The fact that the stereotype threat effect in both investigations (i.e., Keller & Dauenheimer, 2003; our Study 2) was found when participants' math grades were controlled may seem inconsistent with STT. But as revealed by the importance ratings from Study 2 (not available in Keller and Dauenheimer's study), most participants identified highly with the stereotyped (math) domain, which according to STT is a crucial condition for stereotype threat to occur.

Second, not only can the present findings be taken as evidence that middle school girls are affected by stereotype threat at school, it also seems that this phenomenon is even more problematic than once assumed. Indeed, here the girls did not have to take a real math test; they exhibited a performance deficit (related to visual memory) when they were simply erroneously led to believe that they were taking such a test. This finding suggests that schoolgirls are influenced by the negative stereotype about their gender's math ability on any high-pressure test that is correctly or incorrectly viewed as assessing mathematical skills. The task we used is thought to tap a variety of skills (visual–perceptual and visual–spatial) as well as cognitive and metacognitive processes (attention, organization, and strategy use) that are basic components of academic performance (e.g., Kirkwood et al., 2001). Our findings indicate that at least some of these fundamental skills and processes can be temporarily disrupted in girls who simply believe they are taking a math test. The disruption is perhaps due to intrusive thoughts (such as "These exercises are too difficult for me"; see Brown & Josephs, 1999; Cadinu et al., 2005) and the impact such thoughts might have on working memory capacity (Schmader & Johns, 2003; see also Aschcraft & Kirk, 2001) and executive-processing resources (Croizet et al., 2004; Inzlicht, McKay, & Aronson, 2006). These processes, however, were not the focus of our investigation.

Third, whatever its exact nature, the interference associated with the stereotype threat condition in the mixed-gender setting was eliminated in the same-gender setting. This finding is another important if not key contribution of the present research. Once again, there are ample reasons to believe that gender composition has an effect on the emergence of stereotype threat (Inzlicht & Ben-Zeev, 2000, 2003; Sekaquaptewa & Thompson, 2003). So far, however, direct evidence in real classroom environments has been lacking. The elimination of stereotype threat interference in the same-gender setting enhances the ecological validity of past laboratory findings and provides incentive for further investigation in the schools. At first glance, the findings of Study 2 may give the impression that the physical presence of the opposite gender is needed for stereotype threat to occur in schoolgirls taking math tests (or tests perceived as such). In Study 1, however, the girls worked alone, and this setting did not eliminate their performance deficit in the geometry condition. The presence of the opposite gender, then, does not seem to be a necessary condition for the intervention of the negative gender stereotype in a school context, at least among middle school girls. The results obtained in the same-gender condition are hard to reconcile with the claim that, because of its visual–spatial component, women are intrinsically inferior to men on the ROCF recall memory task (e.g., Bennett-Levy, 1984; Vingerhoets, Lannoo, & Wolters, 1998). Although the complex figure used here was an adaptation of the original one, our findings are more consistent with the alternative view that men and women do not in fact differ in the visual–spatial abilities required by this task (see also Berry, Allen, & Schmitt, 1991; Boone, Lesser, Hill-Gutierrez, Berman, & D'Elia, 1993; Gagnon, Awad, Mertens, & Messier, 2003).

Lastly, the nomination findings help understand why the stereotype threat effect was neutralized among girls in the same-gender setting. As expected, for girls in the stereotype threat condition, the presence of only same-gender classmates was associated with a greater tendency (than in the mixed-gender setting) to nominate high-math-ability girls as opposed to both low-math-ability girls and high-math-ability boys. Assuming that positive role models have to be particularly salient for this tendency to occur, one can expect that girls in the same-gender setting will not experience stereotype threat (Marx et al., 2005), which is exactly what we found. And the mediation findings using participants' nominations as a mediator were clearly consistent with this explanation. Furthermore, the girls' tendency to nominate high-math-ability girls was found in every condition but the one in which they underperformed (i.e., geometry test in a mixed-gender setting). This finding can be taken as reasonable evidence that the salience of positive role models is the rule rather than the exception among schoolgirls in the natural context of school, at least with children of this age. Consistent with this argument, social comparison studies in this context have shown that the vast majority of middle school students compare their math grades upward with same-gender classmates (Blanton, Buunk, Gibbons, & Kuyper, 1999; Dumas, Huguet, Monteil, Rastoul, & Neslek, 2005; Huguet, Dumas, Monteil, & Genestoux, 2001), with a beneficial effect on their performance (see also Huguet, Galvaing, Monteil, & Dumas, 1999, for experimental evidence that upward comparison with same-gender peers can facilitate performance). Clearly, on the basis of our data, stereotype threat can be considered as running counter to girls' default mindset because it decreases the salience of counter-stereotypic group members (i.e., high-math-ability girls) to the benefit of stereotypic ones (i.e., low-math-ability girls as well as high-math-ability boys).

The present findings also have several implications for educational practices. One implication concerns the controversy about the merits of single-gender over coed education. Added to the fact that both boys and girls performed better in the same-gender setting than in the mixed-gender setting (Study 2), the elimination of stereotype threat in the former setting may be taken as a serious argument for separating the genders. After all, if such a separation could minimize the deleterious effects of gender stereotypes, why not make use of it? Several points must nevertheless be made here. First of all, although same-gender education may help prevent stereotypes from taking effect downstream (i.e., in testing situations), it is ineffective if not detrimental upstream (i.e., stereotype formation and propagation), which is obviously not satisfactory. As indicated by a myriad of findings in the social-categorization literature (e.g., Cadinu & Rothbart, 1996; Dovidio, Glick, & Rudman, 2005; Gaertner & Insko, 2000; Otten & Moskowitz, 2000; Tajfel & Turner, 1979), putting individuals into separate groups typically strengthens (or even creates) stereotypes rather than reducing them and the consequences they trigger. Clearly, if same-gender education leads to better performance in the classroom but generates tension and discrimination outside, it is undesirable or at the very least should be used with caution (i.e., temporarily rather than on a regular basis). In line with this argument, Rosenthal and Crisp (2006) demonstrated that interventions designed to reduce separation and intergroup boundaries between the two genders undermine stereotype threat in women. Second, our findings show that it is not the sheer presence of men or boys that is problematic. We agree with Inzlicht and Ben-Zeev (2000) that being outnumbered by men or boys typically leads women and girls to increased awareness of their gender-group stereotypes and thus of stereotype threat. However, girls worked alone in Study 1 (i.e., in the absence of any members of the opposite gender), yet this arrangement did not prevent their performance deficit from showing up in the threatening condition. Finally, separating the genders is not the only way to proceed at the practical level. Johns et al. (2005), for instance, showed that teaching students about stereotype threat is an efficient means of reducing its detrimental effects in testing situations. This option is especially attractive, as it may also help people propagate counter-stereotypic views within their social network, including among their own children, which would be effective both downstream and upstream. Additional interventions (for reviews, see Ben-Zeev et al., 2005; Maass & Cadinu, 2003) can be found in Aronson et al. (2002), who showed that exposing students to an incremental view of intelligence can help mitigate the impact of stereotype threat (see also Good et al., 2003). Likewise, the two studies reported here provide further evidence that redefining the context in which a test is taken (in a way that is less threatening) is a powerful solution to this problem (see also Cadinu, Maass, Lombardo, & Frigerio, 2006; Croizet & Claire, 1998; Croizet et al., 2004; Spencer et al., 1999; Steele & Aronson, 1995). Actions that heighten performers' levels of expectation (Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003), promote self-affirmative thoughts prior to test taking (Cohen, Garcia, Apfel, & Master, 2006; Croizet & Després, 2003; see also Spencer, Fein, & Lomore, 2001), or increase the accessibility of positive role models (Blanton, Crocker, & Miller, 2000; Marx et al., 2005; McIntyre, Paulson, & Lord, 2003) also represent valuable options.

Identifying effective interventions and specifying their boundary conditions remains a crucial challenge for both scientists and teachers. Our data show that middle school girls are affected by stereotype threat in quasi-ordinary classroom circumstances, and we have every reason to believe that they will go on being underrepresented in scientific disciplines and careers if this challenge is not taken seriously. We hope that our work will provide new impetus in this direction.

References

Aboud, F. E. (1988). *Children and prejudice.* New York: Basil Blackwell.

Akshoomoff, N. A., & Stiles, J. (1995). Developmental trends in visuospatial analysis and planning: 1. Copying a complex figure. *Neuropsychology, 9,* 364–377.

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.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38,* 113–125.

Aschcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130,* 224–237.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Benbow, C. P. (1988). Sex differences in mathematical-reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral Brain Sciences, 11,* 169–232.

Benbow, C. P., & Benbow, R. M. (1984). Biological correlates of high mathematical reasoning ability. In G. J. Devries, J. P. C. Debruin, H. B. M. Hylings, & M. A. Corner (Eds.), *Sex differences in the brain, the relation between structure and function: Progress in brain research* (Vol. 61, pp. 469–490). Amsterdam: Elsevier.

Benbow, C. P., & Stanley, J. C. (1980, December 12). Sex differences in mathematical ability: Fact or artifact? *Science, 210,* 1262–1264.

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

Bennett-Levy, J. (1984). Determinants of performance on the Rey-

Osterrieth Complex Figure Test: An analysis, and a new technique for single-case assessment. *British Journal of Clinical Psychology, 23,* 109–119.

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: Psychology Press.

Berry, D. T., Allen, R., & Schmitt, F. A. (1991). Rey-Osterrieth Complex Figure: Psychometric characteristics in a geriatric sample. *Clinical Neuropsychologist, 5,* 143–153.

Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology, 76,* 420–430.

Blanton, H., Crocker, J., & Miller, D. T. (2000). The effects of in-group versus out-group social comparison on self-esteem in the context of a negative stereotype. *Journal of Experimental Social Psychology, 36,* 519–530.

Boone, K. B., Lesser, I. A., Hill-Gutierrez, E., Berman, N. G., & D'Elia, L. S. (1993). Rey-Osterrieth complex figure performance in healthy, older adults: Relationship to age, education, sex, and IQ. *Clinical Neuropsychologist, 7,* 22–28.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performances. *Journal of Personality and Social Psychology, 76,* 246–257.

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.

Cadinu, M., Maass, A., Lombardo, M., & Frigerio, S. (2006). Stereotype threat: The moderating role of locus of control beliefs. *European Journal of Social Psychology, 36,* 183–197.

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.

Cadinu, M., & Rothbart, M. (1996). Self-anchoring and differentiation processes in the minimal group setting. *Journal of Personality and Social Psychology, 70,* 661–677.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33,* 669–680.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006, September 1). Reducing the racial achievement gap: A social–psychological intervention. *Science, 313,* 1307–1310.

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24,* 588–594.

Croizet, J. C., & Després, G. (2003, February). *How does stereotype threat undermine performance?* Presentation at the annual meeting of the Society for Personality and Social Psychology, Los Angeles, CA.

Croizet, J. C., Després, G., Gauzins, M. E., Huguet, P., Leyens, J.-P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin, 30,* 721–731.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT–grade and ability–job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220–230.

Dovidio, J. F., Glick, P., & Rudman, L. (Eds.). (2005). *On the nature of prejudice: Fifty years after Allport.* Malden, MA: Blackwell.

Dumas, F., Huguet, P., Monteil, J.-M., Rastoul, C., & Neslek, J. B. (2005). Social comparison in the classroom: Is there a tendency to compare upward in elementary school? *Current Research in Social Psychology, 10,* 166–187.

Eccles, J. S., Adler, T., & Meece, J. L. (1984). Sex differences in achievement: A test of alternate theories. *Journal of Personality and Social Psychology, 46,* 26–43.

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society, 11,* 367–380.

Eccles, J. S., Jacobs, J. E., & Harold, R. E. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues, 46,* 183–201.

Eccles, J. S., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64,* 830–847.

European Research Council. (2003). *Women and science.* Available at http://europa.eu.int/comm/research/sitemap/sitemap.cfm?lang=en#14

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43,* 95–103.

Gaertner, L., & Insko, C. A. (2000). Intergroup discrimination in the minimal group paradigm: Categorization, reciprocation, or fear? *Journal of Personality and Social Psychology, 78,* 77–94.

Gagnon, M., Awad, N., Mertens, V. B., & Messier, C. (2003). Comparing the Rey and Taylor Complex Figures: A test–retest study in young and older adults. *Journal of Clinical & Experimental Neuropsychology, 25,* 878–890.

Goethals, G. R., & Darley, J. (1987). Social comparison theory: Self-evaluation and group life. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 21–47). New York: Springer-Verlag.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology, 24,* 645–662.

Guy, R., & McEwen, B. (1980). *Sexual differentiation of the brain.* Cambridge, MA: MIT Press.

Halpern, D. F. (1992). *Sex differences in cognitive abilities.* Hillsdale, NJ: Erlbaum.

Hedges, L. V., & Nowell, A. (1995, July 7). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45.

Hirschfeld, L. A. (1996). *Race in the making.* Cambridge, MA: MIT Press.

Huguet, P., Brunot, S., & Monteil, J.-M. (2001). Geometry versus drawing: Changing the meaning of the task as a means to change performance. *Social Psychology of Education, 4,* 219–234.

Huguet, P., Dumas, F., Monteil, J.-M., & Genestoux, N. (2001). Social comparison choices in the classroom: Further evidence for students' upward comparison tendency and its beneficial impact on performance. *European Journal of Social Psychology, 31,* 557–578.

Huguet, P., Galvaing, M. P., Monteil, J.-M., & Dumas, F. (1999). Social presence effects in the Stroop task: Further evidence for an attentional view of social facilitation. *Journal of Personality and Social Psychology, 77,* 1011–1025.

Huston, A. C. (1987). The development of sex typing: Themes from recent research. In S. Chess & A. Thomas (Eds.), *Annual progress in child psychiatry and child development* (pp. 168–186). New York: Brunner/Mazel.

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

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.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achievement female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95,* 796–805.

Inzlicht, M., McKay, L., & Aronson, J. (2006). Stigma as ego depletion:

How being the target of prejudice affects self-control. *Psychological Science, 17,* 262–269.

Jacobs, J., & Eccles, J. S. (1992). The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology, 63,* 932–944.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16,* 175–179.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217.

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.

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.

Kirkwood, M. W., Weiler, M. D., Bernstein, J. H., Forbes, P. W., & Waber, D. P. (2001). Sources of poor performance on the Rey-Osterrieth complex figure test among children with learning difficulties: A dynamic assessment approach. *Clinical Neuropsychologist, 15,* 345–356.

Leyens, J. P., Désert, M., Croizet, J.-C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin, 26,* 1189–1199.

Lummis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement in mathematics: A cross-cultural study. *Developmental Psychology, 26,* 254–263.

Maass, A., & Cadinu, M. (2003). Stereotype threat: When minority members underperform. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 14, pp. 243–275). New York: Taylor & Francis.

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting female students' math test performance. *Personality and Social Psychology Bulletin, 28,* 1185–1197.

Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology, 88,* 432–446.

Massey, S. D., Charles, C. Z., Lundy, G. F., & Fischer, M. J. (2003). *The source of the river: The social origins of freshmen at America's selective colleges and universities.* Princeton, NJ: Princeton University Press.

McFarland, L. A., Lev Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16,* 181–205.

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype through salience of group achievement. *Journal of Experimental Social Psychology, 39,* 83–90.

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development, 74,* 498–515.

Mills, C. J., Ablard, K. E., & Stumpf, H. (1993). Gender differences in academically talented young students' mathematical reasoning: Patterns across age and subskills. *Journal of Educational Psychology, 85,* 340–346.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89,* 852–863.

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.

National Science Foundation. (2004). *Women, minorities and persons with disabilities in science and engineering*. Retrieved March 14, 2006, from http://www.nsf.gov/statistics/wmpd/

National Science Foundation. (2006). *Women, minorities and persons with disabilities in science and engineering*. Retrieved March 14, 2006, from http://www.nsf.gov/statistics/wmpd/

Nguyen, H. H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16,* 261–293.

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.

Osterrieth, P. A. (1944). Le test de copie d'une figure complexe [The complex figure copy test]. *Archives de Psychologie, 30,* 206–356.

Otten, S., & Moskowitz, G. B. (2000). Evidence for implicit evaluative in-group bias: Affect-biased spontaneous trait inference in a minimal group paradigm. *Journal of Experimental Social Psychology, 36,* 77–89.

Penner, A. M. (2003). International Gender × Item Difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology, 95,* 650–655.

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16,* 231–259.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57,* 55–71.

Rey, A. (1941). L'examen psychologique dans le cas d'encephalopathie traumatique [Psychological examination of traumatic encephalopathy]. *Archives de Psychologie, 28,* 286–340.

Robinson, N. M., Abbott, R. D., Berninger, V. W., & Busse, J. (1996). The structure of abilities in math-precocious young children: Gender similarities and differences. *Journal of Educational Psychology, 88,* 341–352.

Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin, 32,* 501–511.

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16,* 295–309.

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

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39,* 68–74.

Sells, L. W. (1973). High school mathematics as the critical filter in the job market. In R. T. Thomas (Ed.), *Developing opportunities for minorities in graduate education* (pp. 37–39). Berkeley: University of California Press.

Smith, J. L. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47,* 179–191.

Spencer, S. J., Fein, S., & Lomore, C. D. (2001). Maintaining one's self-image vis-à-vis others: The role of self-affirmation in the social evaluation of the self. *Motivation and Emotion, 25,* 41–64.

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

Stapel, D. A., & Suls, J. (2004). Method matters: Effects of explicit versus implicit social comparisons on activation, behavior, and self-views. *Journal of Personality and Social Psychology, 87,* 860–875.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

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.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M.

Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.
Stricker, L. W., & Bejar, I. I. (2004). Test difficulty and stereotype threat on the GRE general test. *Journal of Applied Social Psychology, 34,* 563–597.
Stricker, L. W., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and sex, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693.
Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–48). Pacific Grove, CA: Brooks/Cole.
Vingerhoets, G., Lannoo, E., & Wolters, M. (1998). Comparing the Rey-Osterrieth and Taylor Complex Figures: Empirical data and meta-analysis. *Psychologica Belgica, 38,* 109–119.
Waber, D. P., Bernstein, J. H., & Merola, J. (1989). Remembering the Rey-Osterrieth Complex Figure: A dual-code, cognitive neuropsychological model. *Developmental Neuropsychology, 5,* 1–15.
Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39,* 456–467.
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.
Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A., Freedman-Doan, K., et al. (1997). Changes in children's competence beliefs and subjective task values across the elementary school years: A three-year study. *Journal of Educational Psychology, 89,* 451–469.
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.

Appendix

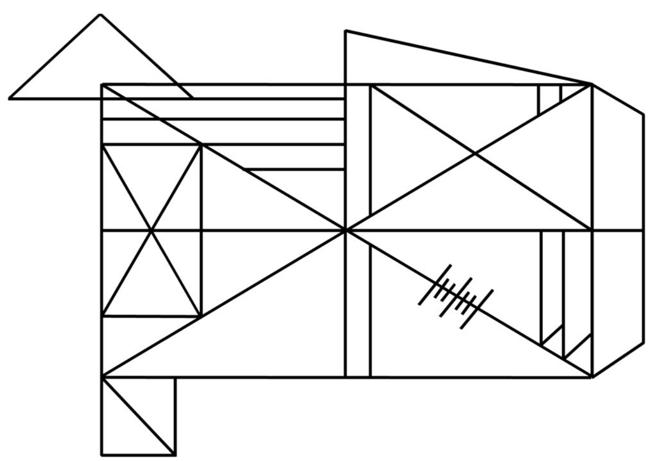

Figure A1. Adapted from the Rey-Osterrieth Complex Figure (ROCF) recall memory task, "Le Test de Copie d'une Figure Complexe [The Complex Figure Copy Test]," P. A. Osterrieth, 1944, *Archives de Psychologie, 30,* pp. 206–356; and "L'Examen Psychologique Dans le Cas D'Encephalopathie Traumatique [Psychological Examination of Traumatic Encephalopathy]," A. Rey, 1941, *Archives de Psychologie, 28,* pp. 286–340.

Received June 29, 2006
Revision received January 23, 2007
Accepted January 25, 2007 ■