Available online at www.sciencedirect.com
ScienceDirect
Journal of Applied Developmental Psychology 29 (2008) 17–28
Journal of Applied Developmental Psychology

Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses ☆

Catherine Good [a,*], Joshua Aronson [b,*], Jayne Ann Harder [c]

[a] Barnard College, Columbia University, United States
[b] New York University, United States
[c] The University of Texas, Austin, United States

Available online 20 November 2007

Abstract

It is well established that negative stereotypes can undermine women's performance on mathematics tests. Despite considerable laboratory evidence for the role of "stereotype threat" in girls' and women's math test performance, the relevance of such findings for the "real world" gender test-score gap remains unclear and debates about causes focus primarily on innate sex differences in cognitive capacity. Reported here are results of a field experiment that tested the usefulness of the stereotype threat formulation for understanding women's performance in upper levels of college mathematics — men and women who are highly motivated and proficient mathematicians and who are in the pipeline to mathematics and science professions. Our primary hypothesis was confirmed. Test performance of women in a stereotype-nullifying presentation of the test in an experimental group was raised significantly to surpass that of the men in the course. In a control group, in which test-takers were given the test under normal test instructions, women and men performed equally. The pattern of results suggests that even among the most highly qualified and persistent women in college mathematics, stereotype threat suppresses test performance.

Keywords: Stereotype threat; Gender stereotypes; Math; Sex differences; Achievement; Adolescence

1. Introduction

How much of the gender gap in math and science achievement is attributable to nature and how much to nurture is a source of concern, debate, and political thin ice for many who have a stake in understanding the issue of why boys and girls appear to differ in mathematical ability. Although some studies have found evidence for biological differences (e.g., Halpern, 1992), there remains considerable debate on the extent to which biology explains the differential

☆ Preparation of this paper was supported by a grant from the National Science Foundation (BCS-0217251) to Catherine Good and Carol S. Dweck, and by a National Science Foundation CAREER award (BCS-9875957) to Joshua Aronson.

* Corresponding authors. Good is to be contacted at Baruch College, CUNY, Vertical Campus Room 8-215, One Bernard Baruch Way, New York, NY 10010. Tel.: 646 207 4413; fax: 646 312 3781. Aronson, New York University, Department of Applied Psychology, 239 Greene Street, 5th Floor, New York, NY 10003, United States.

E-mail addresses: Catherine_Good@baruch.cuny.edu (C. Good), joshua.aronson@nyu.edu (J. Aronson).

0193-3973/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.appdev.2007.10.004

outcomes in math achievement (Benbow & Stanley, 1980; Eccles, Adler, & Meece, 1984; Gottfredson, 2002; Pinker, 2005; Spelke, 2005; Stipek & Gralinski, 1991). The depth and political charge of this debate was underscored recently when then-president of Harvard University, Lawrence Summers, suggested during a speech (Summers, 2005) that one of the primary reasons men vastly outnumber women at the upper rungs of science and mathematics professions is biological differences that result in "different availability of aptitude at the high end." In other words, he argued that women are not significantly hindered merely by "social forces"; rather their limitation is mostly attributable to innate biological differences that make them less capable than men of doing high-level mathematics.

In this paper, we present data that cast doubt on the "different availability of aptitude at the high end" hypothesis. Specifically, we report data from a field study examining the relevance of stereotype threat – an environmental impediment to women's math performance – to the test-score gap between men and women in a context that is at the epicenter of the nature-nurture debate — in the advanced mathematics courses at a major university. Such courses are important because they are the "pipelines" to hard-science careers. For students to enter the fields where women are traditionally under-represented, they must succeed in these courses. The design of the experiment – comparing a group of men and women for whom the stereotype is nullified with a control group who receives the same test under standard testing procedures – permits us to evaluate whether stereotype threat is a significant factor in the test-score gap among these students. Our data show that there are significant (albeit not equal) numbers of talented women in the pipeline to hard-science professions. Moreover, the data strongly suggest that not only are women able to achieve grades that are on par with men's grades in the most difficult mathematics courses, but once disruptive social forces are minimized, women can even surpass men on difficult math tests.

1.1. Background: females' math achievement and stereotype threat

The gap in mathematics performance that has rekindled the nature-nurture debate can be documented at practically every level of schooling, especially when looking at the highest levels of achievers (National Science Foundation, 2006). For example, a higher percentage of males than females perform at or above proficiency in mathematics in the fourth (35% versus 30%), eighth (30% versus 27%), and twelfth (19% versus 14%) grades. Moreover, although the gap between males and females on the SAT has shrunk over the years, males still outperform females by 34 points on this high-stakes test (College Board, 2005). What is more, these early differences in performance may lay the foundation for future career aspirations. For example, although women earn nearly half of all science and engineering doctoral degrees, they are vastly under-represented in the so-called "hard" sciences. Specifically, in 2003 women earned over half of the doctoral degrees in the social and behavioral sciences, yet they earned only 29% of doctoral degrees in the physical sciences, 24% of doctoral degrees in mathematics, and only 19% of doctoral degrees in engineering (National Science Foundation, 2006).

What explains these sex-based performance and participation differences? Although biological factors may represent a potent factor (see Halpern, 1992 for a review), the role of sociocultural forces, such as sex stereotypes, has also been well established (Eccles, 1994; Eccles & Jacobs, 1992; Eccles, Jacobs, & Harold, 1990; Good, Dweck, & Rattan, in preparation-a,b; see Aronson & Steele, 2005; Steele, Spencer, & Aronson, 2002 for reviews). For example, one line of research suggests that parents are important socializers of sex-based achievement differences in mathematics (Eccles, 1994; Eccles & Jacobs, 1992; Eccles, Jacobs, & Harold, 1990). Expectancies set early in childhood by parents may lay the foundation for later underperformance, interest, and participation during adolescence.

Another growing line of research suggests that girls and women suffer negative performance outcomes on math tests, not necessarily because they are socialized by parents to lack ability, but because of their vulnerability to negative stereotypes disseminated in the broader culture. The research shows that when stereotypes are not activated, or if they are nullified by other cues in the environment, girls and women perform better. When negative stereotypes *are* activated and left unchecked, they trigger a number of disruptive psychological processes that undermine test performance (Croizet et al., 2004; Dar-Nimrod & Heine, 2006; Davies, Spencer, Quinn, & Gerhardstein, 2002; Good, Aronson, & Inzlicht, 2003; Inzlicht & Ben-Zeev, 2000; Johns, Schmader, & Martens, 2005; McGlone & Aronson, 2006; McIntyre, Paulson, & Lord, 2003; Schmader, 2002; Schmader & Johns, 2003; Spencer, Steele, & Quinn, 1999). This phenomenon is called "stereotype threat" and has been applied to the underperformance of African-Americans, Latinos, and a variety of other minority groups (for reviews see Aronson & Steele, 2005; Steele, Spencer & Aronson, 2002).

This research has offered a better understanding of the situations that are most likely to lead to underperformance (Good et al., in preparation-a,b; Inzlicht & Ben-Zeev, 2000; Spencer et al., 1999), the age at which females are likely to experience stereotype threat (Good & Aronson, in preparation), and methods of overcoming stereotype threat (Aronson, Fried, & Good, 2002; Cohen, Garcia, & Master, 2006; Good et al., 2003; Johns et al., 2005). For example, researchers have found that certain situations, such as taking a diagnostic test in a domain in which one's group faces negative stereotypes, is sufficient to invoke underperformance on the test (Steele & Aronson, 1995). In addition, anything that makes one's social identity salient – such as indicating one's race prior to taking a diagnostic verbal test – can also lead to stereotype-based underperformance (Steele & Aronson, 1995). Even the sex composition of the testing situation can influence vulnerability to stereotype threat: researchers have found that women's performance on a math test went down as a function of the number of men taking the test in the same room (Inzlicht & Ben-Zeev, 2000). Studies such as these clearly implicate both the social situation and the temporary mindset of the performer in the underperformance of stereotyped groups, such as women in math.

Yet despite considerable empirical support, stereotype threat is not widely accepted as a factor in the mathematics gender gap. One reason for this, we believe, is that much of the research showing that stereotype threat undermines test performance has been conducted in psychology laboratories. As such, results can easily be discounted as irrelevant to test-score gaps existing in the real world. For example, Sackett and his colleagues (Sackett, Hardison, & Cullen, 2004) recently argued that stereotype threat experiments conducted by Steele and Aronson (1995) – which demonstrated large effects of stereotypes on the test scores of African-Americans – were interesting but irrelevant to the test-score gap. In essence, this argument suggests that the experiments have shown that an existing gap can be made larger by inducing stereotype threat but that the studies fail to show that reducing stereotype threat in the real world of the classroom or testing center would narrow the actual gap between black and white students.

Scientific evidence at odds with this argument is accumulating, demonstrating specifically that reducing stereotype threat significantly narrows the gap between stereotyped and non-stereotyped students. For example, a large study conducted by Stricker and Ward (2004) for the Educational Testing Service (ETS) found that simply moving the standard demographic inquiry about test-taker gender to the end of the test (and thus reducing the level of stereotype threat during the test) resulted in significantly higher performance among women who took the AP calculus test. The results showed that if ETS made official this simple change in procedure, 4700 additional female students would receive Advanced Placement credit in calculus each year. Perhaps attesting to the politicized nature of research in this area, Stricker and Ward presented their data as indicating no significant effect of asking about gender on women's performance. Yet in fact, a reanalysis of the ETS data (Danaher & Crandall, in press) showed these effects to be substantial and significant. Other studies have shown the relevance of stereotype threat by showing that interventions that reduce stereotype threat also produce significant reductions of performance gaps relative to control groups (Brown & Day, 2006; Cohen et al., 2006; McGlone & Aronson, 2006). This strongly suggests the role of stereotype threat in extant gaps between men and women in mathematics and minorities and whites in all areas of cognitive ability.

1.2. Developmental trends in stereotype threat

Research investigating the contribution of stereotype threat to women's mathematics performance has focused, for the most part, on women in college-level mathematics (see Steele et al., 2002, for a review), leaving unanswered the question of when stereotype-based underperformance begins to emerge. The few studies that have tackled this question suggest that stereotyped individuals may be vulnerable to stereotype threat as early as elementary school. For example, African-American and Latino children as young as six years old performed more poorly on a cognitive task when that task was described as a measure of ability than when the same task was described as a problem-solving task (McKown & Weinstein, 2003). However, this difference emerged only for those children who were aware of broadly held stereotypes about academic ability. Moreover, race/ethnicity-based stereotypes and sex-based stereotypes do not necessarily operate in the same fashion. For example, stereotype threat has been found to undermine girls' math performance in the sixth grade, but not in the fourth or fifth grades (Good & Aronson, in preparation). Regardless of the specific age or grade level at which negative stereotypes were found to affect performance, these studies illustrate that stereotype threat begins to undermine performance in children's elementary years. And by late adolescence, its cumulative effects can undermine students' career plans (Good et al., in preparation-a), social identities (Good, Dweck, & Aronson, 2007), and performance on high-stakes evaluations.

1.3. Methods of overcoming stereotype threat

In recent years, some researchers have turned their attention toward finding methods of alleviating stereotype threat. One line of research has addressed the underlying message of the stereotype — that stereotyped individuals are inherently limited because of their group membership. This approach has encouraged stereotyped individuals to reject the idea that intelligence is a fixed trait and instead to adopt the mindset that intelligence is a quality that can be increased with hard work and effort. In one study, African-American college students who were exposed to this view received higher grade point averages and later reported that academic achievement was more important to their senses of selves (Aronson, Fried, & Good, 2002). In a related study, college students mentored middle school students and helped them create web-pages advocating the view that intelligence is a malleable capacity (Good, Aronson, & Inzlicht, 2003). At the end of the year, the middle school females who received the malleable message earned higher scores on the state-wide standardized test in math than their female peers who instead received an anti-drug message as part of the intervention. In a similarly remarkable study, Cohen, Garcia and Master (2006) reduced the black–white GPA gap among low-income middle school students by affirming the students' self-concepts (and presumably inoculating them from stereotype threat) at the beginning of the school term.

These techniques, although effective, are not likely to be implemented in standard testing situations such as the SAT, GRE, etc. Furthermore, they are unlikely to be adopted by most mathematics curricula. Thus another aim of this research was to test whether or not stereotype threat could be reduced in a manner that would be easily replicated and implemented in everyday school experiences.

In sum, although stereotype threat has been found to undermine the performance of females in mathematics as early as the sixth grade, most studies investigating the effects of stereotypes on females' math achievement have focused on college-aged women. Moreover, many of these studies have concentrated primarily on females' performance on general mathematics tasks. Although most studies recruit participants who are highly identified with mathematics, or who have performed well on mathematics tasks in the past (for example, Spencer et al., 1999), none have tested whether or not stereotype threat suppresses the scores of females who are enrolled in the most difficult mathematics courses that lay the foundation for future careers in math and science. Consequently, we need to obtain information about those in the pipeline who aspire to be the women who would fill the void described by Summers. For example, Spencer and his colleagues (Spencer et al., 1999) used the psychology subject pool for recruitment of study participants. Although these participants also were screened for high math achievement (for example, earning a B or better in calculus), the population likely included females who had no intention of pursuing a math or math-based major or career. Consequently, these past studies left unaddressed the issue of the role of stereotypes and stereotype threat in math performance in the upper end of the math ability distribution. Most importantly, however, the experiments were conducted in the laboratory and thus their applicability to real-world test-score gaps can be questioned. In sum, the goal of the present study was to test whether the negative effects of stereotype threat extend beyond the laboratory and beyond women's mathematics performance in general to women's underperformance in the most difficult college mathematics courses — those that produce future mathematicians, engineers, and scientists, and those that conceivably might be avoided by women vulnerable to stereotype threat.

2. Method

2.1. Participants and design

Participants were 174 calculus students enrolled in the final course of the most rigorous and fast-paced calculus sequence offered by the university, a large public university in the southwest. This course satisfied degree requirements for mathematics, engineering, and many of the natural sciences such as biology, chemistry, geology, and physics. Because students who enrolled in the course had already successfully navigated the previous semester's "gatekeeper" course, they were well on their way to mathematics, engineering, or science degrees. Moreover, the university offered less rigorous calculus courses, but these were generally taken by students not planning to major in math or science.

Of the 174 students who took part in the study, 100 were male, 57 were female, and 17 did not report their sex. We analyzed data only for those participants who indicated their sex, leaving a final sample of 157 participants. Of these,

91 self-identified as Anglo-American (36 females, 55 males), 31 self-identified as Asian-American (10 females, 21 males), 17 self-identified as Hispanic-American (6 females, 11 males), 7 self-identified as African-American (2 females, 5 males), and 11 either did not report their ethnicity or indicated "other" (3 females, 8 males). In addition, because some ethnic groups had low representation in the calculus class we did not conduct analyses based on ethnicity. We randomly assigned participants to one of two conditions in a 2 (sex) by 2 (test description: stereotype threat, non-threat) factorial design.

2.2. Materials and procedure

We designed a calculus test, aligned with the topics the professors were discussing in their calculus courses, to be used as a practice test for an upcoming course examination. Test items were taken from the Graduate Record Examination (GRE) Mathematics Subject test that represented the same content as covered in the calculus course. These items were pilot tested with 12 advanced calculus students who were enrolled in the same level course as the participants in this study to ensure an adequate level of difficulty for the practice test.

Toward the end of the semester, the calculus professors announced that a practice test would be administered during the discussion sections the following week. The professors explained that the practice test would be similar in content to an upcoming course examination and thus would provide a good indicator of the students' preparedness for the impending examination. The professors offered extra credit on the course examination based on the students' performance on the practice test. In reality, every student who took the practice test received the same number of extra credit points, regardless of their performance. Extra credit was offered to increase the likelihood that students would take the test seriously and perhaps experience it psychologically as they would a regular course examination.

Additional protocols were implemented to replicate the testing conditions of standard course examinations. For example, the sex composition of the practice test reflected that of the course, and teaching assistants were present (as they were during regular course examinations). With these protocols, the alignment of the content of the test with the course topics, the diagnostic nature of the test, and the potential for performance on the practice test to influence final course grades, the testing situation was made as similar as possible to that of the students' regular course examinations.

On the day of the practice test, the teaching assistant introduced the researcher who then told the students the following:

> The material on this practice test covers a range of calculus topics, just like your examination next week will, so your performance on this test will be a good indicator of your readiness for your exam. Your test will be graded by your teaching assistant so that she/he may have a better idea of the areas where each of you may need a little extra help before next week's exam. Your score on this test will be counted as extra credit in this course. Thank you for participating and giving the test a genuine effort.

The researcher then presented an example problem and instructed the students to indicate how confident they were in their answer for each question (1 = low confidence; 5 = high confidence).

Next, the researcher distributed the research packets, which included the consent form, the experimental manipulation, the practice test, a post-measure questionnaire, demographic information questions, and a debriefing section. The packets were randomized. Although the experimenter was knowledgeable about the hypotheses of the study, the experimenter remained blind to condition. Sealed tabs separated each section of the packet, and the researcher instructed the students not to proceed to succeeding sections until told to do so.

Students then read and signed the consent page and then read the manipulation page. The manipulation consisted of a description of the nature of the test. Students in the stereotype threat condition read a statement indicating that the test they were about to take was designed to measure their math abilities:

> For the next 20 min, you will be taking a math test aimed at measuring your mathematical abilities. Why? As you probably know, math skills are crucial to performance in many important subjects in college. Yet surprisingly little is known about the mental processes underlying math ability. This research is aimed at better understanding what makes some people better at math than others. After you finish the test, your teacher will score it. This will enable us to analyze your performance and compare it with other students taking this test.

Diagnostic statements such as these have been shown to invoke stereotype threat in the past (e.g., Steele & Aronson, 1995). Students in the non-threat condition read the same diagnostic information but with the following added paragraph:

> What about gender differences? This mathematics test has not shown any gender differences in performance or mathematics ability. The test has been piloted in many mathematics courses across the nation to determine how reliable and valid the test is for measuring mathematics ability. Analysis of thousands of students' test results has shown that males and females perform equally well on this test. In other words, this mathematics test shows no gender differences.

Telling students that a test has never shown any gender differences in the past has been shown to reduce stereotype threat for women on general mathematics tasks (Quinn & Spencer, 2001; Spencer et al., 1999).

Importantly, the non-threat condition in this study differs from past studies in that the diagnosticity of the test was retained. In past studies, stereotype threat was reduced either by ensuring females that the test was gender-fair while at the same time not explicitly mentioning the diagnostic nature of the test (e.g., Quinn & Spencer, 2001; Spencer et al., 1999), or by explicitly nullifying the assumed diagnosticity of the test (Steele & Aronson, 1995). Because removing the diagnostic nature of course examinations is an unrealistic expectation for regular course examinations or for standardized math testing situations, we wanted to test whether or not stereotype threat could be reduced simply by addressing the specter of gender-based performance differences within the context an explicitly diagnostic examination.

In addition, past studies have induced stereotype threat either by explicitly stating that males do better than females on the test or by explicitly raising the possibility of sex-based performance differences (Spencer et al., study 2, 1999). This, however, is unlikely to be the way that stereotype threat operates in real-world settings. Instead, we relied on a more realistic and common occurrence to create the threat — simply taking a diagnostic math test. Furthermore, although other studies have induced stereotype threat by simply telling participants that they were about to take a math test, the extent to which their performance on this test would be indicative of underlying mathematics ability was not explicitly addressed (e.g., Spencer et al., 1999). Our study, in contrast, explicitly stated that the test was diagnostic of math ability. We believe that these instructions not only represent a stronger and more direct manipulation of stereotype threat, but also generalize more to real testing situations. To summarize, the threat condition explicitly presented the calculus test as diagnostic of ability and made no mention of sex-based performance differences; the non-threat condition also explicitly presented the test as diagnostic of ability but included the added component that the test was gender-fair.

Students were given 20 min to work on the calculus test. This time limit was set for a number of reasons. First, we wanted to replicate the students' classroom testing situation, which uses timed tests. Second, because the test was too long and difficult to complete in 20 min, this added time pressure could further tax working memory, which past research has shown is hindered in stereotype-threatening situations (Schmader & Johns, 2003). After the 20 min had passed, students were instructed to proceed to the post-test questionnaire, demographic section, and debriefing section at their own pace. The post-test questionnaire included a manipulation check in which participants used a 15-point Likert-type scale to respond to the following item: "This test was biased" (1 = not at all; 15 = extremely). The experimenter and teaching assistant monitored the class for the duration of the experiment and collected the research packets as the students left the classroom. At the end of the semester, we obtained students' calculus course grades from the calculus professors.

We predicted that the diagnosticity of the test would create a stereotype-threatening situation for women, and consequently, that they would perform worse than men on the calculus test. We also predicted that framing the diagnostic test as sex-fair would alleviate stereotype threat; thus, we expected women's performance in the non-threat condition to exceed that of the women in the threat condition. We predicted that the manipulations would have no significant effect on the men's performance.

3. Results

3.1. Manipulation check

Because two participants failed to answer the question about whether or not the test was biased, these data were analyzed for 155 participants. The two-way ANOVA performed on the participants' responses to this question revealed

Table 1
Mean (and *SD*) number correct on the test (maximum = 15 items) and course grades (maximum = 4.0) for male and female participants in the non-stereotype threat and the stereotype threat conditions

		Non-stereotype threat		Stereotype threat	
		Female	Male	Female	Male
All Participants					
	# Correct	3.60 (2.00) $n = 25$	2.60 (1.42) $n = 35$	3.13 (1.95) $n = 32$	3.08 (1.47) $n = 65$
	Course Grades	2.88 (.97) $n = 25$	2.55 (1.09) $n = 33$	2.71 (.97) $n = 31$	2.76 (1.01) $n = 59$
Anglo-American Participants					
	# Correct	4.44 (1.90) $n = 16$	2.70 (1.26) $n = 20$	2.85 (1.79) $n = 20$	2.89 (1.41) $n = 35$
	Course Grades	2.81 (1.05) $n = 16$	2.83 (1.04) $n = 18$	2.65 (1.04) $n = 20$	2.68 (1.09) $n = 34$

only an effect of condition, $F(1, 151) = 3.94$, $p = .05$, Cohen's $D = .34$, $r = .17$. This represents a medium effect size. Participants in the non-stereotype threat condition ($M = 1.51$, $SD = 1.38$) reported a significantly lower level of test bias than participants in the stereotype threat condition ($M = 2.27$, $SD = 2.89$). Thus, the manipulation statement indicating that the test was gender-fair reduced the amount of bias that participants in the non-stereotype threat condition perceived.

3.2. Did stereotype threat affect women's calculus performance?

A two-way ANOVA was performed on the number of items participants answered correctly. The main effect of Condition approached significance, $F(1, 153) = 3.47$, $p = .06$, Cohen's $D = .04$, $r = .02$, which was qualified by a Sex × Condition interaction that also approached the level of significance, $F(1, 153) = 2.86$, $p = .09$. Planned comparisons indicated that in the non-threat condition, women outperformed the men, $t(60) = 2.30$, $p = .02$, Cohen's $D = .58$, $r = .28$. This represents a medium effect size (see Table 1 for means and standard deviations).

These differences might not have yielded results that attained the level of significance because sex differences in math performance may exist primarily among Anglo-American students (American Association of University Women, 1998; Hyde, 1994; Hyde, Fennema, & Lamon, 1990). Thus, we conducted the analyses including data obtained for only the Anglo-American participants. Despite the reduced sample size, this analysis yielded a significant main effect of Sex, $F(1, 87) = 6.21$, $p < .02$, Cohen's $D = .44$, $r = .21$, and a significant main effect of Condition, $F(1, 87) = 4.22$, $p < .05$, Cohen's $D = .36$, $r = .18$. These results were qualified by a significant Sex × Condition interaction, $F(1, 87) = 6.75$, $p < .02$ (see Fig. 1). Planned comparisons indicated that the Anglo-American women in the non-threat condition

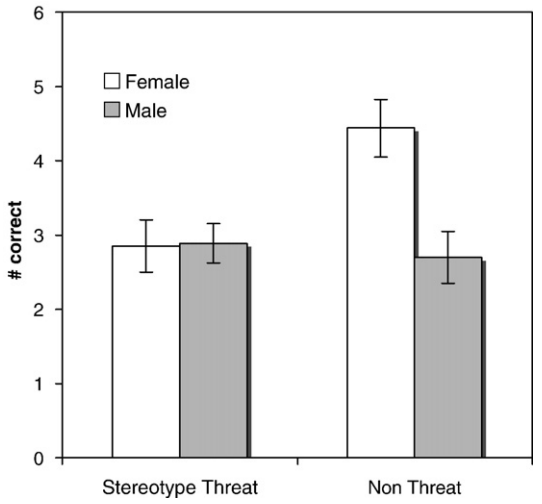

Fig. 1. Anglo-American women's and men's scores on the calculus test as a function of stereotype threat condition. Error bars represent the standard error of the mean.

Table 2
Pairwise comparisons and effect sizes for number correct

	Non-threat females	Non-threat males	Threat females	Threat males
Non-threat females		$t = 3.32$, $p = .001$	$t = 3.03$, $p = .003$	$t = 3.30$, $p = .001$
Non-threat males	$D = 1.08$, $R = .47$		$t = .30$, $p = .76$	$t = .42$, $p = .67$
Threat females	$D = .86$, $R = .40$	$D = .10$, $R = .05$		$t = .082$, $p = .94$
Threat males	$D = .93$, $R = .42$	$D = .14$, $R = .07$	$D = .02$, $R = .01$	

Note: Upper diagonal shows *t*-tests and significance values; Lower diagonal shows Cohen's *D* and effect sizes; Anglo-American participants only.

outperformed Anglo-American participants in every other condition (see Table 2 for pairwise comparisons and effect sizes). Most strikingly, these women not only outperformed women in the stereotype threat condition, but surprisingly, also outperformed *men* in both the non-threat condition and the stereotype threat condition. The manipulation did not significantly affect men's performance.

3.3. Did women in the stereotype threat condition answer fewer problems and were they less accurate in their responses?

To address questions about items attempted and response accuracy, we first conducted a two-way ANOVA on number attempted. This analysis yielded a significant interaction between Sex and Condition, $F(1, 87) = 6.34$, $p < .02$. Planned comparisons indicated that males in the stereotype threat condition ($M = 9.2$, $SD = 3.08$) answered more questions than males in the non-threat condition ($M = 7.00$, $SD = 2.97$), $t(55) = 2.59$, $p < .02$, Cohen's $D = .73$, $r = .34$. In addition, females in the non-threat condition ($M = 9.44$, $SD = 3.39$) answered more questions than males in the non-threat condition ($M = 7.00$, $SD = 2.97$), $t(36) = 2.40$, $p < .02$, Cohen's $D = .77$, $r = .36$. These represent large effect sizes. These were the only comparisons to reach significance (all other ps > .26).

To determine whether reducing stereotype threat enabled females to be more accurate in their responses (see Steele & Aronson, 1995; Inzlicht & Ben-Zeev, 2000), we conducted a two-way ANOVA on accuracy (number correct divided by number attempted). This analysis yielded a significant main effect for Condition, $F(1, 87) = 4.80$, $p < .03$, Cohen's $D = .45$, $r = .22$. Overall, participants in the non-threat condition ($M = .46$, $SD = .23$) were more accurate than participants in the stereotype threat condition ($M = .36$, $SD = .21$). Planned comparisons revealed that women in the non-threat condition ($M = .51$, $SD = .22$) were significantly more accurate than women in the stereotype threat condition ($M = .36$, $SD = .22$), $t(36) = 1.99$, $p < .05$, Cohen's $D = .68$, $r = .32$ and than men in the stereotype threat condition ($M = .36$, $SD = .21$), $t(51) = 2.27$, $p < .03$, Cohen's $D = .70$, $r = .33$. These were the only comparisons to reach significance. The manipulation did not significantly affect men's accuracy.

3.4. Did stereotype threat reduce women's confidence in their responses?

For each test question answered, participants indicated on a scale from one to five (1 = low confidence; 5 = high confidence) their confidence that their response was correct. We averaged these responses to obtain a measure of each participant's overall confidence. The ANOVA on Confidence yielded significant main effects of Sex, $F(1, 87) = 6.98$, $p < .01$, Cohen's $D = .47$, $r = .23$, and Condition, $F(1, 87) = 5.42$, $p < .02$, Cohen's $D = .50$, $r = .24$. Males ($M = 3.51$, $SD = .78$) were more confident than females ($M = 3.14$, $SD = .79$); and participants in the non-threat condition ($M = 3.60$, $SD = .78$) were more confident than participants in the stereotype threat condition ($M = 3.21$, $SD = .79$). These results are consistent with past studies showing women to be less confident in mathematics than their male counterparts, despite the lack of performance differences (American Association of University Women, 1992, 1998; Stipek & Gralinski, 1991).

3.5. Did males and females differ in the grades they earned in the course?

Because the conditions of our study – and in particular, the stereotype threat condition – mimicked the conditions under which the participants took their regular calculus examinations, we wondered whether or not males and females differed in the grades they earned for the course. To answer this question we conducted a one-way ANOVA on Course Grades. This analysis yielded no significant results ($p > .9$), suggesting that males ($M = 2.73$, $SD = 1.07$) and females ($M = 2.72$, $SD = 1.03$) earned similar grades in the course.

3.6. Are course grades accurate predictors of males' and females' performance on the practice test?

Because males and females did not differ in their course grades, we wondered how well course grades predicted males' and females' performance on the practice test in the threat and non-threat conditions. First, to determine the prediction equation, we conducted a regression analysis which yielded the following: Predicted Number Correct = 1.84 + .47 * (Course Grade). Next, we computed the Predicted Number Correct for each participant. Finally, we conducted four paired-samples *t*-tests (comparing males and females in the stereotype threat and non-threat conditions) using Predicted Number Correct and Number Correct to determine whether there were significant differences in the predicted number correct and actual number correct on the practice test. These analyses indicated that for females in the non-threat condition, course grades significantly underpredicted performance on the practice test by 1.28 points ($p < .02$). For all other groups, grades were non-significant overpredictors of performance on the practice test (all ps > .12).

4. Discussion

This study revealed that the effects of stereotype threat are not limited to the typical woman's performance on general mathematics tests. Rather, even women at the upper ends of the ability distribution in college who opt to enroll in the most difficult math courses can be vulnerable to the effects of negative stereotypes. Specifically, framing a difficult calculus test as diagnostic of ability suppressed women's calculus performance. However, ensuring women that the same diagnostic test was free of gender-bias reduced stereotype threat and unleashed their mathematics potential. Not only did the women in the non-threat condition outperform women in the stereotype threat condition, but they also outperformed the men in either testing condition. In addition, women in the non-threat condition were more accurate on the test than women and men in the stereotype threat condition. Thus, test-diagnosticity clearly suppressed women's test performance, both in terms of actual number correct and accuracy. But when the situational threat was removed by assuring women that the test was unbiased, women's performance increased to a level that surpassed their male counterparts, a finding that was both surprising and encouraging.

Why would women outperform their male counterparts in this situation? It is likely that because women are more apt to self-select out of math and science fields early in their educational careers (American Association of University Women, 1998; Eccles, 1994), the women in advanced mathematics classes, such as those in our sample, comprise the most motivated and prepared female mathematics students. Men, on the other hand, may be more likely to pursue math and science careers even if they are less prepared academically to do so. Thus, the female participants in our sample may have been a more select group of math students than the male participants. Unconstrained by stereotype threat in our experimental group, these women performed closer to their latent ability levels, which were higher on average than the males in the course.

Still, if the pool of women enrolled in advanced mathematics courses was more self-selected than the pool of men, why did we not find a female advantage in course grades, as is often the case in mathematics courses (Bridgeman & Wendler, 1991)? One possibility is that the calculus course itself was inherently prone to induce stereotype threat in women. That is, there may have been an ethos of diagnosticity without any mention of gender-fairness or refuting of stereotypes that led to a suppression of females' course grades. In fact, the stereotype of male superiority in calculus was alive and well in these calculus courses. We asked 364 calculus students who they believed was better at calculus – men or women – and who they thought *other people* believed was better at calculus. The women believed that men and women have equal abilities, whereas the men believed that men are superior ($p < .001$). We also found that both men and women believed that *other people* thought men were better than women ($p < .001$) (Good & Aronson, 1998). And given the men's endorsement of male superiority in calculus, this was, in fact, an accurate perception of the existing stereotype in their learning environment.

In addition, stereotype threat may have been fostered by the sex composition of the class (Inzlicht & Ben-Zeev, 2000). In this course, the men outnumbered the women by a margin of two to one, a difference that is quite representative of mathematics departments across the country (for example, see Massachusetts Institute of Technology, 1999; National Science Foundation, 2006). Thus, the course exams, on which students' grades are primarily based, may have been confounded with stereotype threat effects, thus suppressing women's grades by suppressing their test performances.

Indeed, the lack of sex differences in course grades mirrors the lack of sex differences in test performance in the stereotype threat condition. In the non-threat condition, however, course grades significantly underpredicted women's

performance on the practice test. As our pattern of data suggests, if stereotype threat had been removed from the classroom culture, these women very likely would have earned higher grades, perhaps even higher than their male counterparts' grades. In support of this interpretation, consider our finding that grades underpredicted women's test performance, but only in the non-threat condition. In other words, when stereotype threat is alleviated, women may be capable of higher calculus performance than their course grades indicate.

A second possibility for why the females outperformed the males is that females simply could have prepared more for the practice test than the males did. Indeed, research indicates that females do as much, if not more, math homework than males (Eccles & Jacobs, 1986; Mau & Lynn, 2000). Consequently, it may be the case that the females studied more for this practice test and thus, did better on it. This, however, is not a complete explanation for our pattern of results. If females and males entered the class with equal ability but the females simply prepared more than the males, then why did the females not outperform the males in both conditions? That sex differences varied by condition suggests that females' scores were affected by a process other than preparation — namely, stereotype threat.

Whether our manipulation suppressed females' scores in the stereotype threat condition or enhanced them in the non-threat condition cannot be determined. Nevertheless, what remains clear is that when females simply took the diagnostic test – a situation that has suppressed females' scores in past studies – they did worse than when they were assured that the test was sex-fair.

Despite women's better performance in the non-threat condition compared to the stereotype threat condition, our manipulations did nothing to affect their level of confidence. This finding, however, should not be surprising for research has shown that women lack confidence in mathematics even when their performance is high (Linn & Hyde, 1989).

As discussed previously, this study differs from past studies in both the selectivity of the sample used, and in the methods of inducing and reducing vulnerability to stereotype threat. First, the men and women in our sample were enrolled in a calculus course specifically designed for mathematics, engineering, and science majors. This, together with the fact that women earned grades in the course that were no different from the men's grades, created a highly selective group of women who were well poised for successful careers as scientists and mathematicians. No study that we know of has examined the role of stereotype threat with such a select group of aspirants. Although others have used participants who scored well on standardized tests of mathematics (see Quinn & Spencer, 2001; Spencer et al., 1999), these participants generally were recruited from psychology courses or from the psychology subject pool. Their intended majors and career paths, therefore, likely included non-math-based fields. Furthermore, although Quinn and Spencer (2001) recruited highly capable females, they explicitly excluded those who scored above 700 on the SAT-I. In contrast, we drew our sample specifically from courses that were likely to produce future mathematicians and scientists and also demanded high math ability. Thus, the results of our study expand our knowledge of populations that are vulnerable to stereotype threat effects: even women enrolled in the most advanced math courses that prepare students for careers in mathematics and science can experience underperformance due to stereotype threat.

Second, our procedures for inducing stereotype threat mirror the testing procedures in most classrooms, and thus are a truer reflection of real-world stereotype threat effects. Furthermore, simply addressing the fairness of the test while retaining its diagnostic nature presents a more practical method of alleviating stereotype threat that is easily transferable to any testing situation. Past studies have gone to great lengths to inoculate students against the effects of negative stereotypes. In particular, researchers have explicitly taught students about the malleability of intelligence as one method of countering the stereotype's message of limited, fixed ability due to group membership. These labor-intensive methods have included showing students videos that teach about the brain's ability to form new neural connections (Aronson et al., 2002), having students write pen-pal letters espousing the malleable view of intelligence (Aronson et al., 2002), or constructing a mentoring program in which the focus of the mentoring relationship was to teach students about the malleable nature of intelligence (Good et al., 2003). These methods, although highly effective, are unlikely to be implemented in regular mathematics classrooms.

Our study, in contrast, presents a method of reducing stereotype threat that is easily-implemented not only in mathematics classrooms but also in standardized testing situations. Specifically, testing procedures could easily be modified to include a brief statement that the test, although diagnostic of underlying mathematics ability, is sex-fair. Moreover, this easily-implemented procedure could be used with mathematics students in any math class – elementary through graduate school – or testing situation — NAEP, SAT, or GRE, for example. Doing so has the potential to change the current dynamic of gender-based performance differences in mathematics.

Moreover, it is important to identify easily replicable methods for reducing stereotype threat because doing so creates the potential to change not only females' performance in mathematics but also their career aspirations. For

example, research has shown that women who perceived that their college calculus classes conveyed negative stereotypes about women's math abilities reported a lower sense of belonging to the math community – that is, they felt less like accepted members of the math community whose presence was valued – than those who did not perceive a stereotypical climate (Good et al., in preparation-a). Moreover, this threat to their identity as a future mathematician (or scientist) had real consequences for their achievement and career aspirations: when women's sense of belonging was reduced due to their perceptions of a stereotypical environment, they earned lower grades in the course and were less likely to intend to take any more math classes in the future.

Thus, stereotypes can cause individuals enough discomfort to lead them to drop out of the domain and redefine their professional identities. When the domain is something as fundamental as mathematics, domain avoidance essentially shuts the door to potentially lucrative careers in science, engineering, and technology. And as with mathematics performance, the effects of stereotypes on professional identity have roots early in schooling, for it has been found that stereotypes can undermine sense of belonging for girls in math as early as middle school (Good et al., in preparation-b). This has important consequences for girls' identities as future mathematicians and scientists, because it is precisely the middle school years when girls' confidence in and liking of mathematics begins to wane.

The results of this study also have direct implications for the discussion of social forces as a contributing factor to women's underperformance and under-representation in science and mathematics disciplines. These results show quite clearly the role of non-biological factors in the mediation of test performance in the upper strata of college mathematics curriculum. Without denying the link between biology and math abilities (e.g., Halpern, 1992), it is clear that even among a select group of females and males with equivalent grades in high-level mathematics, stereotype threat can prevent women from performing up to their potential. Among these participants, this meant that stereotype threat suppressed their scores to the level of their male counterparts.

Our hope is that educators will incorporate these results into any discussion about the problem of and remedies for the under-representation of women in the upper echelons of mathematics and science professions. Clearly, social forces, such as negative stereotypes, remain at the forefront of the factors that contribute to the lack of women who succeed at the highest levels of mathematics. But with wise attention to situational factors – such as reassuring gender-fair testing – educators can help females at any stage of their mathematics education approach their potential and increase their numbers in mathematics and science professions.

References

American Association of University Women. (1992). *Shortchanging girls, Shortchanging America: A call to action.* Washington, DC: American Association of University Women.

American Association of University Women. (1998). *Gender gaps: Where schools still fail our children.* Executive summary. Washington, DC: American Association of University Women Educational Foundation

Aronson, J., Fried, C., & 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.

Aronson, J., & Steele, C. M. (2005). Stereotypes and the fragility of human competence, motivation, and self-concept. In C. Dweck & E. Elliot (Eds.), *Handbook of competence & motivation* (pp. 436−456). Guilford: New York.

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

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology*, *83*, 275−284.

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's advanced progressive matrices. *Journal of Applied Psychology*, *91*(4), 979−985.

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

College Board. (2005). 2005 College-bound seniors: Total group profile report Retrieved August 1, 2006 from http://www.collegeboard.com

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

Danaher, K. & Crandall, C.S. (in press). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology.*

Dar-Nimrod, I., & Heine, S. J. (2006). Exposure to scientific theories affects women's math performance. *Science*, *314*, 435.

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

Eccles, J. S., Adler, T. F., & 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. (1994). Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement related choices. *Psychology of Women Quarterly*, *18*, 585−609.

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs*, *11*, 367−380.

Eccles, J. S., & Jacobs, J. E. (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.

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

Good, C., & Aronson, J. (1998). Stereotypes in the calculus classroom: Men's and women's perceptions of sex-differences. Unpublished data. The University of Texas at Austin.

Good, C., & Aronson, J. (in preparation). The development of stereotype threat in children. Manuscript in preparation. Barnard College.

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.

Good, C., Dweck, C. S., & Aronson, J. (2007). Social identity, stereotype threat, and self-theories. In A. Fuligni (Ed.), *Social categories, identities and educational participation* (pp. 121−145). New York, NY: Russell Sage Foundation.

Good, C., Dweck, C.S., & Rattan, A. (in preparation-a). The effects of perceiving fixed-ability environments and stereotyping on women's sense of belonging to math. Manuscript in preparation. Barnard College.

Good, C., Dweck, C.S., & Rattan, A. (in preparation-b). Do I belong here? Middle school girls' sense of belonging and achievement in math. Manuscript in preparation. Barnard College.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, *15*, 25−46.

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

Hyde, J. S. (1994). Can meta-analysis make feminist transformations in psychology? *Psychology of Women Quarterly*, *18*, 451−462.

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.

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.

Linn, M., & Hyde, J. (1989). Gender, mathematics, and science. *Educational Researcher*, *18*, 17−27.

Massachusetts Institute of Technology. (1999). MIT Faculty Newsletter, vol. XI, no. 4, March, 1999. Retrieved August 1, 2006 from http://web.mit.edu/fnl/women/women.html#numberwomen

Mau, W., & Lynn, R. (2000). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution & Gender*, *2*, 119−125.

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, *27*, 486−493.

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

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

National Science Foundation. (2006). Science and Engineering Indicators, 2006. Retrieved August 1, 2006 from http://www.nsf.gov/sbe/srs/seind06

Pinker, S. (2005). The science of difference: Sex ed. *The New Republic.* February 14.

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.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American — White difference on cognitive tests. *American Psychologist*, *59*, 7−13.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, *38*, 194−201.

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

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science: A critical review. *American Psychologist*, *60*, 950−958.

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

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., & Aronson, J. (2002). Contending with images of one's group: The psychology of stereotype and social identity threat. In M. Zanna (Ed.), *Advances in Experimental Social Psychology*, *34* (pp. 379−440). San Diego: Academic Press.

Stipek, D., & Gralinski, H. (1991). Gender differences in children's achievement-related beliefs and emotional responses to success and failure in mathematics. *Journal of Educational Psychology*, *83*, 361−371.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*, 665−693.

Summers, L. (2005). Remarks at NBER Conference on Diversifying the Science & Engineering Workforce. Retrieved August 1, 2006 from http://www.president.harvard.edu/speeches/2005/nber.html